DOCTORAT BRETAGNE LOIRE / MATHSTIC

UBO Université de Bretagne Occidentale

b<>com

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE

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

Par
François SALMON

Contrôle des impressions spatiales dans un environnement acoustique virtuel.

Thèse présentée et soutenue à Rennes, le 26 mars 2021
Unité de recherche : Lab-STICC, CNRS, UMR 6285

Rapporteurs avant soutenance :

Mathieu LAVANDIER	Chargé de Recherche, ENTPE, Lyon
Roland BADEAU	Professeur Institut Mines-Télécom, Télécom ParisTech, Paris

Composition du Jury :

Présidente :	Catherine LAVANDIER	Professeur des Universités, Université de Cergy-Pontoise
Examinateurs :	Olivier WARUSFEL	Chargé de Recherche, Ircam, Paris
	Mathieu LAVANDIER	Chargé de Recherche, ENTPE, Lyon
	Roland BADEAU	Professeur Institut Mines-Télécom, Télécom ParisTech, Paris
Dir. de thèse :	Mathieu PAQUIER	Professeur des Universités, UBO, Brest
Co-encadrant :	Etienne HENDRICKX	Maître de Conférences, UBO, Brest
	Nicolas EPAIN	Ingénieur de Recherche, b<>com, Cesson-Sévigné
	Vincent KOEHL	Maître de Conférences, UBO, Brest
Invité(s)		
	Jean-Yves AUBIE	Responsable du laboratoire AMC, b<>com, Cesson-Sévigné

Table des matières

Glossaire **9**

Remerciements **11**

Introduction **13**

I Étude de l'influence de la source sonore et de la vision sur la perception de l'acoustique d'une salle **17**

1 Modélisation et mesures de l'acoustique d'une salle **19**
1.1 Modélisation de l'acoustique d'une salle 20
1.1.1 Caractérisation temporelle. 20
1.1.2 Caractérisation fréquentielle 22
1.2 Mesures perceptives de l'acoustique d'une salle 23
1.2.1 Les méthodes d'évaluation perceptive. 24
1.2.2 Application à la perception de l'acoustique d'une salle 26
1.2.3 Les lexiques d'attributs perceptifs 28
1.3 Mesures physiques de l'acoustique d'une salle 33
1.3.1 Les durées de décroissance. 35
1.3.2 Les mesures de clarté 37
1.3.3 La force sonore 39
1.3.4 Les mesures d'impression spatiale. 40
1.4 Conclusion 47
Bibliographie. 49

2 Expérience I : l'influence de la vision sur la perception de l'acoustique d'une salle **59**
2.1 Les interactions entre modalités sonore et visuelle 59
2.1.1 Localisation de source sonore 60
2.1.2 Perception de la distance 60
2.1.3 Externalisation 61
2.1.4 Impressions spatiales 62
2.1.5 Le but de l'étude 63
2.2 Test perceptif. 63
2.2.1 Stimuli visuels 63
2.2.2 Réponses impulsionnelles de salles 64
2.2.3 Stimuli sonores 65
2.2.4 Binauralisation et head-tracking 66

2.2.5 Procedure 66
2.2.6 Sujets 67
2.3 Résultats 67
2.3.1 Matrices de dissemblance 67
2.3.2 Analyse de la variance 67
2.3.3 Analyse multidimensionnelle 69
2.4 Discussion 70
2.5 Conclusion 71
Bibliographie 72

3 Expérience II : l'influence de la source sonore sur la perception de l'acoustique d'une salle 77
3.1 Introduction 77
3.2 Test perceptif 78
3.2.1 Sources sonores 78
3.2.2 Binauralisation et head-tracking 79
3.2.3 Procédure 79
3.2.4 Sujets 80
3.3 Résultats 80
3.3.1 Matrices de dissimilarité 80
3.3.2 Analyse de la variance 81
3.3.3 Analyse multidimensionnelle 83
3.3.4 Caractérisations objectives des dimensions perceptives 85
3.4 Discussion 86
3.5 Conclusion 87
Bibliographie 87

II La paramétrisation d'une réponse impulsionnelle spatiale pour l'auralisation de l'acoustique d'une salle 89

4 Paramétrisations de réponses impulsionnelles spatiales enregistrées 91
4.1 Spatial Decomposition Method (SDM) 93
4.2 Spatial Impulse Response Rendering (SIRR) 96
4.3 Higher-Order Spatial Impulse Response Rendering (HO-SIRR) 99
4.4 Reverberant Spatial Audio Object (RSAO) 100
4.5 Synthèse paramétrique par Sound Field Analysis (SFA) 104
4.6 Autres méthodes de paramétrisation spatiale 105
4.7 La question de la décorrélation 106
4.7.1 Les différentes méthodes de décorrélation 106
4.7.2 La minimisation de l'emploi de la décorrélation 107
4.8 Conclusion 109
Bibliographie 112

5 Expérience III : l'influence de la résolution spatiale d'une réponse impulsionnelle sur la perception de l'acoustique d'une salle en binaural 117
5.1 Introduction . . . 118
5.2 Manipulation du champ sonore dans le domaine ambisonique . . . 119
5.2.1 Segmentation temporelle . . . 120
5.2.2 Reconstruction mixte du champ sonore . . . 120
5.2.3 Représentation du champ sonore dans le plan . . . 121
5.3 Test perceptif . . . 121
5.3.1 Réponses impulsionnelles de salles . . . 122
5.3.2 Sources sonores . . . 123
5.3.3 Binauralisation et head-tracking . . . 123
5.3.4 Procedure . . . 124
5.3.5 Sujets . . . 125
5.4 Résultats . . . 125
5.5 Discussion . . . 128
5.5.1 Aucune influence observée de l'ordre ambisonique mixte . . . 128
5.5.2 Aucune influence observée de la dimensionalité . . . 128
5.5.3 Influence de l'espace sonore . . . 129
5.6 Conclusion . . . 131
Bibliographie . . . 132

6 Expérience IV : l'influence des résolutions temporelle et fréquentielle de la paramétrisation des premières réflexions sur la perception de l'acoustique d'une salle en binaural 137
6.1 Introduction . . . 138
6.2 Paramétrisation d'une SRIR par matrice de covariance . . . 139
6.2.1 Création de signaux décorrélés . . . 140
6.2.2 Reproduction de la matrice de covariance . . . 142
6.2.3 Reproduction du signal omnidirectionnel . . . 142
6.3 Test perceptif . . . 143
6.3.1 Les résolutions temps-fréquences employées . . . 143
6.3.2 Le calcul des paramètres SDM . . . 144
6.3.3 Génération de l'ancre . . . 144
6.3.4 Création des fichiers binauraux . . . 145
6.3.5 Procédure . . . 145
6.3.6 Sujets . . . 148
6.4 Résultats . . . 148
6.4.1 Outliers . . . 148
6.4.2 Analyse de la variance . . . 148
6.4.3 Comparaisons des niveaux du facteur «méthode» . . . 149
6.4.4 Résultats des interactions avec le facteur «méthode» . . . 150
6.5 Discussion . . . 152
6.5.1 Le niveau de diffusion . . . 152
6.5.2 L'influence des résolutions employées . . . 152
6.5.3 L'influence de l'espace . . . 152
6.5.4 L'influence de la source . . . 153

6.5.5 La méthode SDM 153
6.5.6 L'effet de précédence 154
6.5.7 L'avantage de la paramétrisation par matrice de covariance par rapport à la SDM 155
6.6 Conclusion 156
Bibliographie 157

III Étude du contrôle perceptif des impressions spatiales 161

7 Les origines physiques des impressions spatiales 163
7.1 La base de données GRAP 164
7.1.1 Les notes du RAQI 166
7.1.2 Les réponses impulsionnelles de salle 166
7.2 Ajout de données complémentaires : la génération de SRIRs 168
7.3 Les paramètres acoustiques étudiés 169
7.4 La prédiction de la largeur apparente de source 172
7.4.1 Corrélations entre les notes d'ASW et les paramètres acoustiques 172
7.4.2 Régression linéaire multidimensionnelle 174
7.4.3 Les paramètres acoustiques pertinents pour la classification . 176
7.5 Prédiction de l'enveloppement 182
7.5.1 Corrélations entre les notes d'enveloppement et les paramètres acoustiques 182
7.5.2 Régression linéaire multidimensionnelle 184
7.5.3 Les paramètres acoustiques pertinents pour la classification . 185
7.6 Conclusion 189
Bibliographie 190

8 Expérience V : l'influence de la spatialisation des premières réflexions sur la largeur apparente de source 193
8.1 Introduction 193
8.2 Les modifications spatiales des premières réflexions 194
8.2.1 Amplification directionnelle 195
8.2.2 Déformation angulaire 197
8.2.3 Élargissement 198
8.2.4 Décorrélation des composantes sectorielles 199
8.3 Test Perceptif 201
8.3.1 Réponses impulsionnelles spatiales de salle 201
8.3.2 Sources sonores 202
8.3.3 Création des stimuli sonores 203
8.3.4 Procédure 207
8.3.5 Sujets 209
8.4 Résultats 209
8.4.1 Outliers 209
8.4.2 Analyse de la variance 209
8.4.3 Comparaisons des niveaux du facteur «traitement» 210

8.4.4 Résultats des interactions avec le facteur «traitement» 211
8.5 Discussion 213
8.5.1 Le niveau de diffusion 213
8.5.2 L'efficacité des traitements employés 213
8.5.3 Les variations d'énergie des premières réflexions 214
8.5.4 L'emploi d'un autre indicateur d'ASW 214
8.5.5 L'influence de la source 216
8.5.6 L'influence de l'espace 216
8.6 Conclusion 219
Bibliographie 220

9 Étude du contrôle anisotrope de la réverbération tardive 223
9.1 Le réseau récursif de lignes à retard 226
9.1.1 Le calcul des filtres d'atténuation 227
9.1.2 La modulation de la matrice de mélange 231
9.1.3 Le choix des délais et des périodes de modulation 233
9.2 Le FDN directionnel 234
9.3 Évaluation de la méthode 237
9.3.1 Les paramètres utilisés 238
9.3.2 L'estimation du temps de mélange 238
9.3.3 Résultats 239
9.3.4 Discussion 240
9.4 Conclusion 241
Bibliographie 242

Conclusion 245

Annexes 251

A La mesure d'une réponse impulsionnelle de salle 253
A.1 Balayage sinusoïdal 254
A.2 Extension de la décroissance de l'intensité sonore 256
Bibliographie 258

B La représentation spatiale d'une réponse impulsionnelle multicanale 261
B.1 Système de coordonnées spatiales adopté 261
B.2 Décomposition en harmoniques sphériques 262
B.3 Erreur de troncature 264
B.4 Captation et encodage d'un champ sonore 265
B.4.1 L'encodage des signaux captés 265
B.4.2 Fréquence de repliement spatial 267
Bibliographie 268

C L'écoute binaurale dynamique d'une scène sonore réverbérée 269
C.1 Fonctions de transfert relatives à la tête 269
C.2 Calcul des filtres de décodage binaural 271
C.3 Prise en compte des mouvements de la tête 274

Bibliographie 275

D La simulation de réponses impulsionnelles de salle 279

D.1 La méthode des sources images 280

D.2 Le lancé de particules 281

D.3 La prise en compte de la diffraction 282

Bibliographie 283

E Publications liées à la thèse 285

F The Influence of Vision on Perceived Differences Between Sound Spaces 287

Glossaire

ANOVA *Analysis Of Variance.*
ASW *Apparent Source Width.*
BQI *Binaural Quality Index.*
BRIR *Binaural Room Impulse Response.*
C_{80} Indice de clarté (calculé sur 80 ms).
C_{50} Indice de clarté (calculé sur 50 ms).
COMPASS *COding and Multidirectionnal Parametrization of Ambisonics Sound Scenes.*
D_{50} Indice de définition.
DRR *Direct-to-Reverberant Ratio.*
DirAC *Directional Audio Coding.*
EDC *Early Decay Curve.*
EDT *Early Decay Time.*
FBR *Front-to-Back Ratio.*
FDN *Feedback Delay Network.*
G Force sonore.
G_E Force sonore précoce.
G_{EL} Force sonore latérale précoce.
G_L Force sonore tardive.
GRAP *Ground Truth and Room Acoustical Analysis and Perception.*
IACC *Interaural Cross Correlation.*
J_{LF} Fraction d'énergie latérale.
HO-DirAC *Higher-Order Directional Audio Coding.*
HO-SIRR *Higher-Order Spatial Impulse Response Rendering.*
HMD *Head Mounted Display.*
HRTF *Head-Related Transfer Function.*
HRIR *Head-Related Impulse Response.*
INDSCAL *INdividual Difference SCALing.*
JND *Just Noticeable Difference.*
LEV *Listener Envelopment.*
L_J Force sonore latérale tardive.
LLF *Late Lateral Fraction.*
MDS *Multidimensional Scaling.*
MUSHRA *MUlti-Stimulus with Hidden Reference and Anchor.*
RAQI *Room Acoustical Quality Inventory.*
RSAO *Reverberant Spatial Audio Object.*
RT *Reverberation Time.*
SBT$_s$ *Spatially-Balanced Central Time.*
SDM *Spatial Decomposition Method.*
SFA *Sound Field Analysis.*

SIRR *Spatial Impulse Response Rendering*.
SRIR *Spatial Room Impulse Response*.
$\boldsymbol{T}_s$ Temps central.

Remerciements

Je souhaite en premier lieu remercier mon directeur de thèse, Mathieu Paquier, pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour son écoute et son regard avisé. Je tiens également à remercier Etienne Hendrickx pour son enseignement de la psycho-acoustique puis son accompagnement dans ces travaux de recherche, pour sa clarté et la justesse de ses remarques minutieuses. Un grand merci à Nicolas Epain pour son soutien quotidien, sa pertinence et sa rigueur ainsi que pour la patience dont il a fait preuve en répondant à mes nombreuses questions. Je le remercie également pour la richesse de ses recommandations musicales.

J'adresse mes remerciements à Mathieu Lavandier et Roland Badeau pour avoir accepté d'être rapporteurs de ce manuscrit ainsi qu'à Olivier Warusfel et Catherine Lavandier pour leur participation au jury. Je remercie également Catherine Lavandier pour avoir dirigé mon comité de suivi individuel en compagnie de Thibaut Carpentier que j'associe à ces remerciements.

Ce travail n'aurait pas été possible sans le soutien de l'Institut de Recherche Technologique b<>com et en particulier de Ludovic Noblet et de Jean-Yves Aubié qui m'ont permis de me consacrer sereinement à l'élaboration de ce travail doctoral. Merci à tous les membres de l'équipe AMC pour leurs encouragements et leur bonne humeur qui ont participé a créer un environnement de travail convivial durant ces trois années. Je souhaite également remercier Marc Muller et Quentin George de la société *Aspic Technologies* grâce à qui ces travaux ont débuté.

Merci à mes collègues Cathy Colomes, Jérôme Daniel et Louis Anglionin ainsi qu'à Yaël Laouar et Charles Verron de la société *Noise Makers* pour leur aide dans la réalisation des stimuli utilisés pour les tests perceptifs. Je souhaite également exprimer ma gratitude envers les sujets qui ont pris part à ces tests, trop nombreux pour être mentionnés nommément. Ils se reconnaitront si ils savent à quel point «*il est un air pour qui je donnerais tout Rossini, tout Mozart et tout Weber...*».

Je remercie les personnes m'ayant ouvert les portes de lieux rennais pour l'enregistrement de réponses impulsionnelles de salle : Béatrice Macé, Florence Havy, Robert Langouët, Bernard Heudré et César Olivier. Merci également à Cyrille Dodard pour nous avoir permis d'accéder à la chambre anéchoïque de la société *Cabasse* afin de mesurer l'enceinte utilisée lors des enregistrements.

Mes pensées vont à mes amis de Paris et d'Angers physiquement distants depuis trop longtemps et avec qui il est grand temps que la bamboche reprenne. Merci à mes compagnons Simon, Thomas et Cían, dont les feux à longues portées me guident toujours. Merci à mes formidables parents pour m'avoir toujours soutenu dans mes choix. Merci à Victorine dont la patience et la tendresse nous ont permis de passer ces trois belles années dans la sérénité. Je remercie également Serge et Brigitte pour leur apport nutritif journalier.

Introduction

L'acoustique d'un espace, c'est-à-dire l'ensemble des qualités relatives à sa sonorité, peut modifier considérablement la perception d'un événement sonore qui s'y produit. La perception d'une voix dans une cathédrale est très différente de celle d'une voix dans une salle de bain. Ce phénomène est dû à la réverbération du son. La réverbération correspond à la prolongation d'un phénomène sonore en raison de la réflexion du son sur les surfaces d'un espace complètement ou partiellement clos. Ainsi, sans pour autant être à l'origine de l'excitation sonore, l'espace se manifeste de manière audible, révélant ainsi sa forme, les matériaux qui le constituent ou encore les objets qu'il contient. Notre perception des sources sonores peut être simplement influencée par la présence ou l'absence d'un mur, d'une porte ouverte ou par la température de la pièce. La réverbération du son dans un espace participe à rendre cet environnement présent dans la mesure où elle renseigne l'auditeur sur les caractéristiques de l'environnement qui l'entoure.

Depuis l'apparition de l'enregistrement et de la diffusion sonore, les ingénieurs du son ont employé des effets de réverbération, c'est-à-dire des procédés artificiels visant la reproduction de propriétés acoustiques d'une salle. Ces effets de réverbération permettent de plonger l'auditeur dans un contexte acoustique particulier et de créer des relations entre les sources sonores enregistrées pour mieux les mélanger. Ces procédés ont par exemple consisté à utiliser des dispositifs de vibrations mécaniques employant des ressorts, des plaques métalliques ou à utiliser plus simplement une chambre d'écho spécialement construite. Désormais, les effets de réverbération sont quasi-exclusivement produits par des procédés numériques qui sont plus simples à manipuler et à paramétrer. De nombreuses méthodes existent pour synthétiser numériquement un effet de réverbération de haute qualité et sont utilisées dans des domaines aussi variés que la production musicale, cinématographique, radiophonique et dans le domaine du jeu vidéo.

Les technologies dites «immersives» ont fait l'objet ces dernières années d'un véritable engouement, principalement suscité par la réalité virtuelle et la réalité augmentée. Les formes audiovisuelles complexes liées à ces technologies impliquent de nouveaux enjeux techniques et esthétiques : de nouveaux outils sont nécessaires pour permettre d'autres manières de créer, de montrer, de raconter, de transmettre des sensations et des émotions. En sollicitant l'ouïe et la vue, en engageant le corps et l'attention du spectateur dans l'espace virtuel, les dispositifs immersifs visent à produire des sensations de présence, à réduire l'écart entre espace physique et espace simulé. À cette fin, le rendu sonore binaural est couramment employé pour simuler la présence de sources sonores situées dans toutes les directions de l'espace en créant des indices de localisation sonore. Cette technologie consiste à reproduire les filtrages introduits par la tête et les pavillons des oreilles dépendants de la provenance du son. On peut ainsi donner à entendre à l'auditeur une scène sonore perçue en dehors de sa tête,

dans l'espace environnant et ceci simplement grâce à un casque audio. Pour des raisons pratiques, les filtrages propres à la tête de l'auditeur ne peuvent généralement pas être pris en compte. Des filtres correspondant à ceux d'une autre personne ou d'une tête artificielle sont souvent utilisés. Dans ce cas, le rendu binaural est qualifié de *non-individualisé*. Lorsqu'il est associé à un dispositif de suivi des mouvements de tête, le rendu binaural est dit *dynamique*. En utilisant un tel dispositif de suivi - inclus nativement dans les visiocasques de réalité virtuelle - la scène sonore tourne en accord avec les mouvements de tête de l'auditeur, de la même manière que la scène visuelle.

Avec la production croissante des contenus destinés à la réalité virtuelle, pour lesquels le réalisme et la sensation d'immersion sont primordiaux, une attention particulière doit être portée par les ingénieurs du son à la conception de l'espace sonore et plus précisément à l'usage des effets de réverbération. L'utilisation d'un effet de réverbération contextualise l'écoute, améliore notamment la perception de la distance, l'externalisation (la capacité à percevoir les sons en dehors de sa tête) et la sensation de présence. Pourtant, les outils destinés à créer un effet de réverbération pour la production de contenus immersifs sont très similaires à ceux utilisés pour la vidéo non-360° ou pour des formats sonores 2D tels que la stéréophonie. De nouveaux outils et nouvelles pratiques permettraient aux ingénieurs du son de mieux prendre en considération la dimension spatiale du son immersif.

Dans cette thèse, nous souhaitons étudier les liens entre les propriétés temporelles, spectrales et spatiales d'un espace sonore et des attributs perceptifs - c'est-à-dire des caractéristiques sonores perceptibles - de l'acoustique d'une salle. L'identification de ces liens permettra un contrôle des attributs perceptifs et donc une aide à la conception d'un espace sonore.

Pour cela nous avons identifié deux pré-requis. D'une part, il est nécessaire d'étudier l'influence du contenu audiovisuel sur les attributs perceptifs afin d'établir les facteurs à prendre en compte dans le contrôle d'un effet de réverbération. Plus précisément, nous étudierons comment la présence d'images et la nature de la source sonore peuvent influencer la perception de l'espace sonore. D'autre part, il est nécessaire d'identifier les éléments pertinents du signal sonore à manipuler. En effet, la description d'un espace sonore peut constituer un grand nombre de données et la réduction de ces données peut permettre de simplifier la description de l'acoustique d'une salle et ainsi de mieux appréhender les effets de la modification du signal sonore. Dans un contexte où les ressources en calculs sont limitées, la réduction des données présente également l'avantage de diminuer les opérations nécessaires à la reproduction d'un espace sonore.

Dans ce document, nous étudierons la perception de l'espace sonore selon une reproduction communément employée dans le contexte de la réalité virtuelle : un rendu binaural dynamique non-individualisé. Bien que l'utilisation d'un rendu binaural non-individualisé puisse accroître les erreurs de localisation et des difficultés d'externalisation, ces problèmes sont efficacement atténués par la présence de réverbération dans les signaux sonores et l'usage d'un dispositif de suivi des mouvements de tête. Par ailleurs, nous utiliserons le format ambisonique pour la représentation spatiale des scènes sonores étudiées. Ce format répandu permet la reproduction d'un champ sonore décrit à la position de la tête d'un auditeur et la prise en compte efficace des

mouvements de tête. Pour une description détaillée du rendu binaural adopté dans le but de reproduire l'acoustique d'une salle, le lecteur peut se référer aux annexes A, B et C.

Dans la première partie du document, après avoir exposé la modélisation ainsi que les mesures physiques et perceptives de l'acoustique d'une salle, nous présenterons deux études portant sur l'influence de facteurs jouant un rôle important dans la perception de l'espace :

— Une étude portant sur l'influence de la vision sur la perception de l'acoustique d'une salle. De nombreuses études font état d'une influence significative de l'environnement visuel sur la perception de la position d'une source. Qu'en est-il de la perception de l'acoustique ? La présence des informations visuelles implique-t-elle une attention moindre vis à vis des informations sonores ? Le cas échéant, est-il possible de réduire la précision des informations sonores ? Nous déterminerons si les tests perceptifs liés à l'étude des attributs perceptifs doivent inclure la composante visuelle des espaces étudiés.
— Une étude portant sur l'influence de la source sonore sur la perception de l'acoustique d'une salle. La source sonore joue un rôle fondamental pour révéler les propriétés acoustiques d'une salle. Dans quelle mesure des sources sonores communes influencent-elle cette perception ? Nous chercherons à savoir à quel point la perception d'espaces sonores peut changer d'une source sonore à l'autre. Nous déterminerons si les tests perceptifs liés à l'étude des attributs perceptifs doivent inclure un grand nombre de sources sonores.

Dans la seconde partie du document, nous identifierons les paramètres de réponses impulsionnelles spatiales de salles (SRIRs) - tels que des réflexions spéculaires ou des pentes de décroissance - nécessaires et suffisants à la reproduction fidèle de l'acoustique d'une salle et qui facilitent le contrôle de la reproduction sonore. Dans un premier temps, les méthodes de paramétrisation existantes qui ont identifié de tels paramètres seront décrites. Ces méthodes n'ayant pas déterminé la précision avec laquelle paramétrer des SRIRs dans notre contexte d'écoute, nous présenterons deux études visant à évaluer les résolutions suffisantes pour la paramétrisation :

— Une étude portant sur la résolution spatiale. Nous observerons si il est pertinent d'un point de vue perceptif de reproduire les premières réflexions et la réverbération tardive d'une SRIR avec une précision spatiale réduite en termes d'ordre ambisonique.
— Une étude portant sur la résolution temporelle et fréquentielle de la paramétrisation des premières réflexions. Nous observerons si il est pertinent de reproduire la répartition spatiale de l'énergie sonore des premières réflexions avec des précisions temporelle et fréquentielle réduites.

Dans la dernière partie du document, faute de pouvoir étudier le contrôle de tous les attributs perceptifs récurrents, nous nous concentrerons sur le contrôle de deux attributs relatifs aux impressions spatiales de l'acoustique d'une salle. Le terme «impressions spatiales» désigne dans la littérature deux attributs perceptifs distincts : la

largeur apparente de source et l'enveloppement. Nous tenterons d'abord d'identifier les propriétés acoustiques d'un champ sonore qui sont liées à la perception d'une source large et à la sensation d'enveloppement. Deux études évaluant le contrôle de ces attributs perceptifs seront ensuite présentées :

— Une étude portant sur le contrôle de la largeur apparente de source. Plusieurs méthodes de transformation spatiale permettant d'accroître et de réduire la largeur apparente de source seront évaluées.
— Une étude portant sur le contrôle de la directivité de l'énergie tardive. Une architecture de réverbérateur artificiel basée sur plusieurs réseaux récursifs de lignes à retard sera décrite et sa capacité à reproduire la sensation d'enveloppement sera évaluée.

I

Étude de l'influence de la source sonore et de la vision sur la perception de l'acoustique d'une salle

1

Modélisation et mesures de l'acoustique d'une salle

Ce premier chapitre caractérise l'acoustique d'une salle sous différents aspects. Premièrement, nous décrivons les propriétés physiques du phénomène de réverbération formé par la propagation des ondes sonores dans un espace clos. En particulier, nous détaillons les structures temporelle et fréquentielle qui régissent ce phénomène. Pour ce faire, nous présentons la modélisation d'une salle comme un canal de transmission acoustique entre une source sonore et un récepteur qui peut être décrit par une ou plusieurs réponses impulsionnelles. Deuxièmement, les différents attributs perceptifs couramment utilisés dans la littérature pour caractériser la perception de l'acoustique d'une salle sont présentés. Nous exposons des lexiques élaborés pour uniformiser les termes et définitions employés dans les évaluations perceptives. Enfin, nous présentons plusieurs paramètres acoustiques extraits de réponses impulsionnelles de salle qui sont pertinents pour caractériser des attributs perceptifs. Certains de ces paramètres acoustiques sont décrits dans la norme ISO 3382-1. Celle-ci ne couvre cependant que le cas des salles de spectacle et ne permet pas de couvrir l'ensemble des attributs perceptifs identifiés. De plus, la caractérisation objective des attributs perceptifs reste partielle, notamment pour la largeur apparente de source et l'enveloppement.

Wallace Sabine est considéré comme un précurseur dans le domaine de l'acoustique des salles. Il fut le premier à lier le temps de réverbération - le temps nécessaire à ce que l'intensité sonore décroisse de 60 dB - au volume et aux surfaces absorbantes d'une salle [1]. Au début du siècle dernier, les travaux en acoustique se sont consacrés à l'étude des valeurs optimales du temps de réverbération et de sa dépendance en fréquence. Après les années 50, l'étude de l'acoustique des salles s'est étendue à l'analyse d'autres critères perceptifs ; en témoignent les travaux de référence de Leo Beranek [2]. On y trouve une description complète d'une centaine de salles de concert à travers le monde, des relations entre propriétés physiques et perceptives des salles ainsi qu'un classement de 50 salles selon leur qualité acoustique sur la base d'interviews et de questionnaires remplis par des experts du milieu (acousticiens, chefs

d'orchestres, musiciens, critiques). Les travaux de Beranek se sont largement fondés sur l'expérience d'auditeurs experts ainsi que sur son intuition et son interprétation. Par la suite d'autres études ont permis d'examiner l'acoustique des salles selon des méthodes plus contrôlées et plus rigoureuses. Les nombreuses études réalisées depuis ont mené à l'élaboration d'un standard de l'*International Organization for Standardization* pour la mesure de paramètres acoustiques dans les salles de concert : l'ISO 3382-1 [3]. Cette norme caractérise cinq aspects de la perception de l'acoustique de salles de concert : la réverbérance, la clarté, la force sonore, la largeur apparente de source et l'enveloppement. Nous verrons dans ce chapitre que la norme ISO 3382-1 reste incomplète en termes de paramètres acoustiques pour caractériser la perception de l'acoustique et que d'autres caractérisations physiques ont été établies, notamment pour décrire les impressions spatiales. Par ailleurs, la liste des propriétés perceptives couvertes par la norme ISO 3382-1 n'est pas exhaustive et d'autres aspects récurrents dans la littérature permettent de décrire la perception de l'acoustique d'une salle. Pour mesurer l'influence de l'acoustique sur ces propriétés perceptives, l'enjeu est d'identifier leur origines physiques. Pour cela, il est utile de formaliser à l'aide d'un modèle la manière dont les éléments qui interviennent dans le processus de réverbération s'organisent et interagissent les uns par rapport aux autres.

1.1. Modélisation de l'acoustique d'une salle

1.1.1. Caractérisation temporelle

Le son produit par une source sonore dans un espace clos ne disparaît pas immédiatement après l'arrêt de la source, mais reste audible pendant un certain temps avec une intensité sonore décroissante. La réverbération correspond à cette persistance du son dans un espace en raison de multiples réflexions sur les surfaces constituant l'environnement dans lequel il se propage [4]. Le terme «réflexion» désigne le processus auquel est soumis une onde sonore lorsqu'elle frappe un mur et désigne également le résultat de ce processus. En interagissant ainsi avec le lieu dans lequel il évolue, le son véhicule à l'auditeur des informations relatives à la position de la source, à la géométrie et aux matériaux de la salle.

Dans un espace donné, le trajet entre un émetteur sonore et un récepteur peut être assimilé à un ou plusieurs filtres linéaires invariants dans le temps : une ou plusieurs réponses impulsionnelles de salle. La réverbération peut alors être considérée comme la somme de toutes les réflexions sonores arrivant à un certain point de la salle après que celle-ci a été excitée par un signal sonore impulsif. Une telle réponse est composée d'un grand nombre de réflexions, comme le montre schématiquement la figure 1.1. Ce diagramme simplifié de la réponse impulsionnelle d'une salle, est également appelé «échogramme» ou «diagramme de réflexion».

Il est courant de décomposer cette réponse impulsionnelle en plusieurs régions temporelles pouvant comprendre : le son direct, des réflexions spéculaires précoces et une réverbération tardive [4]. Un signal sonore qui se propage sans réflexions depuis une source vers un auditeur arrive après une certaine durée depuis une direction

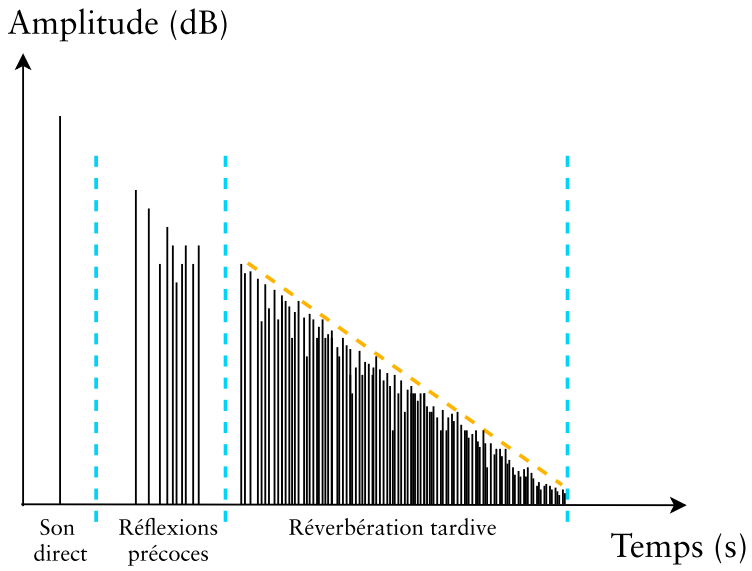


FIGURE 1.1 – Modélisation d'une réponse impulsionnelle de salle en trois régions temporelles distinctes.

précise. Un tel signal est qualifié de son direct. Il est suivi par des réflexions précoces apparaissant de manière sporadique et provenant de directions distinctes déterminées par la géométrie de la pièce. Leurs incidence, retard et amplitude ont une influence importante sur notre perception de l'espace.

En se propageant, ces ondes sonores se réfléchissent sur les surfaces de l'espace clos et sont dispersées dans diverses directions. La densité de réflexion - ou densité d'écho - augmente au fur et à mesure du temps à tel point que les réflexions ne peuvent plus être perçues individuellement. La densité d'écho δ_E, en un point de l'espace d'une salle rectangulaire aux parois rigides, est une fonction quadratique du temps d'arrivée et inversement proportionnelle au volume de la salle V [5] :

$$\delta_E = \frac{\mathrm{d}N_\mathrm{r}}{\mathrm{d}t} = 4\pi \frac{c^3 t^2}{V} \tag{1.1}$$

où N_r est le nombre de réflexions, t est le temps d'arrivée de la réflexion et c est la célérité du son. Cette relation quadratique de la densité d'écho au temps d'arrivée se vérifie également pour des salles de forme arbitraire [4].

A mesure que le temps avance, le champ sonore se constitue d'un grand nombre d'ondes sonores décorrélées, en raison de la diversité des modifications spectrales et temporelles subies par le signal sonore émetteur. Après un certain temps, le champ sonore est dit diffus [6] : il résulte de la superposition d'un grand nombre d'ondes planes, de sorte que toutes les directions de propagation sont également probables et que les relations de phase des ondes sont aléatoires en chaque point donné de l'espace. Dans des situations réelles, du fait des nombreuses irrégularités de formes qui constituent un espace, le champ sonore peut être assez bien approximé par un champ diffus. Néanmoins, il est commun de rencontrer des réverbérations tardives dont la distribution aléatoire de l'énergie n'est pas uniforme dans toutes les directions de l'espace. Dans ce cas, le champ sonore réverbéré est dit anisotrope et peut être caractérisé

par un degré plus ou moins élevé de diffusion que nous nommerons diffusivité dans la suite du document.

Les comportements temporel et spatial d'un champ diffus peuvent être caractérisés par un processus stochastique [7]. Ce sont les caractéristiques statistiques formées par l'ensemble des réflexions qui deviennent pertinentes d'un point de vue perceptif. Les réflexions tardives transportant de moins en moins d'énergie, la mesure h du champ diffus en un point de l'espace peut-être modélisée par un ou plusieurs bruits gaussiens à moyenne nulle ayant une décroissance exponentielle de son énergie [8, 9] :

$$h(t) = b(t)e^{-\alpha t} \tag{1.2}$$

où $\alpha > 0$ et $b(t)$ est un processus gaussien centré résultant d'un grand nombre de contributions indépendantes.

Le temps après lequel le champ sonore devient diffus est nommé temps de mélange. La réverbération tardive - ou queue de réverbération - étant communément assimilée à un champ diffus, le temps de mélange correspond donc au temps de transition entre les réflexions précoces et la réverbération tardive. D'un point de vue perceptif le temps de mélange peut être défini comme le moment où la queue de réverbération ne peut être distinguée de celle mesurée en toute autre position de l'auditeur dans la pièce et/ou d'orientation de l'auditeur [10]. Bien que le temps de mélange dépende de l'espace considéré, des valeurs de l'ordre de 80 ms pour les grandes salles et de 50 ms pour les petites sont couramment utilisées [3].

1.1.2. Caractérisation fréquentielle

D'un point de vue fréquentiel, les ondes sonores sont modifiées différemment selon la fréquence par les matériaux qui constituent l'espace et en raison de la géométrie de la salle des résonances apparaissent à certaines fréquences. On retrouve une segmentation similaire au comportement temporel avec une région fréquentielle basse fréquence où des résonances sont perceptibles individuellement et une région haute fréquence où ce sont des caractérisations statistiques des résonances qui sont représentatives de la perception de la réverbération [8].

Selon la géométrie et les matériaux constituant l'espace dans lequel se propagent des ondes sonores, un ensemble de fonctions - nommées modes - forment une base de l'espace des solutions de l'équation d'onde [4]. La propagation du son dans une salle est déterminée par ces modes propres de la salle. Il s'agit d'ondes stationnaires tridimensionnelles qui peuvent être excitées aux fréquences propres - ou fréquence de résonance - caractéristiques de la salle. Chaque mode propre est associé à une répartition spatiale différente de la pression sonore qui implique parfois d'importantes fluctuations de l'intensité sonore. À basse fréquence, les modes sont isolés et différentiables mais pour une gamme de fréquence plus élevée il n'est pas simple de les distinguer. La densité moyenne des fréquences de résonance δ_f, c'est-à-dire le nombre de fréquences de résonance N_f par Hz à une fréquence f, peut être approchée par la formule suivante pour une salle de forme arbitraire [11] :

$$\delta_f = \frac{\mathrm{d}N_f}{\mathrm{d}f} \approx 4\pi \frac{Vf^2}{c^3} \tag{1.3}$$

où c est la vitesse du son, et V est le volume de l'espace. De manière similaire à la densité d'écho, le densité des fréquences de résonance est proportionnelle à la fréquence au carré et au volume de la salle. Cette approximation est applicable lorsque la longueur d'onde est petite par rapport aux dimensions de la salle et est d'autant plus vrai que la fréquence augmente.

On définit la largeur de bande d'un mode comme l'ensemble des points situés à -3 dB de part et d'autre du «pic» d'énergie situé à la fréquence de résonance. Si la largeur de bande des modes est beaucoup plus grande que l'espacement entre les modes, leur chevauchement ne permettra pas de les distinguer individuellement. Le recouvrement devient trop élevé et conduit à adopter une approche statistique. La fréquence de Schrœder est la valeur au-dessus de laquelle ces modes ne sont plus distinguables individuellement [12]. Elle est déterminée en considérant que des modes ne sont pas identifiables individuellement si l'espacement moyen entre deux modes consécutifs est inférieur à un tiers de la largeur de bande de chaque mode. Cette fréquence limite f_{sch}, calculée d'après la densité moyenne des fréquence de résonance de l'équation (1.3), est donnée par :

$$f_{\text{sch}} \approx 2000\sqrt{\frac{T}{V}} \tag{1.4}$$

où T correspond au temps de réverbération, c'est-à-dire au temps nécessaire à une décroissance de l'intensité sonore de 60 dB après l'extinction d'une source sonore continue. De manière générale, la fréquence de Schrœder est inférieure à 50 Hz dans les salles de grand volume (ex : T = 1.5 s et V = 2000 m^3). Le seuil est si faible que les fréquences fondamentales de la plupart des sources sonores usuelles sont bien plus élevées. Ce n'est que dans les petites salles qu'une partie de la gamme de fréquences usuelles se situe en dessous de f_{sch}. Ce seuil peut par exemple s'élever à 300 Hz (ex : T = 0.7 s et V = 37 m^3).

Ainsi, dans une région particulière du domaine temps-fréquence, délimité par la fréquence de Schrœder et le temps de mélange, la densité fréquentielle et temporelle sont assez élevées pour que la réverbération puisse être décrite de façon stochastique.

1.2. Mesures perceptives de l'acoustique d'une salle

La caractérisation d'un son est liée à la mesure de différents aspects, qui peut être effectuée selon des perspectives très différentes. Un évènement acoustique peut être analysé en employant des capteurs physiques (un ou plusieurs microphones) afin de quantifier une ou plusieurs de ses grandeurs physiques telles que le niveau sonore, le contenu fréquentiel ou encore le temps de réverbération. Il peut également être caractérisé en employant des capteurs sensoriels (les oreilles d'un auditeur) afin de quantifier les sensations auditives que celui-ci éprouve lorsqu'il est confronté à l'événement acoustique. On parle alors d'évaluation perceptive [13].

L'évaluation perceptive peut être d'ordre perceptif ou affectif. Dans le premier cas, le sujet effectue une quantification sensorielle de propriétés auditives du stimulus perçu sans que n'interviennent ses émotions, goûts ou opinions. L'objectif principal est de donner des informations sur la caractéristique du son tel qu'il est perçu par le système auditif. Les caractéristiques du stimulus perçu sont évaluées en termes objectifs. Cette approche postule que la perception de certaines caractéristiques sonores est commune à tous les auditeurs si bien que leur évaluation mène à des quantifications similaires pour un grand nombre de sujets. Il est important que ces caractéristiques soient bien définies et que les sujets aient la même compréhension de leur signification.

Dans le cas de jugements affectifs, l'évaluation perceptive est associée à une opinion personnelle telle que des jugements de préférence ou de gêne. Les données collectées d'un sujet à l'autre peuvent donc être très variables et ce sont les tendances observées au sein d'une population qui sont généralement étudiées. Ce type d'évaluation ne sera pas abordé dans ce document.

1.2.1. Les méthodes d'évaluation perceptive

Une méthode d'analyse sensorielle peut être définie comme «un moyen d'évoquer une réponse de la part des sujets, qui peut être mesurée de telle sorte que les données puissent être analysées et interprétées» [14]. Elle désigne donc une manière de présenter les stimuli à des sujets, la collecte de données et l'interprétation des résultats d'analyse en vue de répondre au questionnement des expérimentateurs. Un grand nombre de méthodes d'évaluation perceptive existe et elles peuvent être différenciées en terme de précision, robustesse, de facilité de mise en œuvre ou d'adéquation avec des situations réelles. Elles peuvent avoir pour but de déterminer ou quantifier un ou plusieurs attributs perceptifs ou dimensions perceptives. Un attribut perceptif désigne une caractéristique perceptible [15]. Il est généralement étroitement associé à un terme caractérisant cette perception. La combinaison d'un terme et d'une définition permet la bonne compréhension et communication de l'attribut.

La notion de dimensions perceptives est également communément utilisées dans les méthodes d'évaluation perceptive. Elle se réfère à la structure perceptive sous-jacente au jugement d'un ensemble de stimuli. En particulier, la dimension désigne un axe résultant d'une analyse statistique multivariée. Ainsi, une dimension peut résulter d'une combinaison de plusieurs attributs selon différentes pondérations et non être uniquement corrélée à un attribut [13].

Parmi les méthodes d'évaluation perceptive se trouvent différentes catégories : les méthodes discriminantes, intégratives et descriptives. Les méthodes discriminantes sont principalement employées pour déterminer s'il existe une différence perceptible entre plusieurs stimuli. Les jugements effectués par les sujets sont soit binaires, soit permettent d'établir une relation d'ordre entre les stimuli sans toutefois évaluer de manière quantitative la distance qui les sépare. Les appréciations des sujets peuvent concerner des attributs perceptifs, des jugements de qualité ou de préférence. Parmi ces méthodes de discrimination se trouvent le test ABX [16], le test duo-trio [17], la comparaison par paire [18] ou encore le test tetrad [19].

Les méthodes intégratives désignent une famille de méthodes dans laquelle les su-

jets jugent la perception d'un ensemble de stimuli selon un unique critère telle que la qualité, la dégradation ou la préférence. Elles emploient généralement une échelle hédonique à 9 points) [20]. Ainsi, une seule variable est jugée par les sujets dans l'évaluation en utilisant une seule échelle de notation. Ces méthodes sont très répandues et leur usage est soumis à des normes bien établies. En particulier, deux standards ITU sont les plus couramment utilisées : la recommandation BS.1534-3 [21] et la recommandation BS.1116-3 [22]. La première, également appelé MUSHRA (test *MUlti-Stimulus with Hidden Reference and Anchor*), est utilisée pour évaluer principalement la dégradation en rapport à une référence. Elle est appliquée aux cas où des artefacts sont présents et clairement audibles. La seconde est une méthode plus précise et employée lors de l'évaluation de situations où les artefacts sont peu audibles. Ces deux méthodes seront abordées plus amplement dans le chapitre 5, 6 et 8.

Les méthodes descriptives visent à identifier et quantifier des attributs ou dimensions perceptives liés à un ensemble de stimuli avec un groupe de sujets entrainés. Dans cette optique, il peut être demandé aux sujets de verbaliser les sensations perçues au moyen d'un vocabulaire individuel ou commun à tous les sujets. Lorsqu'un vocabulaire individuel est employé, le sujet décrit avec ses propres termes les sensations perçues. On parle alors d'*Individual Vocabulary Methods*. Il est possible que les sujets perçoivent les stimuli de la même manière tout en utilisant des termes différents pour exprimer leurs sensations. C'est alors à l'expérimentateur d'effectuer l'analyse linguistique permettant des rapprochements entre les différentes terminologies employées dans le but de déterminer les attributs perceptifs inclus dans l'ensemble de test. On trouve parmi ces méthodes le *Free-Choice Profiling* [23], *le Flash Profile* [24], l'*Individual Vocabulary Profiling* [25] et la *Repertory Grid Technique* [26, 27].

Une des manières de minimiser les incertitudes du langage est de faire en sorte que les sujets s'accordent préalablement sur un vocabulaire précis à employer. On parle alors de *Consensus Vocabulary Methods*. Un ensemble de sujets est utilisé pour définir ou sélectionner un ensemble de termes communs qui décrivent les attributs perceptifs des stimuli considérés. De plus, un ou plusieurs stimuli sonores peuvent être utilisés en guise de référence pour illustrer les termes présents dans le vocabulaire commun. Les sujets utilisent ensuite ce vocabulaire pour juger les stimuli. On compte parmi ces méthodes la *Flavor Profile Method* [28], la *Quantitative Descriptive Analysis* [29] ou la *Semantic Differential Method* [30]. Elles sont considérées comme plus fastidieuses à mettre en œuvre par rapport aux méthodes à vocabulaire individuel, car nécessitent un effort supplémentaire pour obtenir un vocabulaire faisant consensus.

Des approches dites indirectes peuvent également être utilisées dans les méthodes descriptives. Contrairement aux méthodes directes, les méthodes indirectes ne demandent pas aux sujets de juger des attributs spécifiques mais déterminent un espace perceptif multidimensionnel latent directement à partir d'une analyse multivariée. Avec ces méthodes, il n'est donc pas nécessaire de présupposer du nombre ni de la nature des attributs perceptifs en jeu dans le jugement des stimuli. Il est en effet possible d'analyser la perception d'un ensemble de stimuli en dégageant une ou plusieurs dimensions perceptives sous-jacentes qui expliquent leurs disparités. Les dimensions perceptives peuvent être révélées par l'analyse des différences perçues entre les stimuli. Parmi ces méthodes se trouvent l'analyse multidimensionnelle ou *Multidimensional*

Scaling (MDS) [31], *le Free Sorting* [32], *le Projective Mapping* [33], la *Perceptual Structure Analysis* [34, 35].

Ces méthodes semblent intéressantes car n'introduisent pas de biais potentiels associés à l'utilisation d'attributs perceptifs. Les évaluateurs comparent ou trient les stimuli selon ce qu'ils perçoivent. Dans le cas de l'analyse multidimensionnelle, des paires de stimuli sont présentées aux sujets auxquels il est seulement demandé de juger du degré de similarité ou de dissemblance. L'analyse des matrices de dissemblance résultantes fournit les coordonnées de ces stimuli sur chacune des dimensions perceptives supposées sous-tendre leur perception [36]. Cependant, l'un des défis de la MDS consiste à trouver le bon nombre de dimensions perceptives, c'est-à-dire les dimensions qui méritent d'être interprétées. Il incombe au chercheur de trouver la nature de ces dimensions car aucune information additionnelle n'est recueillie auprès des sujets. Cette tâche peut s'avérer complexe et l'interprétation des dimensions varier d'un expérimentateur à l'autre.

Les sujets peuvent différer dans leur manière de juger la dissemblance entre stimuli. Afin de prendre en considération ces différences individuelles, Carrol *et al.* ont élaboré l'*Individual Difference Scaling* (INDSCAL) [37]. Cette méthode est fondée sur l'hypothèse que tous les sujets utilisent les mêmes dimensions perceptives pour juger les stimuli mais qu'ils y appliquent des pondérations différentes. Ainsi l'analyse INDSCAL trouve un groupe de dimensions perceptives communes à tous les sujets, auxquelles sont associées à la fois les coordonnées de chaque stimulus et de chaque sujet. Cette méthode sera abordée plus amplement dans le chapitre 2 et le chapitre 3.

1.2.2. Application à la perception de l'acoustique d'une salle

Ces méthodes d'évaluation perceptive sont des outils précieux pour étudier la perception sonore et plus précisément ici la perception d'un espace sonore. Elles offrent un cadre rigoureux permettant de définir et d'évaluer des attributs perceptifs ou de révéler des dimensions perceptives. Néanmoins, les nombreuses contributions dans ce domaine montrent à plusieurs égards que l'acoustique des salles est un sujet d'étude complexe.

D'abord, un espace sonore ne peut pas être perçu en tant que tel mais seulement comme un milieu de propagation modifiant les propriétés du contenu auditif présenté. Des sources sonores sont nécessaires pour exciter l'espace sonore et le rendre perceptible. Cette excitation produit une réponse complexe de la salle et nous percevons en partie l'espace par la modification du spectre, de l'intensité et de la séquence temporelle de la source sonore. Les détails acoustiques d'un espace sont rarement excités par un grand nombre de sources différentes dans une large gamme de fréquences, d'amplitude ou de positions et ne sont donc pas toujours apparents. Les propriétés perçues d'un espace dépendent donc fortement des propriétés de la source sonore impliquée [38].

Ensuite, la perception a une nature multi-modale. L'interaction image-son est par exemple un mécanisme complexe dont la séparation et le traitement indépendant est commode pour l'analyse mais semble peu réaliste [39]. Plusieurs études ont notamment examiné l'influence de la vision sur l'évaluation de la localisation des sources so-

nores [40–42], la perception de la distance auditive [43–47], l'externalisation [48–52] ou les impressions spatiales [53–55]. Beaucoup de méthodes d'évaluation perceptives se sont appliquées à l'évaluation d'attributs perceptif dans des expériences seulement sonores. Pourtant, il est assez rare d'expérimenter un espace sans apprécier son aspect visuel.

De plus, les propriétés spatiales perçues d'un espace sonore dépendent de l'expérience et des préférences de l'auditeur auquel on s'adresse. La capacité à percevoir des attributs spatiaux dans un stimulus sonore peut varier d'un groupe de sujets à l'autre et il est possible que des différences majeures apparaissent dans leur rapport au phénomène acoustique. Les auditeurs expérimentés acquièrent une capacité à distinguer l'acoustique en tant qu'objet auditif distinct en discriminant dans le champ sonore les caractéristiques propres à la source et à la salle. Il n'est pas évident qu'une cohérence perceptive puisse être dégagée pour tout groupe d'auditeurs : les musiciens, acousticiens, architectes, ingénieurs du son, ou auditeurs inexpérimentés envisagent l'acoustique sous un angle particulier, sans qu'aucun de ces points de vue ne caractérise à lui seul l'acoustique d'une salle. Une sélection aléatoire d'individus ne pourrait produire des résultats représentatifs pour toutes ces catégories d'auditeurs.

A cela s'ajoute le fait que retranscrire une expérience sensorielle par la parole n'est pas un exercice simple pour un sujet. L'analyse du langage pour décrire des sensation est une entreprise complexe pour un expérimentateur : que signifie réellement un mot utilisé dans un contexte spécifique par une personne particulière ? Quel est le contraire de «doux» ? Est-ce «dur», «fort» ou «rugueux» ? Peut-on grouper des termes synonymes qui n'ont pas les mêmes antonymes ? A cette difficulté s'ajoute le fait que les adjectifs pour décrire des impressions sonores peuvent être plus ou moins développés selon la langue utilisée. En général, l'audition n'a pas de vocabulaire propre aussi élaboré que celui de la vision, le toucher ou le goût : on dit d'une mélodie qu'elle est sombre, d'un haut-parleur qu'il est transparent, d'une ambiance qu'elle est feutrée, d'un son qu'il est brillant ou chaud. Pour Blesser [38], ce constat est d'autant plus vrai pour la caractérisation de sensations sonores liées à l'espace et ce manque linguistique peut être expliqué par plusieurs facteurs. D'une part, en étant fondamentalement orienté vers la communication visuelle, notre culture accorde peu d'importance aux sensations auditives et donc à la conscience auditive de l'espace. D'autre part, notre aptitude à percevoir une acoustique de salle dépend de notre mémoire auditive qui est parfois peu fiable. Pour comparer l'acoustique de deux pièces il est généralement nécessaire de se souvenir d'un espace pendant plusieurs minutes ou jours, voire davantage. Il faut compter la plupart du temps sur notre mémoire à long-terme qui, sans pratique ou entrainement, est moins fiable que la mémoire à court-terme. Bien entendu, des simulateurs de réverbération peuvent être employés pour permettre la comparaison directe entre deux espaces différents mais seuls certains professionnels y ont accès et la fidélité de la représentation n'est pas toujours satisfaisante. Pour cette raison, peu d'individus accumulent des expériences sonores spatiales et sont en mesure de les transmettre.

Enfin, les études en laboratoires permettent une évaluation détaillée de l'acoustique d'une salle mais les conditions restent superficielles par rapport au cadre réel d'écoute où l'expérience comporte de nombreux aspects difficiles à reproduire. Néan-

moins, ces études ont fourni des informations précieuses sur la perception de l'acoustique et ont vérifié et étendu des observations faites dans des conditions réelles. De plus, les récents progrès technologiques en acoustique virtuelle permettent de réaliser des expériences en laboratoire avec un plus grand contrôle expérimental.

1.2.3. Les lexiques d'attributs perceptifs

De nombreuses méthodes d'évaluation perceptive ont été employées directement dans des salles de concert ou dans des laboratoires pour élaborer des ensembles d'attributs perceptifs permettant l'évaluation de l'acoustique de salles de concert. Des démarches similaires ont également été entreprises pour l'évaluation d'environnements acoustiques virtuels ou de technologies de spatialisation. Zacharov recense par exemple plusieurs études ayant élaboré ou utilisé des attributs perceptifs dans ces domaines de recherche [56]. Ces études sont répertoriées dans le Tableau 1.1 qui permet d'exposer la variabilité du nombre d'attributs employés d'une étude à l'autre. Les attributs employés furent soit définis *ad hoc* par les expérimentateurs eux-mêmes soit le résultat d'analyses linguistiques appliquées aux verbalisations d'un groupe de sujets. A l'évidence, bien que ces recherches aient recensé un grand nombre d'attributs perceptifs depuis plusieurs décennies, il ne semble pas qu'un ensemble fixe d'attributs perceptifs se soient imposés comme attributs de référence. Aucun véritable consensus ne s'est donc dégagé concernant les attributs à utiliser pour décrire la perception de l'espace sonore. Il est vrai que la signification et la validité des attributs définis d'une étude à l'autre peuvent être remises en cause dans la mesure où 1) certaines définitions sont trop peu précises, 2) des définitions différentes peuvent désigner le même attribut 3) des attributs différents peuvent faire référence à la même définition.

Néanmoins, certains attributs sont récurrents et semblent assez similaires dans leurs définitions [75]. Plutôt que d'utiliser un ensemble d'attributs différents à chaque étude, plusieurs travaux ont depuis entrepris de créer un lexique permettant de décrire les caractéristiques spatiales d'un environnement sonore. Fortes d'une littérature abondante sur le sujet, ces études se nourrissent des efforts linguistiques et des résultats issus de méthodes d'évaluation perceptive variées employées dans des laboratoires ou dans des salles de concert. De tels lexiques sont des instruments d'analyse utiles à l'étude comparée de différentes expérimentations et peuvent servir de base à de futures recherches. Pour élaborer un lexique il est nécessaire de satisfaire autant que possible les facteurs suivants :

— les attributs et leur définition doivent faire consensus ;
— les définitions ne doivent pas être sujettes à ambiguïté ;
— les attributs doivent permettre de donner des évaluations reproductibles ;
— les attributs doivent permettre de discriminer efficacement des espaces sonores ;
— les attributs doivent êtres faiblement corrélés.

Parmi les propositions récentes de lexiques d'attributs perceptifs, se trouvent :

— les travaux de Zacharov *et al.* [76] ayant mené à l'élaboration du lexique nommé *The Sound Wheel* destiné à des champs d'application variés de l'audio spatialisé ;

Source	Référence	Année	Nombre d'attributs
Wilkens	[57]	1975	19
Gabrielsson and Sjögren	[58]	1979	8
Toole	[59]	1985	15
Lavandier	[60]	1989	14
Letowski	[61]	1989	12
Kahle	[62]	1995	7
Mason and Rumsey	[63]	2000	4
Zacharov and Koivuniemi	[64]	2001	12
Rumsey	[65]	2002	20
Berg and Rumsey	[66]	2003	13
Guastavino and Katz	[67]	2004	9
Lorho	[25]	2005	16
Berg and Rumsey	[27]	2006	14
Choisel and Wickelmaier	[34]	2006	8
Sazdov et al	[68]	2007	6
Silzle	[69]	2007	7
Wittek	[70]	2007	18
Pedersen	[71]	2008	∼450
Lokki et al	[72]	2012	60
Le Bagousse et al	[73]	2014	28
Lindau et al	[74]	2014	48
Kaplanis et al	[75]	2014	74

Tableau 1.1 – Liste non exhaustive du nombre d'attributs définis dans diverses études du domaine de l'acoustique des salles de concert, de l'évaluation d'environnement sonore virtuel ou de technologie de spatialisation sonore. D'après Zacharov [56].

— le *Spatial Audio Quality Inventory* (SAQI) proposé par Lindau *et al.* [74] destiné à l'évaluation de technologies de spatialisation sonore utilisées pour la reproduction ou la synthèse d'environnement acoustique ;
— le *Room Acoustical Quality Inventory* (RAQI) proposé par Weinzierl *et al.* [77] destiné à l'évaluation d'acoustique de salles de concert.

Au cours de ces travaux, la méthode du *Focus Group* [78] fût employée. Cette procédure consiste à mettre en œuvre des discussions en groupe supervisées par un ou plusieurs modérateurs. Ceux-ci doivent s'assurer que la parole circule équitablement et restent attentifs à la communication non-verbale. Ces discussions ont permis l'élaboration de vocabulaires faisant consensus grâce à l'expérience pratique et théorique de groupes d'experts.

The Sound Wheel

Une *Sound Wheel* ou roue d'attributs est une représentation visuelle hiérarchique d'un lexique d'attributs perceptifs. Les attributs présentant des similitudes sont placés dans la même catégorie et les catégories similaires sont regroupées. Zacharov *et al.* [76] ont élaboré ce lexique pour favoriser l'évaluation perceptive dans des champs d'application variés de l'audio spatialisé. Les auteurs ont pu extraire dans les 22 études du tableau 1.1, 401 attributs perceptifs ainsi que leur définition. Un nombre conséquent d'entre eux ayant des définitions similaires, une classification sémantique a permis de réduire ce nombre de termes à 50 groupes d'attributs [76]. Dans un second temps, un panel de douze experts entreprirent de passer en revue les groupes ainsi constitués afin de les recombiner puis de les décrire précisément. Cette procédure s'apparente à la méthode de *Focus Group* sans qu'elle soit explicitement mentionnée

dans l'étude. Suite à ces discussions, douze attributs perceptifs en lien avec la perception de l'espace ont été définis. Ils sont reportés dans le tableau 1.2, leur définitions sont disponibles dans la référence [76].

Catégorie	Terme	Sous-terme
Spatial Extent	Depth	
	Width	
	Envelopment	Horizontal Envelopment
		Vertical Envelopment
Localization	Distance	
	Internality	
	Localisability	
Spatial / Timbral	Clarity	
Environment	Reverberance	Level of Reverberance
		Duration of Reverberance

Tableau 1.2 – Attributs perceptifs en lien avec la perception de l'espace sonore sélectionnés dans l'étude de Zacharov *et al.* [76]

Le SAQI

Le SAQI est un lexique comprenant 48 termes décrivant les qualités perceptives permettant de comparer des espaces sonores virtuels entre eux ou à des références. Ce lexique fut établi par 21 experts germanophones de l'acoustique virtuelle. Ces experts ont pris part aux discussions selon un *two-way Focus Group* (les participants étaient séparés en deux groupes : un groupe de discussion avec deux modérateurs et un groupe d'observation). Cinq autres experts n'ayant pas pris part aux discussions ont ensuite passé en revue le vocabulaire établi afin de lever toute ambiguïté sur les termes et leur définition. La traduction en anglais a ensuite été assurée par huit experts bilingues.

Les 48 termes sont classés selon huit catégories : *Timbre*, *Tonalness*, *Geometry*, *Room*, *Time Behavior*, *Dynamics*, *Artifacts* et *General*. L'usage de ce lexique est destiné à l'évaluation perceptive de technologie de spatialisation sonore pour la synthèse d'environnement acoustique. Certains termes désignent des propriétés inhérentes à un dispositif de reproduction et non des attributs perceptifs lié à la perception de l'acoustique (par exemple : *Pitched, Impulsive, Noise-Like artifacts*, *Responsiveness*, *Spatial disintegration*, *etc*...). Seuls des termes applicables à la caractérisation perceptive d'un espace et non au moteur de reproduction sont donc reportés dans le tableau 1.3. Un protocole de test basé sur ce lexique a été élaboré et incorporé à une librairie MATLAB de test perceptif nommée WhisPER [79].

Catégorie	Terme
General	Clarity
	Speech Intelligibility
	Naturalness
	Presence
Room	Level of reverberation
	Duration of reverberation
	Envelopment (by reverberation)
Geometry	Distance
	Depth
	Width
	Height
	Localizability
Timbre	Tone color bright-dark
	High / Mid / Low-frequency tone color
	Sharpness
	Roughness
	Comb filter coloration
	Metalic tone color

Tableau 1.3 – Attributs perceptifs en lien avec la perception de l'espace sonore sélectionnés parmi les termes du SAQI [74]

Le RAQI

Le RAQI est un outil destiné à l'évaluation de la qualité acoustique des salles de spectacles pour la musique et la parole. Il a été développé en rassemblant dans un premier temps des connaissances de 12 experts germanophones spécialisés en acoustique des salles ou en psychoacoustique, issus du monde universitaire ou du conseil en acoustique. Le groupe de discussion a fourni une terminologie complète sous la forme d'une liste de 50 éléments permettant de décrire les aspects pertinents d'environnements tels que des salles de répétition, de conférence, de musique de chambre, de concert symphonique ou des grandes cathédrales. Les résultats issus du groupe de discussion ont ensuite servi de base à un test d'écoute pour évaluer un ensemble de stimuli. Ce test a permis d'étudier les dépendances statistiques des termes choisis par les experts et ainsi d'envisager la réduction du vocabulaire. Le test perceptif a consisté en l'évaluation de 35 salles dont les acoustiques ont été simulées grâce à des modélisations en trois dimensions et au moteur de calcul RAVEN [80]. La plupart de ces modèles correspondait à des salles existantes. Des modèles sans équivalent réel ont été créés afin que tous les attributs prédéfinis puissent être perçus parmi les stimuli du test. Pour chacune des salles, deux positions d'écoutes ont été considérées ainsi que

trois contenus sonores enregistrées dans une chambre anéchoïque : un orchestre symphonique, une trompette solo et un discours dramatique (l'orchestre symphonique n'a cependant été utilisé que pour 25 salles pour des raisons de taille). Le test a été effectué en laboratoire en utilisant une synthèse binaurale dynamique. 190 sujets ont participé à l'évaluation. Chacun d'entre eux a noté 14 stimuli selon les termes élaborés par le groupe de discussion. Le test fut répété avec 88 participants, soit 46% des sujets initiaux, afin de tester la fiabilité de l'évaluation des attributs au cours du temps.

L'analyse factorielle des données du questionnaire a abouti à l'élaboration d'un lexique pouvant comprendre jusqu'à neuf dimensions perceptives. Les termes correspondants ont été sélectionnés sur la base d'un test de fiabilité inter-sujet, intra-sujet et d'indépendance statistique. L'ensemble de ces attributs est rapporté dans le tableau 1.4.

D'un point de vue statistique, la version du RAQI comprenant les 6 premières dimensions est celle garantissant une indépendance suffisante des différents attributs. Sans pouvoir permettre d'identifier une acoustique de manière unique, cette version permet de caractériser une acoustique de manière suffisamment complète pour être en mesure de discriminer les 35 salles du corpus utilisé.

La dimension liée à la qualité perçue peut être considérée comme un aspect de second ordre qui résulte des autres dimensions perceptives. Les jugements de qualité se sont avérés positivement corrélés aux jugement de brillance et de force sonore et négativement corrélés aux jugement liés à la coloration et à l'irrégularité de la décroissance sonore. Cependant ces dimensions n'expliquaient que la moitié de la variance des jugements de qualité. D'autres attributs non mesurés semblent donc également en jeu dans le jugement de l'acoustique d'une salle.

L'étude de la validité statistique du vocabulaire élaboré par les experts est une plus-value par rapport aux lexiques mentionnés précédemment dans la mesure où elle assure qu'avec un nombre restreint de termes on puisse obtenir une caractérisation relativement complète d'un vaste corpus de salles et cohérente d'un sujet à l'autre et au cours du temps. Une telle étude appliquée aux petites salles permettrait d'établir un lexique spécifique à ce type d'espace malheureusement trop souvent écarté des études perceptives en acoustique.

La base de données ayant servi à l'évaluation perceptive a été publiée avec la modélisation de tous les environnements acoustiques, les réponses impulsionnelles de salle et les évaluations perceptives associées [81]. Pour chaque salle, il est possible de mettre en parallèle les stimuli sonores avec les notes moyennes obtenues selon les attributs établis par les experts ou selon les 6 premières dimensions du RAQI. Ainsi, en plus des termes et définitions inclus dans celui-ci ces exemple sonores permettent de préciser la signification des attributs par l'écoute. Par ailleurs, cette base de données peut servir de socle au développement ou à la redéfinition de mesures physiques pour la prédiction des attributs perceptifs du lexique.

Dimension	Terme	Pôles
Quality	Liking	I like it - I don't like it
	Room acoustic suitability	suitable - not suitable
	Ease of listening	difficult - effortless
	Global balance	balanced - unbalanced
Strength	Size	small - large
	Loudness	soft -loud
	Width	small - large
Reverberance	Duration of reverberance	short - long
	Reverberance	dry - reverberant
	Strength of reverberation	weak - strong
	Envelopment by reverberation	weak - strong
Brillance	Brillance	not brillant - very brillant
	Tone color bright / dark	bright - dark
	Treble range characteristic	attenuated - emphasized
Irregular Decay	Flutter echo	none - very strong
	Echo	none - very strong
	Irregularity in sound decay	none - very strong
Coloration	Bouminess	not boomy - very boomy
	Roughness	not rough - very rough
	Comb filter coloration	none - very strong
Clarity	Temporal clarity	clear - blurred
	Spatial transparency	blurred - transparent
	Precision of localization	precise - diffuse
Liveliness	Liveliness	dead - lively
	Spatial precense	low - high
	Dynamic range	small - large
Intimacy	Intimacy	remote - intimate
	Distance	close - distant
	Warmth	cool - warm
Single Items	Metalic tone color	not metallic - very metallic
	Openness	open - constricted
	Attack	soft - crisp
	Richness of sound	low - high

Tableau 1.4 – Les termes du RAQI élaborés suite à l'analyse statistiques de l'évaluation perceptive de salles de spectacle.

1.3. Mesures physiques de l'acoustique d'une salle

Bien que le temps de réverbération soit une mesure incontournable pour décrire une acoustique, il est loin de pouvoir la caractériser entièrement et ne donne qu'une

indication des propriétés acoustiques. Diverses études perceptives ont montré que plusieurs quantités extraites directement d'une réponse impulsionnelle de salle peuvent avoir des corrélations avec des attributs perceptifs. Nous désignerons ces quantités par «paramètres acoustiques». Des rapports d'énergies, pressions acoustiques ou des indices de corrélations binaurales fournissent une description plus complète de l'acoustique d'une salle. Lacatis *et al.* dénombrent dans un bref historique 41 paramètres différents définis depuis les années 50 [82] et quelques uns d'entre eux figurent dans la norme ISO 3382-1 [3] qui vise à standardiser la mesure des paramètres acoustiques d'une salle.

En raison de la variabilité du milieu de propagation, il est difficile d'associer une unique valeur de paramètre à un espace sonore. Le milieu peut faire l'objet de turbulence en raison de tournoiements de masses d'air, de réfraction en raison du déplacement de couches thermiques ou de modification de l'absorption en raison de la présence d'objets ou d'auditeurs dans la salle. Un grand volume implique de longs trajets de propagation des ondes sonores au cours desquelles toutes ces perturbations dynamiques peuvent apparaitre. Deux réponses impulsionnelles mesurées dans un espace sonore sont alors toujours différentes ; sans pour autant que ces différences soient toujours perceptibles. De plus, chaque élément du dispositif de mesure est susceptible de faire varier les résultats d'une mesure à l'autre et pose ainsi la question de la précision et de la reproductibilité. Les effets directionnels de la source peuvent notamment avoir un impact important sur les variations entre deux positions de mesure [83]. Ainsi, la norme ISO 3382-1 [3] décrit plusieurs spécifications en lien avec la procédure de mesure concernant la position et la directivité des sources et des microphones ; la présence d'un auditoire et de musiciens ; le nombre de mesures minimales à effectuer selon le volume de la salle. Ces spécifications permettent d'obtenir des résultats comparables et reproductibles.

Type	Paramètre	Moyenne	*JND*	Valeurs typiques
Durées de décroissance	EDT , T_{30}, T_{20} (s)	500 - 1000 Hz	5%	[1 ; 3]
Mesures de Clarté	D_{50}	500 - 1000 Hz	0.05	[0.3 ; 0.7]
	C_{80}, C_{50} (dB)	500 - 1000 Hz	1	[-5 ; 5]
	T_s (ms)	500 - 1000 Hz	10	[60 ; 260]
Force Sonore	G (dB)	500 - 1000 Hz	1	[-2 ; 10]
Impressions spatiales	J_{LF}, J_{LFC}	125 - 1000 Hz	0.05	[0.05 ; 0.35]
	L_J (dB)	125 - 1000 Hz	NC	[-14 ; 1]
	IACC	NC	0.075	NC

Tableau 1.5 – Résumé des paramètres acoustiques présents dans la norme ISO 3382-1, des bandes d'octave d'après lesquels obtenir une valeur moyenne, des plus petites différences perceptibles et des valeurs typiques rencontrées pour des salles de spectacles. La moyenne sur les bandes d'octave signifie la moyenne arithmétique des valeurs dans les bandes d'octave excepté pour L_J dont le calcul de la moyenne est obtenu en moyennant préalablement les énergies par bande d'octave.

Le tableau 1.5 liste les paramètres acoustiques spécifiés dans la norme qui furent

sélectionnés en raison de leur importance subjective et de leurs facilité de calcul car pouvant être directement obtenus d'après la réponse impulsionnelle. Ces paramètres acoustiques peuvent être classés selon quatre types de mesures : les mesures de durée de décroissance, de force sonore, de clarté et d'impressions spatiales.

1.3.1. Les durées de décroissance

Les durées de décroissance correspondent aux temps exprimés en seconde nécessaires au niveau de pression sonore pour décroitre de 60 dB [1]. Les temps de réverbération sont liés aux propriétés physiques de la salle telles que le volume et les surfaces qui la constituent. La forme de la pièce, les matériaux des murs, l'ameublement influencent les valeurs de temps de réverbération. Deux types de durée de décroissance sont définis : le temps de réverbération (T_{30}, T_{20} ou RT) et la durée de décroissance initiale notée EDT (pour *Early Decay Time*).

Ces paramètres sont calculés d'après la décroissance du niveau de la pression sonore en fonction du temps, après qu'une source sonore s'est arrêtée d'émettre un signal continu. Cette décroissance peut être soit mesurée après l'arrêt soudain de la source ou estimée d'après la méthode de la réponse impulsionnelle intégrée. Cette méthode nécessite le calcul de la courbe de Schrœder, c'est à dire l'intégration de la réponse impulsionnelle de salle élevée au carré et retournée dans le temps [84]. Notons E cette courbe :

$$E(t) = \int_{\infty}^{t} p^2(\tau)\mathrm{d}(-\tau) \tag{1.5}$$

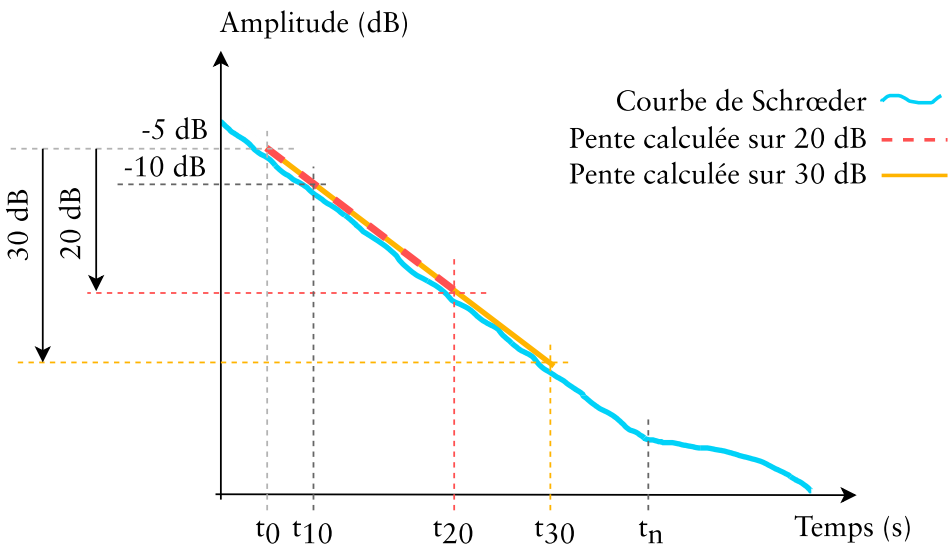


FIGURE 1.2 – Courbe de Schrœder, plages temporelles, plages dynamiques et pentes de décroissances calculée sur 20 et 30 dB de dynamique.

L'EDT correspond à la durée de décroissance calculé d'après une régression linéaire effectuée sur les 10 premiers décibels de décroissance. Le temps de réverbération est calculé grâce à une régression linéaire effectuée sur la courbe de décroissance

entre les niveaux 5 dB et 35 dB inférieur au niveau initial. Lorsque le rapport de signal sur bruit ne le permet pas, cette régression linéaire peut être effectuée sur une dynamique de 20 dB. Les temps de réverbération calculés sur des pentes utilisant 20 et 30 dB de dynamique sont nommés T_{20} et T_{30} respectivement.

La figure 1.2 illustre les différentes plages temporelles et les dynamiques d'après lesquelles les durées de décroissances sont calculées. Une fois le coefficient directeur a de la régression linéaire calculé, la durée de décroissance correspondante T est obtenue d'après la formule suivante :

$$T = \frac{-60}{a} \tag{1.6}$$

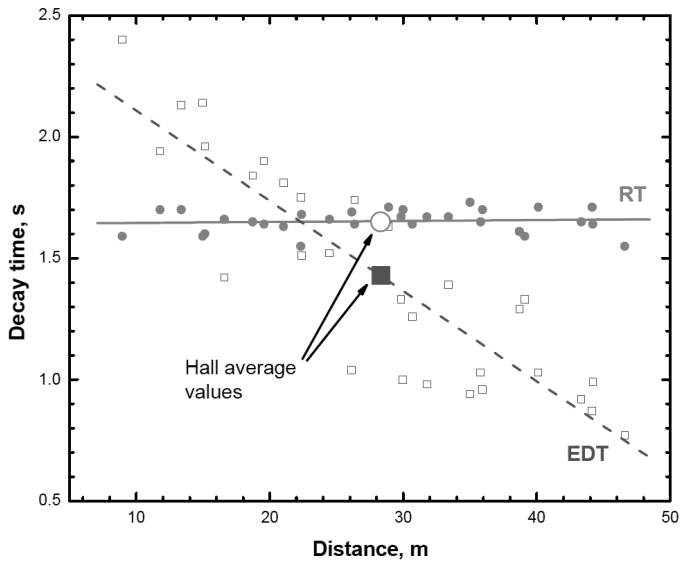


FIGURE 1.3 – Valeurs aux fréquences moyennes de l'EDT et du RT en fonction de la distance entre la source et le microphone dans le Northern Alberta Jubilee Auditorium. D'après Bradley [85].

Les temps de décroissance sont calculés par bande d'octave ou tiers d'octave. Une valeur unique de ces temps est généralement reportée, elle correspond à la moyenne des valeurs calculées par bande d'octave à 500 Hz et 1 kHz ou à la moyenne des 6 valeurs calculées par tiers d'octave de 400 Hz à 1250 Hz.

Bien que les mesures moyennes d'EDT et le temps de réverbération présentent parfois une forte corrélation, la première semble plus sensible à la position de mesure et est plus à même de refléter des particularités de la réverbération d'une salle [85]. Alors qu'une moyenne spatiale du temps de réverbération fournit une information acoustique sur la propriété d'une salle, une unique valeur d'EDT ne peut être obtenue pour décrire une salle particulière. C'est pourquoi l'EDT est jugée comme ayant une importance subjective plus grande que le temps de réverbération. La figure 1.3 présente des valeurs de RT et d'EDT obtenues dans différents positions de mesures d'une salle. On constate une variabilité de l'EDT en fonction de la distance à la source qui n'est pas présente pour le RT. D'après Pelorson [83], lorsque l'on considère des mesures effectuées en différents points d'une salle, les valeurs de corrélation entre les

valeurs de RT et d'EDT peuvent varier de 0.3 à 0.92 selon la salle. La dépendance entre le RT et l'EDT varie donc selon la salle étudiée.

La norme stipule que l'EDT est «subjectivement plus importante et liée à la réverbérance, alors que le temps de réverbération est lié aux propriétés physiques de la salle». La réverbérance désigne l'évaluation subjective du phénomène de réverbération permettant de qualifier un espace de sec ou de réverbérant.

Selon les matériaux qui les constituent, il est possible qu'une grande salle et qu'une salle de taille moyenne aient les mêmes valeurs de durées de décroissance, qui ne sauraient à elles seules caractériser les espaces de manière unique. C'est pourquoi d'autres critères ont été introduits afin de décrire plus finement les caractéristiques de l'acoustique d'une salle.

1.3.2. Les mesures de clarté

Dans un champ sonore réverbéré, l'auditeur est soumis a un grand nombre d'ondes sonores arrivant sous toutes les incidences. Certaines de ces ondes sonores viennent renforcer la perception du signal alors que d'autres nuisent à son intelligibilité.

En utilisant un haut-parleur reproduisant la voix d'un orateur dans une salle, Helmut Haas constata que tant que le décalage entre le signal issu de l'orateur et du haut-parleur reste inférieur à 50 ms, la localisation du son perçue par l'auditoire n'est pas perturbée ; cela même lorsque le niveau du haut-parleur est supérieur de 10 dB à celui de l'orateur [86]. Les réflexions qui se produisent dans un court laps de temps - nommé seuil d'écho - après le son direct ne sont pas perçues comme des événements distincts, mais fusionnent perceptivement avec le son direct. Le seuil d'écho est dépendant de la source sonore employée. Des valeurs oscillant entre 5 ms pour des clics et 50 ms pour de la parole sont présents dans la littérature [87]. La direction d'incidence perçue de l'ensemble des réflexions qui ont fusionné en un unique évènement sonore est déterminée par celle du premier front d'onde (le son direct). On parle alors d'effet de précédence. En raison de leur capacité à fusionner avec le son direct les réflexions précoces peuvent être un atout pour renforcer sa perception et ainsi favoriser l'attribut perceptif nommé clarté.

La mesure de clarté la plus rencontrée dans la littérature correspond au ratio entre les quantités d'énergie sonore portées par les réflexions précoces et par les réflexions tardives. C'est pourquoi plus la valeur de cet indice est élevée meilleure est la clarté. Il est calculé avec une limite temporelle des premières réflexions, t_e, fixée à 50 ou 80 ms selon que le paramètre est destiné à caractériser la clarté de la parole ou de la musique respectivement.

$$\mathrm{C}_{t_e} = 10 \log_{10} \frac{\int_0^{te} p^2(t)\mathrm{d}t}{\int_{t_e}^{\infty} p^2(t)\mathrm{d}t} \tag{1.7}$$

La norme définit également une mesure de la définition, le D_{50}, comme le rapport entre l'énergie précoce et l'énergie totale. Ce paramètre est généralement employé pour décrire la perception de la parole.

$$D_{50} = \frac{\int_0^{0,050s} p^2(t)\mathrm{d}t}{\int_0^{\infty} p^2(t)\mathrm{d}t} \tag{1.8}$$

La norme ISO 3382-3 [88], destinée à la mesure des paramètres acoustiques de bureaux ouverts, contient une mesure plus élaborée de l'intelligibilité de la parole : l'indice de transmission de la parole noté STI (pour *Speech Transmission Index*). Cette mesure est fondée sur la capacité de l'auditeur à détecter des modulations d'amplitude dans le signal de parole [89]. Le STI est fonction du temps de réverbération de l'espace considéré ainsi que le rapport signal sur bruit par bande de fréquence.

Le temps central T_s correspond au centre de gravité de la réponse impulsionnelle au carré mesurée en secondes. Il représente le moment où l'énergie contenue dans la réponse impulsionnelle en amont est égale à l'énergie contenue en aval. Ce paramètre permet de ne pas utiliser de limite temporelle entre les réflexions précoces et tardives.

$$T_s = \frac{\int_0^{\infty} tp^2(t)\mathrm{d}t}{\int_0^{\infty} p^2(t)\mathrm{d}t} \tag{1.9}$$

Selon plusieurs études, les valeurs des mesures de clarté sont assez bien corrélées aux durées de décroissance ce qui suggère qu'elles fournissent également des indications sur la réverbérance [83, 90–93]. Bien qu'elles aient employé un nombre de salles différent, des moyennages spatiaux et fréquentiels différents, ces études rapportent en effet que ces paramètres ne sont pas indépendants ; en atteste le tableau 1.6 qui rassemble les coefficients de corrélation obtenus.

		Bradley [90]	Pelorson [83]	Beranek [91]	Cerda [92]	Polack [93]
RT,	EDT	0.992	0.56	0.99	0.95	0.91
RT,	C_{80}	-0.943	-0.3	-0.84	-0.91	-0.78
EDT,	C_{80}	-0.952	-0.88	-0.88	-0.88	-0.82
Ts,	RT	0.983	0.55	-	0.93	0.85
Ts,	EDT	0.986	0.94	-	0.84	0.84
Ts,	C_{80}	-0.983	-0.95	-	-0.92	-0.95

Tableau 1.6 – Corrélations entre les durées de décroissance et mesures de clarté calculées dans plusieurs salles, selon 6 études différentes.

Dans l'étude de Polack [93], l'analyse statistique d'un ensemble de 14 salles de concerts et théâtres parisiens a été réalisée [93]. Afin de caractériser les salles il considéra les indices suivants : T_{30}, EDT, T_s, G, C_{80}, J_{LF}, BR [1] et TR [2]. Les résultats

1. Bass ratio (rapport entre le temps de réverbération moyens à 125 Hz et 500 Hz sur le temps de réverbération moyens à 500 Hz et 1 kHz).

2. Treble ratio (rapport entre le temps de réverbération moyens à 2 kHz et 4 kHz sur le temps de

d'une analyse par composantes principales a établi que quatre composantes principales contribuaient à 88% à la variance des données et que le T_{30}, EDT, T_s et C_{80} contribuaient fortement à la première composante de l'analyse (responsable de 46% de la variance des données). Ceci atteste d'une forte corrélation entre les durées de décroissance et les mesures de clarté. La force sonore G - définie dans la section suivante - était un facteur discriminant pour distinguer les espaces considérés dans la mesure où elle contribuait fortement à la seconde composante principale qui expliquait 20% de la variance des données.

1.3.3. La force sonore

La force sonore, notée G, est une mesure de la contribution des réflexions d'une salle au champ sonore généré par une source en terme de niveau sonore. Une salle réverbérante procure par exemple une sensation de niveau sonore plus importante qu'une salle absorbante. Elle peut être mesurée comme la différence des niveaux de pression sonore de la réponse impulsionelle de salle et de la réponse mesurée à 10 m de la même source sonore en champ libre. Ainsi,

$$\mathrm{G} = 10\log_{10}\frac{\int_0^\infty p^2(t)dt}{\int_0^\infty p_{10}^2(t)dt} = L_p - L_{p,10} \tag{1.10}$$

avec $p(t)$ la pression sonore instantanée de la réponse impulsionnelle, $p_{10}(t)$ est celle mesurée à 10 m en champ libre, L_p et $L_{p,10}$ les niveaux de pression sonore respectifs de $p(t)$ et $p_{10}(t)$.

Dans le cas où il n'est pas possible d'effectuer la mesure de $L_{p,10}$ dans une chambre anéchoïque suffisamment grande, le niveau sonore à une distance d de la source peut être calculé puis $L_{p,10}$ obtenu de la manière suivante :

$$L_{p,10} = L_{p,d} + 20\log_{10}(\frac{d}{10}) \tag{1.11}$$

Lorsque l'on réalise cette mesure en champ libre il est recommandé d'effectuer des mesures tous les 12.5° autour de la source sonore et de calculer la moyenne des niveaux de pression sonore pour moyenner sa directivité [3].

Des variantes de ce paramètre sont présentes dans la littérature : la force sonore est parfois estimée d'après différentes régions temporelles de la réponse impulsionnelle de salle. La force sonore calculée sur la partie précoce de la réponse impulsionnelle a montré des corrélations avec la largeur apparente de la source [94, 95] ou avec l'enveloppement lorsqu'elle est calculée sur la partie tardive [96, 97]. La force sonore des basses et des aigus a également été employée afin d'appréhender la qualité tonale d'une salle [85, 98].

réverbération moyens à 500 Hz et 1 kHz).

1.3.4. Les mesures d'impression spatiale

Dans la littérature, les impressions spatiales désignent deux attributs perceptifs distincts liés à des aspects différents du champ sonore : la largeur apparente de source notée ASW (pour *Apparent Source Width*) et l'enveloppement noté LEV (pour *Listener Envelopment*). Le premier correspond à l'étendue spatiale de la source et semble être principalement influencé par l'énergie précoce. Le second correspond à l'impression d'être enveloppé par le champ réverbéré et est généralement lié aux propriétés spatiales de la réverbération tardive.

La largeur apparente de source

En raison de l'effet de précédence certaines réflexions précoces ne sont pas perçues comme des échos distincts mais fusionnent perceptivement avec le son direct [87]. Cette agglomération sonore semble provenir de la direction du premier front d'onde, celle du son direct. Lorsque ce sont des réflexions précoces latérales qui fusionnent avec le son direct, elles tendent à flouter la localisation du son et ainsi à élargir la source sonore.

Néanmoins, il semble que même lorsque les réflexions précoces ne fusionnent pas avec le son direct mais sont perçues séparément, elles contribuent également à l'ASW. Dans une étude menée par Barron [99], un test perceptif fût réalisé en considérant une source et une réflexion séparées d'un angle de 40°. L'expérience montre que la réflexion contribue à la sensation de largeur même lorsqu'elle est perçue comme un écho distinct. Il semble cependant que ce phénomène dépende de la direction d'incidence : en considérant une réflexion séparée d'un angle de 135°, Morimoto *et al.* rapportent au contraire que la largeur de source apparaissait plus étroite lorsque la source était perçue comme un écho distinct [100].

Johnson *et al.* [101] ont étudié la région d'incidence produisant une ASW maximum dans le plan horizontal en présence d'un son direct frontal et en utilisant une seule réflexion. Les sujets n'ont pas perçu de différence en terme d'ASW pour une réflexion située entre 40° et 130°. De plus, dans cette région angulaire, l'influence de la réflexion sur l'ASW ne semblait pas être affectée par le retard de la réflexion jusqu'à 30 ms. Toute réflexion contenue dans cette région angulaire et temporelle semble donc produire une ASW maximum.

Plusieurs paramètres acoustiques définis dans la norme ISO 3382-1 [3] sont censés prédire l'ASW : la fraction latérale ($\mathrm{J_{LF}}$), le coefficient de fraction latérale ($\mathrm{J_{LFC}}$) et le coefficient de corrélation interaurale (IACC).

La fraction d'énergie latérale peut être quantifiée grâce aux 80 premières millisecondes après le son direct des réponses impulsionnelles mesurées avec un microphone omnidirectionnel et un microphone bidirectionel pointant dans les directions latérales :

$$\mathrm{J_{LF}} = \frac{\displaystyle\int_{5ms}^{80ms} p_L^2(t)\mathrm{d}t}{\displaystyle\int_{0}^{80ms} p^2(t)\mathrm{d}t} \tag{1.12}$$

où p_L^2 correspond à l'énergie de la réponse impulsionnelle mesurée avec le microphone bidirectionnel.

Étant donné que la directivité d'un microphone bidirectionelle suit un motif en cosinus et que la pression est élevée au carré, la contribution d'une réflexion latérale varie selon le carré du cosinus de l'angle d'incidence. Un autre paramètre, nommé coefficient de fraction latéral, est également défini dans la norme. Il prend en considération la contribution de la réflexion selon le cosinus - et non le cosinus au carré - de l'angle d'incidence :

$$\mathrm{J_{LFC}} = \frac{\displaystyle\int_{5ms}^{80ms} |p_L(t) \cdot p(t)|\,\mathrm{d}t}{\displaystyle\int_{0}^{80ms} p^2(t)\mathrm{d}t} \tag{1.13}$$

Par ailleurs, le coefficient de corrélation intéraural est défini comme le maximum en valeur absolue de la fonction de correlation intéraurale :

$$\mathrm{IACC}_{t_1,t_2} = \max\left|\mathrm{IACF}_{t_1,t_2}(\tau)\right|, \text{ où } -1 \text{ ms} < \tau < 1 \text{ ms} \tag{1.14}$$

avec

$$\mathrm{IACF}_{t_1,t_2}(\tau) = \left[\int_{t_1}^{t_2} p_l(t) \cdot p_\mathrm{r}(t+\tau)\mathrm{d}t\right] / \left[\int_{t_1}^{t_2} p_l^2(t)\mathrm{d}t \int_{t_1}^{t_2} p_r^2(t)\mathrm{d}t\right]^{1/2} \tag{1.15}$$

où p_l est la réponse impulsionnelle à l'entrée du conduit auditif gauche et p_r celle à l'entrée du droit.

Un coefficient nul correspond à un champ diffus et au contraire un coefficient proche de 1 correspond à une image sonore très étroite. L'utilisation du Binaural Quality Index (BQI), proposé par Beranek [102], est également répandue :

$$\mathrm{BQI} = 1 - \mathrm{IACC_{E3}} \tag{1.16}$$

où $\mathrm{IACC_{E3}}$ désigne le coefficient de corrélation interaurale calculé sur l'énergie précoce et moyenné sur les bandes de fréquences à 500 Hz, 1 kHz et 2 kHz.

De Vries *et al.* [103] remettent en cause la pertinence de ces paramètres car ils peuvent présenter de larges fluctuations entre deux points de mesures proches sans pour autant qu'un changement d'ASW ne soit perçu. Les valeurs de *Just Noticeable Difference* (JND) établies dans la littérature leur semblent ainsi trop faibles. Reichardt et Schmidt rapportent par exemple des valeurs de JND comprises entre 0.06 et 0.09 pour une valeur de J_{LF} comprise entre 0.2 et 0.4 [104]. Pour la corrélation interaurale, Cox *et al.* rapportent un JND de 0.075 pour une valeur d'IACC de 0.33 [105]. De plus, il est possible que ces valeurs soient dépendantes de la source sonore et du niveau d'écoute [106, 107].

Pour contrer la variabilité de ces paramètres acoustiques, d'autres ont été proposés : la force sonore latérale précoce $\mathrm{G_{EL}}$ [94, 95], les fractions latérales précoces filtrées par secteur angulaire $\mathrm{B_{LFC}}$ et $\mathrm{B_{LF}}$ [103] ou le modèle auditif pour la perception acoustique de salle RAP [107, 108].

En effet, de Vries *et al.* [103] mettent en lumière le fait que les mesures de largeur apparente de source sont sujettes à des fluctuations en raison d'interférences entre ondes sonores. Étant donné que la perception de l'ASW ne fluctue pas de la même manière, il semble que le système auditif ne soit pas sensible à ces phénomènes d'interférences. Ils proposent alors de calculer $\mathrm{J_{LF}}$ et $\mathrm{J_{LFC}}$ selon des secteurs angulaires plus précis que celui capté par un microphone bidirectionnel car les ondes sonores captées par ce dipôle peuvent se mélanger et ainsi créer des interférences constructives ou destructives. Utiliser des secteurs angulaires plus précis grâce à un filtrage spatial permettraient de limiter ces interférences. Ceci implique de mesurer la réponse impulsionnelle avec un réseau de microphones et non un microphone bidirectionnel pour être capable de filtrer le champ sonore par secteur angulaire. Soient $\mathrm{B_{LF}}$ et $\mathrm{B_{LFC}}$ les fractions latérales précoces filtrées par secteur angulaire :

$$\mathrm{B_{LF}} = \frac{\displaystyle\int_{5ms}^{80ms} \sum_{n=0}^{N} \left(p_i(t)\cos\phi\right)^2 dt}{\displaystyle\int_{0}^{80ms} \sum_{n=0}^{N} p_i(t)^2 dt} \tag{1.17}$$

$$\mathrm{B_{LFC}} = \frac{\displaystyle\int_{5ms}^{80ms} \sum_{n=0}^{N} |p_i(t)||p_i(t)\cos\phi| dt}{\displaystyle\int_{0}^{80ms} \sum_{n=0}^{N} p_i(t)^2 dt} \tag{1.18}$$

avec N le nombre de secteurs angulaires et ϕ désigne l'azimut.

Klockgether et van de Par ont défini des prédicteurs d'impressions spatiales en utilisant des mesures de corrélation interaurale [108]. Cette méthode emploie une modélisation psychoacoustique des limites de la perception spatiale nommée RAP. Elle est directement appliquée au signaux binauraux et non aux réponses impulsionnelles contrairement aux mesures mentionnées précédemment. Les indices binauraux sont extraits des signaux et manipulés en prenant en compte une propriété de la perception spatiale : le système auditif est sensible aux petites variations de la corrélation interaurale lorsque la corrélation est importante et peu sensible aux variations lorsque la corrélation est faible. Une fonction puissance est donc appliquée aux mesures de corrélation interaurale. De plus les signaux binauraux sont préalablement filtrés en utilisant 42 filtres gammatone d'ordre 4. Pour la prédiction de la largeur apparente de source, les coefficients de corrélation interaurale sont calculés et pondérés par une estimation P de la probabilité que la fenêtre temporelle courante contienne le champ direct. Soit $\mathrm{RAP_{ASW}}$ la prédiction de la largeur apparente de la source :

$$\mathrm{RAP_{ASW}} = 1 - \frac{\displaystyle\sum_{n,c} \mathrm{IACC}^4(n,c)P(n,c)}{\displaystyle\sum_{n,c} P(n,c)} \tag{1.19}$$

La probabilité P est estimée d'après le rapport entre le niveau sonore de chaque segment divisé par le niveau sonore du segment précédent. Le rapport permet de déterminer si les segments contiennent du son direct ou si ils sont dominés par la réverbération. L'estimation de la largeur apparente de source s'est montrée plus performante avec le $\mathrm{RAP_{ASW}}$ que le BQI dans un test perceptif mené par Novak et Klockgether [107].

L'enveloppement

L'enveloppement correspond à l'impression subjective d'être entouré par le champ sonore réverbéré. Les propriétés physiques de la réverbération qui influencent cet attribut semblent être la distribution temporelle des réflexions, leur distribution spatiale ainsi que le niveau de la réverbération tardive.

En effet, le rapport entre l'énergie précoce et l'énergie tardive affecte la sensation d'enveloppement [95, 109, 110]. L'enveloppement est inversement proportionnel à ce rapport car c'est l'énergie tardive qui semble le plus participer à l'enveloppement. Par ailleurs, il est également possible d'augmenter la sensation d'enveloppement en augmentant le temps de réverbération du champ sonore. Cependant, Bradley et Soulodre ont montré que l'équilibre entre l'énergie sonore précoce et tardive est plus influent que la variation du temps de réverbération [97].

De plus, les études de Bradley et Soulodre montrent une influence des réflexions latérales tardives sur l'enveloppement [96, 97]. Dans une moindre mesure, les réflexions tardives venant des autres directions de l'espace - devant, derrière ou au dessus de l'auditeur - ont aussi un impact sur l'enveloppement [110–114].

Ces résultats ont permis de définir une mesure de l'enveloppement nommée $\mathrm{L_J}$ (auparavant noté GLL, LG ou $\mathrm{LG}_{80}^{\infty}$), la force de l'énergie latérale tardive :

$$\mathrm{L_J} = 10\log_{10}\frac{\displaystyle\int_{80ms}^{\infty} p_L^2(t)dt}{\displaystyle\int_{0}^{\infty} p_{10}^2(t)dt} \tag{1.20}$$

avec p_L la pression sonore instantanée de la réponse impulsionnelle de salle mesurée avec un microphone bidirectionnel positionné perpendiculairement à la source et p_{10} la réponse impulsionnelle de la source mesurée à 10 mètres de la source en champ libre. Ce paramètre est calculé par bande de fréquence et contrairement aux autres paramètre définis dans la norme, sa valeur moyenne ne résulte pas d'une moyenne arithmétique appliquée aux valeurs de chaque bande de fréquence mais aux quantités d'énergies contenues dans la partie tardive de chaque bande de fréquence :

$$\mathrm{L_{J,moy}} = 10\log_{10}\frac{1}{4}\sum_{i=1}^{4} 10^{\mathrm{L_{J,i}}/10} \tag{1.21}$$

Bien qu'elles ne figurent pas dans la norme, d'autres mesures de l'enveloppement sont communément utilisées : la fraction latérale d'énergie tardive (LLF) [96], la force sonore tardive ($\mathrm{G_{late}}$) [115] et le coefficient de corrélation interaurale de la réverbération tardive ($\mathrm{IACC_{L,3}}$) [116].

LLF est un paramètre similaire à J_{LF} mais calculé sur la partie tardive de la réponse impulsionnelle.

$$\mathrm{LLF} = \frac{\int_{80ms}^{\infty} p_L^2(t)dt}{\int_{80ms}^{\infty} p^2(t)dt} \tag{1.22}$$

où p correspond à la pression sonore instantanée de la réponse impulsionnelle de salle mesurée avec un microphone omnidirectionnel et p_L l'énergie de la réponse impulsionnelle mesurée avec un microphone bidirectionnel.

G_{late} est une mesure de force sonore appliquée à la réverbération tardive.

$$\mathrm{G_{late}} = 10\log_{10}\frac{\int_{80ms}^{\infty} p^2(t)dt}{\int_{0}^{\infty} p_{10}^2(t)dt} \tag{1.23}$$

Ainsi L_J peut s'exprimer en fonction LLF et G_{late} :

$$\mathrm{L_J} = \mathrm{G_{late}} + 10\log_{10}(\mathrm{LLF}) \tag{1.24}$$

En analysant les mesures effectuées dans 17 salles de spectacles, Barron [117] constata que le facteur contribuant le plus à la mesure de L_J est la force sonore tardive ; la variance de la fraction latérale tardive étant beaucoup plus faible.

Le paramètre $\mathrm{IACC_{L,3}}$ désigne le coefficient de corrélation interaurale calculé sur l'énergie tardive d'après l'équation (1.14) avec $t_1 = 0.080$ et $t_2 = 1$. Elle correspond à la moyenne des valeurs obtenues pour les bandes de fréquence centrées sur 500 Hz, 1 kHz et 2 kHz.

Bien que ces paramètres soient définis pour caractériser l'énergie sonore au-delà de 80 ms, les réflexions sonores peuvent également contribuer à la sensation d'enveloppement en-deçà de cette limite [108, 118]. Dick *et al.* rapportent que des modifications apportées à la partie précoce d'une réponse impulsionnelle de salle peuvent influencer la sensation d'enveloppement tandis qu'au delà de 120 ms de telles modifications n'ont qu'un faible impact [118]. D'autres valeurs de l'ordre de 100 ms sont présentes dans la littérature [119, 120]. Soulodre *et al.* [119] proposent également d'utiliser une limite temporelle dépendante de la fréquence (160 ms à 125 Hz jusqu'à 45 ms à 8 kHz). De plus amples études semblent nécessaires pour déterminer avec précision l'influence des différentes régions temporelles et spatiales sur la sensation d'enveloppement.

Ainsi, certaines études soutiennent que les paramètres acoustiques couramment utilisés ne sont pas adaptés à la mesure de l'enveloppement [114, 118, 121]. D'une part, l'usage d'un microphone bidirectionnel dans la mesure de L_J ou LLF sous-estime l'importance de la réverbération provenant de directions non latérales. D'autre part, la segmentation temporelle de 80 ms souvent utilisée dans les mesures courantes semble arbitraire et néglige souvent l'influence des premières réflexions. D'autres paramètres ont ainsi été proposés, tels que le rapport d'énergie avant arrière (FBR) [112], le temps

central spatialement équilibré ($\mathrm{SBT_s}$) [110], la mesure d'enveloppement issue du modèle auditif pour la perception acoustique de salle (RAP) [108].

Morimoto *et al.* ont montré que l'enveloppement croit avec la quantité d'énergie réfléchie provenant de l'arrière de l'auditeur. Le rapport d'énergie avant arrière FBR a donc été introduit pour prédire la sensation d'enveloppement. Il est défini par l'équation suivante :

$$\mathrm{FBR} = 10 \log \left(E_\mathrm{f} / E_\mathrm{b} \right) \tag{1.25}$$

avec E_f et E_b l'énergie contenue respectivement dans la partie frontale et arrière du plan horizontal. L'énergie du son direct n'est pas incluse dans cette équation, de même que les réflexions contenues dans le plan coronal (c'est-à-dire le plan qui coupe à angle droit le plan horizontal et le plan médian et qui contient l'axe interaural).

Hanyu et Kimura ont mis en évidence l'interaction entre le niveau et la distribution angulaire de l'énergie : la contribution de réflexions à l'enveloppement dépend du niveau et de la direction d'arrivée des autres réflexions [110]. Par exemple, l'efficacité des réflexions frontales diffère selon la direction d'arrivée des autres réflexions. La présence de réflexions frontales participe à la sensation d'enveloppement lorsque de l'énergie arrive également par l'arrière ; c'est à dire lorsque l'équilibre spatial avant / arrière augmente. Ainsi une large distribution angulaire des directions d'incidences ainsi qu'une distribution d'énergie équilibrée dans l'espace participent à la sensation d'enveloppement. En d'autres termes, plus l'auditeur est entouré de réflexions de toutes les directions avec un niveau uniforme, plus l'impression d'enveloppement est forte. Pour mesurer cette caractéristique Hanyu et Kimura ont introduit le temps central spatialement équilibré $\mathrm{SBT_s}$. Soit T_si le temps central calculé dans une direction i :

$$T_\mathrm{si} = \frac{\displaystyle\int_0^\infty t p_i^2(t) \mathrm{d}t}{\displaystyle\int_0^\infty p^2(t) \mathrm{d}t} \tag{1.26}$$

avec p_i la pression sonore mesurée dans la direction i et p la pression mesurée grâce à un microphone omnidirectionnel. Les temps centraux sont calculés dans plusieurs directions uniformément réparties dans le plan horizontal. Chaque temps central est ensuite pondéré de telle sorte que les temps centraux calculés dans les directions frontales et arrières soient pondérés par 0.5 et ceux calculés dans les directions latérales par 1. Soit a_i le résultat de cette pondération :

$$a_i = T_\mathrm{Si} \left(1 + |\sin \theta_\mathrm{i}| \right) / 2 \tag{1.27}$$

avec θ_i l'angle entre le plan médian et l'axe de la direction utilisée pour le calcul du temps central. Afin de prendre en considération l'interaction entre les directions d'incidence des réflexions dans leur contribution à l'enveloppement, le $\mathrm{SBT_s}$ est obtenu en pondérant les coefficient a_i par ceux correspondant aux autres directions avec une importance plus ou moins grande selon l'angle qui les sépare.

$$\mathrm{SBT_s} = \left[\sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j \sin\left(\theta_{ij}\right) \right]^{\frac{1}{2}} \tag{1.28}$$

où θ_{ij} est l'angle entre les directions i et j.

Le modèle de perception des impressions spatiales RAP [108] décrit pour le calcul de la largeur apparente de source permet également d'estimer la sensation d'enveloppement. De la même manière, les coefficients de corrélation interaurale des signaux binauraux considérés sont calculés et pondérés par une estimation P de la probabilité que la fenêtre temporelle courante contienne le champ direct. Soit $\mathrm{RAP_{LEV}}$ une estimation de l'enveloppement :

$$\mathrm{RAP_{LEV}} = 1 - \frac{\sum_{n,c} \mathrm{IACC}^4(n,c)(1 - P(n,c))}{\sum_{n,c} P(n,c)} \tag{1.29}$$

Les liens entre largeur apparente de source et enveloppement

Bradley *et al.* ont montré que les composantes précoces et tardives d'une réponse impulsionnelle de salle peuvent avoir des influences contraires sur ces impressions spatiales [95, 96]. Par exemple, accroître l'énergie tardive tend à amoindrir l'ASW et accroître l'énergie précoce tend à amoindrir l'enveloppement. Ils montrent également que l'ASW est moins affecté par la modification du rapport d'énergie précoce et tardive : l'ASW semble être moins masquée par l'énergie tardive que l'enveloppement ne l'est par l'énergie précoce. En raison de ces interactions, les deux attributs perceptifs coexistent dans des proportions différentes selon les salles. Bien que ces deux attributs soient définis et perçus comme des attributs distincts, il semble que l'ASW et l'enveloppement soient intimement liés et parfois difficiles à distinguer.

La mesure de ces attributs emploie un microphone bidirectionnel pour capter la contribution de l'énergie latérale. Cependant l'utilisation d'un réseau de microphones présente l'avantage d'offrir une résolution spatiale plus élevée que la mesure utilisant une directivité en dipôle. Les mesures de réponses impulsionnelles spatiales effectuées à l'aide d'un réseau de microphones peuvent faire l'objet d'une analyse spatiale du champ sonore au moyen de techniques de formation de faisceau (ou *beamforming* en anglais) dans le domaine des harmoniques sphériques. Cette approche permettrait de prendre en considération la contribution d'énergie provenant de directions non-latérales qui semblent participer à la sensation d'enveloppement. De plus, elle permettrait de satisfaire les précisions angulaires nécessaires pour le calcul du $\mathrm{B_{LF}}$, $\mathrm{B_{LFC}}$ ou du $\mathrm{SBT_s}$. Notons néanmoins que lorsqu'une directivité bidirectionnelle est nécessaire, la directivité en basses et hautes fréquences d'un réseau de microphones peut s'éloigner du motif en dipôle, selon le réseau de microphones utilisé [122].

De plus, l'étude des impressions spatiales a principalement employé des champs sonores simulés reproduits sur un nombre limité de haut-parleurs, couvrant ainsi une zone angulaire limitée, et avec parfois un faible nombre de réflexions. C'est par

exemple le cas des études de Barron [123] ou de Bradley et Soulodre [96, 97] à l'origine de la normalisation des paramètres J_{LF}, J_{LFC} et L_J respectivement. Bien qu'ils permettent un examen rigoureux des résultats, ces cas sont moins représentatifs du champ sonore réel vécu dans une salle de concert dont la partie tardive est généralement plus diffuse. Les récents progrès technologiques en acoustique virtuelle permettent la reproduction de champs sonores tout en offrant un contrôle expérimental plus important.

1.4. Conclusion

Plusieurs vocabulaires définis *ad hoc* au moyen de tests subjectifs en situation de concert ou en laboratoire ont permis de déterminer de très nombreux attributs perceptifs dont certains sont récurrents. Ces ensembles d'attributs ont été élaborés selon diverses méthodes d'analyse sensorielle et pour des applications différentes. L'examen de ces études montre que les jugements subjectifs de l'acoustique d'une salle impliquent de nombreux attributs perceptifs. En raison des méthodes d'analyse sensorielle et des stimuli employés, le nombre d'attributs perceptifs en jeu diffère grandement ainsi que les termes et définitions utilisés pour les décrire. Néanmoins, ces termes perceptifs permettent de détailler des expériences sensorielles et d'alimenter une terminologie sonore. Plusieurs études se sont nourries de ces efforts linguistiques et de l'expérience pratique d'experts en spatialisation sonore pour élaborer des lexiques d'attributs perceptifs permettant la description de l'acoustique d'une salle. En particulier, en se basant à la fois sur l'expérience de plusieurs experts et sur une étude statistique permettant d'attester de la qualité de l'ensemble d'attributs élaboré, le RAQI est un instrument d'analyse précieux pour juger de l'acoustique d'une salle de concert. Des démarches similaires doivent être entreprises pour couvrir le cas des salles de plus petits volumes.

La plupart des études perceptives se sont concentrées sur des sources spécifiques (instruments de musique, voix) et n'en ont généralement considéré qu'un nombre limité pour des raisons pratiques. Or, la source est un élément essentiel du champ sonore dans la mesure où l'acoustique d'une salle ne peut pas être perçue en tant que telle mais seulement comme un milieu façonnant les propriétés spatiales, temporelles et spectrales des sources sonores qu'il contient. Seules quelques études ont mentionné l'impact de la source sonore sur la perception de l'espace sans pour autant qu'elle fasse l'objet d'une étude approfondie. Par ailleurs, pour les expérimentations menées en laboratoire, l'impression visuelle de l'espace sonore étudié n'était généralement pas considérée. Dans une situation d'écoute réelle, ces deux éléments importants font partie intégrante de la perception de l'espace. Ils sont alors susceptibles de biaiser les résultats d'une évaluation perceptive si ils ne sont pas correctement pris en compte. Il parait donc important d'établir 1) si l'environnement visuel a une influence significative sur la perception de l'acoustique des salles ; 2) dans quelle mesure des sources sonores usuelles peuvent influencer notre appréciation de l'acoustique des salles. Ces questions font l'objet du chapitre 2 et 3 respectivement.

Conjointement aux études perceptives, de nombreux paramètres acoustiques ont

été élaborés pour caractériser les propriétés acoustiques d'une salle depuis la définition du temps de réverbération par Sabine. Certains se sont imposés en raison de leur pertinence perceptive mais plusieurs critiques peuvent être émises à l'égard de cette normalisation.

D'une part, la caractérisation de la perception d'une salle au moyen de ces paramètres acoustiques parait incomplète. Les attributs perceptifs définis dans les lexiques ne sont pas tous couverts par les paramètres de la norme : la brillance, la coloration ou l'intimité sont par exemple des attributs qui ne sont pas représentés.

D'autre part, certains paramètres acoustiques semblent redondants. Plusieurs d'entre eux semblent prendre en compte les mêmes propriétés acoustiques au vu de leurs corrélations. C'est par exemple le cas des durées de décroissance et des mesures de clarté qui sont fortement corrélées entre elles. De plus, certains paramètres acoustiques présentent une forte variabilité selon la position dans la salle. C'est notamment le cas des mesures de largeur apparente de source. Des alternatives, telles que le B_{LF} ou le B_{LFC}, ont été proposées pour contrecarrer cette variabilité en employant un réseau de microphones.

De surcroit, la limite temporelle définissant la région des premières réflexions et des réflexions tardives fixée dans la norme parait arbitraire et le calcul de cette valeur en cohérence avec l'espace étudié semble plus judicieux, d'autant plus lorsque que l'on considère des petits espaces pour lesquels ces limites temporelles sont susceptibles d'être différentes [75]. Il est d'ailleurs nécessaire de confirmer que les corrélations identifiées entre paramètres acoustiques et attributs perceptifs pour les salles de concert s'appliquent également aux petits espaces. Malheureusement, pour les salles dont le volume est inférieur à $300m^3$, la norme ISO ne se concentre que sur la mesure du temps de réverbération [124] ou sur l'intelligibilité de la parole pour les bureaux ouverts [88].

Enfin, plusieurs sources d'incertitudes persistent concernant les sources et récepteurs employés. Bien que la norme recommande l'usage d'une source omnidirectionnelle pour effectuer les mesures, une source sonore omnidirectionnelle n'est qu'approximativement omnidirectionelle, ce qui peut avoir un impact sur la mesure [125, 126]. On peut d'ailleurs se demander si l'usage d'une source omnidirectionnelle seulement est pertinent dans la mesure où il n'est pas représentatif d'une situation réelle dans laquelle on rencontre le plus souvent des sources sonores ayant des directivités bien différentes. D'autres sources d'incertitudes peuvent provenir des mesures binaurales effectuées pour le calcul des coefficients de corrélation interaurale. Les caractéristiques de l'oreille humaine étant individuelles, la généralisation ou la comparaison de ces paramètres d'une étude à l'autre semble problématique. Les normes définies pour les têtes artificielles [127] spécifient une certaine forme de tête et de torse mais pas de pavillon, un élément physiologique pourtant crucial. Au delà des biais relatifs aux sources et récepteurs employés, le post-traitement des réponses impulsionnelles contient divers degrés de liberté concernant l'implémentation des filtrages et intégrations, qui peuvent influencer significativement les résultats [128, 129].

Malgré ces limites, les mesures physiques et perceptives permettent de caractériser l'acoustique d'une salle sous de multiples aspects. En particulier, la partie III se concentrera sur l'évaluation physique et perceptive de la largeur apparente de source

et de l'enveloppement afin d'étudier le contrôle des impressions spatiales. La capacité des paramètres acoustiques mentionnés dans la section précédente à prédire les impressions spatiales sera notamment évaluée dans le chapitre 7.

Bibliographie

[1] W. C. Sabine, *Collected papers on acoustics*. Harvard university press, 1922.

[2] L. L. Beranek, *Music, acoustics & architecture*. RE Krieger Publishing Company, 1979.

[3] ISO 3382-1, "Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces," International Organization for Standardization, Geneva, CH, Standard, 2009.

[4] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[5] R. Bolt, P. Doak et P. Westervelt, "Pulse statistics analysis of room acoustics," *The Journal of the Acoustical Society of America*, vol. 22, n°. 3, p. 328–340, 1950.

[6] T. Schultz, "Diffusion in reverberation rooms," *Journal of Sound and Vibration*, vol. 16, n°. 1, p. 17–28, 1971.

[7] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, n°. 5, p. 299–306, 1987.

[8] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, n°. 12, p. 1819–1823, 1962.

[9] J.-M. Jot, L. Cerveau et O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," dans *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[10] A. Lindau, L. Kosanke et S. Weinzierl, "Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses," *Journal of the Audio Engineering Society*, vol. 60, n°. 11, p. 887–898, 2012.

[11] P. M. Morse et K. U. Ingard, *Theoretical acoustics*. Princeton university press, 1986.

[12] M. R. Schroeder et K. Kuttruff, "On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima," *The Journal of the Acoustical Society of America*, vol. 34, n°. 1, p. 76–80, 1962.

[13] N. Zacharov, *Sensory evaluation of sound*. CRC Press, 2018.

[14] H. Stone, R. N. Bleibaum et H. A. Thomas, *Sensory evaluation practices*. Academic press, 2012.

[15] ISO 5492, "Sensory analysis vocabulary," International Organization for Standardization, Geneva, CH, Standard, 2008.

[16] W. Munson et M. B. Gardner, "Standardizing auditory tests," *The Journal of the Acoustical Society of America*, vol. 22, n°. 5, p. 675–675, 1950.

[17] S. Purcell, “Duo-trio,” dans *Discrimination Testing in Sensory Science*. Elsevier, 2017, p. 197–207.

[18] ISO 5495, “Sensory analysis methodology : paired comparison test,” International Organization for Standardization, Geneva, CH, Standard, 2005.

[19] J. M. Ennis et V. Jesionka, “The power of sensory discrimination methods revisited,” *Journal of Sensory Studies*, vol. 26, n°. 5, p. 371–382, 2011.

[20] L. V. Jones, D. R. Peryam, L. Thurstone *et al.*, “Development of a scale for measuring soldiers’ food preferences.” *Food research*, vol. 20, p. 512–520, 1955.

[21] ITU-R BS.1534-3, “Method for the subjective assessment of intermediate quality level of audio systems.” International Telecommunication Union, Standard, 2015.

[22] ITU-R BS.1116-3, “Method for the subjective assessment of small impairments in audio systems.” International Telecommunication Union, Standard, 2015.

[23] A. A. Williams et S. P. Langron, “The use of free-choice profiling for the evaluation of commercial ports,” *Journal of the Science of Food and Agriculture*, vol. 35, n°. 5, p. 558–568, 1984.

[24] V. Dairou et J.-M. Sieffermann, “A comparison of 14 jams characterized by conventional profile and a quick original method, the flash profile,” *Journal of food science*, vol. 67, n°. 2, p. 826–834, 2002.

[25] G. Lorho, “Perceived quality evaluation : an application to sound reproduction over headphones,” Thèse de doctorat, Aalto-yliopiston teknillinen korkeakoulu, 2010.

[26] G. Kelly, *The Psychology of Personal Constructs*. Routledge, London ; New York, 1955.

[27] J. Berg et F. Rumsey, “Identification of quality attributes of spatial audio by repertory grid technique,” *Journal of the Audio Engineering Society*, vol. 54, n°. 5, p. 365–379, 2006.

[28] S. Cairncross et L. Sjöström, “Flavor profiles : a new approach to flavor problems,” *Descriptive Sensory Analysis in Practice*, p. 15–22, 1997.

[29] H. Stone, J. Sidel, S. Oliver, A. Woolsey et R. C. Singleton, “Sensory evaluation by quantitative descriptive analysis,” *Descriptive Sensory Analysis in Practice*, vol. 28, p. 23–34, 2008.

[30] C. E. Osgood, G. J. Suci et P. H. Tannenbaum, *The measurement of meaning*. University of Illinois press, 1957, n°. 47.

[31] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis,” *Psychometrika*, vol. 29, n°. 1, p. 1–27, 1964.

[32] S. Chollet, D. Valentin et H. Abdi, “Free sorting task,” *Novel techniques in sensory characterization and consumer profiling*, p. 207–228, 2014.

[33] E. Risvik, J. A. McEwan, J. S. Colwill, R. Rogers et D. H. Lyon, “Projective mapping : A tool for sensory analysis and consumer research,” *Food quality and preference*, vol. 5, n°. 4, p. 263–269, 1994.

[34] S. Choisel et F. Wickelmaier, “Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound,” *Journal of the Audio Engineering Society*, vol. 54, n°. 9, p. 815–826, 2006.

[35] F. Wickelmaier et W. Ellermeier, “Deriving auditory features from triadic comparisons,” *Perception & psychophysics*, vol. 69, n°. 2, p. 287–297, 2007.

[36] I. Borg et P. J. Groenen, *Modern multidimensional scaling : Theory and applications*. Springer Science & Business Media, 2005.

[37] J. D. Carroll et J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition,” *Psychometrika*, vol. 35, n°. 3, p. 283–319, 1970.

[38] B. Blesser et L.-R. Salter, *Spaces speak, are you listening ?: experiencing aural architecture*. MIT press, 2009.

[39] B. E. Stein et M. A. Meredith, *The merging of the senses*. The MIT Press, 1993.

[40] C. E. Jack et W. R. Thurlow, “Effects of degree of visual association and angle of displacement on the “ventriloquism” effect,” *Perceptual and motor skills*, vol. 37, n°. 3, p. 967–979, 1973.

[41] E. Hendrickx, M. Paquier, V. Koehl et J. Palacino, “Ventriloquism effect with sound stimuli varying in both azimuth and elevation,” *The Journal of the Acoustical Society of America*, vol. 138, n°. 6, p. 3686–3697, 2015.

[42] C. W. Bishop, S. London et L. M. Miller, “Visual influences on echo suppression,” *Current Biology*, vol. 21, n°. 3, p. 221–225, 2011.

[43] M. B. Gardner, “Proximity image effect in sound localization,” *The Journal of the Acoustical Society of America*, vol. 43, n°. 1, p. 163–163, 1968.

[44] D. H. Mershon, D. H. Desaulniers, T. L. Amerson et S. A. Kiefer, “Visual capture in auditory distance perception : Proximity image effect reconsidered.” *Journal of Auditory Research*, 1980.

[45] L. Hládek, C. C. Le Dantec, N. Kopco et A. Seitz, “Ventriloquism effect and aftereffect in the distance dimension,” dans *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, n°. 1. Acoustical Society of America, 2013, p. 050042.

[46] E. R. Calcagno, E. L. Abregu, M. C. Eguía et R. Vergara, “The role of vision in auditory distance perception,” *Perception*, vol. 41, n°. 2, p. 175–192, 2012.

[47] H.-J. Maempel et M. Jentsch, “Audio-visual interaction of size and distance perception in concert halls-a preliminary study,” dans *Proc. International Symposium on Room Acoustics (ISRA) Toronto*, 2013.

[48] G. Plenge, “On the problem of “in head localization”,” *Acta Acustica united with Acustica*, vol. 26, n°. 5, p. 241–252, 1972.

[49] A. Neidhardt et N. Knoop, “Investigating the room divergence effect in binaural playback,” 2015.

[50] J. C. Gil-Carvajal, J. Cubick, S. Santurette et T. Dau, "Spatial hearing with incongruent visual or auditory room cues," *Nature Scientific Reports*, vol. 6, p. 37342, 2016.

[51] J. Udesen, T. Piechowiak et F. Gran, "The effect of vision on psychoacoustic testing with headphone-based virtual sound," *Journal of the Audio Engineering Society*, vol. 63, n°. 7/8, p. 552–561, 2015.

[52] S. Werner, F. Klein, T. Mayenfels et K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," dans *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, p. 1–6.

[53] P. Larsson, D. Västfjäll et M. Kleiner, "Auditory-visual interaction in real and virtual rooms," dans *Proceedings of the Forum Acusticum, 3rd EAA European Congress on Acoustics, Sevilla, Spain*, 2002.

[54] D. Cabrera, A. Nguyen et Y. Choi, "Auditory versus visual spatial impression : A study of two auditoria," dans *ICAD 04-Tenth Meeting of the International Conference on Auditory Display*. Georgia Institute of Technology, 2004.

[55] B. N. Postma et B. F. Katz, "The influence of visual distance on the room-acoustic experience of auralizations," *The Journal of the Acoustical Society of America*, vol. 142, n°. 5, p. 3035–3046, 2017.

[56] N. Zacharov et T. H. Pedersen, "Spatial sound attributes—development of a common lexicon," dans *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[57] H. Wilkens, "Mehrdimensionale beschreibung subjektiver beurteilungen der akustik von konzertsälen," Thèse de doctorat, "", 1975.

[58] A. Gabrielsson, "Dimension analyses of perceived sound quality of sound-reproducing systems," *Scandinavian Journal of Psychology*, vol. 20, n°. 1, p. 159–169, 1979.

[59] F. E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society*, vol. 33, n°. 1/2, p. 2–32, 1985.

[60] C. Lavandier, "Validation perceptive d'un modèle objectif de caractérisation de la qualité acoustique des salles," Thèse de doctorat, Le Mans, 1989.

[61] T. Letowski, "Sound quality assessment : concepts and criteria," dans *Audio Engineering Society Convention 87*. Audio Engineering Society, 1989.

[62] E. Kahle, "Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras," Thèse de doctorat, Université du Maine, Le Mans, 1995.

[63] R. Mason et F. Rumsey, "An assessment of spatial performance of virtual home theatre algorithms by subjective and objective methods," *Audio Engineering Society Preprint*, vol. 5137, 2000.

[64] N. Zacharov et K. Koivuniemi, "Unravelling the perception of spatial sound reproduction : Language development, verbal protocol analysis and listener training," dans *Audio Engineering Society Convention 111*. Audio Engineering Society, 2001.

[65] F. Rumsey, “Spatial quality evaluation for reproduced sound : Terminology, meaning, and a scene-based paradigm,” *Journal of the Audio Engineering Society*, vol. 50, n°. 9, p. 651–666, 2002.

[66] J. Berg et F. Rumsey, “Systematic evaluation of perceived spatial quality,” dans *Audio Engineering Society Conference : 24th International Conference : Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[67] C. Guastavino et B. F. Katz, “Perceptual evaluation of multi-dimensional spatial audio reproduction,” *The Journal of the Acoustical Society of America*, vol. 116, n°. 2, p. 1105–1115, 2004.

[68] R. Sazdov, G. Paine et K. Stevens, “Perceptual investigation into envelopment, spatial clarity, and engulfment in reproduced multi-channel audio,” dans *AES 31st International Conference*, 2007.

[69] A. Silzle, “Quality taxonomies for auditory virtual environments,” dans *Audio Engineering Society Convention 122*. Audio Engineering Society, 2007.

[70] H. Wittek, “Perceptual differences between wavefield synthesis and stereophony,” Thèse de doctorat, University of Surrey Surrey, UK, 2007.

[71] T. H. Pedersen, “The semantic space of sounds,” *Delta*, 2008.

[72] T. Lokki, J. Pätynen, A. Kuusinen et S. Tervo, “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles,” *The Journal of the Acoustical Society of America*, vol. 132, n°. 5, p. 3148–3161, 2012.

[73] S. Le Bagousse, M. Paquier, C. Colomes et S. Moulin, “Sound quality evaluation based on attributes-application to binaural contents,” dans *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[74] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman et S. Weinzierl, “A spatial audio quality inventory (saqi),” *Acta Acustica united with Acustica*, vol. 100, n°. 5, p. 984–994, 2014.

[75] N. Kaplanis, S. Bech, S. H. Jensen et T. van Waterschoot, “Perception of reverberation in small rooms : a literature study,” dans *Audio Engineering Society Conference : 55th International Conference : Spatial Audio*. Audio Engineering Society, 2014.

[76] N. Zacharov, T. Pedersen et C. Pike, “A common lexicon for spatial sound quality assessment-latest developments,” dans *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, p. 1–6.

[77] S. Weinzierl, S. Lepa et D. Ackermann, “A measuring instrument for the auditory perception of rooms : The room acoustical quality inventory (raqi),” *The Journal of the Acoustical Society of America*, vol. 144, n°. 3, p. 1245–1257, 2018.

[78] D. W. Stewart et P. N. Shamdasani, *Focus groups : Theory and practice*. Sage publications, 2014, vol. 20.

[79] S. Ciba, A. Wlodarski et H.-J. Maempel, “Whisper–a new tool for performing listening tests,” dans *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[80] D. Schröder et M. Vorländer, “RAVEN : A real-time framework for the auralization of interactive virtual environments,” dans *Forum Acusticum*. Aalborg Denmark, 2011, p. 1541–1546.

[81] D. Ackermann, M. Ilse, D. Grigoriev, S. Lepa, S. Pelzer, M. Vorländer et S. Weinzierl, “A ground truth on room acoustical analysis and perception (GRAP),” 2018.

[82] R. Lacatis, A. Giménez, A. Barba Sevillano, S. Cerdá, J. Romero et R. Cibrián, “Historical and chronological evolution of the concert hall acoustics parameters,” *Journal of the Acoustical Society of America*, vol. 123, n°. 5, p. 3198, 2008.

[83] X. Pelorson, J.-P. Vian et J.-D. Polack, “On the variability of room acoustical parameters : reproducibility and statistical validity,” *Applied Acoustics*, vol. 37, n°. 3, p. 175–198, 1992.

[84] M. R. Schroeder, “New method of measuring reverberation time,” *The Journal of the Acoustical Society of America*, vol. 37, n°. 6, p. 1187–1188, 1965.

[85] J. S. Bradley, “Review of objective room acoustics measures and future needs,” *Applied Acoustics*, vol. 72, n°. 10, p. 713–720, 2011.

[86] H. Haas, “The influence of a single echo on the audibility of speech,” *Journal of the Audio Engineering Society*, vol. 20, n°. 2, p. 146–159, 1972.

[87] R. Y. Litovsky, H. S. Colburn, W. A. Yost et S. J. Guzman, “The precedence effect,” *The Journal of the Acoustical Society of America*, vol. 106, n°. 4, p. 1633–1654, 1999.

[88] ISO 3382-3, “Acoustics-measurement of room acoustic parameters—part 3 : Open-plan offices,” International Organization for Standardization, Geneva, CH, Standard, 2012.

[89] T. Houtgast, H. J. Steeneken et R. Plomp, “Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics,” *Acta Acustica united with Acustica*, vol. 46, n°. 1, p. 60–72, 1980.

[90] J. Bradley, “Auditorium acoustics measures from pistol shots,” *The Journal of the Acoustical Society of America*, vol. 80, n°. 1, p. 199–205, 1986.

[91] L. L. Beranek, “Subjective rank-orderings and acoustical measurements for fifty-eight concert halls,” *Acta Acustica united with Acustica*, vol. 89, n°. 3, p. 494–508, 2003.

[92] S. Cerdá, A. Giménez, J. Romero, R. Cibrián et J. Miralles, “Room acoustical parameters : A factor analysis approach,” *Applied Acoustics*, vol. 70, n°. 1, p. 97–109, 2009.

[93] J.-D. Polack, F. L. Figueiredo et S. Liu, “Statistical analysis of a set of parisian concert halls and theatres,” dans *Acoustics 2012*, 2012.

[94] T. Okano, L. L. Beranek et T. Hidaka, “Relations among interaural cross-correlation coefficient (iacc e), lateral fraction (lf e), and apparent source width (asw) in concert halls,” *The Journal of the Acoustical Society of America*, vol. 104, n°. 1, p. 255–265, 1998.

[95] J. Bradley, R. Reich et S. Norcross, “On the combined effects of early-and late-arriving sound on spatial impression in concert halls,” *The Journal of the Acoustical Society of America*, vol. 108, n°. 2, p. 651–661, 2000.

[96] J. S. Bradley et G. A. Soulodre, “The influence of late arriving energy on spatial impression,” *The Journal of the Acoustical Society of America*, vol. 97, n°. 4, p. 2263–2271, 1995.

[97] J. S. Bradley et G. A. Soulodre, “Objective measures of listener envelopment,” *The Journal of the Acoustical Society of America*, vol. 98, n°. 5, p. 2590–2597, 1995.

[98] G. A. Soulodre et J. S. Bradley, “Subjective evaluation of new room acoustic measures,” *The Journal of the Acoustical Society of America*, vol. 98, n°. 1, p. 294–301, 1995.

[99] M. Barron, “The subjective effects of first reflections in concert halls—the need for lateral reflections,” *Journal of sound and vibration*, vol. 15, n°. 4, p. 475–494, 1971.

[100] M. Morimoto, K. Nakagawa et K. Iida, “The relation between spatial impression and the law of the first wavefront,” *Applied Acoustics*, vol. 69, n°. 2, p. 132–140, 2008.

[101] D. Johnson et H. Lee, “Just noticeable difference in apparent source width depending on the direction of a single reflection,” dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[102] L. Beranek, *Concert halls and opera houses : music, acoustics, and architecture*. Springer Science & Business Media, 2012.

[103] D. de Vries, E. M. Hulsebos et J. Baan, “Spatial fluctuations in measures for spaciousness,” *The journal of the Acoustical Society of America*, vol. 110, n°. 2, p. 947–954, 2001.

[104] W. Reichardt et W. Schmidt, “The audible steps of spatial impression in music performances,” *Acta Acustica united with Acustica*, vol. 17, n°. 3, p. 175–179, 1966.

[105] T. J. Cox, W. J. Davies et Y. W. Lam, “The sensitivity of listeners to early sound field changes in auditoria,” *Acta Acustica united with Acustica*, vol. 79, n°. 1, p. 27–41, 1993.

[106] J. Becker et M. Sapp, “Synthetic soundfields for the rating of spatial perceptions,” *Applied acoustics*, vol. 62, n°. 2, p. 217–228, 2001.

[107] J. Nowak et S. Klockgether, “Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations,” *The Journal of the Acoustical Society of America*, vol. 142, n°. 3, p. 1634–1645, 2017.

[108] S. Klockgether et S. van de Par, “A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception,” *Acta Acustica united with Acustica*, vol. 100, n°. 5, p. 964–971, 2014.

[109] J. Berg et D. Nyberg, “Listener envelopment–what has been done and what future research is needed ?” dans *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[110] T. Hanyu et S. Kimura, “A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections,” *Applied Acoustics*, vol. 62, n°. 2, p. 155–184, 2001.

[111] H. Furuya, K. Fujimoto, C. Y. Ji et N. Higa, “Arrival direction of late sound and listener envelopment,” *Applied Acoustics*, vol. 62, n°. 2, p. 125–136, 2001.

[112] M. Morimoto, K. Iida et K. Sakagami, “The role of reflections from behind the listener in spatial impression,” *Applied Acoustics*, vol. 62, n°. 2, p. 109–124, 2001.

[113] A. Wakuda, H. Furuya, K. Fujimoto, K. Isogai et K. Anai, “Effects of arrival direction of late sound on listener envelopment,” *Acoustical science and technology*, vol. 24, n°. 4, p. 179–185, 2003.

[114] W. Lachenmayr, A. Haapaniemi et T. Lokki, “Direction of late reverberation and envelopment in two reproduced berlin concert halls,” dans *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[115] J. S. Bradley, “Using iso 3382 measures, and their extensions, to evaluate acoustical conditions in concert halls,” *Acoustical science and technology*, vol. 26, n°. 2, p. 170–178, 2005.

[116] T. Hidaka, L. L. Beranek et T. Okano, “Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls,” *The Journal of the Acoustical Society of America*, vol. 98, n°. 2, p. 988–1007, 1995.

[117] M. Barron, “Late lateral energy fractions and the envelopment question in concert halls,” *Applied Acoustics*, vol. 62, n°. 2, p. 185–202, 2001.

[118] D. A. Dick et M. C. Vigeant, “An investigation of listener envelopment utilizing a spherical microphone array and third-order ambisonics reproduction,” *The Journal of the Acoustical Society of America*, vol. 145, n°. 4, p. 2795–2809, 2019.

[119] G. A. Soulodre, M. C. Lavoie et S. G. Norcross, “Objective measures of listener envelopment in multichannel surround systems,” *Journal of the Audio Engineering Society*, vol. 51, n°. 9, p. 826–840, 2003.

[120] D. Griesinger, “Objective measures of spaciousness and envelopment,” dans *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[121] H. Lee, “Apparent source width and listener envelopment in relation to source-listener distance,” dans *Audio Engineering Society Conference : 52nd International Conference : Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.

[122] D. A. Dick et M. C. Vigeant, “A comparison of late lateral energy (GLL) and lateral energy fraction (LF) measurements using a spherical microphone array and conventional methods,” dans *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, 2014.

[123] M. Barron et A. H. Marshall, “Spatial impression due to early lateral reflections in concert halls : the derivation of a physical measure,” *Journal of Sound and Vibration*, vol. 77, n°. 2, p. 211–232, 1981.

[124] ISO 3382-2, “Acoustics—measurement of room acoustic parameters, part 2 : Reverberation time in ordinary rooms,” International Organization for Standardization, Geneva, CH, Standard, 2008.

[125] T. W. Leishman, S. Rollins et H. M. Smith, “An experimental evaluation of regular polyhedron loudspeakers as omnidirectional sources of sound,” *The Journal of the Acoustical Society of America*, vol. 120, n°. 3, p. 1411–1422, 2006.

[126] R. S. Martín, I. Witew, M. Arana et M. Vorländer, “Influence of the source orientation on the measurement of acoustic parameters,” *Acta acustica united with acustica*, vol. 93, n°. 3, p. 387–397, 2007.

[127] International Electrotechnical Commission, “Provisional head and torso simulator for acoustic measurements on air conduction hearing aids,” 1990.

[128] M. Guski et M. Vorländer, “Comparison of noise compensation methods for room acoustic impulse response evaluations,” *Acta Acustica united with Acustica*, vol. 100, n°. 2, p. 320–327, 2014.

[129] B. F. Katz, “International round robin on room acoustical impulse response analysis software 2004,” *Acoustics Research Letters Online*, vol. 5, n°. 4, p. 158–164, 2004.

2

L'influence de la vision sur la perception de l'acoustique d'une salle

Peu d'études ont abordé l'influence des indices visuels sur la perception de l'espace sonore, au-delà de leur influence sur la position perçue de la source sonore. Des études antérieures suggèrent que la perception des réflexions tardives n'est pas affectée par la vision d'une salle, mais seul un nombre limité d'attributs perceptifs a été étudié. Dans ce chapitre, les interactions audiovisuelles sont examinées sans faire d'hypothèses sur le nombre et la nature des dimensions perceptives impliquées dans la perception de l'espace sonore. Nous analysons les résultats d'un test perceptif dans lequel des sujets ont jugé de la dissemblance perçue entre des espaces sonores tout en regardant un stimulus visuel donné. Des comparaisons par paires ont été répétées en utilisant plusieurs conditions visuelles et une condition uniquement sonore. L'expérimentation a été réalisé dans un environnement virtuel en utilisant un visiocasque et un rendu binaural dynamique. Il apparait que la modalité visuelle n'a pas eu d'impact sur les différences perçues entre les espaces sonores. Cette étude a fait l'objet d'une publication dans le journal de l'Audio Engineering Society (Volume 68, Numéro 7/8, p. 522-531, juillet 2020).

2.1. Les interactions entre modalités sonore et visuelle

Diverses études ont défini de nombreux attributs perceptifs liés à l'évaluation d'environnements acoustiques virtuels, de technologies de spatialisation ou de l'acoustique des salles de concert [1]. Certains travaux ont utilisé des salles existantes et ont effectué des études perceptives directement dans ces salles ou dans des laboratoires à l'aide d'enregistrements. Cependant, dans ce dernier cas de figure, on peut affirmer qu'il y a un manque de contrôle expérimental concernant les stimuli présentés : l'im-

pression visuelle de l'espace sonore étudié n'était généralement pas considérée [2–6]. Dans ce chapitre, nous abordons l'influence de la vision sur la perception de l'espace sonore afin d'évaluer si le manque de contrôle expérimental concernant l'environnement visuel a été préjudiciable à de telles études. À cette fin, nous avons utilisé des technologies avancées d'auralisation et de visualisation qui nous permettent d'avoir un meilleur contrôle expérimental sur les stimuli audiovisuels à l'étude.

Plusieurs études montrent une interaction entre les indices auditifs et visuels liés à la perception de l'espace. Certaines études ont notamment examiné l'influence de la vision sur l'évaluation de la localisation des sources sonores [7–9], la perception de la distance auditive [10–14], l'externalisation [15–19] ou les impressions spatiales [20–22].

2.1.1. Localisation de source sonore

Il semble que la vision ait un fort impact sur la perception de la localisation des sources sonores. À la suite des premiers travaux de Jack et Thurlow [7], de nombreuses expériences ont été réalisées pour déterminer si les sujets pouvaient percevoir un stimulus auditif perceptivement unifié avec un objet visuel d'une localisation différente (effet dit ventriloque) [23–27]. Tous sont arrivés à la conclusion qu'une forte attraction visuelle se produit et que l'effet augmente avec la diminution de la différence angulaire entre les positions des stimuli sonores et visuels. Hendrickx *et al.* [8] ont également mis en évidence une plus forte attraction visuelle en élévation qu'en azimut. Bishop *et al.* [9] ont étudié une autre interaction intermodale concernant la localisation de la source : l'impact des indices visuels sur l'effet de précédence. D'après ce phénomène, lorsque le retard entre le son direct et une réflexion est court, la réflexion n'est pas perçue comme un événement sonore distinct. Dans ce cas, la localisation de la source est dominée par la localisation du son direct. Il a été démontré que la force de l'effet de précédence peut être renforcée lorsque l'information visuelle coïncide spatialement et temporellement avec le premier front d'onde, c'est-à-dire avec le son direct. Inversement, l'effet de précédence est réduit lorsque l'information visuelle coïncide spatialement et temporellement avec la réflexion.

2.1.2. Perception de la distance

L'attraction visuelle se produit également avec la perception de la distance sonore, même en présence de multiples indices auditifs [10, 11]. En particulier, Hládek *et al.* [12] ont mesuré les performances de localisation à distance dans une pièce réverbérante sombre en utilisant des salves de bruit qui étaient spatialement congruentes ou non congruentes avec des stimuli visuels (LED). Ils ont rapporté un effet ventriloque en distance : un déplacement vers les stimuli visuels était perçu lorsqu'ils étaient présentés plus près ou plus loin des stimuli auditifs. En outre, Calcagno *et al.* [13] ont signalé que, même lorsque la source sonore était visuellement occultée, les jugements auditifs de distance étaient plus précis lorsque des informations visuelles de l'ensemble de la scène étaient disponibles. Ils ont émis l'hypothèse que des informations visuelles autres que la distance perçue par rapport à la source, telles que la taille de la pièce,

peuvent être utilisées par les auditeurs pour effectuer des jugements de distance auditive.

2.1.3. Externalisation

Plusieurs études montrent également que les aspects visuels ont une influence sur l'externalisation [15–19]. Cette capacité à entendre des événements auditifs hors de la tête en écoute binaurale diminue si la salle de synthèse et la salle d'écoute sont non congruentes, c'est-à-dire quand un décalage est présent entre l'acoustique virtuelle et l'impression visuelle de la salle d'écoute.

Dans une étude menée par Udesen *et al.* [18], des jugements d'externalisation ont été effectués dans deux pièces : l'une congruente et l'autre non congruente avec l'espace sonore virtuel des stimuli binauraux. Bien que les mêmes échantillons sonores binauraux aient été utilisés - seuls les indices visuels différaient - les résultats ont montré des différences significatives entre les environnements de test : les jugements d'externalisation étaient plus faibles dans la condition de non congruence. Les auteurs ont émis l'hypothèse que lorsque les attentes d'un environnement sonore réaliste ne sont pas satisfaites, en particulier lorsqu'une divergence entre les repères visuels et auditifs se produit, le réalisme de la scène sonore est affecté et conduit à une perception intra-crânienne.

Il semble donc que la perception auditive soit sensiblement affectée par les impressions visuelles de la salle d'écoute et les attentes de l'auditeur - qui peuvent être par ailleurs modifiées par le phénomène d'apprentissage [19]. Néanmoins, les résultats d'un test d'externalisation mené par Gil-Carvajal *et al.* [17] ont montré que la modalité auditive avait un impact plus important sur l'externalisation que la modalité visuelle. Dans leur expérience, une réponse impulsionnelle binaurale de salle (BRIR) a été mesurée dans une salle de référence pour créer des stimuli sonores individualisés pour 18 sujets. Les stimuli ont été restitués au casque, les sujet étant présents dans la salle de référence et dans deux autres salles : 1) une salle plus petite et plus réverbérante, 2) une salle plus grande et anéchoïque. La congruence entre la salle de référence et les salles de test différait selon le volume (repères visuels) et le temps de réverbération (repères auditifs). Trois conditions ont été testées :

— une condition de congruence auditive. Le test a été effectué dans l'obscurité et des salves de bruit était diffusées en plus des stimuli sonores pour fournir des indices acoustiques supplémentaires congruents avec la salle d'écoute.
— une condition de congruence visuelle. Les sujets pouvaient voir la pièce et aucun signal auditif n'était diffusé en plus des stimuli sonores.
— une condition de congruence visuelle et auditive. Les sujets pouvaient voir la pièce et des salves de bruit était diffusées en plus des stimuli sonores pour fournir des indices acoustiques supplémentaires congruents avec la salle d'écoute.

Les auteurs ont indiqué que les degrés d'externalisation étaient considérablement réduits lorsque les sujets recevaient des indices auditifs supplémentaires de la salle d'écoute qui ne correspondaient pas à ceux de l'espace sonore virtuel. *A contrario*, les taux d'externalisation n'étaient pas affectés lorsque les sujets pouvaient voir une pièce différente de celle qu'ils entendaient au casque. On peut donc supposer que la

connaissance préalable des caractéristiques acoustiques de l'environnement d'écoute peut avoir un impact plus important sur l'externalisation que les attentes acoustiques basées sur les indices visuels de la pièce. Contrairement aux autres études sur l'externalisation mentionnées, les indices visuels de la salle ne semblent pas avoir eu d'influence sur les jugements d'externalisation. Il est clair que la vision peut biaiser l'audition de multiples façons non triviales et que les liens entre la perception visuelle et auditive ne sont pas entièrement compris.

2.1.4. Impressions spatiales

Des photographies ont été utilisées comme support visuels dans de nombreuses études sur les interactions audiovisuelles multi-modales. Cependant, Larsson *et al.* [20] ont montré que le degré de réalisme visuel affectait l'évaluation des attributs perceptifs liés à une salle d'écoute. En particulier, leur étude a indiqué que les sources sonores étaient considérées comme beaucoup plus larges lorsque le sujet se trouvait dans la pièce réelle ou utilisait un visiocasque noté HMD (pour *Head Mounted Display*) par rapport à des conditions sans repères visuels ou utilisant des photographies fixes. En conséquence, Postma et Katz [22, 28] ont utilisé un dispositif de test avec un haut degré de réalisme visuel - un système CAVE [1] - pour étudier l'influence de la vision sur la perception acoustique de la pièce. Pour la pièce étudiée, ils ont observé une influence significative des repères visuels sur la perception de la distance alors qu'aucune influence significative concernant la largeur apparente de la source ou l'enveloppement n'a été trouvée. On peut émettre l'hypothèse que les différentes configurations de test ou les différences entre les espaces sonores utilisés dans ces études peuvent expliquer les résultats contradictoires obtenus concernant la largeur apparente de la source. D'autres études sont nécessaires pour déterminer si les indices visuels ont un impact sur la perception de l'espace sonore. Alors que la perception de la distance, la localisation et l'externalisation d'une source sonore semblent être influencées par des indices visuels, l'influence de la vision sur d'autres attributs sonores spatiaux reste floue.

Récemment, Schutte et al [29] ont étudié l'influence de la vision sur la réverbérance (*degree of reverberation* dans le texte) au sein d'environnements virtuels audiovisuels. Trois conditions ont été testées : 1) environnements auditifs et visuels congruents, 2) environnements auditifs et visuels non congruents, et 3) condition auditive seule. Ils n'ont constaté aucune influence de la vision sur les réponses des sujets, que l'environnement soit congruent ou non congruent. Cependant, cette étude a porté l'attention sur la réverbérance uniquement et les sujets peuvent donc avoir basé leur jugement sur des caractéristiques limitées, telles que le temps de réverbération ou le rapport entre l'énergie du son direct et l'énergie du champ réverbéré. La perception de l'espace sonore est généralement considérée comme multidimensionnelle et implique d'autres attributs perceptifs tels que la clarté, l'enveloppement, la brillance ou la largeur (voir section 1.2.3). Par conséquent, un protocole de test qui n'oriente pas l'attention vers une caractéristique sonore particulière serait plus en phase avec une perception natu-

1. Un système CAVE (pour *Computer Automatic Virtual Environment*) désigne un espace de réalité virtuelle généralement cubique dans lequel des images sont projetées sur ses parois pour créer un environnement immersif qui permet à l'utilisateur de se déplacer.

relle de l'acoustique et pourrait mettre en évidence des facteurs autres que la réverbérance qui interagiraient avec la modalité visuelle.

2.1.5. Le but de l'étude

Nous avons examiné l'influence de multiples conditions visuelles sur des notes de dissemblance liées à un ensemble de stimuli sonores. Les sujets ont dû évaluer les différences perçues entre plusieurs espaces sonores tout en regardant un même stimulus visuel. Les comparaisons par paires ont été répétées avec différents stimuli visuels (y compris dans une condition uniquement sonore). Nous n'avons donc pas fait d'hypothèses sur le nombre et la nature des dimensions perceptives impliquées dans la perception de l'espace sonore. En outre, nous avons utilisé l'analyse multidimensionnelle (MDS) pour nous assurer que la structure perceptive sous-jacente à la perception des espaces sonores employés n'était pas unidimensionnelle.

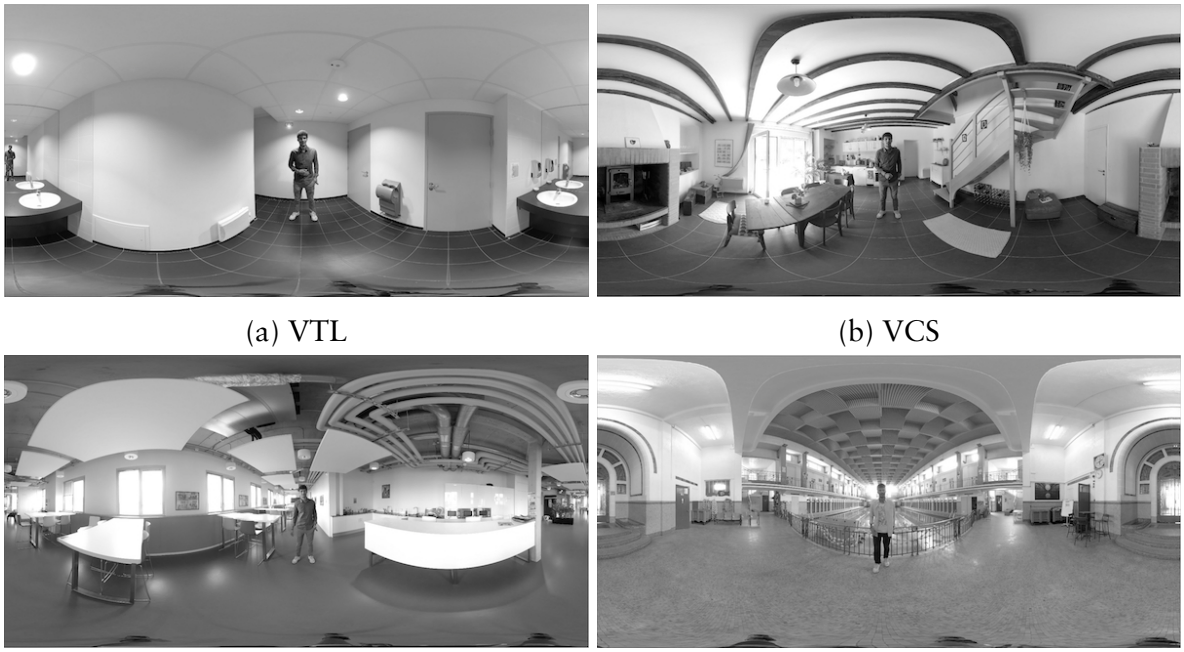
Comme mentionné précédemment, en présence de conflits entre les indices visuels et auditifs dans la perception de l'espace, la vision domine et biaise généralement l'audition [30]. Nous avons mentionné à titre d'exemple, l'impact des impressions visuelles d'une salle d'écoute sur l'externalisation de signaux binauraux ou encore l'apparition d'un effet ventriloque lorsque les stimuli audio et visuel sont discordants en terme de localisation - que ce soit par rapport à leur angle d'incidence ou à leur distance. Ce phénomène pourrait s'expliquer par le fait que le système visuel a une plus grande précision spatiale que le système auditif et est donc plus fiable [8, 31, 32]. De même, on peut émettre l'hypothèse que la vision offre plus d'informations sur une pièce donnée (comme la taille de la pièce [14] par exemple), ce qui peut biaiser la perception de l'espace sonore pour la rendre plus conforme à nos attentes visuelles, surtout si les environnements visuel et auditif sont non congruents. Dans la présente expérimentation, ce phénomène peut conduire à ce que les sujets perçoivent moins de différences lorsqu'ils comparent des espaces sonores tout en regardant un même espace visuel que lorsque les comparaisons sont effectuées sans aucun repère visuel.

Comme l'évaluation des attributs perceptifs spatiaux semble être affectée par le degré de réalisme visuel [20], Schutte *et al.* ont utilisé un HMD pour reproduire l'espace avec une haute fidélité [29]. De même, nous avons utilisé des vidéos à 360 degrés affichées sur un HMD ainsi qu'une reproduction sonore binaurale dynamique pour garantir une qualité de simulation élevée. De plus, les espaces sonores impliqués ont été mesurés avec un réseau dense de microphones pour permettre une reproduction précise des champs sonores.

2.2. Test perceptif

2.2.1. Stimuli visuels

Les stimuli visuels consistaient en une scène simple : un acteur situé à 150 cm de distance et récitant un poème de Gérard de Nerval - *Fantaisie* - dans quatre espaces différents : des toilettes, une cuisine, un réfectoire et une piscine. Les stimuli visuels



(a) VTL (b) VCS

(c) VRF (d) VPS

FIGURE 2.1 – Projection équirectangulaire des différents stimuli visuels. VTL : *Toilettes*, VCS : *Cuisine*, VRF : *Réfectoire*, VPS : *Piscine*. Les volumes des espaces concernés sont respectivement de 23 m^3, 81 m^3, 360 m^3 et 7448 m^3.

seront désignés respectivement par VTL, VCS, VRF et VPS dans les sections suivantes. Une cinquième condition visuelle, à savoir l'absence totale d'image, a été ajoutée. Cette condition sera par la suite appelée V0.

Les stimuli visuels étaient des vidéos 360°, enregistrées en 4K (3840 × 1920 pixels) à 30 images par seconde à l'aide d'une caméra Insta360 Pro. Les vidéos ont couvert tout l'espace en azimut et en élévation. Des captures d'écran des stimuli visuels sont affichées dans la figure 2.1.

Afin d'assurer des différences minimales entre les stimuli par rapport à la performance de l'acteur, un haut-parleur diffusait l'enregistrement anéchoïque de sa voix durant les prises de vue et l'acteur devait réciter le poème de manière synchrone. Ce procédé a permis d'assurer une synchronicité labiale pour tous les stimuli. Ainsi, aucun son n'était enregistré lors de la capture des stimuli visuels. Une étape de post-synchronisation était ensuite nécessaire pour synchroniser le son et la vidéo. Un test informel effectué par les auteurs et des discussions informelles avec les sujets ont suggéré qu'il n'y avait pas de désynchronisation perceptible.

2.2.2. Réponses impulsionnelles de salles

Douze espaces différents ont été utilisés pour créer les stimuli audio, y compris les quatre espaces utilisés pour les stimuli visuels. Ces espaces ont été choisis de façon à couvrir une large gamme de différences perceptibles. Les espaces sonores considérés n'étaient pas seulement limités aux salles de concert ou aux grands volumes, qui sont

privilégiés par une partie de la littérature de l'acoustique des salles, mais comprenaient également des lieux de la vie quotidienne. Les différentes mesures présentées dans le tableau 2.1 illustrent cette hétérogénéité. Les pièces considérées couvrent une large gamme de temps de décroissance et présentent divers rapports d'énergie pour des temps de décroissance similaires.

Tableau 2.1 – Abréviations, durées de décroissance, rapport champ direct / champ réverbéré (DRR) et mesures de clarté des espaces sonores utilisés dans le test perceptif. Les mesures ont été réalisées aux fréquences médiums. Les valeurs minimales et maximales sont affichées en bleu.

Espace	Abréviation	RT (s)	EDT (s)	DRR (dB)	C_{80} (dB)	D_{50} (%)	T_s (ms)
Box	BOX	0.38	0.31	-3.73	16.07	85.87	45.3
Toilettes	TLT	0.41	0.33	-2.13	15.14	82.62	41.4
Salle de réunion	REU	0.43	0.35	0.98	15.22	89.90	43.9
Petite salle de concert	SCP	0.46	0.31	6.08	16.6	96.19	33.4
Salle de classe	CLA	0.55	0.40	3.67	15.53	93.17	37.5
Cuisine	CUI	0.60	0.51	-2.56	10.19	81.28	51.0
Refectoire	RFC	0.83	0.65	2.61	11.85	87.40	43.1
Salle de concert moyenne	SCM	0.85	0.27	8.16	18.37	96.88	30.1
Piscine	PSC	1.91	0.72	3.74	11.95	90.13	41.2
Halle	HLL	2.58	0.55	9.11	16.36	96.57	33.3
Église	EGL	3.56	2.17	6.55	9.84	88.75	58.1
Cathédrale	CAT	6.55	2.48	6.85	12.72	92.78	59.3

Les SRIRs ont été mesurées avec l'antenne sphérique de microphone Eigenmike EM32 et une enceinte Genelec 8040. L'enceinte était positionnée à la même distance du microphone (150 cm) dans chaque pièce. Un signal à balayage sinusoïdal de 10 secondes fût utilisé durant la mesure. Les mesures de réponses impulsionnelles ont ensuite été débruitées selon la procédure décrite par Cabrera *et al.* [33]. Afin de compenser la réponse en fréquence de l'enceinte dans les mesures de SRIRs, un filtre à réponse impulsionnelle finie de 128 échantillons a été appliqué aux mesures. La correction fréquentielle fût appliquée entre 60 Hz et 16 kHz. Le filtre fut calculé d'après une mesure de la réponse impulsionnelle de l'enceinte effectuée dans une chambre anéchoïque avec un microphone omnidirectionnel situé dans l'axe de l'enceinte. Enfin, les réponses impulsionnelles issues des capsules du microphone sphérique furent converties en ambisonique à l'ordre 4. Les procédés employés pour la mesure, le débruitage et l'encodage des SRIRs sont décrits en détail dans l'annexe A et B.

2.2.3. Stimuli sonores

Une seule source sonore a été considérée dans ce test : une voix masculine enregistrée dans une chambre anéchoïque récitant le poème *Fantaisie*. Le signal enregistré a

été convolué avec les SRIRs correspondant aux 12 espaces sonores. Nous avons donc obtenu 12 stimuli ambisoniques d'ordre 4. Une égalisation du volume sonore a été effectuée subjectivement par les expérimentateurs avant le test perceptif (en accord avec la recommandation AES [34]), de manière à ce que le volume sonore perçu reste le même d'un stimuli sonore à l'autre. La durée totale des stimuli était de 20 s (durée de la source sonore) + 6,5 s (temps de réverbération des SRIRs les plus longs) = 26,5 s. Les stimuli ont été échantillonnés à une fréquence de 48000 Hz avec une résolution de 24 bits.

2.2.4. Binauralisation et head-tracking

Les signaux ambisoniques ont été convertis en signaux binauraux en utilisant le plug-in VST *Binaural Decoder* disponible en *open-source* [35]. Cet outil utilise des filtres de décodage binaural calculés selon le procédé exposé en annexe C.2 d'après 2702 HRTFs mesurées par une tête artificielle Neumann KU 100 [36].

L'utilisation d'HRTFs non individualisées peut accroître les occurrences de confusion avant-arrière [37] ou produire des localisations intra-crâniennes [38, 39]. Néanmoins, ces problèmes sont efficacement atténués grâce à un procédé de suivi des mouvements de tête (*head-tracking*) [40, 41] et par la présence de réverbération dans les signaux sonores [42, 43]. De plus, Begault *et al.* ont rapporté que les HRTF individualisées n'offraient pas d'avantage en termes de précision de localisation et d'externalisation pour la synthèse binaurale des stimuli vocaux reproduits dans le plan horizontal [42].

Le suivi des mouvements de tête a été réalisé en utilisant le HMD et le plugin de rotation ambisonique *ambix_rotator_o7* [44]. La latence du dispositif de suivi du HMD utilisé était de 22 ms [45].

2.2.5. Procedure

Le test perceptif consistait en des comparaisons par paires successives. A chaque appréciation, les sujets devaient noter la dissemblance entre deux stimuli sonores. Chaque paire de stimuli impliquait deux des 12 espaces sonores présentés plus haut. Chaque jugement de dissemblance était effectué tout en regardant une vidéo 360 synchronisée avec le son. Le même stimulus visuel était utilisé pour comparer deux stimuli sonores. Les évaluations de dissemblance pour toutes les comparaisons par paires ont été effectuées dans les cinq conditions visuelles différentes. Un total de 66 jugements pour chaque condition visuelle a donc été effectué par chacun des sujet. Au total, les sujets devaient évaluer 330 paires de stimuli sonores dans des conditions visuelles différentes : les paires de stimuli testés comprenaient soit des espaces sonores qui étaient tous les deux non congruents avec l'espace visuel, soit des espaces sonores dont un seul était congruent avec l'espace visuel.

Pour chaque paire, il fut demandé aux participants de quantifier la différence perçue entre les stimuli en déplaçant un curseur sur une échelle de 100 points dont les extrémités étaient étiquetées «identique» (0) et «très différent» (100). Aucune étiquette ou graduation intermédiaire n'étaient présentes afin d'éviter tout biais indési-

rable [46, 47]. Les sujets ont reçu l'instruction expresse de ne pas fermer les yeux pendant toute la durée de l'expérience et avaient la possibilité d'écouter les deux stimuli de manière répétée et de passer de l'un à l'autre en cours de lecture. Les différentes paires, ainsi que le stimulus visuel associé, ont été présentés dans un ordre aléatoire. L'ordre des stimuli sonores au sein de chaque paire était également aléatoire. Ces randomisations étaient différentes pour chaque sujet.

Le moteur de lecture, l'interface graphique et l'acquisition des données furent réalisés à l'aide de Unity et Max. L'expérience a été réalisée en réalité virtuelle à l'aide d'un HMD HTC Vive et d'un casque Sennheiser HD 650.

L'expérience consistait en deux sessions de test d'une heure chacune sur des jours différents. Avant chaque session, les participants se sont familiarisés avec les 12 stimuli sonores pendant cinq minutes afin d'appréhender la diversité des stimuli inclus dans le test.

2.2.6. Sujets

Onze sujets âgés en moyenne de 23 ans ont participé au test perceptif. Ils étaient tous étudiants du Master Image & Son de l'Université de Brest et étaient donc formés à l'écoute critique en raison des enseignements reçus en prise de son, montage et mixage. Aucun d'entre eux n'a déclaré de perte auditive connue et aucun n'avait d'expérience en Réalité Virtuelle.

2.3. Résultats

2.3.1. Matrices de dissemblance

Les notes de dissemblance forment une matrice de dissemblance de dimension 12×12 (pour les 12 conditions sonores) pour chaque sujet et pour chaque condition visuelle. L'ordre des stimuli sonores dans chaque paire ayant été randomisé, les matrices de dissemblance sont symétriques. Les matrices moyennées sur tous les sujets sont présentées en figure 2.2 pour les cinq conditions visuelles. Ces figures paraissent similaires et les matrices présentent d'importants coefficients de corrélation de Pearson (> 0.94, $p < 0.001$). Cela suggère que les indices de vision ont une influence relativement faible sur les différences perçues entre les espaces.

2.3.2. Analyse de la variance

Comme aucun point d'ancrage intermédiaire n'a été utilisé sur l'échelle de dissemblance, les résultats des sujets ont été normalisés en utilisant le score z, qui traduit les notes comme l'écart à la moyenne des scores de la population en terme de fraction d'écart-type. Le résultat normalisé d'un sujet z_i s'écrit :

$$z_i = \frac{(x_i - \bar{x}_i)}{s_i} \cdot s + \bar{x} \tag{2.1}$$

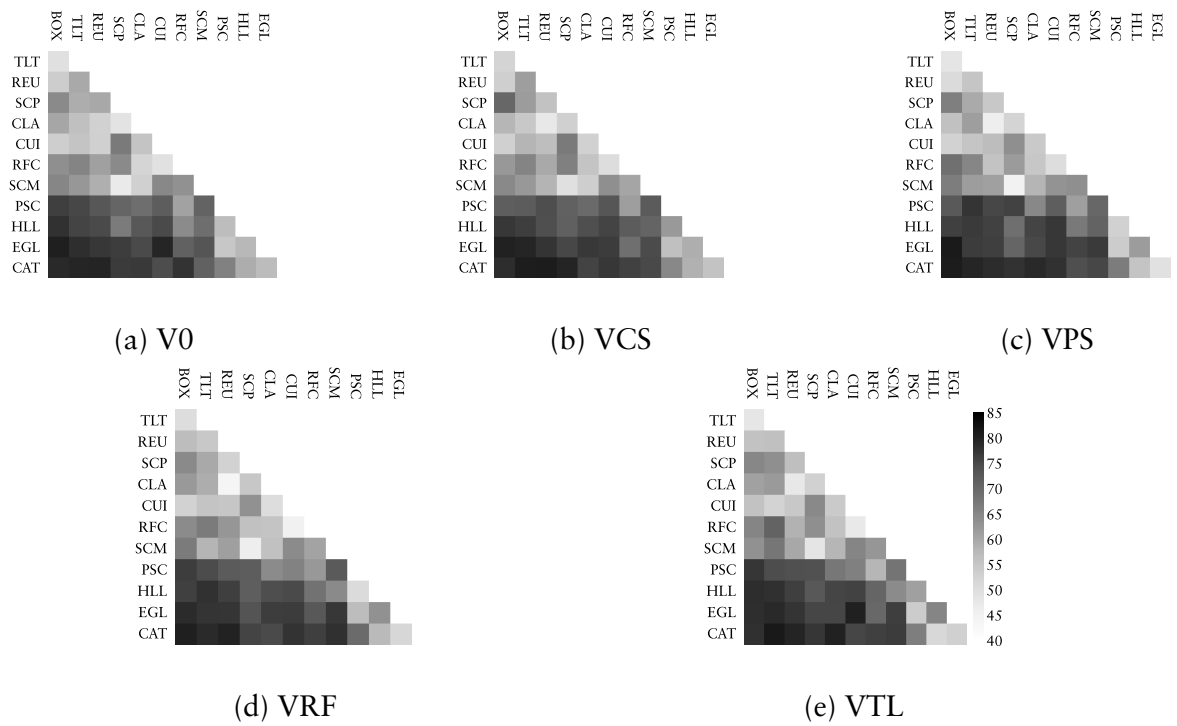


FIGURE 2.2 – Matrices de dissemblance pour les cinq conditions visuelles du test perceptif. Chaque carré représente la note de dissemblance entre deux espaces. L'échelle de gris fait référence à la moyenne des notes de dissemblance à travers les sujets. Une couleur plus sombre signifie une plus grande dissemblance.

Facteur	df	SS	MS	F	p	$\eta^2(\%)$
V	4	334	84	1.6	0.193	0.06
P	65	335374	5160	41.1	< 0.001	60.57
V * P	260	13762	53	1.14	0.067	2.48

Tableau 2.2 – Résultats de l'ANOVA. V : condition visuelle, P : paire d'espaces sonores, SS : somme des carrés, MS : moyenne des carrés, F : valeur f, p : valeur p, η^2 : proportion de variance expliquée.

où x_i correspond à la note donnée par le sujet i, $\bar{x}_i$ est la note moyenne du sujet i, $\bar{x}$ est la note moyenne de tous les sujets, s_i est l'écart-type des notes du sujet i et s est l'écart-type des notes de tous les sujets.

Afin d'effectuer une analyse de la variance (ANOVA), l'hypothèse de normalité doit être remplie : les résidus des observations appartenant à la même cellule (une combinaison de niveaux de variables indépendantes) doivent être normalement distribués [48]. Un test de Shapiro-Wilk de normalité des résidus a été effectué sur chaque cellule à un niveau de signification de 5% et a rejeté l'hypothèse nulle d'une distribution normale pour 46 cellules seulement sur 330, soit 13.9% des cellules. La plupart des études statistiques indiquent que l'ANOVA peut être robuste à ce genre de violation [49].

Les résultats ne contenaient pas de valeurs aberrantes (*outliers*), les résidus studentisés des cellules étant compris entre -3 et +3 [50]. Les données ont donc été sou-

mises à une ANOVA à mesures répétées avec les variables indépendantes suivantes : «condition visuelle» V (5) × «paire d'espaces sonores» P (66). Le test de sphéricité de Mauchly a indiqué que l'hypothèse de sphéricité était validée pour chaque variable indépendante ainsi que pour leurs interactions. Les résultats sont présentés dans le tableau 2.2 : seul l'effet des paires d'espaces sonores a eu une influence significative sur les scores de dissemblance. Aucun effet simple de la variable indépendante «condition visuelle» n'a été observé [F(4, 40) = 1.6, p = 0.193]. La valeur de significativité associée à l'interaction entre les conditions visuelles et les paires d'espaces sonores n'était pas suffisamment faible pour être significative [F(260, 2600) = 1,142, p = 0,067], et le pourcentage de variance imputable à cette interaction n'était que de 2,48%.

2.3.3. Analyse multidimensionnelle

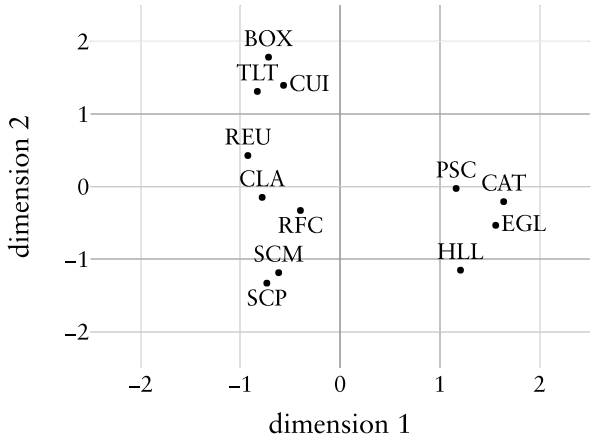


FIGURE 2.3 – Espace perceptif bidimensionnel résultant de l'analyse INDSCAL. Les variances relatives aux dimensions 1 et 2 sont respectivement de 0,757 et 0,096.

Étant donné qu'aucune influence des indices visuels n'a été trouvée dans l'expérience, les matrices de dissemblance ont été moyennées sur les différentes conditions visuelles. L'analyse multidimensionnelle (MDS) a été utilisée pour révéler la structure perceptive sous-jacente aux jugements de dissemblance entre stimuli. La MDS fournit les coordonnées de ces stimuli sur une ou plusieurs dimensions perceptives [51]. En particulier, l'analyse INDSCAL [52] trouve un groupe de dimensions perceptives qui sont communes à tous les sujets et fournit les coordonnées correspondantes de chaque stimulus. L'analyse fournit également les coordonnées de chaque sujet sur ces dimensions afin d'examiner avec quelles pondérations des dimensions perceptives ils ont effectué leur jugement. L'un des défis de ce type d'analyse consiste à trouver le bon nombre de dimensions perceptives, c'est-à-dire les dimensions qui méritent d'être interprétées. L'analyse de la variance expliquée en fonction de la dimension donne un indice sur le nombre de dimensions à considérer : il est intéressant de maximiser la variance expliquée de la solution jusqu'à ce qu'elle augmente de moins de 5% par dimension ajoutée [53].

L'analyse a révélé que l'essentiel de la variance peut être expliqué par deux di-

mensions. La première dimension explique la plupart des différences perçues entre les espaces sonores avec 75,7 % de la variance expliquée, tandis que la deuxième dimension explique presque 10 % de la variance. La représentation des 12 stimuli sonores dans cet espace perceptif est illustrée dans la figure 2.3.

Des analyses statistiques supplémentaires ont été effectuées afin de trouver des corrélations entre les dimensions perceptives obtenues et les mesures acoustiques. Vingt-deux paramètres acoustiques - dont certains sont définis dans la norme ISO 3382-1 [54] - ont été calculés à partir des 12 SRIRs mesurées sur l'ensemble du spectre et aux fréquences médium. La plupart d'entre eux ont été calculés avec le logiciel AURORA [55].

La première dimension s'est avérée bien corrélée au logarithme des temps de réverbération aux fréquences médium avec un coefficient de corrélation de Pearson de 0,966. On peut supposer que l'attribut perceptif correspondant est la réverbérance étudiée par Schutte *et al.* [29]. La mesure D_{50} semble expliquer la deuxième dimension perceptive avec un coefficient de corrélation de Pearson de 0,926. Le D_{50} mesure le rapport entre l'énergie précoce et l'énergie totale de la réponse impulsionnelle de la pièce. Cet indice est une mesure de clarté particulièrement adaptée pour caractériser l'intelligibilité de la parole [54]. Une autre mesure de clarté, le temps central T_s, est également bien corrélé à la seconde dimension avec un coefficient de Pearson de 0,896.

2.4. Discussion

L'analyse de la variance a révélé que les conditions visuelles n'avaient pas d'influence significative sur les différences perçues entre les espaces sonores. Ainsi, quelle que soit la divergence entre les conditions visuelles et auditives, qu'il y ait un contenu visuel ou non, la modalité visuelle ne change pas les différences perçues entre les espaces sonores.

Les résultats de la présente expérience sont en accord avec les conclusions de Schutte *et al.* [29] selon lesquelles la vision n'a pas d'impact significatif sur la perception de la réverbérance. Comme la présente expérience n'a fait aucune hypothèse sur les dimensions impliquées dans la perception de l'espace sonore, elle suggère que la vision n'a pas seulement un faible impact sur la réverbérance, mais aussi plus généralement sur la perception globale des espaces sonores.

Une analyse multidimensionnelle a été effectuée pour déterminer si d'autres dimensions que la réverbérance étaient impliquées pour distinguer les espaces sonores. Elle a révélé que la structure sous-jacente à la perception des espaces sonores était multidimensionnelle. En plus de la réverbérance, les auditeurs semblent avoir évalué des caractéristiques liées à la clarté. Cependant, nous pouvons supposer que la deuxième dimension perceptive est liée à la source sonore utilisée (car liée à l'intelligibilité de la voix) et que l'utilisation d'autres sources sonores pourrait avoir révélé des caractéristiques perceptives différentes. En outre, il est possible que la gamme des attributs spatiaux auditifs couverts par les salles considérées n'ait pas été suffisamment large. En d'autres termes, les sujets n'ont peut-être pas pu évaluer d'autres attributs tels que l'enveloppement ou la largeur apparente de la source, car les diffé-

rences entre les stimuli sonores pour ces attributs n'étaient pas suffisantes. De plus, la plupart des espaces sonores considérés avaient un DRR positif (ceci est dû au fait que la source sonore était située à 1,5 m du sujet) et pour les SRIRs ayant le DRR le plus élevée, la partie réverbérante peut ne pas avoir été assez forte par rapport au son direct pour mettre en évidence des différences dans la perception de certains attributs. Différentes distances pourraient être utilisées pour déterminer si un DRR élevé a eu un impact sur les indices de dissemblance, c'est-à-dire s'il existe un seuil au-dessus duquel les différences de réverbération sont masquées par le son direct. Par conséquent, des études supplémentaires devraient être réalisées en utilisant d'autres sources sonores et espaces sonores pour confirmer nos résultats.

Le nombre limité de dimensions perceptives impliquées dans cette analyse est dû au fait que seulement 12 espaces sonores ont été étudiés, le processus de comparaisons par paires étant chronophage. D'autres protocoles, tels que les tâches de classification libre, auraient permis d'utiliser une quantité beaucoup plus importante de stimuli sonores et donc de découvrir potentiellement d'autres dimensions. Cependant, le but de l'étude était plutôt d'examiner l'influence de la vision sur les différences perçues entre les espaces sonores, plutôt que de déterminer toutes les dimensions perceptives impliquées. À cette fin, la comparaison par paire est une meilleure option car c'est une méthode plus précise pour discriminer des espaces qu'une tâche de classification libre [56].

2.5. Conclusion

Dans ce chapitre, nous avons étudié l'effet de la vision sur la perception de l'acoustique de salles en utilisant des comparaisons par paires de stimuli sonores sous de multiples conditions visuelles. Les résultats ont montré que la perception visuelle d'une salle n'affectait pas les différences perçues entre les espaces sonores étudiés. Cette étude a été réalisée en utilisant des notes de dissemblances sans présumer du nombre et de la nature des dimensions perceptives impliquées : une analyse multidimensionnelle a révélé que principalement deux dimensions - la réverbérance et la clarté - étaient prises en compte par les sujets pour distinguer les espaces sonores. Après les études de Schutte [29] et de Postma [28], la présente étude apporte donc de nouvelles preuves de l'absence d'influence visuelle sur la perception auditive de l'espace.

Ainsi, dans la suite de ce document, nous avons fait le choix de ne pas prendre en compte l'aspect visuel des espaces utilisés pour étudier la perception de l'acoustique des salles. En particulier, les informations visuelles ne seront pas utilisées dans la deuxième partie du document où nous étudierons la réduction des données utiles à la reproduction de l'acoustique d'une salle. En effet, puisque la présence d'un environnement visuel ne semble pas induire une attention moindre vis-à-vis des informations sonores, elle ne pourra pas permettre une réduction de la précision des signaux sonores.

Bibliographie

[1] N. Zacharov, C. Pike, F. Melchior et T. Worch, “Next generation audio system assessment using the multiple stimulus ideal profile method,” dans *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, p. 1–6.

[2] T. Lokki, J. Pätynen, A. Kuusinen et S. Tervo, “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles,” *The Journal of the Acoustical Society of America*, vol. 132, n°. 5, p. 3148–3161, 2012.

[3] C. Lavandier, “Validation perceptive d’un modèle objectif de caractérisation de la qualité acoustique des salles,” Thèse de doctorat, Le Mans, 1989.

[4] R. Mason et F. Rumsey, “An assessment of spatial performance of virtual home theatre algorithms by subjective and objective methods,” *Audio Engineering Society Preprint*, vol. 5137, 2000.

[5] W. Lachenmayr, A. Haapaniemi et T. Lokki, “Direction of late reverberation and envelopment in two reproduced berlin concert halls,” dans *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[6] D. A. Dick et M. C. Vigeant, “An investigation of listener envelopment utilizing a spherical microphone array and third-order ambisonics reproduction,” *The Journal of the Acoustical Society of America*, vol. 145, n°. 4, p. 2795–2809, 2019.

[7] C. E. Jack et W. R. Thurlow, “Effects of degree of visual association and angle of displacement on the “ventriloquism” effect,” *Perceptual and motor skills*, vol. 37, n°. 3, p. 967–979, 1973.

[8] E. Hendrickx, M. Paquier, V. Koehl et J. Palacino, “Ventriloquism effect with sound stimuli varying in both azimuth and elevation,” *The Journal of the Acoustical Society of America*, vol. 138, n°. 6, p. 3686–3697, 2015.

[9] C. W. Bishop, S. London et L. M. Miller, “Visual influences on echo suppression,” *Current Biology*, vol. 21, n°. 3, p. 221–225, 2011.

[10] M. B. Gardner, “Proximity image effect in sound localization,” *The Journal of the Acoustical Society of America*, vol. 43, n°. 1, p. 163–163, 1968.

[11] D. H. Mershon, D. H. Desaulniers, T. L. Amerson et S. A. Kiefer, “Visual capture in auditory distance perception : Proximity image effect reconsidered.” *Journal of Auditory Research*, 1980.

[12] L. Hládek, C. C. Le Dantec, N. Kopco et A. Seitz, “Ventriloquism effect and aftereffect in the distance dimension,” dans *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, n°. 1. Acoustical Society of America, 2013, p. 050042.

[13] E. R. Calcagno, E. L. Abregu, M. C. Eguía et R. Vergara, “The role of vision in auditory distance perception,” *Perception*, vol. 41, n°. 2, p. 175–192, 2012.

[14] H.-J. Maempel et M. Jentsch, “Audio-visual interaction of size and distance perception in concert halls-a preliminary study,” dans *Proc. International Symposium on Room Acoustics (ISRA) Toronto*, 2013.

[15] G. Plenge, “On the problem of “in head localization”,” *Acta Acustica united with Acustica*, vol. 26, n°. 5, p. 241–252, 1972.

[16] A. Neidhardt et N. Knoop, “Investigating the room divergence effect in binaural playback,” 2015.

[17] J. C. Gil-Carvajal, J. Cubick, S. Santurette et T. Dau, “Spatial hearing with incongruent visual or auditory room cues,” *Nature Scientific Reports*, vol. 6, p. 37342, 2016.

[18] J. Udesen, T. Piechowiak et F. Gran, “The effect of vision on psychoacoustic testing with headphone-based virtual sound,” *Journal of the Audio Engineering Society*, vol. 63, n°. 7/8, p. 552–561, 2015.

[19] S. Werner, F. Klein, T. Mayenfels et K. Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” dans *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 2016, p. 1–6.

[20] P. Larsson, D. Västfjäll et M. Kleiner, “Auditory-visual interaction in real and virtual rooms,” dans *Proceedings of the Forum Acusticum, 3rd EAA European Congress on Acoustics, Sevilla, Spain*, 2002.

[21] D. Cabrera, A. Nguyen et Y. Choi, “Auditory versus visual spatial impression : A study of two auditoria,” dans *ICAD 04-Tenth Meeting of the International Conference on Auditory Display*. Georgia Institute of Technology, 2004.

[22] B. N. Postma et B. F. Katz, “The influence of visual distance on the room-acoustic experience of auralizations,” *The Journal of the Acoustical Society of America*, vol. 142, n°. 5, p. 3035–3046, 2017.

[23] P. Bertelson et M. Radeau, “Cross-modal bias and perceptual fusion with auditory-visual spatial discordance,” *Perception & psychophysics*, vol. 29, n°. 6, p. 578–584, 1981.

[24] D. H. Warren, R. B. Welch et T. J. McCarthy, “The role of visual-auditory “compellingness” in the ventriloquism effect : Implications for transitivity among the spatial senses,” *Perception & Psychophysics*, vol. 30, n°. 6, p. 557–564, 1981.

[25] J. Lewald et R. Guski, “Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli,” *Cognitive brain research*, vol. 16, n°. 3, p. 468–478, 2003.

[26] M. T. Wallace, G. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan et J. A. Schirillo, “Unifying multisensory signals across time and space,” *Experimental Brain Research*, vol. 158, n°. 2, p. 252–258, 2004.

[27] C. R. André, E. Corteel, J.-J. Embrechts, J. G. Verly et B. F. Katz, “Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3d video and wave field synthesis,” *International journal of human-computer studies*, vol. 72, n°. 1, p. 23–32, 2014.

[28] B. N. Postma et B. F. Katz, “Influence of visual rendering on the acoustic judgements of a theater auralization,” dans *Proceedings of Meetings on Acoustics 173EAA*, vol. 30, n°. 1. ASA, 2017, p. 015008.

[29] M. Schutte, S. D. Ewert et L. Wiegrebe, “The percept of reverberation is not affected by visual room impression in virtual environments,” *The Journal of the Acoustical Society of America*, vol. 145, n°. 3, p. EL229–EL235, 2019.

[30] S. E. Guttman, L. A. Gilroy et R. Blake, “Hearing what the eyes see : Auditory encoding of visual temporal sequences,” *Psychological science*, vol. 16, n°. 3, p. 228–235, 2005.

[31] D. Alais et D. Burr, “The ventriloquist effect results from near-optimal bimodal integration,” *Current biology*, vol. 14, n°. 3, p. 257–262, 2004.

[32] P. W. Anderson et P. Zahorik, “Auditory/visual distance estimation : accuracy and variability,” *Frontiers in psychology*, vol. 5, p. 1097, 2014.

[33] D. Cabrera, D. Lee, M. Yadav et W. L. Martens, “Decay envelope manipulation of room impulse responses : Techniques for auralization and sonification,” dans *Proceedings of Acoustics*, 2011.

[34] AES20-1996 (s2008), “AES recommended practice for professional audio : Subjective evaluation of loudspeakers,” Audio Engineering Society, Rapport technique, 2008.

[35] D. Rudrich *et al.*, “IEM plug-in suite,” *University of Music and Performing Arts, Graz, Austria : Institute of Electronic Music and Acoustics.*, 2019, accessed : 2020-08-22. [En ligne]. Disponible : https://plugins.iem.at/

[36] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[37] E. M. Wenzel, M. Arruda, D. J. Kistler et F. L. Wightman, “Localization using nonindividualized head-related transfer functions,” *J. Acoust. Soc. Am.*, vol. 94, p. 111–123, 1993, https ://doi.org/10.1121/1.407089.

[38] S. M. Kim et W. Choi, “On the externalization of virtual sound images in headphone reproduction : A Wiener filter approach,” *J. Acoust. Soc. Am.*, vol. 117, p. 3657–3665, 2005, https ://doi.org/10.1121/1.1921548.

[39] D. R. Begault et E. M. Wenzel, “Headphone localization of speech,” *Hum. Fac. Erg. Soc.*, vol. 35, p. 361–376, 1993, https ://doi.org/10.1177/001872089303500210.

[40] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz et C. de Boishéraud, “Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis,” *J. Acoust. Soc. Am.*, vol. 141, p. 3678–3688, 2017a, https ://doi.org/10.1121/1.4978612.

[41] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz et C. de Boishéraud, “Improvement of externalization by listener and source movement using a “binauralized” microphone array,” *J. Audio Eng. Soc.*, vol. 65, p. 589–599, 2017b, https ://-doi.org/10.17743/jaes.2017.0018.

[42] D. R. Begault, E. M. Wenzel et M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *J. Audio Eng. Soc.*, vol. 49, p. 904–916, 2001.

[43] D. R. Begault, “Perceptual effects of synthetic reverberation on three-dimensional audio systems,” *J. Audio Eng. Soc.*, vol. 40, p. 895–904, 1992.

[44] M. Kronlachner, “Plug-in suite for mastering the production and playback in surround sound and ambisonics,” *Gold-Awarded Contribution to AES Student Design Competition*, 2014. [En ligne]. Disponible : http://www.matthiaskronlachner.com/?p=2015

[45] D. C. Niehorster, L. Li et M. Lappe, “The accuracy and precision of position and orientation tracking in the htc vive virtual reality system for scientific research,” *i-Perception*, vol. 8, n°. 3, p. 2041669517708205, 2017.

[46] E. C. Poulton et S. Poulton, *Bias in quantifying judgements*. Taylor & Francis, 1989.

[47] S. Zielinski, F. Rumsey et S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, n°. 6, p. 427–451, 2008.

[48] H. Keselman, J. C. Rogan, J. L. Mendoza et L. J. Breen, “Testing the validity conditions of repeated measures f tests.” *Psychological Bulletin*, vol. 87, n°. 3, p. 479, 1980.

[49] K. Weinfurt, “Repeated measures analysis : Anova, manova, and hlm,” *Reading and Understanding More Multivariate Statistics*, 10 2000.

[50] R. D. Cook et S. Weisberg, *Residuals and influence in regression*. New York : Chapman and Hall, 1982.

[51] I. Borg et P. J. Groenen, *Modern multidimensional scaling : Theory and applications*. Springer Science & Business Media, 2005.

[52] J. D. Carroll et J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition,” *Psychometrika*, vol. 35, n°. 3, p. 283–319, 1970.

[53] W. L. Martens et N. Zacharov, “Multidimensional perceptual unfolding of spatially processed speech i : Deriving stimulus space using indscal,” dans *Audio Engineering Society Convention 109*. Audio Engineering Society, 2000.

[54] ISO 3382-1, “Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces,” International Organization for Standardization, Geneva, CH, Standard, 2009.

[55] S. Campanini et A. Farina, “A new audacity feature : room objective acustical parameters calculation module,” 2009.

[56] E. Parizet et V. Koehl, “Application of free sorting tasks to sound quality experiments,” *Applied Acoustics*, vol. 73, n°. 1, p. 61–65, 2012.

3

L'influence de la source sonore sur la perception de l'acoustique d'une salle

Le chapitre précédent suggère une absence d'influence de la vision sur la perception sonore de l'acoustique d'une salle. Dans ce chapitre, nous étudions un autre facteur important de la perception de l'espace. Cette étude explore comment la nature d'une source sonore peut influencer la perception de l'acoustique d'une salle. Pour cela, un test perceptif a été réalisé afin de comparer de multiples espaces sonores selon différentes conditions de source. Les résultats montrent que la nature de la source sonore peut avoir une forte influence sur la perception de l'espace sonore et que l'utilisation de plusieurs sources sonores peut permettre de mettre en lumière différentes dimensions perceptives liées à l'espace sonore. Cette étude a fait l'objet d'une présentation orale à la seconde conférence internationale «Headphone Technology» organisée par l'Audio Engineering Society en août 2019 à San Francisco.

3.1. Introduction

Parmi les nombreuses études ayant étudié la perception de l'espace sonore, seules quelques études ont mentionné l'impact de la source sonore sur la perception de l'espace sans pour autant qu'elle fasse l'objet d'une étude approfondie. En particulier, dans le domaine de l'acoustique des salles de concerts, Khale [1] rapporte que la pièce musicale, en terme de composition ou orchestration, peut significativement influencer certains attributs perceptifs. Dans le domaine de la reproduction audio spatialisée, Guastavino et Katz [2] rapportent que la perception des caractéristiques spatiales et spectrales d'un système de reproduction multicanal dépend du type de source à diffuser (musique, ambiance intérieure, ambiance extérieure). Pour Gabrielson [3], de nombreuses et complexes interactions sont possibles entre les contenus diffusés et le

système de reproduction si bien qu'il est difficile de séparer les effets dûs au système et au contenu sur les résultats d'un test perceptif.

La source sonore a nécessairement un impact sur la perception d'un espace sonore dans la mesure où les détails acoustiques d'un espace ne sont pas perceptibles sans source. Comme une source lumineuse est nécessaire pour illuminer visuellement une architecture, des sources sonores sont nécessaires pour «illuminer», ou exciter, l'espace sonore afin de le rendre perceptible. Les espaces sonores sont rarement excités par un grand nombre de sources différentes dans une large gamme de fréquences, d'amplitudes et de positions et leur propriétés acoustiques ne sont donc pas toujours apparentes. Lorsque l'on étudie la perception de caractéristiques acoustiques propres à un espace sonore, on peut alors se demander si une grande variété de sources est nécessaire afin de couvrir les principaux attributs perceptifs d'un espace. Il est important d'estimer dans quelle mesure une source sonore usuellement utilisée peut influencer ces attributs, c'est à dire d'étudier l'ampleur des variations de la perception d'une source sonore à l'autre.

Dans cette optique, nous avons exploré les structures perceptives multidimensionnelles associées à la perception d'un ensemble d'espaces sonores sous différentes conditions de sources. Cette étude s'inscrit dans le cadre d'une écoute de contenus immersifs utilisant la reproduction binaurale dynamique. Les sources choisies pour le test consistaient en des extraits de voix ou de musique et les espaces sélectionnés ne couvraient pas seulement les salles de concert et les grands volumes privilégiés dans la littérature mais également d'autres types d'espaces et de tailles plus modestes.

3.2. Test perceptif

Un test perceptif a été réalisé afin de vérifier si les dimensions perceptives utilisées par les sujets pour comparer les salles étaient les mêmes selon la source utilisée. Quatre sources sonores ont été utilisées. Le protocole du test consistait en des comparaisons par paires successives. Pour chaque source sonore, douze espaces ont été comparés deux à deux : chaque paire de stimuli était constituée d'un signal sonore monophonique convolué avec deux réponses impulsionnelles spatiales de salles (SRIRs) différentes. Ainsi chaque sujet a effectué 4×66 jugements.

Les SRIRs employées étaient les mêmes que celles utilisées dans le chapitre précédent. Le lecteur peut se référer à la section 2.2.2 pour la description de ces espaces et de la méthode d'acquisition employée.

3.2.1. Sources sonores

Le choix des signaux sonores utilisés dans le test s'est tourné vers des sources susceptibles d'être rencontrées dans une situation d'écoute ordinaire et non vers des sources plus particulières telles que des sinusoïdes ou bruits jugées moins écologiques. Le test perceptif était constitué des sources suivantes :

— Une voix d'homme récitant quatre phrases, extraite de la base de donnée *Harvard Psychoacoustic Sentences* [4] enregistrée par *Acoustical Design Collabo-*

rative Ltd, dans une chambre anéchoïque ;
— Une voix de soprano chantant l'aria *Nella pace del mesto riposo* d'après Maria Stuarda, composée par Gaetani Donizetti et enregistrée dans une chambre anéchoïque.
— L'enregistrement d'un instrument percussif - des bongos - issu de bibliothèque *General Series 6000 Sound FX* incluse dans la sonothèque *Sound Ideas* (numéro 6027-03).
— Une scène de fiction composée de trois sources : une voix féminine, un téléphone qui sonne et des coups portés à une porte. L'enregistrement de voix fut effectué dans un studio d'enregistrement.

Dans les sections suivantes ces sources seront respectivement désignées par les termes « Voix », « Soprano », « Bongos » et « Fiction ». Les durées de ces signaux sonores étaient comprises entre 10 s et 16 s.

Chaque source fût ensuite convoluée avec les 12 SRIRs, donnant ainsi 48 stimuli encodés en ambisonique d'ordre 4. Une égalisation du volume sonore a été effectuée subjectivement par les expérimentateurs avant le test perceptif (en accord avec la recommandation AES [5]), de manière à que le volume sonore perçu reste le même d'un stimulus sonore à l'autre. Les stimuli diffusés étaient échantillonnés à 48 Hz et quantifiés à 24 bits.

3.2.2. Binauralisation et head-tracking

Les stimuli furent diffusé au travers d'un casque audio Sennheiser HD 650. Pour les mêmes raisons évoquées en section 2.2.4, nous avons fait le choix d'utiliser des HRTFs non individualisées. Les signaux ambisoniques furent encodé en signaux binauraux grâce au plug-in VST nommé *Binaural Decoder* [6]. Cet outil utilise des filtres de décodage binaural calculés selon le procédé exposé en annexe C.2 d'après 2702 HRTFs mesurées par une tête artificielle Neumann KU 100 [7].

Le *head-tracking* fût assuré par le dispositif électronique nommé hedrot [8] associé au plug-in de rotation ambisonique *ambix_rotator_o7* [9]. La latence moyenne du système de suivi ainsi mis en place s'élevait à 48.1 ms (SD = 5.3 ms). Cette valeur reste en deçà du seuil de détectabilité fixé à 60 ms pour des auditeurs experts [10].

3.2.3. Procédure

Pour chaque source sonore, les sujets ont effectué des comparaisons par paires entre les 12 espaces sonores, soit 66 évaluations de dissemblance. Ainsi, un total de 4×66 comparaisons furent évaluées par chaque sujet. Pour chaque paire, il fut demandé aux participants de quantifier la différence perçue entre les espaces sonores en déplaçant un curseur sur une échelle de 100 points dont les extrémités étaient étiquetées «identique» (0) et «très différent» (100). Aucune étiquette ou graduation intermédiaire n'étaient présentes afin d'éviter tout biais indésirable [11, 12]. Les sujets avaient la possibilité d'écouter les deux stimuli en boucle et de changer de stimulus sonore en cours de lecture. Le moteur de lecture, l'interface graphique et l'acquisition des données furent réalisés sous le logiciel Max [13].

L'expérimentation consistait en deux sessions de test d'une heure réalisées sur deux jours différents. Avant chaque session, les participants se familiarisaient pendant cinq minutes avec une sélection de stimuli illustrant la diversité des sources et espaces présents dans le test.

Les différentes paires étaient présentées aux sujets selon un ordre aléatoire. L'ordre des stimuli au sein de chaque paire était également aléatoire. Ces ordonnancements aléatoires étaient différentes pour chaque sujet.

3.2.4. Sujets

22 sujets âgés en moyenne de 23 ans ont pris part au test perceptif. Ils étaient tous étudiants du Master Image & Son de l'Université de Brest et étaient donc déjà formés à l'écoute critique en raison des enseignements reçus en prise de son, montage et mixage. Aucun d'entre eux n'a déclaré de perte auditive connue.

3.3. Résultats

3.3.1. Matrices de dissimilarité

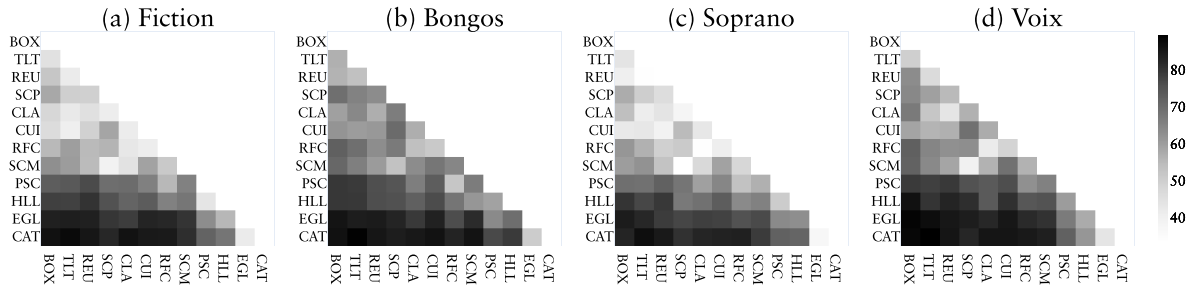


FIGURE 3.1 – Matrices de dissemblances obtenues pour les quatre sources sonores. Chaque carré représente la note de dissemblance entre deux espaces moyennés sur tous les sujets. L'échelle de couleur représente la valeur de cette note. Une couleur plus sombre correspond à une grande dissemblance.

Pour chaque sujet et source sonore, les évaluations des dissemblances par paires fournirent des matrices de dissemblance de dimension N $\times$ N où N est le nombre d'espaces sonores. Étant donné que l'ordre des stimuli dans chaque paire n'était pas différencié, les matrices obtenues sont symétriques. Les matrices correspondant à chaque source ont été moyennées sur l'ensemble des sujets et sont représentées en figure 3.1. Les notes de dissemblances étaient comprises entre 33.28 et 89.29. Le coefficient de corrélation de Pearson entre ces matrices était relativement grand (> 0.902, $p < 0.001$). Cependant, deux des matrices - (b) et (d) - paraissaient plus sombres (contenaient des notes de dissemblance plus importantes). Visuellement, une zone foncée et une zone claire peuvent être distinguées à travers les différentes sources. Ces zones correspondent à deux groupes d'espaces :

— « Cathédrale », « Église », « Hall » et « Piscine ».

— « Box », « Toilettes », « Salle de réunion », « Petite salle de concert », « Salle de classe », « Cuisine », « Réfectoire » et «Salle de concert moyenne ».

Étant donné que cette distinction semble correspondre à un écart en termes de volume, le premier groupe sera par la suite désigné comme étant les «grands volumes» et le second comme étant les «petits volumes». La figure 3.1 montre en effet que les dissemblances sont plus grandes pour des paires qui associent un grand volume et un petit volume que pour des paires dont les espaces sont uniquement compris dans un des deux groupes.

3.3.2. Analyse de la variance

Pour quantifier l'influence de la source sonore sur les dissemblance perçues, une ANOVA à mesures répétées fût réalisée selon les facteurs suivantes : sources sonores (4 niveaux) × paires d'espaces (66 niveaux). Les résultats ont été normalisés par rapport à la moyenne et à l'écart-type en utilisant la normalisation du score z [14].

Un test de Shapiro-Wilk de normalité des résidus a été effectué sur chaque cellule à un niveau de signification de 5% et a rejeté l'hypothèse nulle d'une distribution normale pour 59 cellules seulement sur 264, soit 22.3% des cellules. La plupart des études statistiques indiquent que l'ANOVA peut être robuste à ce genre de violation [15], surtout si la taille de l'échantillon est supérieure à 15 observations par cellule [16]. Avec 22 observations par cellule, nous avons considéré que l'utilisation de l'ANOVA pour l'analyse statistique était toujours légitime.

L'examen de chaque cellule n'a pas révélé de valeurs aberrantes d'après le critère de l'écart absolu à la médiane [17]. De plus, le test de sphéricité de Mauchly a indiqué que l'hypothèse de sphéricité était remplie pour chaque variable indépendante ainsi que pour leur interaction.

Facteur	df	SS	MS	F	p	$\eta^2(\%)$
S	3	76812	25604	68.9	<0.001	6.4
P	65	1044229	16065	89.4	<0.001	87.4
S * P	195	73453	377	4.6	<0.001	6.2

Tableau 3.1 – Résultats issus de l'ANOVA. S : sources, P : paires, SS : somme des carrés, MS : moyenne des carrés, F : valeur f, p : valeur p, η^2 : proportion de variance expliquée.

Le tableau 3.1 montre que les deux variables indépendantes ont eu une influence significative sur les notes de dissemblance et qu'il existe une interaction significative entre ces deux variables. Concernant la taille d'effet, notons que la majeure partie de la variances des résultats est expliquée par les paires d'espaces sonores. La variable S et l'interaction S*P représentent respectivement 6.4% et 6.2% de la variance, ce qui signifie que seulement 12.6% de la variance peut être expliquée par des facteurs liés à la source.

Cependant, il est probable que les effets liés à la source soient atténués par l'important contraste entre les grands et les petits volumes. Le tableau 3.2 montre le pour-

	Paires d'espaces considérées			
Facteur	Toutes	Mixtes	Grands	Petits
S	6.4	8.0	9.1	39.1
P	87.4	85.8	80.7	51.1
S * P	6.2	6.2	10.2	9.8

Tableau 3.2 – Pourcentage de variance (η^2) expliquée par chaque facteur en considérant toutes les paires (Toutes), les paires comprenant des grands volumes (Grands), Les paires comprenant des petits volumes (Petites) et les paires constituées d'un grand et d'un petit volume (Mixtes). S : sources, P : paires.

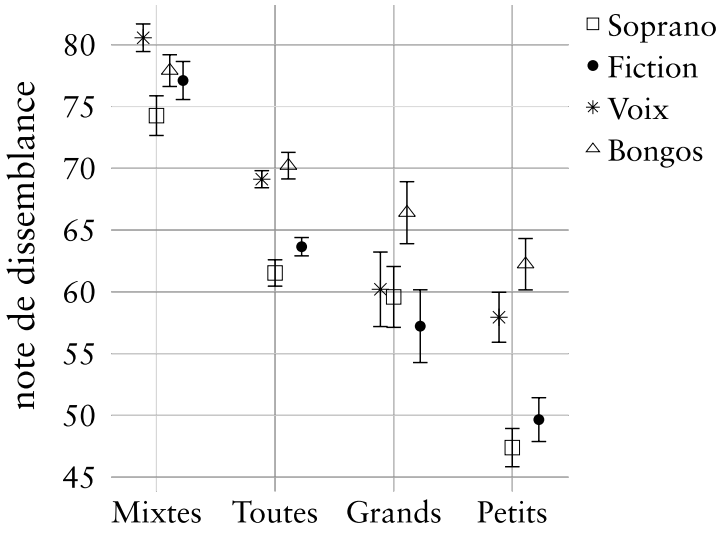


FIGURE 3.2 – Notes de dissemblance moyennes avec les intervalles de confiance à 95% associés pour les quatre sources et les quatre ensembles de données. «Toutes» : toutes les paires considérées, «Grands» : paires limitées aux grands espaces uniquement, «Petits» : paires limitées aux petits espaces uniquement, «Mixtes» : paires mélangeant petits et grands espaces uniquement.

centage de variance expliquée par chaque facteur pour différents sous-ensembles de paires :

— «Toutes» pour lequel toutes les paires sont considérées ;
— «Mixtes» pour lequel seules les paires composées d'un petit et d'un grand volume sont considérées ;
— «Grands», pour lequel seules les paires impliquant uniquement de grands volumes sont considérées ;
— «Petits», pour lequel seules les paires composées de petits volumes sont considérées.

On distingue dans ce tableau que lorsque l'on considère uniquement les paires de grands volumes, le pourcentage de variance expliqué par la source reste faible. Si l'on considère uniquement les paires de petits espaces, le pourcentage de variance imputable à la source augmente considérablement. Dans ce cas, les facteurs liés à la source (S + S * P) expliquent près de la moitié de la variance (48.9 %). Par conséquent, les sources sonores semblaient avoir une influence beaucoup plus grande sur les scores

de dissemblance lors de la comparaison de petits volumes.

La figure 3.2 représente les notes de dissemblance moyennes avec les intervalles de confiance à 95% associés pour chaque source selon les quatre ensembles de paires. Les différences entre les notes de dissemblance semblent être atténuées lorsque l'on compare des espaces très différents. Plus les notes moyennes de dissemblance sont faibles, plus les différences entre les sources sont importantes.

3.3.3. Analyse multidimensionnelle

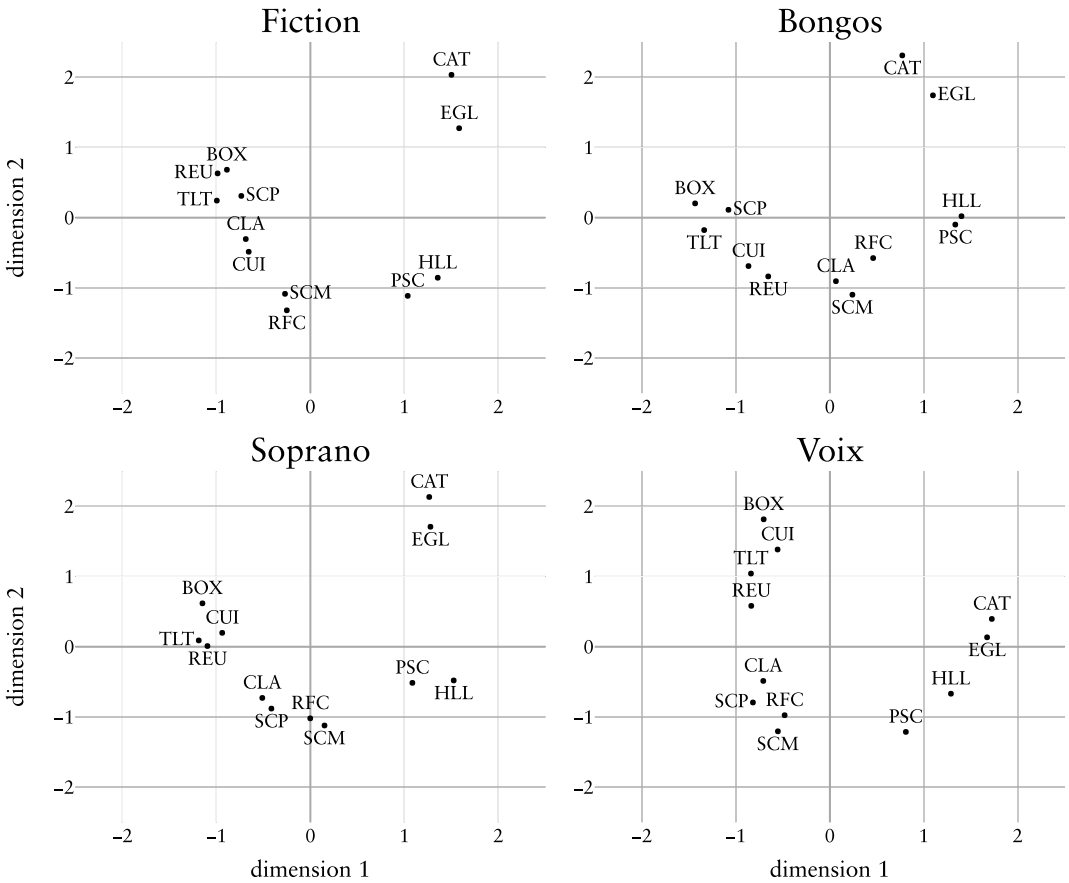


FIGURE 3.3 – Espaces bidimensionnels résultants de l'analyse INDSCAL pour les quatre source sonores.

Dans cette étude, l'objectif n'est pas de trouver toutes les dimensions perceptives en jeu lors de l'écoute des espaces sonores, mais plutôt d'examiner si le choix de la source sonore influence le nombre, la nature et la pondération des principales dimensions perceptives en jeu. Des analyses INDSCAL [18] ont été réalisées sur les matrices de dissemblance pour comparer les structures perceptives en rapport aux quatre sources sonores.

Pour chaque source sonore, l'analyse INDSCAL a révélé que l'espace perceptif associé était bidimensionnel. La figure 3.3 représente les coordonnées des stimuli dans l'espace perceptif. Pour les quatre sources, on constate de nettes différences entre les distributions des stimuli. Dans l'ensemble, les deux groupes d'espaces définis -

petits et grands volumes - sont clairement identifiables bien que le groupe des grands volumes puisse être scindé en deux sous-groupes avec cathédrale et église d'un coté (RT > 3.56 s) et piscine et halle de l'autre (RT < 2.58 s).

Stimulus	Dimension 1	Dimension 2
Bongos	32.9	30.7
Soprano	52.4	23.3
Fiction	63.8	11.2
Voix	63.6	9.3

Tableau 3.3 – Pourcentage de variance des notes de dissemblances attribué à chacune des deux dimensions pour les quatre sources.

Le tableau 3.3 rapporte l'importance des deux dimensions pour chaque source sonore. De grandes variations sont présentes le long des deux axes : de 32.9% à 63.8% de variance expliquée pour les premières dimensions et de 9.3% à 30.7% pour les secondes dimensions. Alors que la plupart des différences perçues entre les espaces sonores peuvent être attribuées à la première dimension de «Fiction», «Voix» et «Soprano», les deux dimensions perceptives de «Bongos» semblent avoir la même importance.

		Dimension 1				Dimension 2			
		FI	BO	FE	MA	FI	BO	FE	MA
Dimension 1	FI		0.88	0.969	0.986	-	-	-	-
	BO			0.931	0.808	-	-	-	-
	FE				0.918	-	-	-	-
	MA					-	-	-	-
Dimension 2	FI						-	0.864	-
	BO							0.858	-
	FE								-
	MA								

Tableau 3.4 – Coefficient de corrélation de Pearson entre les dimensions perceptives associées aux sources sonores. Seuls les coefficients supérieurs à 0.8 sont rapportés (valeur $p < 0.001$ dans tous les cas). FI : «Fiction», BO : «Bongos», FE : «Soprano», MA : «Voix».

Le tableau 3.4 rassemble les plus grands coefficients de corrélation (> 0.7) obtenus entre les différentes dimensions perceptives. Les premières dimensions sont bien corrélées entre elles, ce qui indique qu'un critère perceptif commun a été utilisé pour distinguer les espaces sonores. De plus, «Bongos», «Fiction» et «Soprano» semblent partager leur deuxième dimension perceptive.

Il est intéressant de noter que l'espace perceptif obtenu pour la source «Voix» est visuellement assez similaire à celui obtenu pour la source sonore du chapitre précédent

- désignée «Fantaisie» - dans lequel les mêmes SRIRs ont été employées. De plus, les quantités de variance expliquées par leurs deux dimensions sont similaires : 75.7% et 63.6% de variance expliquée pour la première dimension de «Voix» et «Fantaisie» respectivement ; 9.6% et 9.3% de variance expliquée pour la seconde dimension respectivement.

3.3.4. Caractérisations objectives des dimensions perceptives

		$T_{30,mid}$	$\log(T_{30,mid})$	L_J	$IACC_{E,mid}$	T_s	D_{50}
Dimension 1	FI	0.850	0.975	-	-	-	-
	BO	-	0.84	-	-	-	-
	FE	0.771	0.934	-	-	-	-
	MA	0.895	0.974	-	0.737	-	-
Dimension 2	FI	-	-	-0.734	-	-	-
	BO	0.867	0.748	-	0.774	-	-
	FE	0.728	-	-	0.777	-	-
	MA	-	-	-	-	0.894	0.762

Tableau 3.5 – Coefficient de corrélation de Pearson entre les dimensions perceptives associées aux sources sonores et les mesures des paramètres d'acoustiques de salles. Seules les coefficients supérieurs à 0.7 sont rapportés (valeur $p < 0.001$ dans tous les cas). FI : «Fiction», BO : «Bongos», FE : «Soprano», MA : «Voix».

Dans cette section, les dimensions perceptives révélées par MDS sont comparées aux mesures acoustiques afin de pouvoir les caractériser. Vingt-deux mesures de paramètres acoustiques, dont des mesures définies dans la norme ISO 3382-1 [19], ont été dérivées de chacune des douze SRIRs. Ces mesures ont été calculées aux fréquences médiums et sur l'ensemble du spectre des SRIRs. La plupart des paramètres ont été calculés avec le logiciel AURORA [20].

Le tableau 3.5 rassemble les plus grands coefficients de corrélation (> 0.7) obtenus entre les dimensions perceptives des sources sonores et les mesures acoustiques. Les premières dimensions se sont avérées être bien corrélées aux temps de réverbération, bien que ce soit le logarithme des temps de réverbération qui apparaisse comme le meilleur prédicteur.

Alors que les secondes dimensions de «Fiction», «Bongos» et «Soprano» étaient corrélées entre elles, aucune corrélation commune avec les dimensions physiques n'a été trouvées. Cependant, certaines de ces dimensions se sont révélées être corrélées à des impressions spatiales telles que le coefficient de corrélation croisée interaural de l'énergie précoce ($IACC_{E,mid}$) ou la force sonore latérale tardive (L_J). Par exemple, la seconde dimension perceptive de «Bongos» est corrélée à $IACC_{E,mid}$ avec un coefficient de corrélation de Pearson de 0.774. L'$IACC_{E,mid}$ est couramment utilisé pour caractériser la largeur de source apparente tandis que L_J est lié à la sensation d'enveloppement [19].

Des mesures de clarté telles que T_s et D_{50} semblent expliquer la seconde dimension

perceptive de «Voix» avec des coefficients de corrélation de Pearson de 0.894 et 0.762 respectivement.

3.4. Discussion

Les résultats de cette expérience mettent en évidence l'influence significative de la source sonore sur la perception de la réverbération. Bien que les effets liés à la source aient été atténués par le grand contraste entre les grands et les petits espaces, l'influence de la source sonore est mise en valeur lorsque l'on considère uniquement les petits espaces. Ce phénomène n'a pas pu être mis en évidence en considérant seulement de grands volumes. En effet, le pourcentage de variance expliqué par la source reste faible lorsque l'on considère ce groupe d'espaces. Les dissemblances entre espaces devaient être trop importantes dans ce cas, comme en attestent les valeurs de temps de réverbération comprises entre 1.9 s et 6.54 s. Les différences entre espaces sont tellement importantes qu'elles ont réduit l'influence de la source.

Ainsi, suite à cette expérience, deux hypothèses sont possibles : les facteurs dépendants des propriétés de la source deviennent prédominants lorsque 1) le temps de réverbération est faible ou 2) lorsque le temps de réverbération n'est pas très différent d'un espace sonore à un autre.

Il est intéressant de noter que «Bongos» a conduit aux notes de dissemblance les plus élevées. Les nombreux transitoires compris dans cet instrument percussif peuvent avoir révélé des caractéristiques fréquentielles sur une large partie du spectre audible et ainsi aidé les sujets à discriminer les espaces sonores.

L'influence de la source sonore est également confirmée par les variations de la nature et l'importance de certaines dimensions perceptives d'une source à l'autre. Un critère de perception semble être commun à toutes les sources sonores testées. Cette dimension étant fortement corrélée au temps de réverbération, on peut postuler qu'elle correspondait à la réverbérance. Bien que cette dimension perceptive explique une grande partie de la variance mesurée, son poids peut varier considérablement d'une source à l'autre. Les secondes dimensions des espaces perceptifs étant faiblement corrélées, il n'a pas été possible de conclure qu'un deuxième critère perceptif commun était utilisé pour toutes les sources. Cependant, les impressions spatiales se sont avérées être liées aux secondes dimensions de «Fiction», «Bongos» et «Soprano», tandis que des mesures de clarté étaient bien corrélées à la seconde dimension de «Voix». Des corrélations avec les mesures de clarté ont également été identifiées pour le signal de voix utilisé dans le chapitre précédent. Il semble donc que les dimensions perceptives utilisées lors de la comparaison des stimuli étaient les mêmes pour ces deux sources sonores de même nature.

Dans cette étude, l'objectif n'était pas de découvrir toutes les dimensions perceptives en jeu lors de l'écoute de l'acoustique d'une salle. À cette fin, une grande quantité de stimuli sonores serait nécessaire et d'autres méthodes telles que la tâche de classification libre seraient plus commode car la comparaison par paire reste très chronophage. Nous avons plutôt examiné comment les espaces perceptifs changent d'une source sonore à une autre et à cette fin, une comparaison par paire est une méthode plus précise [21].

3.5. Conclusion

Dans ce chapitre nous avons étudié la perception des espaces sonores selon l'emploi de différentes sources dans un contexte de diffusion binaurale dynamique. L'expérience a montré que la source sonore a une forte influence sur la perception de l'acoustique d'une salle, à l'exception d'un attribut particulier : la réverbérance. Cette caractéristique perceptive a permis de discriminer les espaces sonores quelle que soit la nature de la source employée. Cependant, son importance n'était pas la même d'une source à l'autre. De plus, la nature de la seconde caractéristique perceptive variait d'une source à l'autre. Pour certaines sources les impressions spatiales semblent avoir été utilisées en plus de la réverbérance pour discriminer les espaces sonores. Pour d'autres, ce sont les mesures de clarté qui semblent avoir été utilisées. Cette étude montre donc qu'au-delà de la réverbérance - dont le jugement semble être robuste d'une source sonore à l'autre - plusieurs sources sont nécessaires pour mettre en évidence les caractéristiques perceptives d'un espace sonore. L'usage d'un instrument percussif semble donner plus d'importance aux attributs perceptifs autres que la réverbérance.

L'étude de la perception de l'acoustique des salles nécessite donc une grande variété de sources sonores. Néanmoins, cette contrainte peut augmenter considérablement la durée d'un test perceptif et donc compliquer sa mise en œuvre. Si l'usage de plusieurs sources sonores n'est pas possible faute de temps, cette étude nous informe que les résultats obtenus ne sont valables que pour les sources utilisées et que toute extrapolation des résultats à d'autres types de sources doit être considérée avec précautions.

Bibliographie

[1] E. Kahle, "Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras," Thèse de doctorat, Université du Maine, Le Mans, 1995.

[2] C. Guastavino et B. F. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, vol. 116, n°. 2, p. 1105–1115, 2004.

[3] A. Gabrielsson, "Dimension analyses of perceived sound quality of sound-reproducing systems," *Scandinavian Journal of Psychology*, vol. 20, n°. 1, p. 159–169, 1979.

[4] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, p. 225–246, 1969.

[5] AES20-1996 (s2008), "AES recommended practice for professional audio : Subjective evaluation of loudspeakers," Audio Engineering Society, Rapport technique, 2008.

[6] D. Rudrich *et al.*, "IEM plug-in suite," *University of Music and Performing Arts, Graz, Austria : Institute of Electronic Music and Acoustics.*, 2019, accessed : 2020-08-22. [En ligne]. Disponible : https://plugins.iem.at/

[7] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[8] A. Baskind, J. Messonnier, J. Lyzwa et M. Aussal, “Hedrot–open-source head tracker,” 2017. [En ligne]. Disponible : https://abaskind.github.io/hedrot/

[9] M. Kronlachner, “Plug-in suite for mastering the production and playback in surround sound and ambisonics,” *Gold-Awarded Contribution to AES Student Design Competition*, 2014. [En ligne]. Disponible : http://www.matthiaskronlachner.com/?p=2015

[10] D. S. Brungart, A. J. Kordik et B. D. Simpson, “Effects of headtracker latency in virtual audio displays,” *J. Audio Eng. Soc.*, vol. 54, n°. 1-2, p. 32–44, 2006.

[11] E. C. Poulton et S. Poulton, *Bias in quantifying judgements*. Taylor & Francis, 1989.

[12] S. Zielinski, F. Rumsey et S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, n°. 6, p. 427–451, 2008.

[13] Cycling’74, “Max 8,” 2019. [En ligne]. Disponible : https://cycling74.com

[14] E. Kreyszig, “Advanced engineering mathematics, 10th eddition,” 2009.

[15] K. Weinfurt, “Repeated measures analysis : Anova, manova, and hlm,” *Reading and Understanding More Multivariate Statistics*, 10 2000.

[16] S. B. Green et N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[17] C. Leys, C. Ley, O. Klein, P. Bernard et L. Licata, “Detecting outliers : Do not use standard deviation around the mean, use absolute deviation around the median,” *Journal of Experimental Social Psychology*, vol. 49, n°. 4, p. 764–766, 2013.

[18] J. D. Carroll et J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of “eckart-young” decomposition,” *Psychometrika*, vol. 35, n°. 3, p. 283–319, 1970.

[19] ISO 3382-1, “Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces,” International Organization for Standardization, Geneva, CH, Standard, 2009.

[20] S. Campanini et A. Farina, “A new audacity feature : room objective acustical parameters calculation module,” 2009.

[21] E. Parizet et V. Koehl, “Application of free sorting tasks to sound quality experiments,” *Applied Acoustics*, vol. 73, n°. 1, p. 61–65, 2012.

II

La paramétrisation d'une réponse impulsionnelle spatiale pour l'auralisation de l'acoustique d'une salle

4

Paramétrisations de réponses impulsionnelles spatiales enregistrées

Dans le but de contrôler le rendu sonore de l'acoustique d'une salle selon des attributs perceptifs, il est nécessaire d'identifier les éléments pertinents à modifier dans le signal sonore. Dans ce chapitre, nous passons en revue des méthodes qui ont identifié des éléments pertinents de réponses impulsionnelles de salles qui permettent notamment : la simplification des traitements nécessaires au rendu, une reproduction de l'acoustique sur tout type de dispositif de restitution, une amélioration de la résolution spatiale ou une facilité du contrôle perceptif. Certaines méthodes décrivent les réponses impulsionnelles de salles avec un faible nombre de paramètres tels que des directions d'incidence et des coefficients de diffusivité ou des réflexions spéculaires associées à des pentes de décroissance par bande d'octave. Parmi ces paramètres, certains permettent d'anticiper facilement les conséquences de leur modification sur le rendu de l'acoustique. Néanmoins, l'emploi d'une plus faible quantité de données pour reproduire une acoustique peut introduire des différences significatives avec l'acoustique d'origine, notamment en raison des procédés de décorrélation utilisés pour la reproduction.

Auraliser un espace sonore consiste à créer des fichiers sonores audibles à partir de données numériques liées à ses caractéristiques physiques [1]. Dans cette optique, il est possible de simuler physiquement le phénomène de propagation des ondes sonores en calculant une solution approximative de l'équation d'onde. À cette fin, plusieurs méthodes numériques ont été développées mais possèdent généralement des coûts de calcul prohibitifs. Grâce à la représentation ambisonique, il est également possible de reconstruire une approximation physique précise d'un champ sonore en utilisant une antenne sphérique de microphones. D'autres approches tentent de reproduire un champ sonore réverbéré, en ne prenant en compte que ses aspects les plus pertinents

du point de vue perceptif.

La modélisation présentée en section 1.1, communément employée pour décrire une réponse impulsionnelle de salle, contient plusieurs segments temporels jouant des rôles différents dans la perception d'un espace sonore. Les directions d'incidence des réflexions précoces, leur retard et leur amplitude affectent notamment la localisation du son direct, apportent une coloration spectrale et influencent la clarté, la présence, les dimensions perçues de la salle ou encore la largeur apparente de la source [2]. Dans la partie tardive de la réverbération, les réflexions ne peuvent être perçues individuellement et l'auditeur ne perçoit qu'un champ diffus dont seules les caractéristiques statistiques semblent être pertinentes d'un point de vue perceptif. Le champ diffus affecte notamment la perception de la réverbérance, de la clarté, des dimensions de la pièce, de la distance à la source ou encore de l'enveloppement [2]. Ainsi, il semble nécessaire de reproduire séparément les réflexions spéculaires pour restituer fidèlement la partie précoce de la SRIR et il semble suffisant, dans la partie tardive, de synthétiser un champs diffus ayant des propriétés statistiques similaires au champ sonore mesuré que l'on souhaite reproduire [3]. C'est pourquoi, dans différents moteurs de réverbération artificielle, seules certaines propriétés de la réverbération tardive, pertinentes d'un point de vue perceptif, sont restituées : le temps de réverbération par bande de fréquence, la densité d'écho ou la densité modale [4, 5]. Les caractéristiques qui décrivent le comportement spatial, fréquentiel ou temporel d'éléments d'une SRIR sont nommés paramètres. L'emploi de tels paramètres présente de nombreux avantages parmi lesquels se trouvent :

— La réduction des données. L'ensemble des échantillons d'une réponse impulsionnelle spatiale peut représenter une grande quantité de données et l'emploi de paramètres décrivant des ensembles d'échantillons ou des propriétés statistiques d'échantillons peut réduire considérablement cette quantité.
— La simplification du traitement. Dans le but de générer un champ sonore réverbéré, il est commun d'utiliser la réponse impulsionnelle spatiale d'une salle afin de filtrer une source sonore anéchoïque. Bien qu'il produise un rendu sonore de qualité, ce procédé est coûteux en terme de calcul. Plutôt que d'effectuer une convolution avec la SRIR, il est possible selon la paramétrisation d'utiliser d'autres approches basées par exemple sur l'emploi de réseaux de lignes à retard et de filtrage à réponse impulsionnelle infinie qui permettent des économies de ressources en termes de calcul.
— La flexibilité de la reproduction. La paramétrisation offre la possibilité d'adapter la diffusion d'un champ sonore réverbéré à tout dispositif de restitution. Les paramètres peuvent par exemple représenter des caractéristiques spatiales telles que les directions d'incidence de réflexions sonores, la corrélation interaurale, l'énergie du champ diffus, *etc*... La spatialisation de la ou des sources sonores selon la SRIR peut alors être réalisée en cohérence avec le système sonore à disposition.
— L'amélioration de la résolution spatiale. Du fait d'un nombre trop faible de microphones employés pour la mesure d'une SRIR, l'adaptation des signaux captés à un système de restitution constitué d'un grand nombre de haut-parleurs n'est pas toujours satisfaisante. Une étape d'analyse des directions

d'incidence des réflexions peut permettre d'établir une représentation spatiale plus précise que celle décrite par les signaux de mesures. De plus, la partie diffuse d'une SRIR peut être restituée selon un processus stochastiques pour chaque canal de diffusion et ainsi réduire de trop fortes corrélations présentes dans les signaux qui décrivent une SRIR.

— La facilité du contrôle perceptif. Il est parfois désirable de modifier l'acoustique d'une scène sonore réverbérée, dans le but de satisfaire une intention esthétique. Plutôt que d'agir sur chaque échantillon de la SRIR, il paraît plus pertinent de modifier conjointement ses caractéristiques temporelles, fréquentielles et spatiales selon des paramètres décrivant des ensembles d'échantillons. La paramétrisation de la SRIR permet de modifier aisément le rendu de la réverbération selon des critères bas niveaux (position des réflexions, densité d'écho, pentes de décroissance, *etc*...) ou selon des critères de haut niveaux (réverbérance, enveloppement, largeur apparente de source, *etc*...) si l'influence de ces paramètres sur la perception est connue.

Pour réaliser la paramétrisation d'une SRIR, il est nécessaire d'identifier les éléments pertinents d'un point de vue perceptif permettant de caractériser son comportement spatial, fréquentiel et temporel et de les estimer. Dans le but d'établir un état de l'art sur le sujet, ce chapitre recense différents procédés qui ont été établis pour paramétrer une SRIR. Les méthodes d'analyse et de synthèse présentées dans ce chapitre permettent l'extraction de paramètres capables de restituer un espace sonore aux qualités perceptives comparables à celles d'un enregistrement d'origine sur un dispositif de restitution quelconque.

Bien que ce document traite de l'auralisation au casque d'écoute, la plupart des méthodes présentées dans ce chapitre s'appliquent à une diffusion multicanale sur enceintes. Les signaux multicanaux résultants peuvent cependant être diffusés au casque en considérant chaque haut parleur comme une source ponctuelle virtuelle pour laquelle des filtres de binauralisation peuvent être employés. Les procédés de synthèse peuvent également être adaptés à la méthode de rendu binaurale présentée en annexe C.2. Cette adaptation ne sera pas évoquée dans ce chapitre dont l'intérêt porte principalement sur l'identification et l'estimation des paramètres de SRIR.

Après avoir passé en revue des méthodes employées pour la paramétrisation de SRIR, nous évoquerons les méthodes paramétriques appliquées à des scènes sonores au sens large puis nous aborderons les limites liées à l'emploi de la décorrélation et présenterons une méthode qui minimise son utilisation.

4.1. Spatial Decomposition Method (SDM)

Cette méthode proposée par Tervo *et al.* [6] s'applique aux signaux issus d'un ensemble de microphones formant une antenne microphonique compacte. Pour que la méthode puisse s'appliquer, il est nécessaire qu'un microphone omnidirectionnel soit situé au centre de l'antenne ou qu'il soit possible de le créer virtuellement d'après les autres microphones.

La SDM analyse la réponse impulsionnelle spatiale dans le domaine temporel, à chaque pas de temps, selon une courte fenêtre glissante. Dans cette fenêtre, le champ

sonore est décomposé en une unique onde plane - supposée correspondre à une réflexion - ayant pour amplitude la pression issue du microphone omnidirectionnel central. Ainsi, la paramétrisation consiste à calculer trois valeurs par échantillon : l'azimut et l'élévation de la réflexion dans l'espace et sa valeur en amplitude.

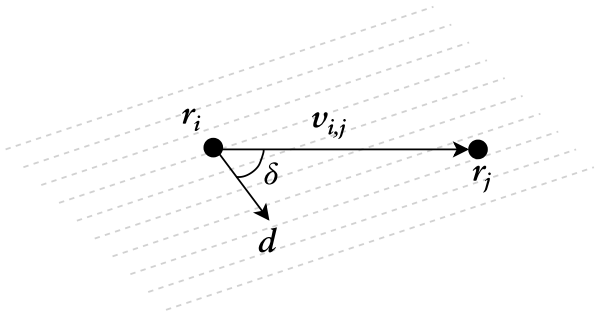


FIGURE 4.1 – Représentation schématique de la propagation d'une onde plane mesurée en deux points de mesure.

Pour calculer la direction d'une onde plane d'après des signaux microphoniques, considérons la propagation d'une onde plane jusqu'aux positions r_i et r_j des microphones i et j de l'antenne, représentée en figure 4.1. La différence de temps d'arrivée à ces positions est donnée par :

$$\tau_{i,j} = \frac{\|\boldsymbol{r}_j - \boldsymbol{r}_i\|_2}{c} \cos(\delta). \tag{4.1}$$

Soit $\boldsymbol{d}$ le vecteur de norme unitaire colinéaire à la direction de propagation, la différence de temps peut s'exprimer :

$$\tau_{i,j} = \frac{1}{c} \|\boldsymbol{r}_j - \boldsymbol{r}_i\|_2 \|\boldsymbol{d}\|_2 \cos(\delta) = \frac{1}{c} (\boldsymbol{r}_j - \boldsymbol{r}_i)^{\mathrm{T}} \boldsymbol{d} = \frac{1}{c} \boldsymbol{v}_{i,j}{}^T \boldsymbol{d} \tag{4.2}$$

où $(.)^T$ désigne l'opérateur de transposition.

La SDM repose sur le calcul du vecteur opposé à $\boldsymbol{d}$, c'est à dire pointant vers la source sonore. Pour ce faire, il est nécessaire d'estimer les différences de temps d'arrivée de l'onde sonore aux microphones. Considérons deux signaux microphoniques de pression sonore p_i et p_j. L'estimation du temps de retard $\hat{\tau}_{i,j}$ entre ces deux signaux est calculée d'après la corrélation croisée ; une mesure de la similarité des signaux selon le décalage temporel de l'un à l'autre. Afin d'estimer le retard avec précision, c'est-à-dire en fraction d'échantillon, le maximum de la fonction de corrélation croisée est approché par une gaussienne dont l'ajustement des paramètres fournit une estimation précise du retard [7]. Soit $R_{ij}(\tau)$ la corrélation croisée calculée selon une fenêtre temporelle de N échantillons. Le retard estimé $\hat{\tau}_{i,j}$ est donné par :

$$\hat{\tau}_{i,j} = \underset{\tau}{\operatorname{argmax}} \{R_{ij}(\tau)\} \tag{4.3}$$

avec

$$R_{ij}(\tau) = \sum_{k=0}^{N} p_i(k) p_j(k+\tau) \,. \tag{4.4}$$

D'après l'équation (4.2), une fois l'estimation du temps de retard $\hat{\tau}_{i,j}$ effectuée, il est possible de calculer la direction d'incidence $\hat{\boldsymbol{d}}$ selon la formule :

$$\hat{\boldsymbol{d}} = -c\, \boldsymbol{v}_{i,j} \hat{\tau}_{i,j} \tag{4.5}$$

Dans le cas d'une antenne de M microphones, une estimation au sens des moindres carrés de la direction d'arrivée $\hat{\boldsymbol{d}}$ à l'instant t est donnée par l'équation suivante [8] :

$$\hat{\boldsymbol{d}}(t) = -c\, \boldsymbol{V}^{\dagger} \boldsymbol{u}(t)\ , \tag{4.6}$$

avec $\boldsymbol{u}$ le vecteur des différences de temps d'arrivée aux microphones,

$$\boldsymbol{u}(t) = [\hat{\tau}_{1,2} \quad \hat{\tau}_{1,3} \quad \dots \quad \hat{\tau}_{M-1,M}]^T, \tag{4.7}$$

$\boldsymbol{V}$ la matrice des différences de positions des microphones exprimées en coordonnées cartésiennes,

$$\boldsymbol{V} = [\boldsymbol{v}_{1,2} \quad \boldsymbol{v}_{1,3} \quad \dots \quad \boldsymbol{v}_{M-1,M}]^T. \tag{4.8}$$

et $\boldsymbol{V}^{\dagger}$ est la pseudo-inverse de $\boldsymbol{V}$.

Le calcul des directions d'incidence peut également être effectué en utilisant les signaux ambisoniques d'ordre 1 issus d'une antenne sphérique de microphones [9]. Dans ce cas, le calcul du vecteur d'intensité acoustique est préféré (cf section 4.2).

Pour une bonne estimation des directions d'incidence, la taille de la fenêtre glissante d'analyse doit être plus grande que le temps nécessaire à une onde sonore pour traverser l'antenne, étant donné que calcul des directions d'incidence résulte de l'intercorrélation des signaux de l'antenne microphonique. De plus, la taille de la fenêtre glissante d'analyse doit être plus grande que la période d'oscillation de la fréquence minimale pour laquelle une direction est estimée. En revanche, la probabilité que plusieurs événements sonores soient présents augmente avec la longueur de la fenêtre d'analyse. Dans ce cas, la direction d'incidence estimée pour l'échantillon considéré peut être erronée.

A mesure que le temps augmente, l'analyse des directions d'incidence tend à fournir des estimations aléatoires en raison des innombrables réflexions contenues dans la réverbération tardive. La méthode crée donc naturellement un processus stochastique en cohérence avec les propriétés spatiales du champ diffus : une probabilité égale de la propagation sonore dans toutes les directions associée à des rapports de phase aléatoire entres les ondes sonores incidentes. La perception de la réverbération tardive étant sensible aux caractéristiques statistiques de sa distribution spatiale, il n'est pas nécessaire de reproduire avec précision les directions d'incidence des réflexions tardives. Néanmoins, lorsque la densité d'écho augmente, la modification rapide des directions d'incidence introduit des changements brusques dans la spatialisation de réflexions, ce qui conduit à un excès de transitoires à large bande dans les réponses impulsionnelles synthétisées. La partie tardive de la réverbération contient alors une énergie trop importante dans les hautes fréquences dont la décroissance ne respecte pas celle de la réponse impulsionnelle spatiale d'origine [6]. Une correction des pentes

de décroissance par bande de fréquences peut être appliquée en cohérence avec les pentes de décroissance du signal omnidirectionnel [10].

Pour adapter la réponse impulsionnelle spatiale sur un dispositif multicanal, chaque réflexion équivalente calculée par pas de temps est affectée à un ou plusieurs canaux de diffusion selon sa direction d'incidence et son amplitude. La SDM fournit alors une réponse impulsionnelle par canal de diffusion et permet de spatialiser un son anéchoïque en le convoluant à celles-ci. La figure 5.3 résume le principe d'analyse et de synthèse de la méthode.

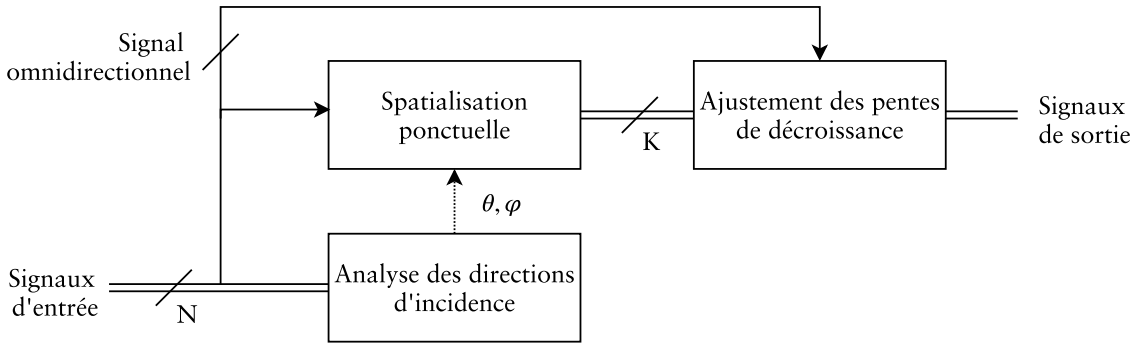


FIGURE 4.2 – Processus d'analyse et de synthèse de la SDM.

4.2. Spatial Impulse Response Rendering (SIRR)

La méthode d'analyse du SIRR est fondée sur l'estimation des directions d'incidence des réflexions et de la diffusivité contenue dans une réponse impulsionnelle spatiale encodée en ambisonique à l'ordre 1 [11]. L'estimation de ces paramètres est réalisée selon une résolution temporelle et fréquentielle proche de celle de l'audition humaine :

— Dans le domaine temporel, plusieurs études sur l'effet de précédence ont identifié des valeurs d'intervalles temporels - nommés seuils d'écho - en dessous desquelles il est difficile de distinguer plusieurs événements sonores. Des valeurs de seuils d'écho comprises entre 5 ms pour des clics et 50 ms pour de la parole sont données dans la littérature [12].

— Dans le domaine fréquentiel, si deux sons sont suffisamment proches en fréquence, un seul événement auditif est perçu (par exemple sous la forme d'un battement). La capacité de séparation entre deux sons distincts en fréquence - la résolution fréquentielle - n'est pas constante sur tout le spectre audible. La largeur des régions fréquentielles dans lesquelles deux sons proches en fréquence fusionnent perceptivement - les bandes critiques - dépend de la fréquence. Ces bandes critiques peuvent être modélisées par des filtres ayant une réponse en fréquence rectangulaire dont la largeur de bande est notée ERB (pour *Equivalent Rectangular Bandwith*) [13]. Avec cette approximation, le contenu fréquentiel peut être analysé en utilisant de tels filtres dont l'ERB est fonction de leur fréquence centrale selon une relation quasi logarithmique.

Ainsi les auteurs de la méthode font l'hypothèse que dans une plage temporelle de quelques millisecondes et dans une bande critique, seule une direction d'incidence peut être détectée. Dans cette région temps-fréquence le champ sonore est alors assimilé à une seule onde plane associée à un champ diffus.

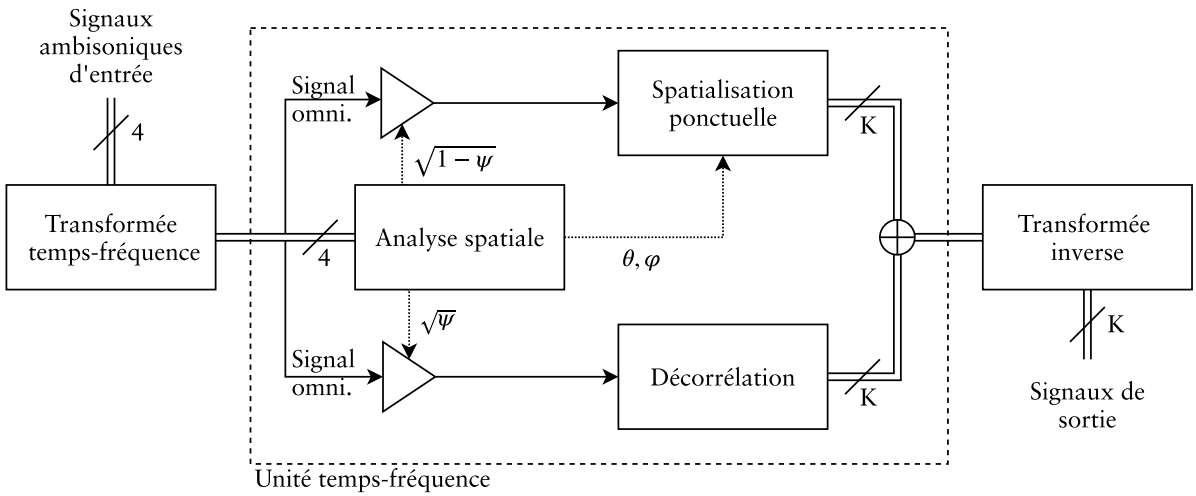


FIGURE 4.3 – Schéma de principe du processus d'analyse et de synthèse du SIRR.

Le processus d'analyse est représenté en figure 4.3. Deux paramètres sont extraits de la réponse impulsionnelle spatiale par région temps-fréquence : une direction d'incidence et un coefficient de diffusivité. Ces valeurs sont obtenues d'après le calcul du vecteur d'intensité instantanée et de la proportion d'énergie sonore oscillante.

Le vecteur d'intensité instantanée est défini comme le produit entre la pression sonore p et le vecteur de vélocité particulaire $\boldsymbol{u}$ [14]. La moyenne temporelle du vecteur d'intensité instantanée caractérise l'intensité du flux acoustique et est nommé intensité active $\boldsymbol{i_a}$. Dans le domaine fréquentiel l'intensité active s'écrit :

$$\boldsymbol{i_a}(\omega) = \Re\{p(\omega)\boldsymbol{u}(\omega)^*\}, \tag{4.9}$$

où * désigne le conjugué complexe et $\Re(.)$ la partie réelle. La direction d'incidence (θ, φ) d'une réflexion équivalente dans la région temps-fréquence considérée, correspond à la direction du vecteur opposé à l'intensité active. Dans le cas d'un encodage idéal de la réponse impulsionnelle en ambisonique, les signaux ambisoniques d'ordre 1 peuvent être utilisés pour calculer ce vecteur. Le vecteur de vélocité particulaire $\boldsymbol{u}$ est lié au gradient de pression du champ sonore donné par les composantes d'ordre 1 et la pression sonore au point de mesure est fournie pas la composante omnidirectionnelle d'ordre 0 [15] :

$$p(\omega) = b_{0,0}(\omega) \text{ et } \boldsymbol{u}(\omega) = -\frac{1}{\rho c\sqrt{3}} \begin{bmatrix} b_{1,1}(\omega) \\ b_{1,-1}(\omega) \\ b_{1,0}(\omega) \end{bmatrix}. \tag{4.10}$$

où les composantes $b_{l,m}$ sont normalisées en N3D.

En pratique, ce calcul peut être effectué dans une bande de fréquence pour laquelle l'estimation des composantes ambisoniques d'ordre 1 est correcte.

L'indice de diffusivité ψ est calculé d'après le rapport entre l'intensité active et la densité d'énergie acoustique E [14] :

$$\psi = 1 - \frac{\|\boldsymbol{i_a}(\omega)/c\|}{E(\omega)} \tag{4.11}$$

avec

$$E(\omega) = \frac{1}{2}\rho\|\boldsymbol{u}(\omega)\|^2 + \frac{1}{2}\frac{1}{\rho c^2}|p(\omega)|^2 \,. \tag{4.12}$$

Le rapport $\|\boldsymbol{i_a}\|/E$ correspond à la vitesse de l'énergie qui se propage au point de mesure, et dont la valeur est comprise entre 0 et c [16]. Lorsque cette vitesse est égale à la célérité c, toute l'énergie au point de mesure se propage. Une valeur plus faible implique qu'une partie de l'énergie oscille localement. La proportion de l'énergie sonore qui se propage est donc donnée par $\|\boldsymbol{i_a}\|/cE$ et ψ représente la proportion d'énergie qui oscille localement. L'indice ψ est nul lorsque que le champ sonore est constitué d'une onde plane car dans ce cas la pression et la vélocité sont reliés par [14] :

$$\|\boldsymbol{u}(\omega)\| = \frac{|p(\omega)|}{\rho.c} \,, \tag{4.13}$$

et la densité d'énergie acoustique E est alors égale à $\|\boldsymbol{i_a}\|/c$. Dans le cas d'un champ parfaitement diffus, la moyenne temporelle du vecteur d'intensité instantanée - le vecteur d'intensité active - est nulle et l'indice ψ est maximal.

Cependant cet indice de diffusivité caractérise mal un champ sonore contenant plusieurs ondes planes. En particulier, en présence de deux sources situées dans des directions opposées et ayant des puissances égales, l'indice de diffusivité correspond à celui d'un champ sonore parfaitement diffus ; quelle que soit la corrélation entre les signaux des sources. Certaines réflexions spéculaires peuvent alors être incluses dans la composante diffuse [17].

Pour restituer la réponse impulsionnelle spatiale sur un dispositif d'écoute, une fois la direction d'incidence et la diffusivité calculées par région temps-fréquence, des réponses impulsionnelles sont synthétisées pour K canaux de diffusion. La synthèse s'effectue en employant la même décomposition temps-fréquence que celle utilisée pour l'estimation des paramètres. Le signal omnidirectionnel issu de la mesure est le seul contenu spatialisé sur le dispositif. L'indice de diffusivité pondère l'amplitude des composantes directionnelles et diffuses de manière à conserver l'énergie originale du signal omnidirectionnel. Notons $\boldsymbol{y}_{\text{dir}}$ et $\boldsymbol{y}_{\text{diff}}$ les vecteurs de K composantes directives et diffuses des réponses impulsionnelles synthétisées. Pour une fenêtre temporelle, ces vecteurs s'écrivent dans le domaine fréquentiel :

$$\boldsymbol{y}_{\text{dir}}(\omega) = \sqrt{1-\psi}\boldsymbol{g}(\theta,\varphi)p(\omega) \tag{4.14}$$

et

$$\boldsymbol{y}_{\text{diff}}(\omega) = \sqrt{\frac{\psi}{K}}\mathcal{D}\left[\boldsymbol{1}_K p(\omega)\right] \tag{4.15}$$

où $\boldsymbol{g}(\theta, \varphi)$ correspond aux K gains appliqués au signal omnidirectionnel pour le spatialiser ponctuellement sur le dispositif, d'après l'algorithme *Vector-Based Amplitude Panning* (VBAP). $\boldsymbol{1}_K$ désigne la copie du signal omnidirectionnel sur les K canaux et $\mathcal{D}[.]$ l'opération de décorrélation de ce signal sur les canaux. La décorrélation peut être effectuée en appliquant des transformations aléatoires du spectre de phase dans chaque fenêtre temporelle ou par convolution, en utilisant de courtes salves de bruit à décroissance exponentielle ayant un spectre d'amplitude plat [15].

Les composantes directives et diffuses sont ensuite sommées et une transformée temps-fréquence inverse est appliquée afin de générer une réponse impulsionnelle par canal de diffusion. La SRIR mesurée est adaptée au dispositif d'enceintes et est opérationnelle pour effectuer la convolution d'un son anéchoïque à spatialiser.

4.3. Higher-Order Spatial Impulse Response Rendering (HO-SIRR)

La méthode présentée dans la section précédente utilise les signaux ambisoniques jusqu'à l'ordre 1 seulement et ne profite donc pas pleinement de la résolution spatiale offerte par une antenne sphérique de microphones formée par un grand nombre de capteurs. Le HO-SIRR est une extension du SIRR pour traiter des réponses impulsionnelles spatiales d'ordre élevé [18].

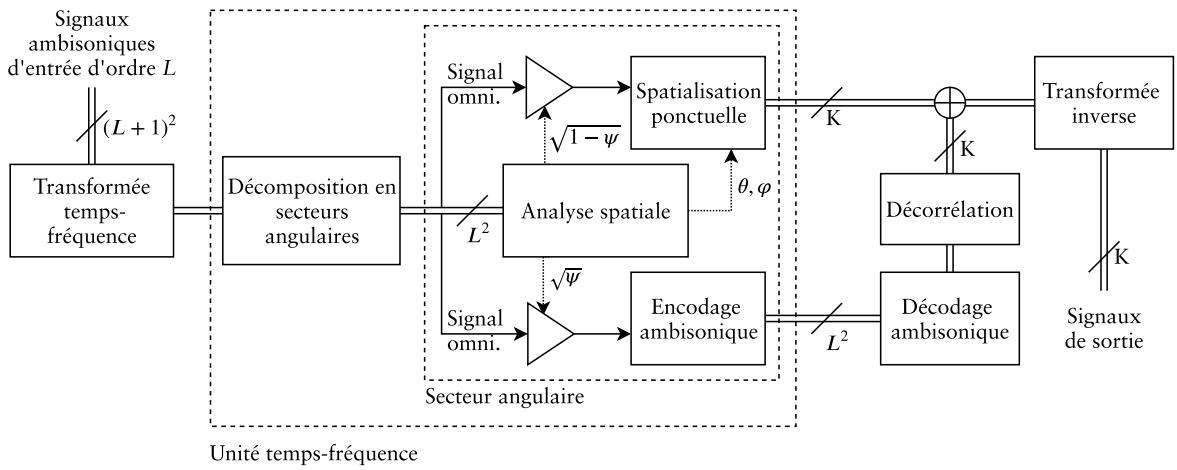


FIGURE 4.4 – Schéma de principe du processus d'analyse et de synthèse du HO-SIRR.

Comme l'illustre la figure 4.4, le principe du traitement des signaux d'ordre supérieur reste le même que celui du SIRR mais repose sur une décomposition préalable du champ sonore en secteurs angulaires. Les directions de ces secteurs sont réparties uniformément sur la sphère et leur nombre est d'autant plus grand que l'ordre ambisonique des signaux d'entrée est important. Dans un secteur angulaire s, la pression

sonore et le vecteur de vitesse particulaire sont calculés, fournissant ainsi un vecteur d'intensité active local $\boldsymbol{i}_{\mathrm{a},s}$:

$$\boldsymbol{i}_{\mathrm{a},s}(\omega) = \Re\left[p_s(\omega)\boldsymbol{u}_s^*(\omega)\right], \quad \text{avec} \quad \begin{bmatrix} p_s(\omega) \\ \boldsymbol{u}_s(\omega) \end{bmatrix} = \boldsymbol{W}_s\boldsymbol{b}(\omega). \tag{4.16}$$

$\boldsymbol{W}_s$ est un matrice de pondération des composantes ambisoniques $\boldsymbol{b}(\omega)$ permettant l'obtention des vecteurs particulaires et de la pression sonore dans le secteur s [19]. Le calcul des paramètres θ, φ et ψ s'effectue de la même manière que précédemment mais leur estimation est dans ce cas plus robuste car moins sensible aux potentiels interférences liées aux contributions de réflexions issues d'autres secteurs de l'espace. Cependant, la diffusivité mesurée pour un secteur angulaire particulier peut également être maximale en présence de deux sources opposées et de même énergie ; bien que cette configuration soit moins probable que dans le cas précédent [17].

Les signaux sonores de chaque secteur angulaire sont ensuite spatialisés ponctuellement grâce aux gains $\boldsymbol{g}$ calculés d'après l'algorithme VBAP et sommés pour former les composantes directives des réponses impulsionnelles synthétisées :

$$\boldsymbol{y}_{\mathrm{dir}}(\omega) = \sum_{s=1}^{S} \sqrt{\frac{1-\psi_s}{S}}\boldsymbol{g}\left(\theta_s, \phi_s\right) p_s(\omega). \tag{4.17}$$

Pour la partie diffuse, les pressions sonore locales sont encodées en ambisonique puis les contributions de chaque secteur sont sommées pour former un vecteur de composantes ambisoniques $\boldsymbol{b}_{\mathrm{diff}}$ représentant le champ diffus :

$$\boldsymbol{b}_{\mathrm{diff}}(\omega) = \sum_{s=1}^{S} \sqrt{\frac{\psi_s}{S}}\boldsymbol{y}_s p_s(\omega), \tag{4.18}$$

où $\boldsymbol{y}_s$ est le vecteur des harmoniques sphériques correspondant à la direction du secteur angulaire s. Les signaux diffus alimentant les haut-parleurs sont ensuite obtenus en décodant le vecteur $\boldsymbol{b}_{\mathrm{diff}}$ dans les directions des haut-parleurs et en décorrélant les signaux résultant dans la fenêtre temporelle considérée :

$$\boldsymbol{y}_{\mathrm{diff}}(\omega) = \mathcal{D}\left[\boldsymbol{D}_K\boldsymbol{b}_{\mathrm{diff}(\omega)}\right] \tag{4.19}$$

où $\boldsymbol{D}_K$ désigne la matrice de décodage ambisonique calculée pour les K haut-parleurs du système de reproduction. Contrairement au SIRR, cette méthode de rendu permet de recréer une réverbération tardive anisotropique ; c'est-à-dire dont les propriétés varient avec la direction. Les composantes diffuses sont en effet obtenues en sommant les signaux sonores extraits de chaque secteur angulaire pondérés par l'indice de diffusivité estimé dans les secteurs correspondants.

4.4. Reverberant Spatial Audio Object (RSAO)

La méthode RSAO modélise une réponse impulsionnelle de salle comme un ensemble de réflexions précoces ponctuelles associées à un filtre représentant la réverbé-

ration tardive [20]. En introduisant une segmentation temporelle entre les réflexions cette modélisation paramétrique s'approche de la description générique d'une réponse impulsionnelle de salle exposée dans la section 1.1 ; contrairement aux méthodes présentées dans les sections précédentes. Cette méthode s'applique à un ensemble de signaux microphoniques issus d'une antenne de microphones compacte ou à des signaux ambisoniques [21].

Le son direct et les réflexions précoces sont caractérisées par quatre paramètres : le retard après le son direct, la direction d'incidence, le niveau et le spectre.

Pour déterminer les temps de retard des réflexions, une sélection des pics d'énergie contenus dans la réponse impulsionnelle est réalisée en utilisant l'algorithme *Dynamic Programming Projected Phase-Slope Algorithm* (DYPSA) [22, 23]. Cet algorithme calcule pour chaque échantillon de la réponse impulsionnelle le retard de groupe moyen dans une fenêtre glissante d'analyse. La figure 4.5 représente une réponse impulsionnelle associée au retard de groupe moyen calculé dans une fenêtre glissante de 128 échantillons. La détection de réflexions est associée au franchissement d'un seuil par le retard de groupe et seules les réflexions supérieures à un seuil d'énergie en dessous du son direct sont conservées. Ces seuils sont fixés de manière heuristique. La figure 4.5 rend compte du caractère arbitraire de la détermination de ces seuils qui permettent de ne prendre en compte qu'une partie des réflexions spéculaires sans pouvoir juger de la pertinence perceptive de cette sélection.

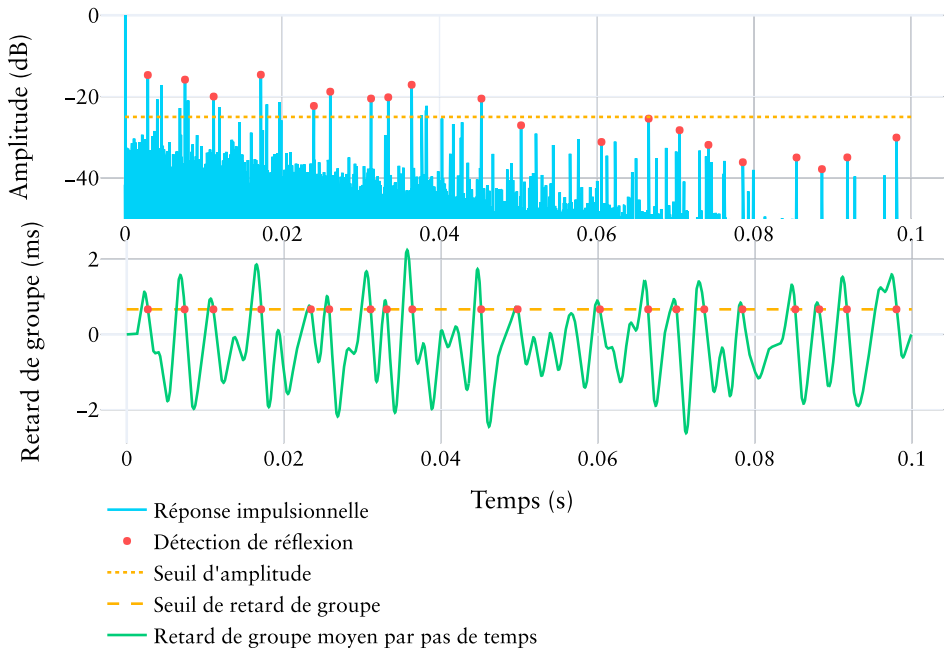


FIGURE 4.5 – Amplitude et retard de groupe d'une réponse impulsionnelle simulée dont le rapport signal-sur-bruit est de 6 dB. Le seuil d'amplitude (-25 dB) et le seuil de retard de groupe (32 échantillons) permettant la sélection des réflexions spéculaires sont indiqués en pointillés.

D'après Begault *et al.* [24] le seuil en amplitude au dessus duquel les réflexions sont perçues dépend du temps de retard. Selon leur étude, pour un stimulus de voix, les réflexions sont inaudibles lorsque leur amplitude est inférieure de 21 dB par rapport au son direct pour un retard de 3 ms et inférieur de 30 dB pour un retard de 15 à

30 ms. Cependant ces seuils dépendent du signal réverbéré et sont 5 à 8 dB plus faibles pour des salves de bruit. Les seuils de détection dépendent également de la direction d'incidence des réflexions et sont d'autant plus faibles que la différence latérale avec le son direct est élevée.

Par ailleurs la sélection des réflexions ne prend pas en compte l'effet de précédence qui implique une dominance de certaines réflexions en fonction de leurs retard, formalisée notamment par Morimoto *et al.* [25].

Les directions d'incidence sont déterminées en estimant la direction contenant le maximum d'énergie aux instants déterminés par l'analyse du retard de groupe. L'évaluation de la répartition énergétique du champ sonore dans l'espace est effectuée par formation de faisceaux (*beamforming*). Considérons un vecteur $\boldsymbol{x}$ constitué des signaux issus de M microphones à une fréquence de pulsation ω, la valeur fréquentielle $d(\omega, \theta, \varphi)$ issue d'une formation de faisceau dans la direction θ, φ est donnée par :

$$d(\omega, \theta, \varphi) = \boldsymbol{w}(\omega, \theta, \varphi)^H \boldsymbol{x}(\omega) \tag{4.20}$$

où $\boldsymbol{w}(\omega)$ est un vecteur de pondération des signaux microphoniques et $(.)^H$ désigne l'opérateur adjoint. Dans le cas d'une formation de faisceau de type *delay-and-sum* (DSB) - adoptée par la méthode - ce vecteur s'écrit [26] :

$$\boldsymbol{w}(\omega, \theta, \varphi) = \frac{1}{M} \boldsymbol{a}(\omega, \theta, \varphi) \tag{4.21}$$

avec $\boldsymbol{a}(\omega, \theta, \varphi)$ le vecteur évalué à la pulsation ω contenant les fonctions de transfert entre une onde plane issue de la direction (θ, φ) et chaque microphone. Pour analyser le contenu énergétique du champ sonore dans S directions régulièrement réparties dans l'espace, le vecteur $\boldsymbol{d}(\omega) = [d(\omega, \theta_1, \varphi_1) \ldots d(\omega, \theta_S, \varphi_S)]^T$ des valeurs fréquentielles issues des S formations de faisceau, peut se calculer de la manière suivante :

$$\boldsymbol{d}(\omega) = \mathbf{W}(\omega)^H \boldsymbol{x}(\omega) \tag{4.22}$$

avec

$$\mathbf{W}(\omega) = \begin{pmatrix} \boldsymbol{w}(\omega, \theta_1, \varphi_1) \\ \vdots \\ \boldsymbol{w}(\omega, \theta_S, \varphi_S) \end{pmatrix}. \tag{4.23}$$

Nous appellerons «carte d'énergie», la représentation du module au carré du vecteur $\boldsymbol{d}$ selon l'azimut et l'élévation de S directions uniformément réparties dans l'espace.

Une fois déterminée la direction de l'espace contenant le maximum d'énergie à un instant donnée, l'enveloppe du spectre du signal issue de la formation de faisceau dans cette direction est approchée par 8 coefficients LPC (*Linear Predictive Coding*) [27].

Concernant la réverbération tardive, celle-ci est modélisée comme un processus gaussien dont l'intensité sonore possède une décroissance exponentielle. Quatre paramètres sont extraits d'une des réponses impulsionnelles de l'antenne pour le caractériser :

— le temps de mélange, estimé d'après les dimensions de la salle dans laquelle a été effectuée la mesure [28] ou d'après un calcul de densité d'échos [29] ;
— une pente d'apparition du champ diffus, correspondant à une pente croissante débutant de la première réflexion jusqu'au temps de mélange. Ceci permet une transition avec la partie précoce en introduisant de l'énergie diffuse avant la réverbération tardive ;
— les niveaux au temps de mélange et les durées de décroissance de la réverbération tardive par bande de fréquence.

La réverbération tardive est restituée au moyen d'un filtre à réponse impulsionnelle finie monodimensionnel. Les copies de ce filtre pour chaque canal de diffusion sont ensuite décorrélées entre elles en employant des filtres passe-tout à phase aléatoire. Les procédés d'analyse et de synthèse décrits dans cette section sont résumés dans la figure 4.6.

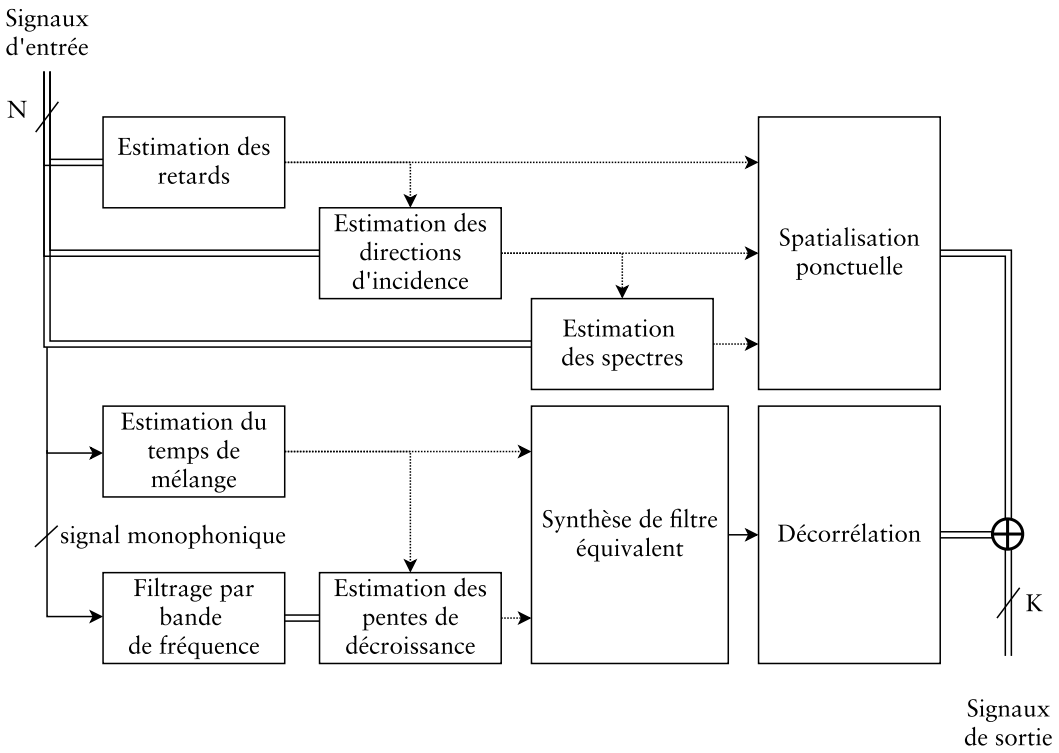


FIGURE 4.6 – Schéma de principe du processus d'analyse et de synthèse de la méthode RSAO.

Coleman *et al.* [21] mettent en évidence la difficulté d'estimer le mixing time et le niveau de la réverbération tardive ainsi que de concevoir des filtres de décorrélation qui n'ont pas d'effet spectral significatif sur la décroissance de la réverbération. Pour plusieurs critères objectifs tels que le RT, l'EDT, le C_{50} et l'IACC, les réverbérations synthétisées selon cette paramétrisation fournissent des valeurs comprises dans les bornes des seuils de discrimination (JND) référencés par la norme ISO 3382-1 [30]. Cette paramétrisation ne permet cependant pas de reproduire des SRIR contenant des échos tardifs ou une réverbération tardive anisotrope car ces cas de figure ne sont pas couverts par la modélisation employée de la SRIR.

4.5. Synthèse paramétrique par Sound Field Analysis (SFA)

Cette méthode paramétrique proposée par Stade *et al.* [31, 32] utilise les signaux ambisonique d'une SRIR et la mesure binaurale de la réponse impulsionnelle de salle (BRIR). Contrairement aux précédentes, la méthode se concentre exclusivement sur la restitution binaurale de la SRIR. De la même façon que pour la RSAO, l'analyse des réflexions précoces et de la partie tardive de la SRIR sont effectuées selon deux processus distincts : la segmentation temporelle employée pour l'analyse de la partie tardive est différente de celle utilisée pour l'analyse des réflexions spéculaires. En revanche, l'analyse et la synthèse des réflexions spéculaires ne sont pas restreintes à la partie précoce de la réverbération. De même, la composante diffuse est générée sur toute la durée de la réponse impulsionnelle. Ainsi il est possible avec cette méthode de restituer des réflexions spéculaires pouvant avoir lieu dans la partie tardive de la SRIR. Le principe d'analyse et de synthèse de la méthode est représenté figure 4.7.

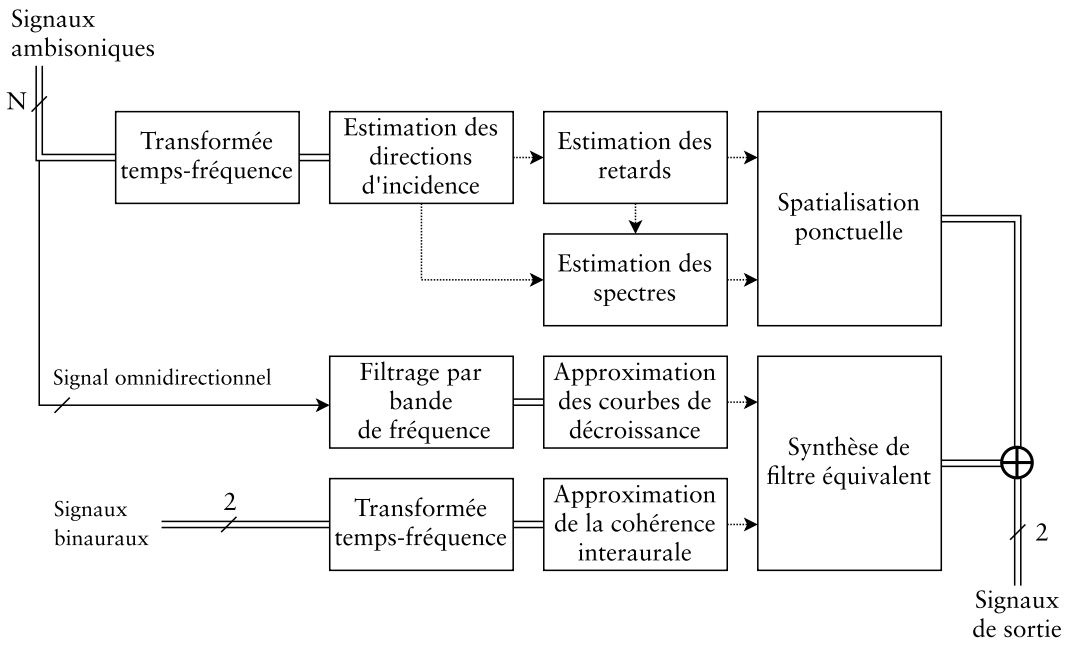


FIGURE 4.7 – Processus d'analyse et de synthèse de la paramétrisation par SFA.

Comme pour la méthode précédente, le retard, la direction, le niveau et le spectre de chaque réflexion spéculaire détectée sont calculés. Pour cela, l'analyse de la SRIR est réalisée par intervalle de temps dans une bande de fréquence donnée. L'estimation des directions d'incidences des réflexions précoces est évaluée en calculant une carte d'énergie d'après l'équation (4.22). Une réflexion spéculaire apparaît alors comme un maximum local d'énergie devant excéder un certain seuil pour être considéré comme une réflexion. Dans ce cas, son spectre est caractérisé par un filtre à réponse impulsionnelle finie qui est calculé d'après la formation de faisceau effectuée dans la direction correspondante.

Pour décrire la partie diffuse, trois paramètres sont extraits de la réponse im-

pulsionnelle binaurale : la corrélation interaurale, l'EDC et l'énergie par bande de fréquence. Un exemple d'implémentation [31] comprend des approximations polynomiales d'ordre 6 de l'EDC obtenues dans 32 bandes de fréquence. Des approximations polynomiales d'ordre 12 de la corrélation croisée intéraurale dans le domaine fréquentiel y sont également calculées par intervalle de temps régulier. Si aucune BRIR n'est disponible, la corrélation interaurale peut être calculée selon des valeurs théoriques déterminées pour un champ parfaitement diffus par Cool *et al.* [33]. Deux bruits gaussiens centrés décorrélés sont ensuite traités pour respecter ces paramètres et la composante diffuse ainsi synthétisée est ajoutée à la partie spéculaire de la réverbération, en veillant à ce que celle-ci n'apparaisse pas avant la première réflexion.

Dans une étude perceptive menée par Stade *et al.* [34], l'ajout d'une composante diffuse dans la partie précoce a produit une similarité plus importante avec une référence mesurée que lorsque des réflexions spéculaires seules étaient considérées. Pour certaines situations, aucune amélioration significative des résultats n'a même été observée avec l'ajout de réflexions spéculaires à une composante diffuse précoce. La présence d'une composante diffuse dans la partie précoce semble donc primordiale.

4.6. Autres méthodes de paramétrisation spatiale

Plusieurs méthodes proposent d'analyser le contenu spatial de scènes sonores afin d'être en mesure d'en améliorer la qualité spatiale ou d'adapter le contenu à des dispositifs de reproduction quelconques. Il est intéressant de les mentionner dans la mesure où les principes d'analyse en jeu s'apparentent à ceux qui peuvent être employés dans la détection de réflexions sonores pour paramétrer une SRIR. Les contenus à identifier ne sont simplement pas restreints à des réflexions mais concernent plus généralement tout type d'évènement sonore ayant un caractère spatial.

Parmi les méthodes de paramétrisation spatiale de scènes sonores se trouve DirAC, une adaptation du SIRR décrit en section 4.2 aux signaux ambisoniques représentant une scène sonore au premier ordre [35]. Une version adaptée au traitement de signaux d'ordre élevé, HO-DirAC, fut également développée [36] .

On compte également le *Spatial Audio Scene Coding* (SASC) [37, 38]. Dans cette méthode un vecteur de coordonnées cartésiennes $\boldsymbol{v}$ est utilisé comme métadonnée de spatialisation associée à chaque bande fréquentielle et segment temporel des signaux analysés. La direction du vecteur - calculé d'après le vecteur de Gerzon [39] - indique la direction d'incidence de la composante principale de la scène sonore et son module est fonction du niveau de diffusivité de la scène : une quantité $(1 - |\boldsymbol{v}|)$ du signal à spatialiser est décorrélée sur l'ensemble des canaux de restitution.

La *High-Angular Plane-wave Expansion* (HARPEX) [40] est une autre méthode de reproduction paramétrique de scène sonore. L'analyse directionnelle est effectuée par région temps-fréquence, dans laquelle le champ sonore est décomposé en deux ondes planes dont les directions constituent les paramètres de spatialisation. A la manière de la méthode SDM, cette méthode n'emploie pas d'estimateur de diffusivité, ni de synthèse de champ diffus.

On trouve également le *COding and Multidirectional Parametrization of Ambisonic Sound Scenes* (COMPASS) [41], qui dans une région temps-fréquence modélise

le champs sonore comme résultant d'une somme de Q composantes directives et d'un champ diffus. Contrairement aux méthodes mentionnées jusqu'à présent, qui n'estiment qu'au plus deux directions d'incidence dans une fenêtre d'analyse, COMPASS procède à l'estimation du nombre d'ondes planes constituant la scène au moyen de l'algorithme SORTE [42]. Les directions d'incidence sont déterminées en analysant la carte d'énergie estimée d'après la méthode MUSIC [43]. Le champ diffus est ensuite calculé comme le champ résiduel de la scène soustraite des composantes directionnelles. Ainsi, l'ensemble des signaux analysés est utilisé pour la reproduction de la scène sonore. Si nécessaire la décorrélation des signaux issus du décodage du champ résiduel est ensuite réalisée.

4.7. La question de la décorrélation

Dans de nombreux cas de figure, la synthèse d'une SRIR ou d'une scène sonore d'après une paramétrisation requiert une étape de décorrélation. Ce traitement est nécessaire lorsque le nombre de signaux synthétisés ou de signaux d'entrée est inférieur à celui requis pour la diffusion multicanale. En effet si un même signal est dupliqué sur différents canaux de diffusion, on observe alors des colorations spectrales et distorsions spatiales en raison d'une corrélation trop importante. Ceci d'autant plus que l'on s'éloigne du centre du dispositif. La décorrélation est alors employée pour synthétiser des signaux à partir des signaux existants qui soient incohérents entre eux et avec les signaux existants.

On retrouve un tel processus dans les méthodes du SRIR, DirAC, SASC pour lesquelles seule le signal sonore omnidirectionnel est spatialisé sur le dispositif de restitution afin de créer un champ diffus. On rencontre également ce traitement pour la méthode RSAO où un filtre équivalent à la partie tardive de la SRIR est utilisé pour créer le champ diffus.

La décorrélation n'est pas seulement utile lorsque l'on souhaite spatialiser un signal monodimensionnel pour créer un champ diffus. Dans le cadre du HO-DirAC ou COMPASS, les signaux décodés sur un dispositif d'enceintes d'après des signaux ambisoniques d'ordre faible sont généralement trop corrélés. On rencontre souvent le cas en basse fréquence qui sont généralement d'ordre faible en raison du faible rayon des antennes sphériques utilisées par rapport au longueurs d'ondes de cette région fréquentielle.

4.7.1. Les différentes méthodes de décorrélation

Pour résoudre les problème liés à la trop forte corrélation des signaux à spatialiser, la décorrélation doit être effectuée tout en évitant d'introduire d'autres artefacts perceptibles. Pour ce faire, différents procédés ont été mis au point :

— Le panning d'amplitude. Dans le cas d'un champ diffus, les directions d'incidence se comportent de manière stochastique en fonction du temps et de la fréquence. La spatialisation rapide dans des directions aléatoires de plusieurs bandes de fréquence d'un signal sonore est une manière de créer un version

décorrélée du signal l'original [44]. Cette approche peut cependant introduire des artefacts audibles à hautes fréquences en raison du changement rapide de localisation des bandes de fréquence qui crée des transitoires.

— Décorrélation par convolution. Une autre technique de décorrélation consiste à concevoir des filtres de décorrélation spécifiques à chaque canal de diffusion. On trouve différents types de filtres dans la littérature : des filtres passe-tout [45], des salves de bruit rectangulaires [46], à décroissance exponentielle ayant un spectre plat [47] ou des filtres de bruit coloré comme proposé dans la paramétrisation par SFA (section 4.5). La diffusion par convolution permet de contrôler la décorrélation dans la mesure où il est possible de modifier la longueur des filtres employés . Cependant des filtres trop longs peuvent influencer le temps de réverbération d'une SRIR. De plus, l'ensemble de ces filtrages perçus au centre du dispositif peut mener à une coloration du signal en raison de la sommation cohérente de ces convolutions fixes.

— Randomisation de phase. Une autre méthode de décorrélation consiste à randomiser la phase du signal à décorréler. La randomisation de phase est effectuée par intervalle de temps régulier en s'assurant que la magnitude du signal décorrélé corresponde à celle d'origine. Les signaux diffusés par les haut-parleurs sont des bruits décorrélés dont l'enveloppe temps-fréquence est similaire au signal à décorréler. Ceci permet d'éviter les artefact spectraux résultant de la sommation cohérente de filtrages fixes mentionnés dans la méthode précédente. Cependant, si la taille de la fenêtre temporelle pour laquelle est effectuée la correction du spectre est trop courte, les basses fréquences peuvent être mal reproduites. Pulkki et Merimaa proposent une méthode hybride dans laquelle une randomisation de phase est effectuée à hautes fréquences et la spatialisation aléatoire du *panning* d'amplitude est effectuée à basses fréquences [48].

— Randomisation du retard de groupe. Boueri et Kyriakakis [49] proposent également d'appliquer différents retards à différentes bandes critiques du signal pour en créer un version décorrélée.

Malheureusement, quelle que soit la méthode de décorrélation, aucune ne peut préserver la qualité du son pour tout type de source sonore. La décorrélation induit notamment une dispersion temporelle des transitoires au détriment de la qualité sonore. Une possibilité pour atténuer les distorsions causées par la décorrélation consiste à commuter le type de décorrélation à appliquer en fonction du type de signal ou d'exclure les transitoires des processus de décorrélation [50].

4.7.2. La minimisation de l'emploi de la décorrélation

Lorsque plusieurs signaux d'entrée sont utilisés pour spatialiser une SRIR ou une scène paramétrée, il est également possible de limiter l'usage de la décorrélation en tirant profit de la décorrélation obtenue après le décodage de ces signaux sur le dispositif de restitution. En effet, Vilkamo *et al.* [51] proposent une méthode permettant d'injecter des signaux décorrélés dans une proportion minimale : uniquement pour compléter les composantes incohérentes manquantes dans les signaux d'entrée.

L'avantage de cette approche est de prendre en considération ce qui est déjà réa-

lisé par les signaux d'entrée en terme de direction dominante et de diffusivité. Les signaux sont d'abord modifiés pour satisfaire les caractéristiques spatiales imposées par la paramétrisation dans une région temps-fréquence considérée. Puis, si ces caractéristiques ne sont pas atteintes, la quantité de signaux décorrélés manquante est ajoutée. Les caractéristiques spatiales à satisfaire sont décrites par une matrice de covariance. La matrice de covariance d'un ensemble de signaux contient les énergies de chacun des signaux ainsi que les relations de phase et d'énergie. La matrice de covariance contient donc tous les indices spatiaux de la scène sonore.

Politis *et al.* [52] ont adapté cette optimisation du rendu sonore pour la restitution binaurale d'une scène paramétrée selon la méthode HO-DirAC. Soit $\boldsymbol{x}$ le vecteur des valeurs fréquentielles des signaux binauraux issus du décodage binaural des signaux ambisoniques d'entrée. La matrice de covariance $\boldsymbol{C_x}$ correspondante est donnée par :

$$\boldsymbol{C_x} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] \,, \tag{4.24}$$

où ^{H} désigne l'opérateur adjoint et $\mathbb{E}[.]$ correspond à l'espérance mathématique donnée ici par la moyenne temporelle.

Soit $\boldsymbol{C}_y$ une matrice de covariance cible contenant les caractéristiques interaurales spécifiées par la paramétrisation. De la même manière qu'une paramétrisation effectuée par la méthode du HO-SIRR présenté en section 4.3, le HO-DirAC estime dans une région temps-fréquence et un secteur angulaire s une direction d'incidence (θ_s, φ_s), un indice de diffusivité ψ_s et une énergie E_s. D'après ces paramètres, il est possible de calculer la matrice de covariance cible des signaux binauraux. Les matrices de covariance $\mathbf{C}_{\text{dir}}^{(s)}$ et $\mathbf{C}_{\text{diff}}^{(s)}$ des composantes directives et diffuses respectivement s'écrivent :

$$\mathbf{C}_{\text{dir}}^{(s)} = (1-\psi_s)E_s\boldsymbol{h}(\theta_s,\varphi_s)\boldsymbol{h}(\theta_s,\varphi_s)^H \tag{4.25}$$

et

$$\mathbf{C}_{\text{diff}}^{(s)} = \psi_s E_s \boldsymbol{U} \tag{4.26}$$

où $\boldsymbol{h}(\theta_s, \varphi_s)$ est le vecteur contenant les HRTFs gauche et droite pour la direction (θ_s, φ_s). $\boldsymbol{U}$ est une matrice permettant de distribuer l'énergie diffuse sur les canaux gauche et droite. Elle s'exprime sous la forme :

$$\boldsymbol{U} = \begin{bmatrix} \alpha & c_{\text{bin}} \\ c_{\text{bin}} & \beta \end{bmatrix} \tag{4.27}$$

où c_{bin} correspond à la corrélation interaurale calculée d'après l'ensemble des d'HRTFs mesurées ou d'après un modèle théorique et α, β correspond à la proportion d'énergie de la composante diffuse du canal gauche et droit.

En faisant l'hypothèse que les composantes diffuses sont décorrélées d'un secteur à l'autre et que les composantes directives sont décorrélées des composantes diffuses, la matrice de covariance cible dans la région temps-fréquence considérée est alors :

$$\boldsymbol{C}_y = \sum_{s=1}^{S} \mathbf{C}_{\text{dir}}^{(s)} + \mathbf{C}_{\text{diff}}^{(s)}. \tag{4.28}$$

Cette matrice contient donc les différences interaurales de niveau, de phase et la corrélation interaurale déterminée par l'analyse paramétrique.

Pour calculer les signaux binauraux $\boldsymbol{y}$ de sortie, l'objectif est de trouver une matrice $\boldsymbol{M}$ et des signaux $\boldsymbol{r}$ de sorte que :

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x} + \boldsymbol{r}\,. \tag{4.29}$$

La transformation $\boldsymbol{M}\boldsymbol{x}$ vise à reproduire le mieux possible les énergies et corrélation imposées par $\boldsymbol{C}_y$ en mélangeant les signaux d'entrée sans employer de décorrélation. L'expression de $\boldsymbol{M}$ est donnée par Valkimo *et al.* [51]. Si la matrice de covariance cible ne peut pas être satisfaite de cette façon, des signaux $\boldsymbol{r}$ issus de la décorrélation des signaux d'entrée sont ajoutés de sorte que :

$$\boldsymbol{C}_r = \boldsymbol{C}_y - \boldsymbol{M}\boldsymbol{C}_x\boldsymbol{M}^H \tag{4.30}$$

Ainsi plutôt que d'ajouter une quantité $\propto \sqrt{\psi}$ de signaux décorrélés, tel que décrit dans le HO-SIRR ou le HO-DirAC, la proportion d'énergie décorrélée est calculée de manière à compléter les composantes diffuses manquantes dans le signal d'entrée. La décorrélation est alors employée dans une moindre mesure. Cette approche peut être employée en utilisant d'autres paramétrisations exposées dans les sections précédentes.

4.8. Conclusion

Le tableau 4.1 rassemble les différents paramètres utilisés par les méthodes présentées dans ce chapitre et les principes d'analyse permettant leur estimation. Selon la méthode, la réponse impulsionelle spatiale est analysée d'après deux approches différentes :

— Une analyse locale du champs sonore à la manière des méthodes de paramétrisation de scènes sonores. Les réponses impulsionnelles spatiales sont décrites dans une fenêtre temporelle ou temps-fréquence restreinte. Pour ces méthodes, le champ sonore peut être localement assimilé à une ou plusieurs directions d'incidence parfois associé à un indice de diffusivité. Figurent parmi ces méthodes le SIRR, HO-SIRR et la SDM. Les procédés d'analyse des directions d'incidence employés par les méthodes de paramétrisation de scènes sonores telles que SASC, HARPEX ou COMPASS pourraient également s'appliquer à la paramétrisation de réponses impulsionnelles spatiales.

— Une analyse générique des éléments de la réponse impulsionnelle spatiale issue de la modélisation générique de la réverbération présentée en section 1.1. La réponse impulsionnelle spatiale est décomposée en un nombre limité de sources secondaires représentant les réflexions spéculaires associées à une composante diffuse. Les procédés d'estimation ont une résolution temporelle et

Méthode	Paramètres	Outils d'analyse
SDM [6]	Direction d'incidence et signal sonore omnidirectionnelle à chaque instant.	Calcul des différences de temps d'arrivée des réflexions aux microphones d'après les valeurs d'intercorrélation des signaux.
SIRR [11]	Direction d'incidence, coefficient de diffusivité et signal sonore omnidirectionnel par région temps-fréquence.	Calcul du vecteur d'intensité active et de la densité d'énergie acoustique.
HO-SIRR [11]	Direction d'incidence, coefficient de diffusivité et signal sonore pour plusieurs secteurs angulaires par région temps-fréquence.	Calcul du vecteur d'intensité active et de la proportion d'énergie oscillante dans chaque secteur angulaire par filtrage spatial.
RSAO [20]	- Réflexions précoces : temps d'arrivée, direction d'incidence, niveau et spectre de chaque réflexion jusqu'au temps de mélange.	Détection de réflexions par franchissement de seuil d'amplitude et du retard de groupe. Estimation de la direction d'incidence et du spectre par formation de faisceaux.
	- Réverbération tardive : temps de mélange, énergie au temps de mélange et durée de décroissance par bande de fréquence	Calcul de la densité d'écho et de l'EDC par bande d'octave.
SFA [32]	- Réflexions spéculaires : temps d'arrivée, direction d'incidence, niveau et spectre de chaque réflexion	Détection de réflexions par franchissement de seuil d'amplitude. Estimation de la direction d'incidence et du spectre par formation de faisceaux.
	- Composante diffuse : courbes de décroissance, énergies par bandes de fréquences. Cohérence interaurale par région temps-fréquence dans la partie tardive.	Approximations polynômiales de l'EDC, calcul de corrélations croisées.

Tableau 4.1 – Comparaisons de méthodes utilisées pour la paramétrisation de réponses impulsionnelles mesurées.

fréquentielle distinctes selon la nature spéculaire ou diffuse des paramètres. Cette approche est adoptée par les méthodes RSAO et SFA. Ces méthodes sont proches et ne diffèrent que dans les régions temporelles destinées à l'estimation du champ diffus et des réflexions spéculaires. Il est intéressant de constater que l'énergie diffuse est présente dans la partie précoce de la réverbération dans les deux cas, contrairement au modèle générique. Les tests perceptifs menés par Stade *et al.* [32] nous informent même que cette propriété est nécessaire en binaural pour reproduire une réverbération proche de la réverbération mesurée.

Les méthodes présentées dans ce chapitre ne sont pas exemptes de biais et produisent parfois des artefacts audibles. En particulier, l'estimateur de diffusivité du SIRR et HO-SIRR peut mener à des estimations de diffusivité erronées et la SDM nécessite une correction haute fréquence des pentes de décroissance des filtres synthétisés. La méthode RSAO détermine de façon arbitraire les réflexions spéculaires à

conserver - au même titre que la paramétrisation par SFA - et ne prend pas en compte l'apparition de réflexions tardives ni la directionalité éventuelle de la réverbération tardive. La paramétrisation par SFA permet de modéliser un champ sonore anisotropique mais pour une direction d'écoute seulement. A cela peuvent s'ajouter les biais liés aux méthodes de calcul du temps de mélange et de formation de faisceaux. Notons de plus que les différentes estimations des directions d'incidence ne prennent pas en compte l'effet de précédence. Cependant d'après ce phénomène une réflexion peut en masquer une autre plus tardive même si cette dernière possède une amplitude plus importante. Enfin, excepté la SDM, toutes ces méthodes doivent employer des étapes de décorrélation qui peuvent introduire des artefacts temporels, spatiaux ou spectraux. C'est pourquoi dans les tests perceptifs menés avec ces paramétrisations - lorsqu'ils existent - on retrouve des différences significatives avec les scènes sonores de référence simulées [6, 18]. La méthode de traitement introduite par Vilkamo *et al.* [51] qui vise à quantifier la proportion de signaux décorrélés manquante par le calcul de leur matrice de covariance peut néanmoins permettre de minimiser ce phénomène.

Dans notre cas d'application, la paramétrisation de réponses impulsionnelles spatiales permet de contrôler le rendu de la réverbération selon des éléments du signal pertinents d'un point de perceptif. En effet, grâce aux paramètres extraits après analyse, il est possible d'appliquer des transformations spatiales telles que des rotations, zoom, filtrage de régions de l'espace, compression, expansion de l'image spatiale [21, 53]. Bien que ces transformations soient possibles dans le domaine ambisonique, la paramétrisation permet de les appliquer avec une résolution spatiale plus importante. De plus, le caractère plus ou moins diffus du champ sonore peut être modifié selon les valeurs des indices de diffusivité ou d'énergie par bande de fréquence. Il est alors possible de modifier le rendu de la réverbération selon des critères bas-niveau pour satisfaire une intention esthétique. Néanmoins, avec une représentation locale par région temps-fréquence, un très grand nombre de paramètres est extrait et il semble difficile d'appréhender l'influence de chacun de ces paramètres sur les impressions spatiales de l'image sonore. Si la paramétrisation est effectuée pour le contrôle perceptif de la réverbération, il semble que la description générique permette une modification plus intelligible du profil de réverbération. La réponse impulsionnelle spatiale est représentée par un nombre limité de réflexions spéculaires se comportant comme des sources secondaires dont le retard, l'amplitude et le spectre sont modifiables ainsi que des pentes de décroissance par bande de fréquence. Le rendu de la réverbération peut être contrôlé en modifiant ces quelques éléments, mais pour être capable de satisfaire des attributs perceptifs tels que ceux définis par le RAQI (*Width, Size, Roughness, Temporal Clarity, Liveliness, Spatial Presence, Intimacy, etc...*), il est nécessaire d'étudier leurs relations. L'étude des liens avec la largeur apparente de source et de l'enveloppement fait notamment l'objet de la partie III. Avant cela, dans les chapitres 5 et 6 suivants, nous étudions les résolutions spatiale puis temporelle et fréquentielle avec lesquelles il est nécessaire de définir ces paramètres.

Bibliographie

[1] M. Vorländer, *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[2] N. Kaplanis, S. Bech, S. H. Jensen et T. van Waterschoot, "Perception of reverberation in small rooms : a literature study," dans *Audio Engineering Society Conference : 55th International Conference : Spatial Audio*. Audio Engineering Society, 2014.

[3] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith et J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, n°. 5, p. 1421–1448, 2012.

[4] M. R. Schroeder, "Natural sounding artificial reverberation," *Journal of the Audio Engineering Society*, vol. 10, n°. 3, p. 219–223, 1962.

[5] J.-M. Jot, L. Cerveau et O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," dans *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[6] S. Tervo, J. Pätynen, A. Kuusinen et T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, n°. 1/2, p. 17–28, 2013.

[7] L. Zhang et X. Wu, "On cross correlation based-discrete time delay estimation," dans *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4. IEEE, 2005, p. iv–981.

[8] T. Pirinen, "Confidence scoring of time delay based direction of arrival estimates and a generalization to difference quantities," *Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication ; 854*, 2009.

[9] M. Frank et F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," *Fortschritte der Akustik, DAGA*, 2016.

[10] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech et T. Lokki, "Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array," *Journal of the Audio Engineering Society*, vol. 63, n°. 11, p. 914–925, 2015.

[11] J. Merimaa et V. Pulkki, "Spatial impulse response rendering i : Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, n°. 12, p. 1115–1127, 2005.

[12] R. Y. Litovsky, H. S. Colburn, W. A. Yost et S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, n°. 4, p. 1633–1654, 1999.

[13] B. R. Glasberg et B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, n°. 1-2, p. 103–138, 1990.

[14] F. J. Fahy, "Sound intensity," 1989.

[15] J. Merimaa *et al.*, *Analysis, synthesis, and perception of spatial sound : binaural localization modeling and multichannel loudspeaker reproduction*. Helsinki University of Technology, 2006.

[16] G. Schiffrer et D. Stanzial, “Energetic properties of acoustic fields,” *The Journal of the Acoustical Society of America*, vol. 96, n°. 6, p. 3645–3653, 1994.

[17] N. Epain et C. T. Jin, “Spherical harmonic signal covariance and sound field diffuseness,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, n°. 10, p. 1796–1807, 2016.

[18] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger et M. Marschall, “Higher-order spatial impulse response rendering : Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution,” *Journal of the Audio Engineering Society*, vol. 68, n°. 5, p. 338–354, 2020.

[19] A. Politis et V. Pulkki, “Acoustic intensity, energy-density and diffuseness estimation in a directionally-constrained region,” *arXiv preprint arXiv :1609.03409*, 2016.

[20] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi et F. Melchior, “Object-based reverberation for spatial audio,” *Journal of the Audio Engineering Society*, vol. 65, n°. 1/2, p. 66–77, 2017.

[21] P. Coleman, A. Franck, D. Menzies et P. J. Jackson, “Object-based reverberation encoding from first-order ambisonic RIRs,” dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[22] M. Brookes, P. A. Naylor et J. Gudnason, “A quantitative assessment of group delay methods for identifying glottal closures in voiced speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n°. 2, p. 456–466, 2006.

[23] L. Remaggi, P. Jackson et P. Coleman, “Estimation of room reflection parameters for a reverberant spatial audio object,” dans *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[24] D. R. Begault, B. U. McClain et M. R. Anderson, “Early reflection thresholds for virtual sound sources,” dans *Proc. 2001 Int. Workshop on Spatial Media*, 2001.

[25] M. Morimoto, K. Nakagawa et K. Iida, “The relation between spatial impression and the law of the first wavefront,” *Applied Acoustics*, vol. 69, n°. 2, p. 132–140, 2008.

[26] B. D. Van Veen et K. M. Buckley, “Beamforming : A versatile approach to spatial filtering,” *IEEE assp magazine*, vol. 5, n°. 2, p. 4–24, 1988.

[27] D. O’Shaughnessy, “Linear predictive coding,” *IEEE potentials*, vol. 7, n°. 1, p. 29–32, 1988.

[28] A. Lindau, L. Kosanke et S. Weinzierl, “Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses,” dans *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[29] J. S. Abel et P. Huang, “A simple, robust measure of reverberation echo density,” dans *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[30] ISO 3382-1, “Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces,” International Organization for Standardization, Geneva, CH, Standard, 2009.

[31] P. Stade et J. M. Arend, "Perceptual evaluation of synthetic late binaural reverberation based on a parametric model," dans *Audio Engineering Society Conference : 2016 AES International Conference on Headphone Technology*. Audio Engineering Society, 2016.

[32] P. Stade, J. M. Arend et C. Pörschmann, "Perceptual evaluation of synthetic early binaural room impulse responses based on a parametric model," dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[33] R. K. Cook, R. Waterhouse, R. Berendt, S. Edelman et M. Thompson Jr, "Measurement of correlation coefficients in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 27, n°. 6, p. 1072–1077, 1955.

[34] P. Stade, J. Arend et C. Pörschmann, "A parametric model for the synthesis of binaural room impulse responses," dans *Proceedings of Meetings on Acoustics 173EAA*, vol. 30. ASA, 2017.

[35] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, n°. 6, p. 503–516, 2007.

[36] V. Pulkki, A. Politis, G. Del Galdo et A. Kuntz, "Parametric spatial audio reproduction with higher-order b-format microphone input," dans *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[37] M. M. Goodwin et J.-M. Jot, "Analysis and synthesis for universal spatial audio coding," dans *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[38] M. Goodwin et J.-M. Jot, "Spatial audio scene coding," dans *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[39] M. A. Gerzon, "General metatheory of auditory localisation," dans *Audio Engineering Society Convention 92*. Audio Engineering Society, 1992.

[40] S. Berge et N. Barrett, "High angular resolution planewave expansion," dans *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, 2010, p. 6–7.

[41] A. Politis, S. Tervo et V. Pulkki, "Compass : Coding and multidirectional parameterization of ambisonic sound scenes," dans *Audio Engineering Society Convention 144*, 04 2018.

[42] Z. He, A. Cichocki, S. Xie et K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, n°. 11, p. 2006–2021, 2010.

[43] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, n°. 3, p. 276–280, 1986.

[44] G. Potard et I. Burnett, "Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays," dans *Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.

[45] M. A. Gerzon, "Signal processing for simulating realistic stereo images," dans *Audio Engineering Society Convention 93*. Audio Engineering Society, 1992.

[46] G. S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, n°. 4, p. 71–87, 1995.

[47] M. J. Hawksford et N. Harris, “Diffuse signal processing and acoustic source characterization for applications in synthetic loudspeaker arrays,” dans *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

[48] V. Pulkki et J. Merimaa, “Spatial impulse response rendering ii : Reproduction of diffuse sound and listening tests,” *Journal of the Audio Engineering Society*, vol. 54, n°. 1/2, p. 3–20, 2006.

[49] M. Bouéri et C. Kyriakakis, “Audio signal decorrelation based on a critical band approach,” dans *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

[50] A. Kuntz, S. Disch, T. Bäckström et J. Robilliard, “The transient steering decorrelator tool in the upcoming mpeg unified speech and audio coding standard,” dans *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[51] J. Vilkamo, T. Bäckström et A. Kuntz, “Optimized covariance domain framework for time–frequency processing of spatial audio,” *Journal of the Audio Engineering Society*, vol. 61, n°. 6, p. 403–411, 2013.

[52] A. Politis, L. McCormack et V. Pulkki, “Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing,” dans *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, p. 379–383.

[53] A. Politis, T. Pihlajamäki et V. Pulkki, “Parametric spatial audio effects,” *York, UK, September*, 2012.

5

L'influence de la résolution spatiale d'une réponse impulsionnelle sur la perception de l'acoustique d'une salle en binaural

Le chapitre précédent a permis de détailler plusieurs méthodes de paramétrisation de réponses impulsionnelles spatiales permettant plusieurs avantages : une économie de données, une économie de calcul et un contrôle du rendu sonore de l'acoustique d'une salle. Dans ce chapitre, nous cherchons à déterminer si il est pertinent d'un point de vue perceptif de reproduire les caractéristiques spatiales d'une SRIR avec précision. Nous considérons ici exclusivement le contexte d'un rendu binaural dynamique non-individualisé. Des scènes sonores générées avec des SRIRs ambisoniques encodées à l'ordre 4 ont été comparées à des scènes générées par des versions d'ordre 1 et 2 de ces SRIRs, avec ou sans réduction de la résolution en élévation (reproduction 3D vs. 2D). Un test perceptif a révélé que : 1) la réduction de l'ordre ambisonique des SRIRs a eu une légère influence ou aucune influence sur la perception selon les salles considérées ; 2) la réduction de la dimensionnalité (3D vs. 2D) des SRIRs n'a entraîné aucune différence perçue pour tous les espaces sonores. Ces résultats suggèrent que, selon la salle et pour une écoute non-individualisée, le nombre de canaux ambisoniques requis pour la reproduction de la réverbération pourrait être considérablement réduit, ce qui permettrait une économie de calcul et d'utilisation de la mémoire.

5.1. Introduction

Dans une contribution récente, Lee *et al.* [1] ont étudié la perception de dégradations spatiales et spectrales liées à l'encodage ambisonique dans le contexte d'un rendu binaural statique non-individualisé. En utilisant plusieurs enregistrements multicanaux, deux approches de rendu binaural ont été comparées. D'une part, un rendu binaural de référence basé sur la convolution des signaux avec les HRIRs mesurées dans la direction des canaux de l'antenne de microphones utilisée. D'autre part, un rendu binaural des enregistrements préalablement encodés en ambisonique aux ordres 1 à 5. Un test perceptif a établi que les stimuli binauraux issus de décodages de signaux ambisoniques d'ordre 2 à 5 n'étaient pas significativement différents entre eux et étaient jugés proches de la référence en termes de qualité spatiale et spectrale. En particulier, pour deux enregistrements (un orchestre et un orgue), les qualités spatiale et spectrale des encodages ambisoniques aux ordres 2 et 3 respectivement n'étaient pas significativement différentes de celles des signaux multicanaux. Les auteurs ont suggéré que le rendu binaural d'un grand nombre de sources sonores proches dans le temps serait moins sujet à une dégradation perçue de la qualité sonore que pour un petit nombre de sources séparées spatialement et temporellement. Il semble donc que pour certaines scènes sonores présentant des caractéristiques spatiales complexes, un ordre ambisonique usuel (< 5) suffise à un rendu binaural satisfaisant. On peut dès lors se demander quel niveau de précision est nécessaire pour reproduire fidèlement une scène sonore particulière - l'acoustique d'une salle - également constituée d'un grand nombre de sources sonores proches dans le temps : des réflexions. En d'autres termes, si il est possible de réduire l'ordre ambisonique ou la dimensionnalité de sa représentation (3D vs. 2D) sans réduire la qualité sonore perçue de manière significative. Ce résultat permettrait de simplifier les calculs nécessaires au rendu de la scène réverbérée. En effet, celui-ci utilise généralement des méthodes basées sur la convolution dont les coûts de calculs sont élevés. Pour permettre une efficacité de calcul et d'utilisation de la mémoire, une partie de la représentation ambisonique d'une SRIR seulement pourrait être prise en compte.

D'une part, on peut se demander quel niveau de précision - en termes d'ordre ambisonique - est nécessaire pour reproduire fidèlement l'acoustique d'une salle. Il serait pertinent de savoir si les différents segments temporels d'une SRIR doivent être aussi bien définis que le son direct. Étant donné que l'intensité sonore évolue selon une décroissance exponentielle et avec une diffusivité croissante, des ordres ambisoniques plus faibles pourraient être suffisants pour reproduire les parties réverbérantes. Cependant, la réduction de l'ordre ambisonique d'une scène sonore peut avoir un impact significatif sur la qualité spatiale et spectrale de sa reproduction. Plusieurs études ont comparé les performances de localisation lors de l'utilisation de différents ordres ambisoniques pour restituer des scènes sonores sur enceintes ou au casque d'écoute [2–6]. Les résultats indiquent que la précision de la localisation dépend de l'ordre ambisonique, les ordres supérieurs entraînant une meilleure localisation. De plus, la réduction de l'ordre ambisonique implique une externalisation réduite [7, 8] et a un impact sur le contenu spectral des signaux ambisoniques [9]. Cependant, les résultats préliminaires de Engel et al. [10] suggèrent que la reproduction d'une SRIR avec un ordre

ambisonique supérieur ou égal à 1 n'améliore pas la perception des espaces sonores - en supposant que le son direct soit rendu avec suffisamment de précision. D'autres études sont cependant nécessaires pour conclure sur la résolution spatiale minimale requise pour reproduire avec précision une SRIR, car une seule pièce, une seule source sonore et un nombre limité de sujets ont été employés dans cette étude [10].

D'autre part, on peut se demander si la résolution spatiale en élévation d'une SRIR peut être réduite sans entraîner une réduction de la précision spatiale perçue. La perception de sources sonores situées en élévation est moins précise par rapport à la localisation dans le plan azimutal. Les sources dans le plan azimutal peuvent être localisées grâce à des différences interaurales de temps et de niveau, alors que la perception de l'élévation repose principalement sur des indices spectraux. Lorsque des HRTFs non-individualisées sont utilisées, ces indices spectraux sont généralement gravement altérés [11] et la perception de l'élévation ne correspond plus à l'expérience d'écoute quotidienne de l'individu. De plus, ces indices peuvent être altérés avec un ordre ambisonique faible [12–14]. Pour une diffusion sur enceintes, Power et al. [4] rapportent que la perception de sources en élévation était altérée avec une réduction de l'ordre ambisonique et qu'un décodage ambisonique d'ordre 3 ne permettait pas la localisation de sources en élévation avec suffisamment de précision par rapport à une spatialisation ponctuelle. Les caractéristiques spectrales au-dessus de 4 kHz sont des indices de localisation pour la perception en élévation, cependant la précision de la reconstruction du champ sonore au-dessus d'une certaine fréquence dépend de l'ordre ambisonique employé (cf annexe B.3). Ainsi, le décodage ambisonique implique des limitations de la perception de l'élévation. En somme, les indices d'élévation étant fortement altérés par le décodage ambisonique et la non-individualisation du rendu binaural, leur restitution dans ce contexte peut ne pas être pertinente pour la reproduction d'une scène réverbérée.

Cette étude traite de l'impact de la représentation d'une SRIR dans le domaine ambisonique sur la perception d'espaces sonores en utilisant un rendu binaural non-individualisé et au maximum une représentation ambisonique d'ordre 4. Un test perceptif a été effectué pour : 1) déterminer l'ordre ambisonique minimal permettant que les différentes parties de la réverbération soient reproduites sans diminution significative de la qualité perçue ; 2) déterminer si la réduction de la résolution en élévation a un impact sur la perception d'une scène réverbérée pour une source sonore dans le plan horizontal.

5.2. Manipulation du champ sonore dans le domaine ambisonique

Pour étudier l'impact sur la perception d'une modification de leur représentation ambisonique, plusieurs SRIRs ont été manipulées en utilisant une représentation mixte : alors que la région temporelle correspondant au son direct est restée inchangée, les régions temporelles correspondant aux réflexions précoces et tardives ont été représentées avec des ordres ambisoniques inférieurs soit avec toutes les composantes

ambisoniques (reproduction 3D), soit en utilisant les composantes sectorielles [1] seulement (reproduction 2D).

5.2.1. Segmentation temporelle

Pour reproduire le son direct et les composantes précoce et tardive d'une SRIR en utilisant différents ordres ambisoniques, il est d'abord nécessaire de trouver les limites temporelles de ces segments.

La région temporelle du son direct

Le temps d'arrivée du son direct était considéré comme le temps correspondant au maximum d'énergie du signal ambisonique d'ordre 0. Nous avons considéré que les réflexions qui se produisaient moins de 5 ms après le son direct n'étaient pas perçues comme des événements distincts mais étaient fusionnées avec le son direct. En effet, dans le cas de stimuli brefs tels que les clics, plusieurs études ont rapporté une valeur minimale du seuil d'écho de 5 ms [15–18]. Cette valeur critique a été choisie, même si les sources sonores utilisées dans le test perceptif n'étaient pas des clics.

Par conséquent, dans les sections suivantes, aucune modification n'a été apportée aux 5 premières millisecondes des SRIRs. Puisque cette région ne contient pas seulement le son direct, elle sera par la suite désignée comme la composante initiale plutôt que son direct.

Détermination du temps de mélange

Le temps de transition entre la région temporelle des réflexions précoces et tardives correspond au temps de mélange. La méthode introduite par Götz et al. [19] a été utilisée pour le déterminer. Cet estimateur effectue une analyse de la diffusivité à partir d'une SRIR mesurée en calculant la variation temporelle du vecteur d'intensité active [20]. Avec cette méthode, l'estimation du temps de mélange prend en compte les propriétés spatio-temporelles du champ sonore et est plus précise que des estimations utilisant un modèle géométrique [21, 22] ou une réponse impulsionnelle de salle monodimensionnelle qui tient uniquement compte des caractéristiques temporelles [22–25].

5.2.2. Reconstruction mixte du champ sonore

La représentation mixte du champ sonore peut être obtenue en tronquant les signaux ambisoniques des composantes précoce et tardive des SRIRs à des ordres plus faibles. Soit L_I, L_E et L_R les ordres ambisoniques des composantes initiale, précoce et tardive, respectivement. Le signal binaural $\hat{x}(k)$ s'écrit alors :

$$\hat{x}(k) = \hat{\boldsymbol{w}}_{L_I}^T(k)\boldsymbol{b}_{L_I}(k) + \hat{\boldsymbol{w}}_{L_E}^T(k)\boldsymbol{b}_{L_E}(k) + \hat{\boldsymbol{w}}_{L_R}^T(k)\boldsymbol{b}_{L_R}(k)\ , \tag{5.1}$$

1. Les harmoniques sphériques sectorielles correspondent aux harmoniques sphériques de degré $|m| = l$. Pour ces harmoniques, quelque soit l'azimut, la phase ne change pas de signe selon l'élévation. Les réflexions sont traitées de la même manière quel que soit le signe de leur élévation.

où $\boldsymbol{b}_{L_I}(k)$, $\boldsymbol{b}_{L_E}(k)$ et $\boldsymbol{b}_{L_R}(k)$ sont les coefficients d'expansion des composantes initiale, précoce et tardive du champ sonore, respectivement, et $\hat{\boldsymbol{w}}_{L_i}(k)$ sont les filtres de rendu optimisés calculés à l'ordre L_i.

5.2.3. Représentation du champ sonore dans le plan

En plus d'évaluer l'influence de l'ordre ambisonique, nous avons cherché à savoir si la réduction de la résolution en élévation a une influence sur la perception de l'espace sonore. À cette fin, les représentations ambisoniques des SRIRs ont été modifiées pour obtenir des SRIRs alternatives sans indices d'élévation : seules les composantes ambisoniques sectorielles ont été retenues. Pour s'assurer que le rendu binaural d'une SRIR et son équivalent 2D transmettent les mêmes informations spectrales, nous avons utilisé un filtre d'égalisation à phase linéaire $g(k)$ calculé de manière à ce que :

$$|g(k)|^2 \sum_{l=0}^{L} \sum_{|m|=l} \left|\hat{w}_{l,m}^{2D}(k)\right|^2 = \sum_{l=0}^{L} \sum_{m=-l}^{l} \left|\hat{w}_{l,m}^{3D}(k)\right|^2 \tag{5.2}$$

Ainsi, les SRIRs 2D ont été convoluées avec un filtre FIR de 64 échantillons pour corriger les artefacts spectraux. Le retard introduit par le filtrage à phase linéaire a été compensé de sorte que le son direct d'une SRIR 2D ait le même temps d'arrivée que pour la SRIR originale. Étant donné la courte durée du filtre à phase linéaire employé, nous avons considéré que la modification induite de la structure temporelle des SRIRs était négligeable.

5.3. Test perceptif

Un test perceptif a été réalisé pour étudier comment le changement de la résolution spatiale d'une SRIR peut influencer la perception de l'espace sonore. À cette fin, les SRIRs ont été manipulées à l'aide des méthodes présentées dans la section précédente. Le test consistait à comparer un stimulus d'ordre mixte à un signal de référence entièrement rendu en binaural à l'aide des filtres de décodage d'ordre 4. Le protocole utilisé était la méthode de «doublement aveugle à triple stimulus et référence dissimulée» recommandée par l'ITU-R BS.1116 [26].

Variable		Nombre
Espaces		5
Sources		2
Ordres ambisoniques mixtes		3
Dimensionalités (2D vs. 3D)		2
Nombre de jugements	=	60

Tableau 5.1 – Variables indépendantes du test.

Un pré-test informel impliquant 8 sujets a permis de définir les variables indépendantes du test et d'estimer l'ordre ambisonique maximum à employer. Le tableau 5.1

Stimuli	Dim	L_I	L_E	L_R
S1	2D	4	1	1
S2	3D	4	1	1
S3	2D	4	2	1
S4	3D	4	2	1
S5	2D	4	2	2
S6	3D	4	2	2
REF	3D	4	4	4

Tableau 5.2 – Résolutions spatiales des stimuli sonores. Le stimulus sonore désigné par REF désigne la référence qui a été comparée aux autres stimuli. Dim : dimensionnalité, L_I : ordre ambisonique de la composante initiale, L_E : ordre ambisonique des premières réflexions, L_R : ordre ambisonique de la réverbération tardive.

indique les variables indépendantes utilisées dans le test perceptif. Dix scènes sonores (deux sources × cinq espaces) ont été déclinées dans six représentations ambisoniques différentes (trois ordres ambisoniques mixtes × deux dimensionalités). Ainsi, un total de 60 comparaisons ont été réalisées par chaque sujet.

Comme indiqué dans le tableau 5.2, la variable d'ordre ambisonique mixte comprenait les configurations suivantes : 1) encodage à l'ordre 1 pour les composantes précoce et tardive, 2) encodage à l'ordre 2 pour les composantes précoce et tardive, 3) encodage à l'ordre 2 pour la composante précoce et à l'ordre 1 pour la composante tardive.

5.3.1. Réponses impulsionnelles de salles

Pour créer les stimuli sonores utilisés dans le test d'écoute, des SRIRs sont été enregistrées dans différents espaces sonores. Ces espaces ont été choisis afin de couvrir un large éventail de salles différentes. Le tableau 5.3 énumère les salles impliquées dans le test perceptif. Les différents temps de réverbération affichés illustrent cette hétérogénéité. Il est intéressant de considérer non seulement les grands espaces mais aussi les petits pour étudier une influence possible du type de salle.

Spaces	Abréviation	V(m^3)	d(m)	RT(s)
Toilettes	TLT	23	0.71	0.41
Cuisine	CUI	81	0.82	0.60
Réfectoire	RFC	360	3.03	0.83
Piscine	PSC	7448	8.62	1.91
Escalier	ESC	333	0.18	3.68

Tableau 5.3 – Espaces sonores utilisés dans le test perceptif. V : Volume de l'espace sonore approché par un parallélépipède, d : distance au mur le plus proche, RT : temps de réverbération.

Les SRIRs ont été mesurées à l'aide de l'antenne sphérique de microphones Eigenmike EM32 [27] et d'une enceinte Genelec 8040. L'enceinte était positionnée à la même distance du microphone (3 m) dans chaque salle. Un signal à balayage si-

nusoïdal logarithmique de 10 secondes a été utilisé pour les mesures. Les réponses impulsionnelles ont ensuite été débruitées selon la procédure décrite par Cabrera et al. [28]. Afin de compenser la réponse en fréquence de l'enceinte dans les mesures de SRIRs, un filtre à réponse impulsionnelle finie de 128 échantillons a été appliqué aux mesures. La correction fréquentielle a été appliquée entre 60 Hz et 16 kHz. Le filtre a été calculé à partir d'une mesure de la réponse impulsionnelle de l'enceinte effectuée dans une chambre anéchoïque avec un microphone omnidirectionnel situé dans l'axe de l'enceinte. Enfin, les réponses impulsionnelles issues des capsules du microphone sphérique ont été converties en ambisonique à l'ordre 4. Les procédés employés pour la mesure, le débruitage et l'encodage des SRIRs sont décrits en détail dans l'annexe A et B.

Espace	TLT	CUI	RFC	PSC	ESC
Temps de mélange (ms)	81	127	84	145	152

Tableau 5.4 – Temps de mélange des espaces sonores utilisés dans l'expérience.

Les cinq SRIRs ont été segmentées en trois régions temporelles (composantes initiale, précoce et tardive), d'après la méthode décrite en section 5.2.1, afin que différents ordres ambisoniques puissent être utilisés pour le rendu binaural. Les temps de mélange estimés des cinq espaces utilisés pour séparer les réflexions précoces de la réverbération tardive figurent dans le tableau 5.4. De même, des versions bidimensionnelles des SRIRs segmentées ont été obtenues d'après la méthode décrite en section 5.2.3 en éliminant les composantes ambisoniques non sectorielles.

5.3.2. Sources sonores

Deux sources sonores ont été utilisées dans le test :

— Une voix masculine récitant *Fantaisie*, un poème de Gérard de Nerval, enregistrée dans une chambre anéchoïque ;

— Un instrument de percussion - un cajon - enregistré dans une chambre anéchoïque ;

Dans les sections suivantes, ces sources sonores seront désignées respectivement par «voix» et «cajon». Les durées des signaux des sources sonores étaient respectivement de 9 s et 11 s. Chaque source sonore a été convoluée avec les SRIRs mesurées pour obtenir cinq scènes ambisoniques d'ordre 4 (une par espace sonore) qui par la suite serviront de référence.

5.3.3. Binauralisation et head-tracking

Les stimuli ont été diffusés au moyen d'un casque Beyerdynamic DT-990 Pro. Les signaux ambisoniques ont été convertis en signaux binauraux en utilisant le plug-in VST *open-source BinauralDecoder* [29] qui intègre des HRTFs dérivés de mesures d'une tête artificielle Neumann KU 100 [30]. Ce décodeur met en œuvre les méthodes de rendu binaural de Schörkhuber *et al.* [31] et Zaunschirm *et al.* [32] dont il est

question en annexe C.2. Ces filtres permettent de conserver une réponse spectrale et une décorrélation interaurale comparables à celles obtenues avec les HRTFs mesurées pour une scène sonore diffuse.

Comme les différentes régions temporelles des SRIRs étaient codées avec des ordres différents, un décodeur binaural a été utilisé pour chaque région temporelle selon l'ordre correspondant. Les signaux décodés ont ensuite été additionnés pour créer les signaux binauraux.

Le head-tracking a été réalisé en utilisant la solution *hedrot* [33] ainsi que le plugin *ambix_rotator_o7* [34]. Lors de l'écoute de stimuli d'ordre mixte utilisant une SRIR 2D, le head-tracking a été effectué selon une rotation en azimut uniquement, de sorte que la réverbération reste dans le plan horizontal.

5.3.4. Procedure

La méthode de doublement aveugle à triple stimulus et référence dissimulée a été choisie pour ce test. Pour chaque jugement, trois stimuli («A», «B» et «X») étaient présentés aux sujets. La référence - décodée d'après une représentation ambisonique d'ordre 4 - était toujours présente sous le stimulus «X». La référence cachée et le stimulus d'ordre mixte étaient également présents, mais attribués de manière aléatoire à «A» et «B» à chaque jugement. Les sujets avaient la possibilité d'écouter les stimuli de façon répétée et de passer de l'un à l'autre à leur gré. Une méthode de commutation quasi-instantanée a été utilisée avec un total de 40 ms pour le fondu de sortie, le changement de lecture et le fondu d'entrée.

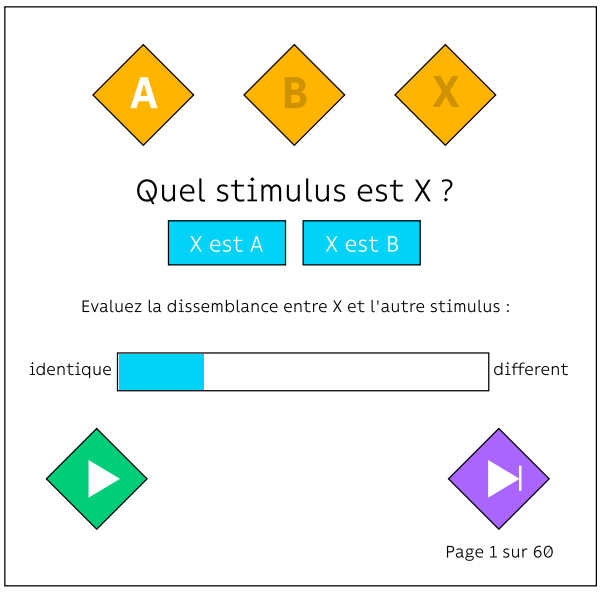


FIGURE 5.1 – Interface graphique du test perceptif.

Pour chaque appréciation, les participants devaient désigner lequel de «A» et «B» était la référence cachée et évaluer la différence entre la référence et l'autre stimulus en déplaçant un curseur sur une échelle continue de 100 points dont les extrémités étaient étiquetées «identique» (0) et «différent» (100). La figure 5.1 montre l'interface

graphique utilisée pour le test. La lecture, les commandes et la saisie des données ont été contrôlées à l'aide de Max 8 [35].

Une égalisation de l'intensité sonore a été effectuée subjectivement par les expérimentateurs avant le test perceptif (conformément à [36]), de sorte que l'intensité sonore perçue reste la même d'une évaluation de dissemblance à l'autre. Le niveau global de lecture a également été fixé subjectivement à une valeur confortable et réaliste qui correspond au niveau d'une expérience d'écoute quotidienne. Le niveau était le même pour tous les sujets.

La recommandation ITU-R BS.1116 suggère d'utiliser une échelle de dégradation à cinq niveaux, néanmoins la quantification de la «gène» employée dans cette échelle ne semblait pas pertinente pour désigner les stimuli d'ordre mixte utilisés. Aucune étiquette intermédiaire ni graduation n'a donc été placée sur l'échelle pour éviter les biais dus à une distribution non linéaire des étiquettes et pour prévenir toute distorsion dans la distribution des données (avec des étiquettes intermédiaires, des groupes de notes peuvent en effet apparaître autour des valeurs où se trouvent les étiquettes [37, 38]). Deux étiquettes ont été utilisées aux extrémités du curseur conformément à la recommandation ITU-R BS.1116 [26].

Avant le début formel du test, une phase d'entraînement a permis aux sujets de se familiariser avec l'environnement de test, le processus de notation et les stimuli d'ordre 1, dans le but d'appréhender l'étendue des différences rencontrées.

Le test consistait en une seule séance d'une heure et demie (pauses comprises). Pendant la durée du test, des changements dans les stratégies de notation utilisées par les sujets ont pu se produire et de la fatigue a pu apparaitre. L'ordre aléatoire de présentation des stimuli permet de réduire ces problèmes [39]. De plus, nous avons tenté de réduire l'effet de la fatigue en demandant aux sujets de faire une pause de 5 minutes tous les 20 jugements.

5.3.5. Sujets

22 sujets ont participé au test d'écoute. Cinq étaient des étudiants du Conservatoire de musique de Rennes, 17 étaient des étudiants en Master du cours Image & Son de l'Université de Brest et étaient donc déjà formés à l'écoute critique en raison des enseignements reçus en prise de son, montage et mixage. Aucun d'entre eux n'a déclaré de perte auditive connue.

5.4. Résultats

Les notes de dissemblance ont été utilisées comme données d'entrée pour l'analyse statistique. Comme l'indique la recommandation ITU-R BS.1116 [26], les notes ont été considérées comme négatives lorsque les sujets ne parvenaient pas à identifier correctement la référence cachée. Lorsqu'on utilise une telle méthode, si la moyenne des notes de dissemblance liées à un stimulus d'ordre mixte est significativement différente de zéro, cela signifie que les sujets ont perçu une différence entre la référence et le stimulus d'ordre mixte considéré.

Comme aucun point d'ancrage intermédiaire n'a été utilisé sur l'échelle de dissemblance, les résultats ont été normalisés par rapport à la moyenne et à l'écart-type en utilisant la normalisation du score z comme recommandé par l'ITU-R BS.1116 [26].

Il est possible que la formation des sujets aux métiers du son d'une part et la formation musicale d'autre part ait eu une influence sur les jugements de dissemblance. N'ayant pas suffisamment de sujets musiciens afin de créer une variable inter-sujet pour l'analyse de la variance, une analyse par classification hiérarchique a été effectuée pour détecter une possible influence. Les coefficients de corrélation entre les notes de dissemblance ont été calculés pour toutes les paires de sujets. Ces notes ont ensuite été soumises à une analyse de classification hiérarchique basée sur l'algorithme du plus proche voisin [40]. Comme les deux catégories de sujets ne faisaient pas partie de groupes séparés, aucune distinction entre les musiciens et les sujets experts n'a été faite dans l'analyse ultérieure.

Les résidus studentisés étant compris entre -3 et +3, les données récoltées ne contenaient pas de valeurs aberrantes (*outliers*) [41]. Pour pouvoir légitimement réaliser une ANOVA à mesures répétées, l'hypothèse de normalité doit être remplie : les résidus des observations appartenant à la même combinaison de niveaux de variables indépendantes doivent être normalement distribués [42]. Un test de Shapiro-Wilk de normalité sur les résidus studentisés a été effectué sur chaque cellule avec un niveau de signification de 5% et a rejeté l'hypothèse nulle d'une distribution normale pour 5 cellules seulement sur 60. La plupart des études statistiques indiquent que l'ANOVA peut être robuste à ce genre de violation [43], surtout si la taille de l'échantillon est supérieure à 15 observations par cellule [44]. Avec 22 observations par cellule, nous avons considéré que l'utilisation de l'ANOVA pour l'analyse statistique était toujours légitime.

Facteur	df	SS	MS	F	p
SO	1	4412	4412	3.197	0.088
SP	4	22157	5539	5.542	***0.001***
OR	2	1507	753	0.608	0.549
DM	1	0.909	0.909	0.001	0.974
SO * SP	4	4531	1132	1.829	0.131
SO * OR	2	3443	1721	1.511	0.233
SP * OR	8	8690	1086	1.092	0.371
SO * DM	1	2410	2410	2.316	0.143
SP * DM	4	889	222	0.238	0.916
OR * DM	2	2410	2410	2.316	0.143

Tableau 5.5 – Résultats de l'ANOVA. SO : source, SP : espace, OR : ordre ambisonique mixte, DM : dimensionalité, df : degré de liberté, SS : somme des carrés, MS : moyenne des carrés, F : valeur f, p : valeur p.

Les données ont été soumises à une ANOVA à mesures répétées avec les variables indépendantes suivantes : «espace sonore» (5) × «source sonore» (2) × «ordre am-

bisonique mixte» (3) × «dimensionnalité» (2). Le test de sphéricité de Mauchly a indiqué que l'hypothèse de sphéricité était remplie pour chaque variable indépendante ainsi que pour leurs interactions. Les résultats sont présentés dans le tableau 5.5. Par souci de lisibilité, seuls les résultats correspondant aux effets principaux et aux interactions simples sont affichés (les interactions à trois et à quatre facteurs n'étaient pas significatives). Un seul effet principal était significatif : la variable indépendante «espace sonore» [F(4, 84) = 5,542 ; p = 0,001].

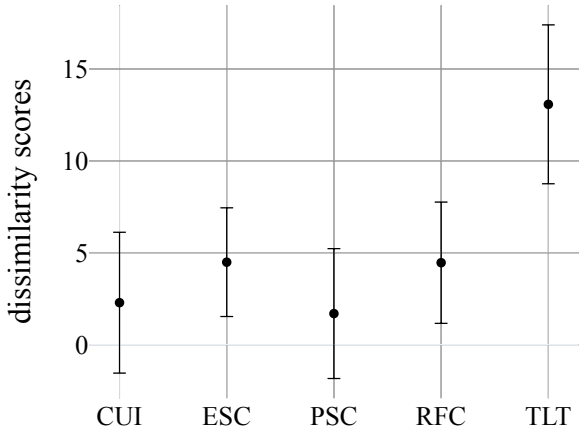


FIGURE 5.2 – Notes de dissemblance moyennes avec les intervalles de confiance à 95% associés pour les cinq espaces.

Espace	t	df	p	Taille d'effet
CUI	1.239	263	0.217	0.08
ESC	2.469	263	0.014	0.15
PSC	0.912	263	0.363	0.06
RFC	2.613	263	0.009	0.16
TLT	5.757	263	0.000	0.35

Tableau 5.6 – Résultats des t-tests t à échantillon unique pour chaque espace. t : valeur t, df : degré de liberté, p : valeur p, Taille d'effet : coefficient *d* de Cohen. Une valeur p < 0,05 signifie que les notes moyennes de dissemblance étaient significativement différentes de zéro et donc que les sujets ont pu percevoir la différence entre les stimuli d'ordre mixte et la référence. Ces valeurs apparaissent en bleu.

La figure 5.2 affiche les notes moyennes de dissemblance et l'intervalle de confiance à 95% pour chaque espace sonore. Des tests LSD de Fisher avec correction de Bonferroni ont été effectués pour les différents niveaux de la variable indépendante «espace». Aucune différence significative n'a été obtenue entre les notes associées à *Cuisine*, *Escalier*, *Piscine* et *Réfectoire*. Un espace en particulier s'est distingué : les notes liées à *Toilettes* étaient significativement supérieures à celles liées à *Cuisine*, *Escalier* et *Piscine* (p = 0.035, p = 0.020, p = 0.044 respectivement). Néanmoins, aucune différence significative n'a été obtenue entre les notes associées à *Toilettes* et *Réfectoire* (p = 0.067).

L'ANOVA permet uniquement de déterminer si les sujets ont évalué de la même manière les différences entre les stimuli d'ordre mixte et la référence d'une condition à l'autre. Pour déterminer si les sujets ont été capables de faire la distinction entre les stimuli d'ordre mixte et la référence, une analyse plus approfondie est nécessaire. L'ANOVA ayant déterminé que les sujets évaluaient différemment les stimuli sonores en fonction de l'espace sonore, les données ont été soumises à des t-tests à échantillon unique afin d'évaluer si les notes moyennes de dissemblance liées à chaque espace sonore étaient significativement différentes de zéro. Les résultats des tests sont affichés dans le tableau 5.6. Ces tests ont révélé que les moyennes des notes obtenues pour *Cuisine* et *Piscine* n'étaient pas significativement différentes de zéro, ce qui suggère que les sujets n'étaient globalement pas capables de percevoir une différence entre le stimulus d'ordre 4 et les stimuli d'ordre mixte pour ces deux espaces. Même si les notes de dissemblance étaient significativement différentes de zéro pour trois espaces sonores (*Escalier*, *Réfectoire* et *Toilettes*), les tailles d'effet d observées peuvent être considérées comme faibles ($d < 0.2$) pour tous les espaces sonores, excepté pour l'espace *Toilettes* pour lequel la taille d'effet est moyenne ($d < 0.5$), d'après les conventions établies par Cohen [45].

5.5. Discussion

5.5.1. Aucune influence observée de l'ordre ambisonique mixte

Une ANOVA à mesures répétées a révélé que la représentation ambisonique utilisée pour les stimuli d'ordres mixtes n'avait aucune influence significative sur la dissemblance perçue entre la référence cachée et les stimuli d'ordres mixtes. Il semble donc que les sujets aient évalué les différences entre la référence d'ordre 4 et les stimuli d'ordre mixte de la même manière pour tous les ordres mixtes testés alors qu'on aurait pu s'attendre à des notes de dissemblance plus élevées pour les ordres mixtes plus faibles.

5.5.2. Aucune influence observée de la dimensionalité

Il semble que les informations fournies par les composantes ambisoniques non sectorielles n'aient pas été discriminantes, puisque l'utilisation de SRIRs 2D ou 3D n'a pas entraîné de différences significatives dans les évaluations. Plusieurs raisons peuvent expliquer ce phénomène : 1) La précision de la localisation en élévation est faible par rapport à la localisation dans le plan horizontal [46]. Il est possible que les indices d'élévation n'aient pas été suffisamment saillants pour être discriminants dans les parties réverbérantes des SRIR considérées. 2) En utilisant des signaux ambisoniques d'ordre 1 et 2, l'espace sonore mesuré peut être reproduit aux oreilles de l'auditeur avec une erreur de reconstruction de 4% jusqu'à environ 640 Hz et 1280 Hz respectivement [47, 48]. Les indices d'élévation étaient peut-être déjà tellement altérés que la réduction de la résolution en élévation n'a pas eu d'impact sur la perception des réflexions. 3) La perception de l'élévation peut avoir été gravement altérée par

l'utilisation d'un rendu binaural non-individualisé, qui ne peut pas reproduire correctement les indices spectraux pour chaque sujet. 4) Le son direct étant dans le plan horizontal, il est possible que les sujets n'aient pas utilisé les indices d'élévation pour comparer les stimuli, car leur attention n'était pas concentrée sur cette zone particulière. Des études supplémentaires sont nécessaires pour évaluer l'impact de chacun des facteurs mentionnés ci-dessus, en utilisant une quantité plus importante d'espaces sonores avec des informations d'élévation saillantes et des sources sonores situées en dehors du plan horizontal.

5.5.3. Influence de l'espace sonore

Le seul effet significatif sur les notes de dissemblance était l'espace sonore. En particulier, le plus petit espace (*Toilettes*) a conduit aux plus grandes dissemblances perçues avec la référence cachée. Cet espace avait à la fois le plus petit volume et le temps de réverbération le plus court. Nous pouvons supposer que la présence d'un nombre limité de réflexions spéculaires d'amplitudes élevées, a aidé les sujets à faire la différence entre les stimuli d'ordre mixte et la référence.

Pour les espaces sonores *Réfectoire*, *Escalier* et *Toilettes*, les sujets ont correctement identifié la référence cachée par rapport aux stimuli d'ordre mixte. Ainsi, pour trois espaces sonores sur cinq, les résultats ne sont pas conformes à l'étude préliminaire d'Engel et al. [10], qui indiquait que l'utilisation d'ordres ambisoniques supérieurs au premier ordre n'améliore pas la qualité de la spatialisation. Il semble que l'utilisation d'ordres ambisoniques plus élevés puisse améliorer, bien que légèrement ($d \leq 0.35$, cf. tableau 5.6), la précision de la reproduction de l'espace sonore.

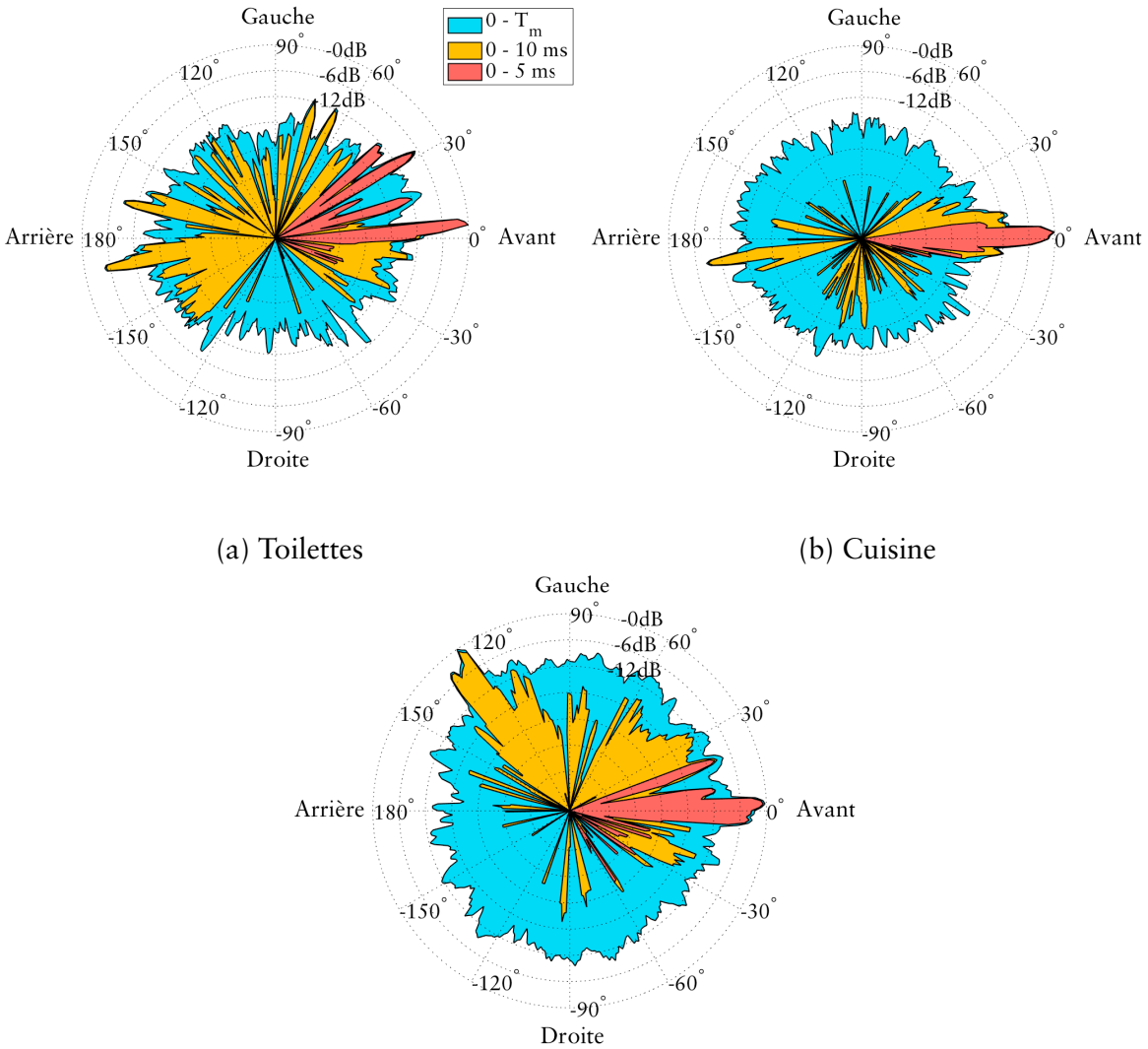
Espace	T_{30}	T_s	C_{80}	D_{50}	DRR	G_E
Cuisine	0.62	35.5	8.75	77.50	-1.89	32.17
Escalier	3.67	173.9	-1.20	32.20	-6.08	37.87
Piscine	1.97	38.1	8.34	76.69	*2.74*	28.37
Réfectoire	0.90	31.7	8.57	79.21	*2.09*	28.86
Toilettes	0.40	22.4	13.85	86.30	-3.42	32.71

Tableau 5.7 – Paramètres acoustiques liés aux espaces sonores : le temps de réverbération (T_{30}), le temps central (T_s), les rapports d'énergies précoces et tardives (D_{50} et C_{80}), le rapport champ direct champ réverbéré (DRR) et la force sonore de l'énergie précoce (G_E). Les valeurs remarquables apparaissent en bleu.

Le tableau 5.7 répertorie plusieurs paramètres acoustiques qui décrivent le contenu des SRIRs. Ces valeurs donnent un éclairage sur les différences significatives observées entre les notes liées à *Toilettes* et aux autres espaces sonores.

Les mesures de D_{50} et C_{80} sont les plus élevées pour et le temps central T_s le plus faible pour *Toilettes*. Cela confirme que la SRIR correspondante avait principalement de l'énergie dans sa partie précoce.

La figure 5.3 montre la décomposition spatiale de trois SRIRs selon différentes régions temporelles en utilisant la méthode de décomposition spatiale (SDM) introduite



(a) Toilettes

(b) Cuisine

(c) Escalier

FIGURE 5.3 – Diagrammes polaires de l'énergie de SRIRs dans le plan horizontal calculés d'après la SDM [49]. Trois régions temporelles sont affichées : jusqu'à 5 ms, jusqu'à 10 ms, jusqu'au temps de mélange (T_m).

par Tervo et al. [49]. Cette méthode décompose les SRIR en ondes planes pour des fenêtres temporelles courtes successives. Les diagrammes polaires représentent l'énergie accumulée au cours des premières 5 ms, 10 ms et jusqu'au moment du temps de mélange, en fonction de la direction.

La figure 5.3a associé à *Toilettes*, montre que les réflexions spéculaires semblent avoir de plus grandes amplitudes. Alors que plusieurs réflexions sont supérieures à -12 dB par rapport au son direct dans différentes directions pour *Toilettes*, une seule réflexion (à -170°) est supérieure à ce seuil pour *Cuisine*. La grande différence des notes de dissemblance liées à ces espaces peut s'expliquer par les amplitudes et la localisation des réflexions spéculaires qui peuvent avoir donné beaucoup plus d'indices de localisation dans *Toilettes*.

De même, la différence des notes de dissemblance entre *Toilettes* et *Réfectoire* ou *Piscine* peut être due au fait que l'amplitude des réflexions spéculaires n'était pas assez élevée dans ces espaces. Toutes les réflexions spéculaires avaient une amplitude inférieure de 18 dB à celle du son direct. De plus, étant donné les rapports élevés entre énergie du son direct et énergie réverbérée pour ces deux espaces, il est possible que l'énergie du son direct ait masqué les premières réflexions des SRIRs.

Alors que dans *Escalier* le microphone était très proche d'un mur, ce qui était censé produire de fortes réflexions précoces très discriminantes, cet espace n'était pas aussi discriminant que *Toilettes*. En particulier, la figure 5.3c montre qu'une réflexion précoce est aussi élevée que le son direct à 125°. La force sonore de l'énergie précoce est la plus élevée mais son rapport entre l'énergie précoce et l'énergie tardive est la plus faible. Compte tenu de la longue durée de réverbération, on peut supposer que les sujets étaient sensibles à l'importante énergie tardive qui a pu masquer certaines informations précoces.

5.6. Conclusion

Dans ce chapitre, les SRIRs ont été manipulées dans le domaine ambisonique pour étudier la perception de scènes sonores dont la partie réverbérée a une résolution spatiale plus faible que le son direct. L'étude s'est placée dans un contexte d'écoute communément rencontré en utilisant un rendu binaural non-individualisé et des signaux ambisoniques allant jusqu'à l'ordre 4. Des filtres de décodage binauraux optimisés ont été utilisés de manière à reproduire précisément les indices spectraux et la corrélation interaurale en champ diffus de HRTFs mesurées.

Un test perceptif a montré que le décodage binaural à un ordre inférieur ou égal à 2 d'une réverbération ambisonique peut avoir une légère influence ou pas d'influence sur la perception de l'acoustique par rapport à un décodage binaural d'ordre 4. Cette influence dépend de l'espace sonore considéré. Il semble que les espaces présentant des réflexions spéculaires importantes, ainsi qu'un temps de réverbération court et un faible rapport entre l'énergie directe et l'énergie réverbérée, peuvent entraîner des différences perceptibles.

De plus, dans cette étude, la réduction de la dimensionnalité des SRIRs en élévation n'a pas eu d'impact sur la perception de l'acoustique des salles. Ce résultat peut être attribué au fait que notre perception de réflexions sonores situées en élévation est moins précise que notre perception de réflexions dans le plan azimutal. Il peut également être dû à l'utilisation d'un encodage ambisonique d'ordre limité (ordre 4) et à l'usage d'un rendu binaural non-individualisé qui affectent la perception de l'élévation.

Les résultats présentés dans ce chapitre ont des implications sur le coût de calcul et l'utilisation de la mémoire nécessaires à la reproduction de SRIRs puisque, pour plusieurs SRIRs, la réduction de 25 à 3 canaux ambisoniques n'a pas entraîné de différences perceptives importantes voire même significatives.

Bibliographie

[1] H. Lee, M. Frank et F. Zotter, “Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings,” dans *Audio Engineering Society Conference : 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[2] M. Frank, F. Zotter et A. Sontacchi, “Localization experiments using different 2d ambisonics decoders,” dans *25th Tonmeistertagung-VDT International Convention, Leipzig*, 2008.

[3] S. Braun et M. Frank, “Localization of 3d ambisonic recordings and ambisonic virtual sources,” dans *1st International Conference on Spatial Audio,(Detmold)*, 2011.

[4] P. Power, W. Davies, J. Hirst, C. Dunn *et al.*, “Localisation of elevated virtual sources in higher order ambisonic sound fields,” *Proceedings of the Institute of Acoustics*, 2012.

[5] S. Bertet, J. Daniel, E. Parizet et O. Warusfel, “Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources,” *Acta Acustica united with Acustica*, vol. 99, n°. 4, p. 642–657, 2013.

[6] L. Thresh, C. Armstrong et G. Kearney, “A direct comparison of localization performance when using first, third, and fifth ambisonics order for real loudspeaker and virtual loudspeaker rendering,” dans *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[7] G. Reardon, G. Zalles, A. Genovese, P. Flanagan et A. Roginska, “Evaluation of binaural renderers : Externalization,” dans *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[8] E. Miller et B. Rafaely, “The role of direct sound spherical harmonics representation in externalization using binaural reproduction,” *Applied Acoustics*, vol. 148, p. 40–45, 2019.

[9] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf et B. Rafaely, “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution,” *The Journal of the Acoustical Society of America*, vol. 133, n°. 5, p. 2711–2721, 2013.

[10] I. Engel, C. Henry, S. V. A. Garí, P. W. Robinson, D. Poirier-Quinot et L. Picinali, “Perceptual comparison of ambisonics-based reverberation methods in binaural listening,” dans *EAA Spatial Audio Signal Processing Symposium*, Paris, France, sept. 2019, p. 121–126. [En ligne]. Disponible : https://hal.archives-ouvertes.fr/hal-02275174

[11] E. Wenzel, M. Arruda, D. Kistler et F. Wightman, “Localization using nonindividualized head-related transfer functions,” *The Journal of the Acoustical Society of America*, vol. 94, p. 111–23, 08 1993.

[12] P. Damaske et B. Wagener, “Directional hearing tests by the aid of a dummy head,” *Acta Acustica united with Acustica*, vol. 21, n°. 1, p. 30–35, 1969.

[13] J. Blauert, *Spatial hearing : the psychophysics of human sound localization*. MIT press, 1997.

[14] M. Gorzel, G. Kearney et F. Boland, “Investigation of ambisonic rendering of elevated sound sources,” dans *Audio Engineering Society Conference : 55th International Conference : Spatial Audio*. Audio Engineering Society, 2014.

[15] E. D. Schubert et J. Wernick, “Envelope versus microstructure in the fusion of dichotic signals,” *The Journal of the Acoustical Society of America*, vol. 45, n°. 6, p. 1525–1531, 1969.

[16] R. L. Freyman, R. K. Clifton et R. Y. Litovsky, “Dynamic processes in the precedence effect,” *The Journal of the Acoustical Society of America*, vol. 90, n°. 2, p. 874–884, 1991.

[17] X. Yang et D. W. Grantham, “Echo suppression and discrimination suppression aspects of the precedence effect,” *Perception & psychophysics*, vol. 59, n°. 7, p. 1108–1117, 1997.

[18] R. Y. Litovsky, H. S. Colburn, W. A. Yost et S. J. Guzman, “The precedence effect,” *The Journal of the Acoustical Society of America*, vol. 106, n°. 4, p. 1633–1654, 1999.

[19] P. Götz, K. Kowalczyk, A. Silzle et E. A. Habets, “Mixing time prediction using spherical microphone arrays,” *The Journal of the Acoustical Society of America*, vol. 137, n°. 2, p. EL206–EL212, 2015.

[20] J. Ahonen et V. Pulkki, “Diffuseness estimation using temporal variation of intensity vectors,” dans *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, p. 285–288.

[21] P. Rubak et L. G. Johansen, “Artificial reverberation based on a pseudo-random impulse response ii,” dans *Audio Engineering Society Convention 106*. Audio Engineering Society, 1999.

[22] T. Hidaka, Y. Yamada et T. Nakagawa, “A new definition of boundary point between early reflections and late reverberation in room impulse responses,” *The Journal of the Acoustical Society of America*, vol. 122, n°. 1, p. 326–332, 2007.

[23] J. S. Abel et P. Huang, “A simple, robust measure of reverberation echo density,” dans *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[24] R. Stewart et M. Sandler, “Statistical measures of early reflections of room impulse responses,” dans *Proc. of the 10th int. conference on digital audio effects (DAFx-07), Bordeaux, France*, 2007, p. 59–62.

[25] G. Defrance, L. Daudet et J.-D. Polack, “Using matching pursuit for estimating mixing time within room impulse responses,” *Acta Acustica united with Acustica*, vol. 95, n°. 6, p. 1071–1081, 2009.

[26] Rec ITU-R, “Bs. 1116-3,” *Methods for the subjective assesment of small impairments in audio systems,” International Telecommunication Union-Radiocommunication Sector*, 2015.

[27] MH Acoustics LLC, “Eigenmike em32 microphone array,” accessed 2020-08-22. [En ligne]. Disponible : https://mhacoustics.com/products

[28] D. Cabrera, D. Lee, M. Yadav et W. L. Martens, “Decay envelope manipulation of room impulse responses : Techniques for auralization and sonification,” dans *Proceedings of Acoustics*, 2011.

[29] D. Rudrich *et al.*, “IEM plug-in suite,” *University of Music and Performing Arts, Graz, Austria : Institute of Electronic Music and Acoustics.*, 2019, accessed : 2020-08-22. [En ligne]. Disponible : https://plugins.iem.at/

[30] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[31] C. Schörkhuber, M. Zaunschirm et R. Höldrich, “Binaural rendering of ambisonic signals via magnitude least squares,” dans *Proceedings of the DAGA*, vol. 44, 2018, p. 339–342.

[32] M. Zaunschirm, C. Schörkhuber et R. Höldrich, “Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint,” *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3616–3627, 2018.

[33] A. Baskind, J. Messonnier, J. Lyzwa et M. Aussal, “Hedrot–open-source head tracker,” 2017. [En ligne]. Disponible : https://abaskind.github.io/hedrot/

[34] M. Kronlachner, “Plug-in suite for mastering the production and playback in surround sound and ambisonics,” *Gold-Awarded Contribution to AES Student Design Competition*, 2014. [En ligne]. Disponible : http://www.matthiaskronlachner.com/?p=2015

[35] Cycling’74, “Max 8,” 2019. [En ligne]. Disponible : https://cycling74.com

[36] AES20-1996 (s2008), “AES recommended practice for professional audio : Subjective evaluation of loudspeakers,” Audio Engineering Society, Rapport technique, 2008.

[37] S. Zielinski, F. Rumsey et S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, n°. 6, p. 427–451, 2008.

[38] E. C. Poulton et S. Poulton, *Bias in quantifying judgements*. Taylor & Francis, 1989.

[39] D. Schwarz, G. Lemaitre, M. ARAMAKI et R. Kronland-Martinet, “Effects of Test Duration in Subjective Listening Tests,” dans *International Computer Music Conference (ICMC)*, H. Timmermans, édit., Hans Timmermans. Utrecht, Netherlands : HKU University of the Arts Utrecht, HKU Music and Technology, sept. 2016, p. 515–519. [En ligne]. Disponible : https://hal.archives-ouvertes.fr/hal-01427340

[40] L. Kaufman et P. Rousseeuw, “Finding groups in data : an introduction to cluster analysis, vol. 5,” 1990.

[41] R. D. Cook et S. Weisberg, *Residuals and influence in regression*. New York : Chapman and Hall, 1982.

[42] H. Keselman, J. C. Rogan, J. L. Mendoza et L. J. Breen, “Testing the validity conditions of repeated measures f tests.” *Psychological Bulletin*, vol. 87, n°. 3, p. 479, 1980.

[43] K. Weinfurt, “Repeated measures analysis : Anova, manova, and hlm,” *Reading and Understanding More Multivariate Statistics*, 10 2000.

[44] S. B. Green et N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[45] J. Cohen, “A power primer.” *Psychological bulletin*, vol. 112, n°. 1, p. 155, 1992.

[46] G. Kearney et T. Doyle, “Height perception in ambisonic based binaural decoding,” dans *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[47] D. B. Ward et T. D. Abhayapala, “Reproduction of a plane-wave sound field using an array of loudspeakers,” *IEEE Transactions on speech and audio processing*, vol. 9, n°. 6, p. 697–707, 2001.

[48] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *Journal of the Audio Engineering Society*, vol. 53, n°. 11, p. 1004–1025, 2005.

[49] S. Tervo, J. Pätynen, A. Kuusinen et T. Lokki, “Spatial decomposition method for room impulse responses,” *Journal of the Audio Engineering Society*, vol. 61, n°. 1/2, p. 17–28, 2013.

6

L'influence des résolutions temporelle et fréquentielle de la paramétrisation des premières réflexions

La réduction de la résolution spatiale d'une réponse impulsionnelle spatiale de salle présente l'avantage de réduire le nombre de ses échantillons, de réduire les calculs nécessaires au rendu sonore et simplifie le contrôle perceptif. Dans le chapitre précédent, nous avons vu qu'il est possible de réduire cette résolution en introduisant une légère différence voire aucune différence audible selon l'espace. Dans ce chapitre, nous poursuivons l'analyse des résolutions avec lesquelles il semble suffisant de définir les paramètres d'une SRIR. Parmi les méthodes de paramétrisation présentées au chapitre 4, il n'existe pas de consensus sur les résolutions fréquentielle et temporelle avec lesquelles définir les paramètres associés aux réflexions précoces. Afin d'étudier ces résolutions, nous proposons une méthode de paramétrisation basée sur le calcul de matrices de covariance qui permet de reproduire la répartition de l'énergie sonore des réflexions précoces dans plusieurs régions temps-fréquence. Des scènes sonores générées avec des SRIRs ambisoniques encodées à l'ordre 4 ont été comparées à des scènes générées par des versions paramétrées d'ordre 1 de ces SRIRs. Un test perceptif a révélé que, parmi les méthodes évaluées, la méthode de paramétrisation employant une résolution temporelle de 5 ms dans quatre bandes de fréquence était la plus proche des scènes sonores de référence. Les résultats obtenus avec cette méthode n'était pas significativement différents de ceux obtenus avec la SDM - une paramétrisation communément employée - tout en réduisant le nombre de paramètres nécessaires.

6.1. Introduction

L'ensemble des échantillons d'une réponse impulsionnelle spatiale de salle (SRIR) peut constituer une grande quantité d'échantillons et l'emploi de données décrivant ses caractéristiques spatiales, fréquentielles et temporelles peut réduire considérablement cette quantité. Ces données sont des paramètres de la SRIRs. Dans le but de modifier le rendu sonore d'une auralisation, plutôt que d'agir sur chaque échantillon d'une SRIR, il est plus simple d'appréhender la modification de paramètres tels que la diffusivité, la position de réflexions spéculaires ou des pentes de décroissance de l'énergie pour contrôler la reproduction du champ sonore.

Parmi les méthodes permettant d'extraire des paramètres d'une SRIR se trouvent notamment la RSAO [1] et la paramétrisation SFA [2]. Ces méthodes analysent une SRIR en accord avec un modèle générique de réverbération qui associe des réflexions spéculaires précoces à un champ diffus tardif. Elles ne spécifient pas de critère permettant de connaitre le nombre de premières réflexions ni la résolution fréquentielle nécessaires pour reproduire fidèlement une SRIR. La détermination des réflexions se fait de manière heuristique en fixant des seuils de détection. Begault *et al.* [3] ont établi des seuils d'audibilité de réflexions en fonction de leur retard, de leur amplitude et de leur direction d'incidence. Néanmoins, de plus amples études doivent être menées pour établir la résolution temporelle avec laquelle analyser et reconstruire l'énergie précoce d'une SRIR.

D'autres méthodes de paramétrisation telles que le SIRR [4], HO-SIRR [5] ou la SDM [6] sont également utilisées. Ces procédés analysent les caractéristiques spatiales d'une SRIR dans des fenêtres temps-fréquence réduites de l'ordre de quelques millisecondes et de dizaines de Hertz. La SDM repose sur l'hypothèse que le champ sonore mesuré peut être représenté comme une succession d'événements acoustiques distincts, chacun d'entre eux associé à une direction d'incidence et une amplitude. Cette méthode a été utilisée pour l'analyse et l'auralisation de champs sonores dans des salles de concert [7, 8] ou des habitacles de voiture [9, 10].

Récemment, Garí *et al.* [11] ont employé cette méthode pour auraliser un espace sonore en binaural dans un contexte de réalité augmentée. La SDM a été choisie afin de reproduire des réponses impulsionnelles binaurales (BRIRs) dont les caractéristiques spatiales sont facilement manipulables grâce aux paramètres extraits. Les résultats d'un test perceptif montrent que la perception au casque d'une source dans un espace auralisé était plausible par rapport à la diffusion de cette même source *in situ* au moyen de haut-parleurs. Néanmoins, La diffusion binaurale de la source était perçue significativement différente de la diffusion sur haut-parleur.

Bien qu'elle constitue une méthode d'auralisation produisant des résultats crédibles, la description du champ sonore par la SDM représente une quantité de données importante car elle consiste en un vecteur contenant le signal omnidirectionnel de la SRIR associé à une matrice indiquant les directions d'incidence pour chacun des échantillons. De plus, les paramètres caractérisent le comportement spatial de la SRIR sur l'ensemble du spectre et non par bande de fréquence, ce qui limite le contrôle spatial.

Dans ce chapitre, nous souhaitons étudier la reproduction de SRIRs selon une

autre méthode de paramétrisation. L'approche proposée permet de réduire le nombre de paramètres nécessaires et un contrôle des caractéristiques spatiales par bande de fréquence. Cette méthode est basée sur l'emploi de matrices de covariance calculées dans plusieurs régions temps-fréquence. La matrice de covariance des signaux ambisoniques contient les énergies et les relations de phase et d'énergie entre chaque signal. Elle permet ainsi de décrire les caractéristiques spatiales du champ sonore. Nous présenterons dans un premier temps la méthode de paramétrisation proposée. Plusieurs SRIRs ont été reproduites avec cette méthode et avec la SDM. Afin d'évaluer la dissemblance entre les signaux reproduits et les signaux issus de SRIRs non paramétrées, un test perceptif a été mis en œuvre.

Les résultats du chapitre 5 nous informent que la réduction de l'ordre d'encodage de la réverbération peut avoir un influence significative sur la perception de l'acoustique selon l'espace. Néanmoins, cette influence étant faible, nous avons choisi d'adopter un moteur de rendu binaural reproduisant le son direct à l'ordre 4 et la réverbération à l'ordre 1, qui présente l'avantage de réduire considérablement les calculs nécessaires à l'auralisation.

6.2. Paramétrisation d'une SRIR par matrice de covariance

Soit $\boldsymbol{B}$ la matrice des $N = (L+1)^2$ signaux ambisoniques représentant les premières réflexions d'une SRIR dans le domaine des harmoniques sphériques jusqu'à l'ordre L. L'estimation de la matrice de covariance $\boldsymbol{C}$ est donnée par :

$$\boldsymbol{C} = \frac{1}{K-1}\boldsymbol{B}\boldsymbol{B}^H \ , \tag{6.1}$$

où ^{H} désigne l'opérateur adjoint et K désigne le nombre d'échantillon des signaux ambisoniques.

Plutôt que de conserver l'ensemble des signaux $\boldsymbol{B}$ pour décrire les premières réflexions d'une SRIR nous proposons de seulement conserver : 1) le signal omnidirectionnel, qui contient la structure temporelle et fréquentielle des premières réflexions captées dans toutes les directions, 2) les matrices de covariance des signaux ambisoniques dans plusieurs régions temps-fréquence. La figure 6.1 représente le spectrogramme d'un signal omnidirectionnel segmenté en différentes régions temps-fréquence auxquelles des matrices de covariance sont associées.

La figure 6.2 représente les opérations à effectuer pour générer des signaux ambisoniques ayant le même signal omnidirectionnel que les signaux d'origine et la même matrice de covariance que les signaux d'origine dans chaque segment temps-fréquence. Pour ce faire, des signaux décorrélés sont créés à partir du signal omnidirectionnel puis mélangés de manière à reproduire la covariance des signaux d'origine. Une matrice de rotation est ensuite appliquée aux signaux générés pour que le signal omnidirectionnel corresponde à celui d'origine.

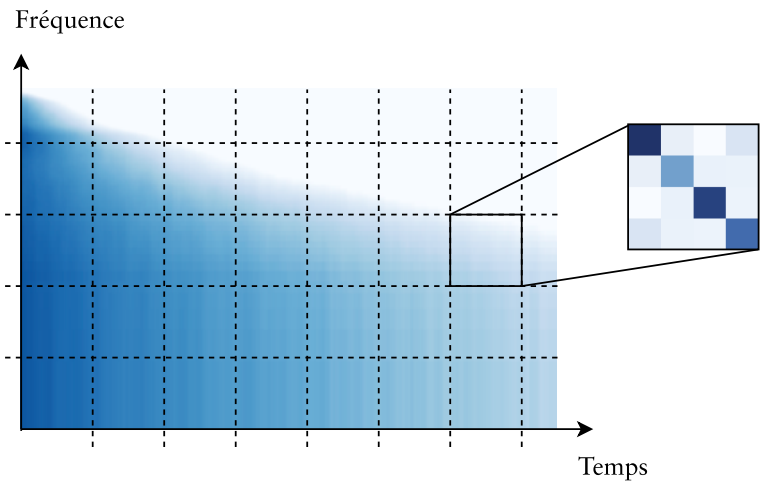


FIGURE 6.1 – Segmentation temps-fréquence du signal omnidirectionnel auquel sont associées des matrices de covariance d'ordre 1 pour chaque segment.

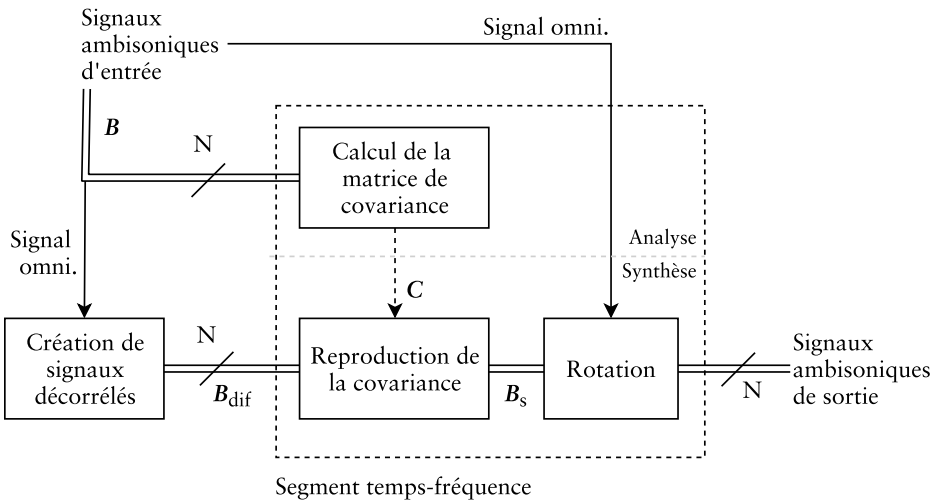


FIGURE 6.2 – Processus d'analyse et de synthèse proposé.

6.2.1. Création de signaux décorrélés à partir du signal omnidirectionnel

Une première étape consiste à générer des signaux ambisoniques d'après le signal omnidirectionnel d'origine. Pour ce faire, on utilise un procédé de décorrélation appelé *panning* d'amplitude. Ce procédé consiste à décomposer le signal en bande de fréquence et à distribuer les composantes fréquentielles selon des directions d'incidence déterminées de manière stochastique variables ou non dans le temps [12]. Plutôt que d'utiliser des directions aléatoires, Pihlajamäki *et al.* [13] préconisent d'utiliser une séquence déterministe de Halton [14] qui permet de couvrir uniformément la gamme des directions d'incidence possibles même avec un faible nombre de directions. Par ailleurs, l'emploi d'une séquence déterministe présente l'avantage de produire les mêmes signaux d'une génération à l'autre. Comme toute méthode de décorrélation,

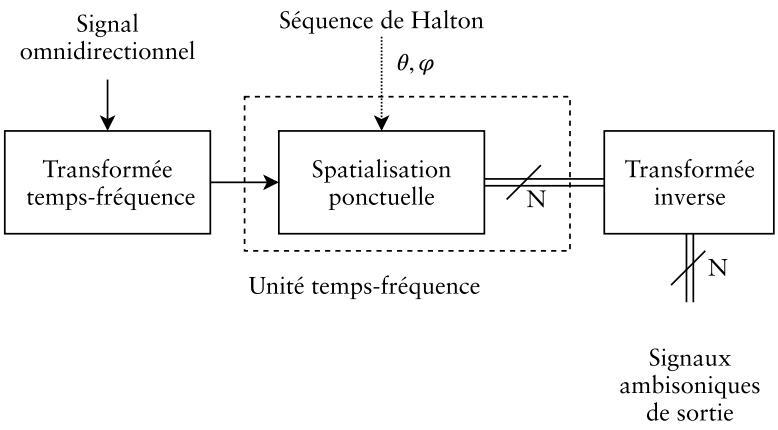


FIGURE 6.3 – Processus de génération de signaux décorrélés d'après le signal omnidirectionnel.

cette approche peut cependant introduire des artefacts audibles à haute fréquence en raison du changement rapide de localisation des bandes de fréquence qui crée des transitoires. Les artefacts introduits sont plus ou moins audibles selon le signal sonore utilisé [13].

La figure 6.3 représente le processus de générations des signaux décorrélés par *panning* d'amplitude. Comme recommandé dans l'étude de Pihlajamäki *et al.* [13], le signal omnidirectionnel est décomposé dans le domaine temps fréquence d'après une transformée de Fourier à court-terme (STFT) de 1024 points puis chaque bande de fréquence est distribuée dans l'espace de la plus basse à la plus haute en utilisant la séquence de Halton.

Les signaux ambisoniques obtenus, notés $\boldsymbol{B}_{\text{pan}}$, ne sont pas parfaitement décorrélés entre eux. Nous souhaitons «blanchir» ces signaux, c'est-à-dire appliquer une matrice $\boldsymbol{T}_1$ de manière à ce que la matrice de covariance $\boldsymbol{C}_{\text{dif}}$ des signaux résultants soit égale à la matrice identité $\boldsymbol{I}_N$. Soient $\boldsymbol{C}_{\text{pan}}$ la matrice de covariance des signaux ambisoniques $\boldsymbol{B}_{\text{pan}}$ et $\boldsymbol{V}_{\text{pan}}$ la matrice des vecteurs propres associés aux valeurs propres $(\lambda_{\text{i, pan}})_{1\leq i\leq N}$ de $\boldsymbol{C}_{\text{pan}}$. La décomposition en valeurs propres de la matrice $\boldsymbol{C}_{\text{pan}}$ s'écrit :

$$\boldsymbol{C}_{\text{pan}} = \boldsymbol{V}_{\text{pan}}^H \operatorname{diag}(\Lambda_{\text{pan}})\boldsymbol{V}_{\text{pan}} \tag{6.2}$$

avec le vecteur des valeurs propres $\Lambda_{\text{pan}} = [\lambda_{1,\text{pan}}\ \lambda_{2,\text{pan}}\ \ldots\ \lambda_{\text{N,pan}}]$. La matrice de covariance $\boldsymbol{C}_{\text{dif}}$ des signaux décorrélés $\boldsymbol{B}_{\text{dif}}$ doivent vérifier l'équation suivante :

$$\boldsymbol{C}_{\text{dif}} = \frac{1}{K-1}\boldsymbol{B}_{\text{dif}}\boldsymbol{B}_{\text{dif}}^H = \boldsymbol{I}_N \tag{6.3}$$

avec $\boldsymbol{B}_{\text{dif}} = \boldsymbol{T}_1\boldsymbol{B}_{\text{pan}}$, on obtient :

$$\boldsymbol{T}_1\frac{1}{K-1}\boldsymbol{B}_{\text{pan}}\boldsymbol{B}_{\text{pan}}^H\boldsymbol{T}_1^H = \boldsymbol{I}_N \tag{6.4}$$

soit

$$\boldsymbol{T}_1\boldsymbol{C}_{\text{pan}}\boldsymbol{T}_1^H = \boldsymbol{I}_N \tag{6.5}$$

$$\boldsymbol{T}_1\boldsymbol{T}_1^H = \boldsymbol{C}_{\text{pan}}^{-1} \tag{6.6}$$

Il existe une infinité de solutions à cette équation et une solution simple consiste à choisir l'expression suivante :

$$\boldsymbol{T}_1 = \boldsymbol{V}_{\text{pan}}^H \operatorname{diag}(\Lambda_{\text{pan}})^{-\frac{1}{2}}\boldsymbol{V}_{\text{pan}} \tag{6.7}$$

6.2.2. Reproduction de la matrice de covariance des signaux d'origine

Nous souhaitons calculer les signaux $\boldsymbol{B}_{\text{s}}$ dont la matrice de covariance $\boldsymbol{C}_{\text{s}}$ est égale à la matrice de covariance des signaux d'origine $\boldsymbol{C}$. Pour ce faire, une matrice $\boldsymbol{T}_2$ est appliquée aux signaux décorrélés $\boldsymbol{B}_{\text{dif}}$. La matrice de covariance des signaux de sortie $\boldsymbol{C}_{\text{s}}$ doit vérifier l'équation suivante :

$$\boldsymbol{C}_{\text{s}} = \frac{1}{K-1}\boldsymbol{B}_{\text{s}}\boldsymbol{B}_{\text{s}}^H = \boldsymbol{C} \tag{6.8}$$

avec $\boldsymbol{B}_{\text{s}} = \boldsymbol{T}_2\boldsymbol{B}_{\text{dif}}$, on obtient :

$$\boldsymbol{T}_2\frac{1}{K-1}\boldsymbol{B}_{\text{dif}}\boldsymbol{B}_{\text{dif}}^H\boldsymbol{T}_2^H = \boldsymbol{C} \tag{6.9}$$

soit

$$\boldsymbol{T}_2\boldsymbol{T}_2^H = \boldsymbol{C} \tag{6.10}$$

Il existe une infinité de solutions à cette équation et une solution simple consiste à choisir l'expression suivante :

$$\boldsymbol{T}_2 = \boldsymbol{V}^H \operatorname{diag}(\Lambda)^{\frac{1}{2}}\boldsymbol{V} \tag{6.11}$$

où $\boldsymbol{V}$ est la matrice des vecteurs propres associés aux valeurs propres $(\lambda_{\text{i}})_{1\leq i\leq N}$ de $\boldsymbol{C}$ et Λ est le vecteur des valeurs propres. À ce stade, la matrice de covariance des signaux d'origine est reproduite. Les signaux de synthèse $\boldsymbol{B}_s$ s'expriment :

$$\boldsymbol{B}_s = \boldsymbol{T}_2\boldsymbol{T}_1\boldsymbol{B} \tag{6.12}$$

6.2.3. Reproduction du signal omnidirectionnel d'origine

Nous souhaitons que les signaux omnidirectionnels des signaux ambisoniques d'origine $\boldsymbol{B}$ et des signaux ambisoniques de sortie notés $\boldsymbol{B}_{\text{s'}}$ soient égaux. Pour cela, une matrice de rotation $\boldsymbol{R}_K$ de dimensions $K \times K$ (où K est le nombre d'échantillons du signal omnidirectionnel) est appliquée à $\boldsymbol{B}_{\text{s}}$:

$$\boldsymbol{B}_{\text{s'}} = \boldsymbol{B}_{\text{s}}\boldsymbol{R}_K \tag{6.13}$$

Une telle matrice peut être calculée d'après le procédé proposé par Aguilera et Pérez [15]. En procédant ainsi, la matrice de covariance des signaux de sortie $\boldsymbol{B}_{\text{s'}}$ est la même que celle des signaux $\boldsymbol{B}_{\text{s}}$ et donc de la matrice $\boldsymbol{C}$.

6.3. Test perceptif

La paramétrisation présentée dans la section précédente a été appliquée à la partie précoce de plusieurs SRIRs selon quatre résolutions temps-fréquence différentes. Un test perceptif a été réalisé afin d'évaluer les dégradations perçues selon les résolutions employées. Le protocole de test était inspiré de la méthodologie MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*) [16]. Pour chaque espace auralisé, les stimuli sonores issus des SRIRs modifiées étaient comparés au stimulus de référence issu de la SRIR d'origine.

Cinq espaces sonores ont été utilisés pour créer les stimuli (une cuisine, une cage d'escalier, une piscine, un réfectoire et des sanitaires) ainsi que deux sources sonores (une voix et un instrument percussif). Pour chaque espace et chaque source, cinq stimuli sonores ont été générés : quatre stimuli issus de la paramétrisation proposée (selon quatre résolutions temps-fréquence différentes) et un stimulus issu de la paramétrisation SDM. En plus de ces stimuli, une référence cachée et une ancre étaient jugées par les sujets. Avec cinq espaces et deux sources sonores, le test perceptif comprenait donc un total de 70 stimuli.

Les stimuli ont été générés en utilisant les SRIRs et les sources sonores employées dans le chapitre 5. Le lecteur peut se référer à la section 5.3.1 pour la description des espaces sonores et de la méthode employée pour l'acquisition des SRIRs ainsi qu'à la section 5.3.2 pour la description des sources sonores. À la manière de la segmentation temporelle adoptée au chapitre 5, les cinq SRIRs ont été segmentées en trois régions temporelles (composantes initiale, précoce et tardive), d'après la méthode introduite par Götz *et al.* [17]. Les temps de mélange estimés des cinq espaces utilisés pour séparer les réflexions précoces de la réverbération tardive figurent dans le tableau 5.4.

Nous poursuivons dans ce chapitre l'étude initiée au chapitre précédent concernant les résolutions spatiales, temporelles et fréquentielles nécessaires à la reproduction de SRIRs. En raison des résultats du chapitre 5, nous avons choisi d'adopter un rendu binaural des composantes réverbérantes de la SRIR à un ordre faible. Ainsi, les parties précoces des SRIRs modifiées et la réverbération tardive ont été encodées en ambisonique à l'ordre 1 et le son direct était encodé à l'ordre 4.

Les SRIRs utilisées pour créer les stimuli de référence étaient entièrement encodées à l'ordre 4. Nous avons choisi d'adopter cette résolution spatiale élevée pour comparer les dégradations perçues entre un rendu sonore ayant la plus grande qualité possible et un rendu sonore paramétrique qui doit s'en approcher.

6.3.1. Les résolutions temps-fréquences employées

Deux résolutions fréquentielles et deux résolutions temporelles ont été adoptées pour paramétrer les SRIRs. Les matrices de covariances ont été calculées selon quatre ou huit bandes de fréquence et dans des fenêtres temporelles de 5 ms ou sur l'ensemble des premières réflexions. Ces valeurs ont été choisies suite à des pré-tests informels. Le tableau 6.1 recense les résolutions temps-fréquence employées. Ces paramétrisations seront par la suite désignées FXTX où F et T désignent à la résolution fréquentielle et temporelle respectivement, $X = 1$ désigne la résolution la plus faible et $X = 2$ la plus

haute.

Abréviation	Bandes de fréquence	Segmentation temporelle
F1T1	4 bandes de fréquence. Bornes fréquentielles (Hz) : [0,1000], [1000, 2000], [2000, 4000], [4000, 24000]	L'ensemble des premières réflexions
F2T1	8 bandes de fréquence. Fréquences centrales (Hz) : [125, 250, 500, 1000, 2000, 4000, 8000, 16000]	L'ensemble des premières réflexions
F1T2	4 bandes de fréquence. Bornes fréquentielles (Hz) : [0,1000], [1000, 2000], [2000, 4000], [4000, 24000]	Toutes les 5 ms
F2T2	8 bandes de fréquence. Fréquences centrales (Hz) : [125, 250, 500, 1000, 2000, 4000, 8000, 16000]	Toutes les 5 ms

Tableau 6.1 – Récapitulatif des résolutions adoptées pour la paramétrisation des SRIRs du test perceptif.

6.3.2. Le calcul des paramètres SDM

De la même manière que dans l'étude de Garí *et al.* [11], la fenêtre d'analyse SDM a été réglée sur la plus petite taille possible, et les vecteurs de directions d'incidence résultants ont été lissés en utilisant une moyenne glissante de 16 échantillons (avec une fréquence d'échantillonnage de 48 kHz). La méthode a été appliquée sur les 32 signaux microphoniques issus de la mesure de réponse impulsionnelle effectuée avec l'EigenMike [18].

D'après Tervo *et al.* [6], les signaux issus d'une antenne de microphones ouverte sont plus à même d'être analysés par la méthode. Néanmoins une antenne sphérique rigide peut être utilisée si : 1) elle est compacte, 2) il est possible de créer un signal omnidirectionnel d'après les signaux enregistrés, 3) au moins quatre microphones qui ne sont pas répartis sur le même plan sont utilisés. Ces conditions sont remplies dans notre cas.

6.3.3. Génération de l'ancre

L'ancre d'un test MUSHRA est un stimuli sonore dont la qualité est délibérément faible et qui est utilisée comme un point de référence basse dans l'interprétation des résultats. Afin de générer une ancre pour chaque configuration, c'est-à-dire pour chaque source dans chaque espace, la partie précoce des SRIRs a subi une transformation simple qui consiste à supprimer les corrélations entre signaux ambisoniques. En d'autres termes, sur l'ensemble de la partie précoce aucune directionnalité particulière n'est favorisée. Pour ce faire, il est nécessaire que la matrice de covariance $\boldsymbol{C}$ des signaux ambisoniques de la partie précoce notés $\boldsymbol{B}$ soit proportionnelle à la matrice identité. Soit $\boldsymbol{B}_{\text{anc}}$ les signaux ambisoniques résultants. $\boldsymbol{B}_{\text{anc}}$ s'écrit :

$$\boldsymbol{B}_{\text{anc}} = \boldsymbol{T}\boldsymbol{B} \tag{6.14}$$

avec

$$\boldsymbol{T} = \boldsymbol{V}^H \operatorname{diag}(\Lambda)^{-\frac{1}{2}} \boldsymbol{V} \tag{6.15}$$

où $\boldsymbol{V}$ est la matrice des vecteurs propres associés aux valeurs propres $(\lambda_\mathrm{i})_{1 \leq i \leq N}$ de la matrice de covariance $\boldsymbol{C}$ des signaux d'origine et Λ est le vecteur des valeurs propres.

6.3.4. Création des fichiers binauraux

Les signaux ambisoniques ont été convertis en signaux binauraux en utilisant des filtres de décodage binaural calculés d'après des mesures d'une tête artificielle Neumann KU 100 [19] selon les méthodes de rendu binaural de Schörkhuber *et al.* [20] et Zaunschirm *et al.* [21] dont il est question en annexe C.2. Comme les différentes régions temporelles des SRIRs étaient encodées avec des ordres différents (ordre 4 pour le son direct et ordre 1 pour les parties précoces et tardives), un décodeur binaural a été utilisé pour chaque région temporelle selon l'ordre correspondant. Les signaux décodés ont ensuite été additionnés pour créer les signaux binauraux.

6.3.5. Procédure

Le méthode d'évaluation perceptive utilisée pour ce test perceptif était inspirée de la méthode MUSHRA [16]. Pour chaque source et chaque espace - soit 10 configurations - 7 stimuli étaient présentés aux sujets en plus de la référence : une référence cachée, une ancre et les 5 stimuli issus des paramétrisations spatiales. La référence était décodée en binaural d'après une représentation ambisonique d'ordre 4. Les stimuli étaient attribués de manière aléatoire à 6 boutons de lecture étiquetés « Son i » avec $i \in [1, 7]$. Les sujets avaient la possibilité d'écouter les stimuli de façon répétée et de passer de l'un à l'autre à leur gré. Un slider horizontal permettait de régler la boucle de lecture sur une région temporelle particulière.

Pour chacune des 10 configurations, les participants devaient évaluer la différence entre la référence et les stimuli en déplaçant un curseur sur une échelle continue de 0 à 100 points dont les extrémités étaient étiquetées « Très différent» (0) à «Identique» (100). Aucune étiquette intermédiaire ni de graduation n'a été placée sur l'échelle pour éviter les biais dus à une distribution non linéaire des étiquettes et pour prévenir toute distorsion dans la distribution des données (avec des étiquettes intermédiaires, des groupes de notes peuvent en effet apparaître autour des valeurs où se trouvent les étiquettes [22, 23]). Les différentes configurations ont été présentées dans un ordre aléatoire.

En raison du contexte sanitaire, le test perceptif a été réalisé sur internet à l'aide de l'application javascript *webMUSHRA* [24]. La figure 6.4 montre l'interface graphique utilisée pour le test. Pour des raisons pratiques, les mouvements de tête des auditeurs n'ont pas été pris en compte dans le test. Avant de débuter le test, les participants se sont familiarisés avec l'interface utilisateur en effectuant l'évaluation des stimuli associés à la source *Voix* dans l'espace *Cuisine*. Le test consistait en une seule session d'environ 40 minutes.

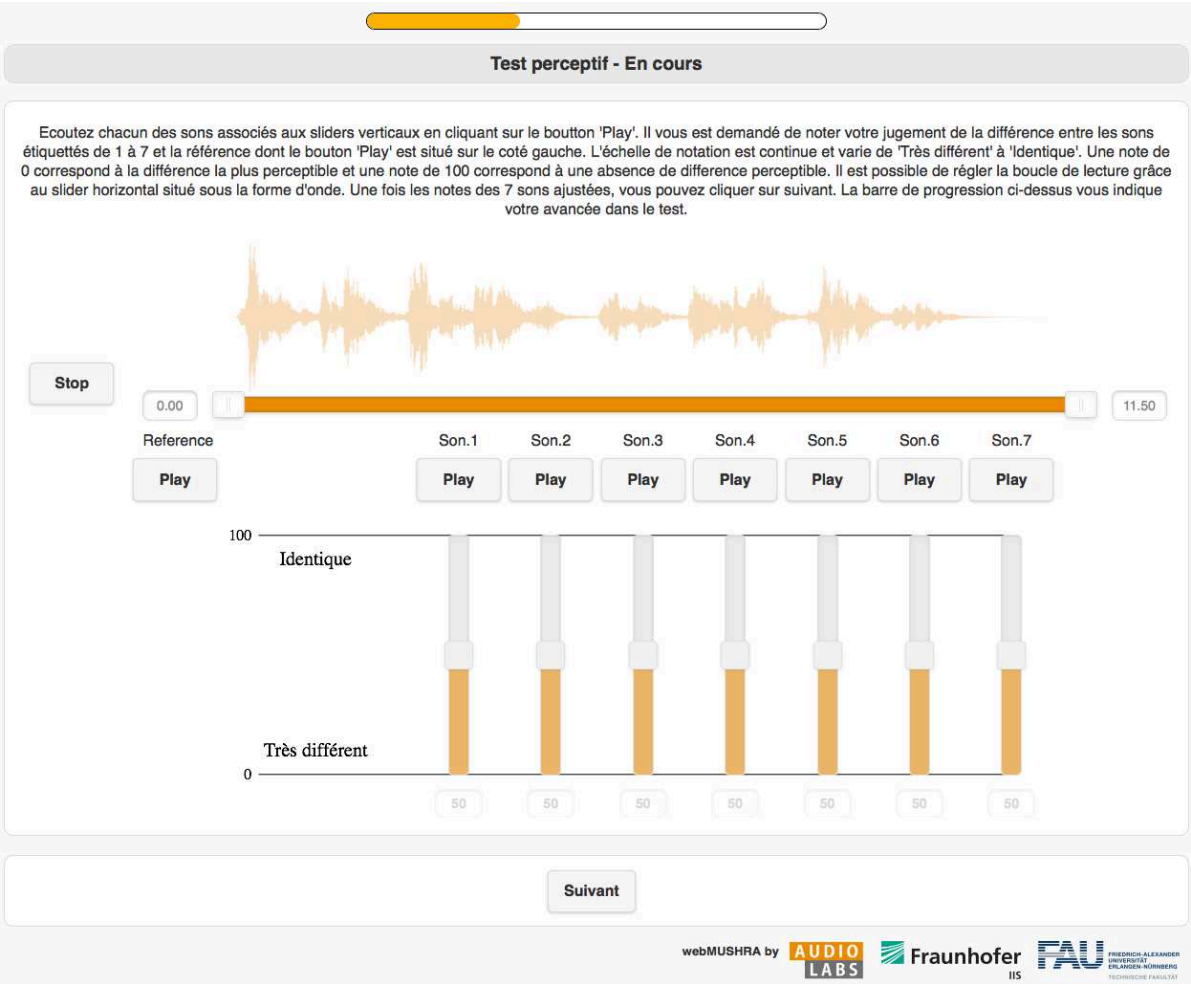


FIGURE 6.4 – Interface graphique du test perceptif.

Réaliser un test sur internet réduit le contrôle expérimental du test perceptif. Des biais peuvent apparaître en raison de l'équipement utilisé et du volume sonore employé. En effet, ce n'est pas seulement l'appréciation des sujets qui varie mais également la restitution des stimuli en raison de la réponse fréquentielle des casques utilisés et de la variation de niveau. Afin de minimiser ces biais, il était demandé aux participants de réaliser le test en possession d'un casque professionnel et nous avons demandé de renseigner la marque du casque utilisé pour vérifier si les réponses fréquentielles étaient accidentées ou non. Au-delà du timbre, la coloration spectrale du casque peut également modifier l'externalisation. Néanmoins, cette influence peut être considérée comme négligeable par rapport à la dégradation induite par l'usage d'HRTFs non-individualisées et de l'absence de dispositif de suivi des mouvements de tête. D'autres biais sont liés au bruit environnant ou à l'implication des sujets. Le temps passé pour effectuer le test était mesuré pour juger de l'attention des sujets et il était demandé de réaliser le test dans un environnement calme.

La question du volume sonore

Avant de procéder au test, il était nécessaire de calibrer le volume sonore du système d'écoute des participants. Ce réglage s'est effectué en plusieurs étapes d'après un protocole de calibration imaginé par Stéphane Pigeon [1] :

— Dans un premier temps, il était demandé d'écouter au casque un premier son de calibration. Ce son consistait en un enregistrement binaural réalisé par un individu de ses deux mains frottées devant son nez. Les participants devaient retirer leur casque et reproduire le même son en frottant leurs mains devant leur nez. Le son de référence ainsi reproduit possède un niveau d'environ 65 dB SPL. Les sujets devaient répéter ce processus tout en ajustant le volume de sortie de leur ordinateur ou carte son de manière à ce que le niveau du son de calibration corresponde à celui qu'ils produisent. Il leur était suggéré de fermer les yeux pour accroitre leur concentration. Ainsi, les sujets ont fixé le volume d'écoute de manière à ce qu'il corresponde au niveau sonore d'une source de référence.

— Dans un second temps, il était demandé de conserver le casque audio sur la tête et de ne pas toucher le volume sonore de l'ordinateur ou carte son. Un second son de calibration était diffusé. Il consistait en une sinusoïde de 1 kHz possédant un faible niveau $l = -61.13$ dBFS. Un slider de l'interface graphique permettait aux sujets de modifier un gain g_l (compris entre 0 et 1) appliqué au son de calibration. Les sujets devaient régler sa valeur au niveau minimum à partir duquel ils percevaient ce son de faible niveau. Ce procédé permet de détecter le niveau de diffusion sonore en décibel $\epsilon = l + 20\log 10(g_l)$ correspondant au seuil d'audibilité à 1 kHz des sujets. Le niveau de diffusion du test était ensuite fixé à $\epsilon + 60$ dB.

— Enfin, un stimulus était diffusé pour présenter le niveau de diffusion calculé pour la suite du test. Ce niveau n'était pas destiné à être modifié, cependant si il n'était pas confortable pour les sujets, il leur était demandé de le réajuster.

L'accroissement de l'externalisation

Aucun dispositif de suivi des mouvements de tête fiable n'a pas pu être mis en place pour ce test perceptif. Néanmoins, l'utilisation d'un dispositif de suivi des mouvements de la tête présente l'avantage de réduire à la fois les ambiguïtés de localisation et d'améliorer l'externalisation [25–27].

Que la scène sonore soit fixe par rapport aux mouvements de tête ou non, l'externalisation d'une source sonore est minimale lorsqu'elle est localisée dans les directions avant ou arrière et maximale lorsqu'elle est située sur les cotés [28, 29]. Afin d'accroitre la sensation d'externalisation, les scènes sonores ont été tournées de +45° et -45° dans le plan horizontal. Ces directions ont été préférées à une rotation de 90° jugée trop éloignée d'une configuration usuelle d'écoute et pouvant provoquer une gène pour les sujets en raison des grandes différences de niveau des signaux binauraux. Les stimuli ont été présentés aux sujets selon une rotation de -45° ou de 45° de manière aléatoire.

1. https://hearingtest.online

6.3.6. Sujets

24 sujets âgés en moyenne de 24 ans ont pris part au test perceptif. En dehors de deux sujets - l'un ayant une formation musicale et l'autre étant ingénieur en traitement du signal audio - les participants étaient étudiants du Master Image & Son de l'Université de Brest ou de l'ENS Louis-Lumière. Ils étaient donc déjà formés à l'écoute critique en raison des enseignements reçus en prise de son, montage et mixage. Aucun d'entre eux n'a déclaré de perte auditive connue.

6.4. Résultats

Comme aucun point d'ancrage intermédiaire n'a été utilisé sur l'échelle de notation, les résultats ont été normalisés par rapport à la moyenne et à l'écart-type en utilisant la normalisation du score z [30].

6.4.1. Outliers

Les données ont été soumises à une ANOVA à mesures répétées avec les variables indépendantes suivantes : «source» (2) × «espace» (5) × «méthode» (7). L'analyse des observations appartenant à la même combinaison de niveaux de variables indépendantes a montré que les données récoltées contenaient des valeurs aberrantes (*outliers*). Les résidus studentisés des résultats étaient supérieurs à 3 en valeur absolue dans certaines cellules pour 4 sujets. L'ANOVA étant sensible aux valeurs aberrantes nous avons fait le choix d'écarter ces sujets. Parmi ces 4 sujets, trois ont réalisé le test en moins de 18 minutes (le temps moyen passé par les sujets était de 28 minutes). Un autre sujet a été exclu de l'analyse pour avoir rapporté qu'il avait effectué le test dans un environnement bruyant. Parmi les 19 sujets dont les réponses ont été utilisées pour l'analyse, 5 ont réajusté le volume sonore calculé à l'étape de calibration.

6.4.2. Analyse de la variance

Facteur	df	SS	MS	F	p
SO	1	65202	65202	110.917	<0.001
SP	4	138828	34707	33.010	<0.001
MT	6	219130	36521	80.036	<0.001
SO * SP	4	2107	526	0.649	0.630
SO * TM	6	27794	4632	13.453	<0.001
SP * TM	24	39786	1657	4.615	<0.001

Tableau 6.2 – Résultats de l'ANOVA. SO : source, SP : espace, MT : méthode, df : degré de liberté, SS : somme des carrés de type III, MS : moyenne des carrés, F : valeur f, p : valeur p.

Pour pouvoir légitimement réaliser une ANOVA à mesures répétées, l'hypothèse de normalité doit être remplie : les résidus des observations appartenant à la même combinaison de niveaux de variables indépendantes doivent être normalement distribués [31]. Un test de Shapiro-Wilk de normalité sur les résidus studentisés a été effectué sur chaque cellule avec un niveau de signification de 5% et a rejeté l'hypothèse nulle d'une distribution normale pour 7 cellules sur 70. La plupart des études statistiques indiquent que l'ANOVA peut être robuste à ce genre de violation [32], surtout si la taille de l'échantillon est supérieure à 15 observations par cellule [33]. Avec 19 observations par cellule, nous avons considéré que l'utilisation de l'ANOVA pour l'analyse statistique était toujours légitime.

L'hypothèse de sphéricité signifie que les variances des différences entre toutes les paires possibles de combinaisons de variables indépendantes doivent être égales. Le test de sphéricité de Mauchly a indiqué que l'hypothèse de sphéricité était remplie par les variables indépendantes ainsi que pour leurs interactions sauf pour la variable indépendante «espace». La valeur F correspondante a donc dû être corrigée en utilisant la correction de Greenhouse-Geisser [32]. Les résultats sont présentés dans le tableau 6.2. L'analyse a révélé une influence significative des variables indépendantes «source» [F(1, 18) = 110.917 ; p = <0.001], «espace» [F(4, 72) = 33.010 ; p = <0.001] et «méthode» [F(6, 108) = 80.036 ; p = <0.001] ainsi que pour les interactions «source» × «méthode» [F(6, 108) = 13.453 ; p = <0.001] et «espace» × «méthode» [F(24, 432) = 4.615 ; p = <0.001].

6.4.3. Comparaisons des niveaux du facteur «méthode»

La figure 6.5 représente les notes moyennes de similarité et l'intervalle de confiance à 95% pour les différents niveaux de la variable indépendante «méthode». Des tests LSD de Fisher avec correction de Bonferroni ont été effectués pour déterminer les différences significatives entre les méthodes de paramétrisation. Les résultats sont présentés dans le tableau 6.3. Pour toutes les méthodes de paramétrisation les stimuli sonores ont été jugés significativement différents de la référence ($p < 0.001$). Les seules méthodes non significativement différentes entre elles sont Ancre et F1T1 ($p = 0.05$), F1T1 et F2T1 ($p = 0.554$), F2T1 et SDM ($p = 1.000$), F1T2, F2T2 et SDM ($p > 0.070$).

Méthode	Ancre	F1T1	F2T1	F1T2	F2T2	SDM	REF
Ancre	-	0.05	0.001	<0.001	<0.001	<0.001	<0.001
F1T1		-	0.554	<0.001	<0.001	<0.001	<0.001
F2T1			-	0.011	0.004	1.000	<0.001
F1T2				-	1.000	0.484	<0.001
F2T2					-	0.070	<0.001
SDM						-	<0.001

Tableau 6.3 – Valeurs p des tests LSD de Fisher avec correction de Bonferroni. Les valeurs significatives sont reportées en gras. REF désigne la référence.

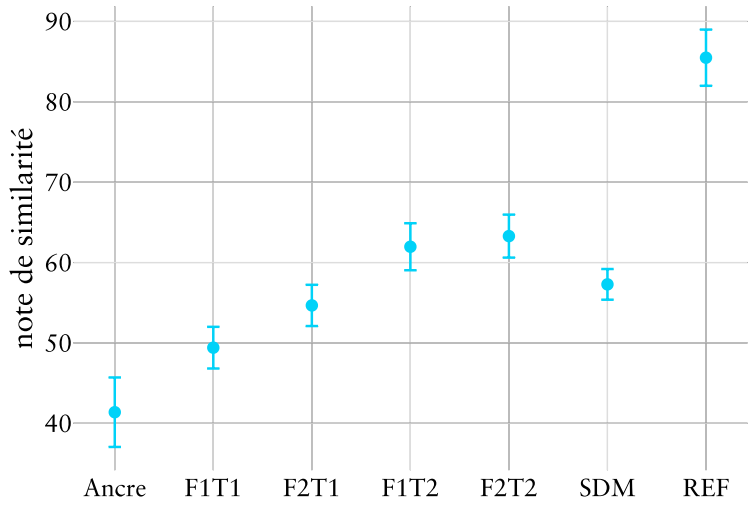


FIGURE 6.5 – Notes moyennes de similarité et intervalles de confiance à 95% associés pour les 7 méthodes de paramétrisation.

6

6.4.4. Résultats des interactions avec le facteur «méthode»

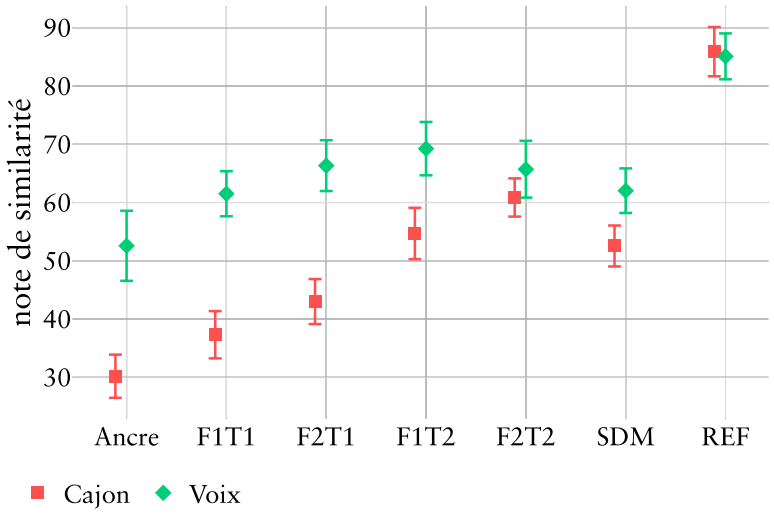


FIGURE 6.6 – Notes moyennes de similarité et intervalles de confiance à 95% associés pour les 7 méthodes selon les deux sources sonores.

La figure 6.6 affiche les notes moyennes de similarité et l'intervalle de confiance à 95% pour chaque méthode de paramétrisation selon la source sonore. Ces notes apparaissent plus faibles et plus étendues lorsque la source *Cajon* était utilisée : la moyenne minimale et maximale s'élevaient à 30.18 et 60.86 pour *Cajon* contre 52.66 et 69.24 pour *Voix*.

La figure 6.7 affiche les notes moyennes de similarité et l'intervalle de confiance à 95% pour chaque méthode selon l'espace sonore. Des tests LSD de Fisher avec correction de Bonferroni ont été effectués pour déterminer :

1. si les notes moyennes des différents traitements étaient significativement différentes de la méthode SDM pour chaque espace ;
2. si les notes moyennes des paramétrisations étaient significativement différentes de la référence cachée pour chaque espace.

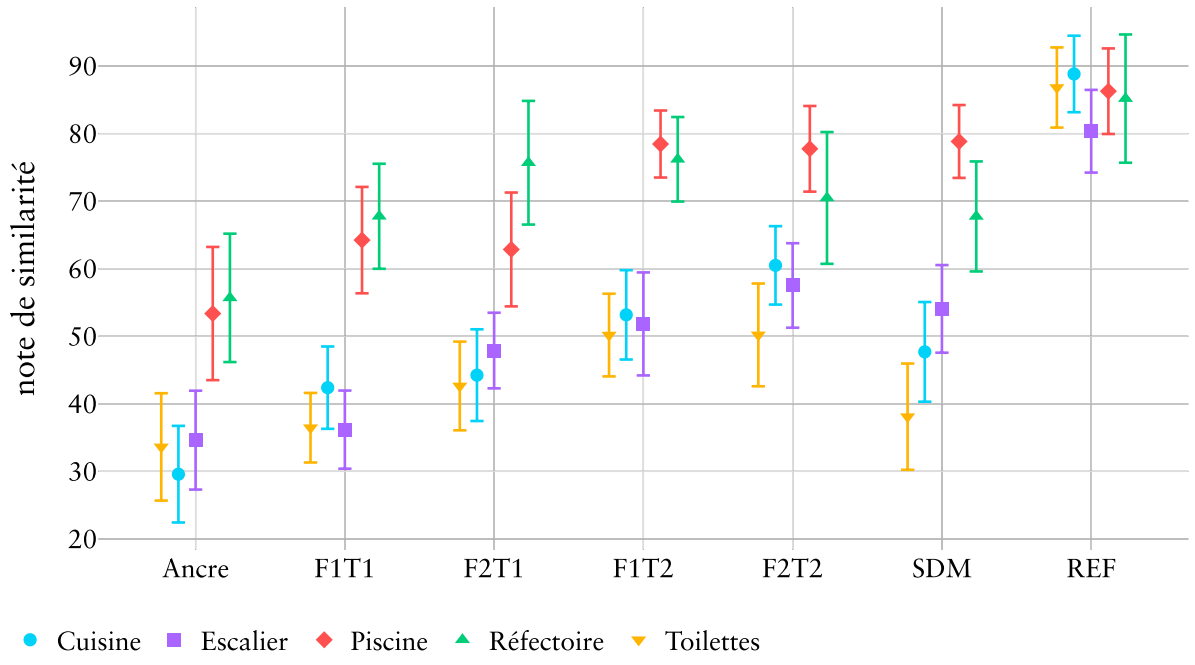


FIGURE 6.7 – Notes moyennes des notes de similarité et intervalles de confiance à 95% associés pour les 7 méthodes selon les 5 espaces sonores.

Méthode	Cuisine	Escalier	Piscine	Réfectoire	Toilettes
F1T1	1.63	1.87	0.67	0.51	2.26
F2T1	1.49	1.35	0.70	-	1.50
F1T2	1.43	1.00	-	-	1.44
F2T2	1.08	0.85	-	-	1.32
SDM	1.45	1.05	-	0.66	1.91

Tableau 6.4 – Coefficient d de Cohen, caractérisant la taille d'effet, pour les méthodes significativement différentes de la référence cachée, selon les cinq espaces sonores. Les valeurs supérieures à 0.8 apparaissent en rouge (effet large) et les valeurs inférieures en vert (effet moyen).

D'une part, on observe une différence significative entre les notes associées à la méthode F1T1 et la SDM pour les espaces *Escalier* ($p < 0.001$) et *Piscine* ($p = 0.033$) ainsi qu'entre les notes associées à l'ancre et la SDM pour les espaces *Cuisine* ($p = 0.027$), *Escalier* ($p = 0.007$) et *Piscine* ($p = 0.003$). D'autre part, les stimuli n'ont pas été perçus significativement différents de la référence cachée dans quatre cas de figure : pour les méthodes F1T2, F2T2 et SDM de l'espace *Piscine* ($p > 1.000$ dans tous les cas) et pour les méthodes F2T1 et F1T2, F2T2 dans l'espace *Réfectoire* ($p > 1.000$ dans tous les cas). Le tableau 6.4 recense les effets de taille obtenus dans les autres

cas en utilisant le coefficient d de Cohen [34]. D'après les conventions établies par Cohen, la taille d'effet est considérée comme faible lorsque d = 0.2, moyen lorsque d = 0.5 et large lorsque d = 0.8. Les effets de tailles sont larges pour les espaces *Cuisine*, *Escalier* et *Toilettes*.

6.5. Discussion

6.5.1. Le niveau de diffusion

Sept des 24 participants ont modifié le niveau de diffusion calculé à l'étape de calibration avant de réaliser le test perceptif. Il était précisé que ce niveau n'était pas destiné à être modifié et qu'il ne devait être ajusté qu'en cas d'inconfort. Malheureusement, les modifications du niveau appliquées par les sujets n'ont pas été enregistrées. Nous ne pouvons donc pas rendre compte de l'écart entre le niveau calculé et le niveau jugé par les sujets comme étant confortable.

Plutôt que de permettre la modification du niveau de diffusion, une nouvelle étape de calibration ou une mise en relation avec l'expérimentateur à ce stade auraient été des options plus judicieuses. Le biais expérimental lié au niveau sonore de diffusion à sans doute été limité grâce à la méthode de calibration mais n'a pas pu être éliminé. Il vient s'ajouter aux biais liés aux équipements audio employés, au bruit environnant et à l'implication des sujets.

6.5.2. L'influence des résolutions employées

La dissemblance perçue entre les SRIRs paramétrées et la référence évolue en cohérence avec la résolution temporelle et fréquentielle des paramétrisations employées. Néanmoins, ces résolutions étaient trop faibles pour être en mesure de reproduire la référence de manière imperceptible.

Il est intéressant de constater que l'augmentation de la résolution fréquentielle des paramétrisations n'a pas amélioré la qualité de la reproduction. En effet, les notes des méthodes F1T1 et F2T1 ne sont pas significativement différentes entre elles ($p = 0.554$) ni les notes des méthodes F1T2 et F2T2 ($p = 1.000$). En revanche, l'augmentation de la résolution temporelle a permis une réduction significative de la différence perçue avec la référence.

6.5.3. L'influence de l'espace

Dans l'étude du chapitre 5 portant sur la résolution spatiale des SRIRs, les mêmes espaces ont été utilisés. Une influence significative de la réduction de l'ordre ambisonique avait été observée pour les espaces *Escalier*, *Réfectoire* et *Toilettes*. Néanmoins, la réduction de la résolution spatiale observée dans le chapitre 5 ne permet pas ici d'expliquer les différences observées entre les espaces.

D'après le tableau 6.4, deux groupes d'espaces peuvent être constitués : les espaces présentant de fortes dissemblances avec la référence (*Cuisine*, *Escalier*, *Toilettes*) et les

Espace	T_{30}	C_{80}	D_{50}	DRR	G_E
Cuisine	0.62	8.75	77.50	-1.89	32.17
Escalier	3.67	-1.20	32.20	-6.08	37.87
Piscine	1.97	*8.34*	76.69	*2.74*	*28.37*
Réfectoire	0.90	*8.57*	79.21	*2.09*	*28.86*
Toilettes	0.40	13.85	86.30	-3.42	32.71

Tableau 6.5 – Paramètres acoustiques liés aux espaces sonores : le temps de réverbération (T_{30}), les rapports d'énergies précoces et tardives (D_{50} et C_{80}), le rapport champ direct champ réverbéré (DRR) et la force sonore de l'énergie précoce (G_E). Les valeurs remarquables apparaissent en bleu.

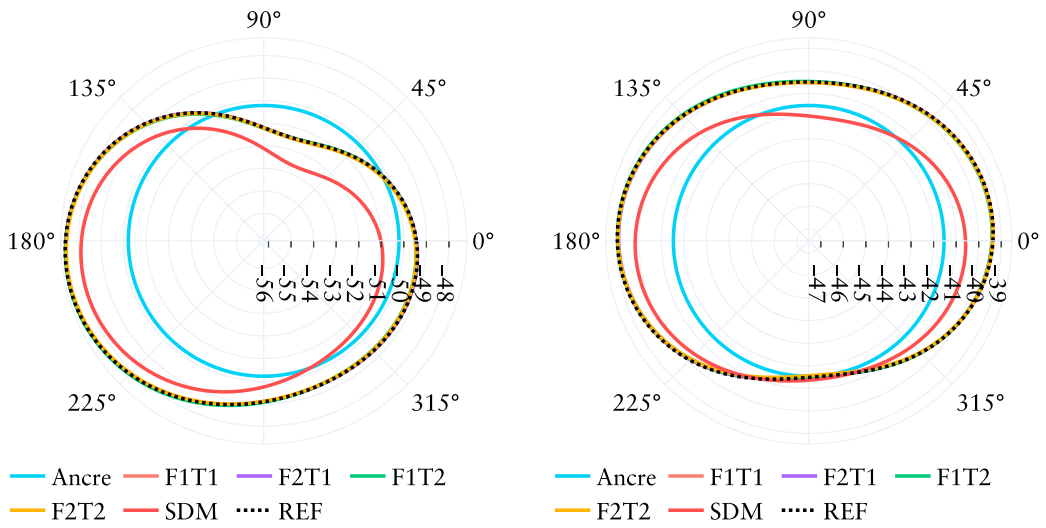
espaces présentant des dissemblances modérées (*Piscine*, *Réfectoire*). Le tableau 6.5 rassemble les paramètres acoustiques des différents espaces. On observe en effet que le DRR des espaces *Piscine* et *Réfectoire* sont les plus élevés et que l'énergie précoce des SRIRs associées sont les plus faibles. Un effet de masquage par le son direct - et par l'énergie tardive pour l'espace *Piscine* - a pu apparaitre. En d'autres termes, il est possible que la modification des premières réflexions ait moins été perçue pour ces espaces car la partie modifiée du signal était en proportion plus faible.

6.5.4. L'influence de la source

Les notes associées à la source *Cajon* ont été significativement plus faibles que celles associées à la source *Voix*. La source contient de nombreux transitoires qui peuvent avoir mis en lumière des artefacts dans une bande fréquentielle plus large qu'avec la source *Voix*. De plus, des changements dans la structure temporelle des SRIRs peuvent avoir été mieux perçus grâce à la nature impulsionnelle du signal percussif. Par ailleurs, pour générer les SRIRs d'après les matrices de covariance, une étape de création de signaux décorrélés est nécessaire. Or, d'après Pihlajamäki *et al.* [13], cette méthode peut introduire des artefacts perceptibles selon la source sonore utilisée ; en particulier pour les signaux comprenant des événements sonores soudains ou impulsionnels.

6.5.5. La méthode SDM

La figure 6.8 représente le diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions des SRIRs de *Piscine* et *Toilettes*. Avec la paramétrisation proposée en section 6.2, la directivité du champ sonore est respectée en raison du respect des matrices de covariance par région temps-fréquence. Avec la SDM, l'énergie du champ sonore correspond à celle de la composante omnidirectionnelle des signaux ambisoniques. Contrairement à la méthode utilisant des matrices de covariance, l'énergie des autres composantes ambisoniques n'est pas reproduite. Dans nos cas de figure, ceci se traduit par une énergie précoce plus faible des stimuli issus de la SDM. Dans le cas de l'espace *Toilettes*, l'énergie du champ sonore



(a) Piscine (b) Toilettes

FIGURE 6.8 – Diagrammes de directivité dans le plan horizontal de l'énergie des premières réflexions des SRIRs associées à *Piscine* et *Toilettes* encodées à l'ordre 1 et modifiées selon les 6 méthodes de paramétrisation.

est moins importante de 2 dB à 90°. Cette dissymétrie crée une concentration d'énergie audible des premières réflexions vers la droite en comparaison à la référence, de la même manière que pour l'ancre. Ceci semble expliquer la non significativité entre les notes associées à ces deux paramétrisations.

6.5.6. L'effet de précédence

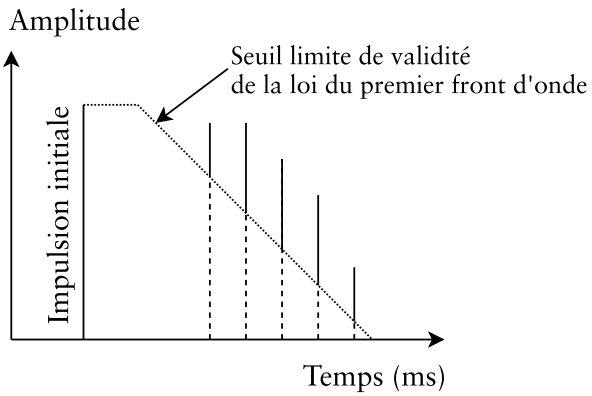


FIGURE 6.9 – Masquage temporel induit par la loi du premier front d'onde.

Bien que les directivités des champs sonores soient très proches entre les stimuli issus des méthodes de paramétrisation par matrice de covariance et la référence - comme illustré en figure 6.8 - des différences sont perceptibles en termes de direction d'incidence. Outre l'intensité des réflexions, les retards des réflexions jouent égale-

ment un rôle important dans la localisation. Selon la loi du premier front d'onde, la direction d'incidence perçue de réflexions qui se produisent dans un court laps de temps correspond à celle du premier front d'onde qui arrive aux oreilles d'un auditeur. Les réflexions ne sont pas perçues comme des événements distincts, mais fusionnent perceptivement. La figure 6.9 illustre le seuil d'audibilité de réflexions émises après une impulsion réflexion d'après Morimoto [35]. Selon Haas [36], cet effet peut se produire même lorsque le niveau d'une réflexion est supérieur de 10 dB au signal du front d'onde initial. L'interaction temporelle entre les réflexions est un phénomène complexe qui dépend à la fois de l'énergie du signal d'origine, des réflexions suivantes et de leur retard. Bien que ce phénomène ait été étudié principalement pour la localisation du son direct, il est possible qu'il se produise également localement au sein des premières réflexions et qu'il influence la perception de leur incidence. La figure 6.10 représente le canal gauche de la partie précoce de BRIRs de référence et celles générées d'après les méthodes SDM et F2T2. Les réflexions spéculaires sont visibles avec des variations en termes de retard et d'énergie. La détermination de réflexions spéculaires pour reproduire la structure temporelle d'une SRIR viendrait ajouter de l'information lors de la reconstruction. Cependant, comme nous l'avons vu dans le chapitre 4, l'identification de réflexions spéculaires est souvent effectuée en utilisant des seuils d'amplitude sans prendre en compte la loi du premier front d'onde. Il semble qu'une meilleure compréhension des mécanismes en jeu est nécessaire pour reproduire convenablement la directivité perçue des réflexions.

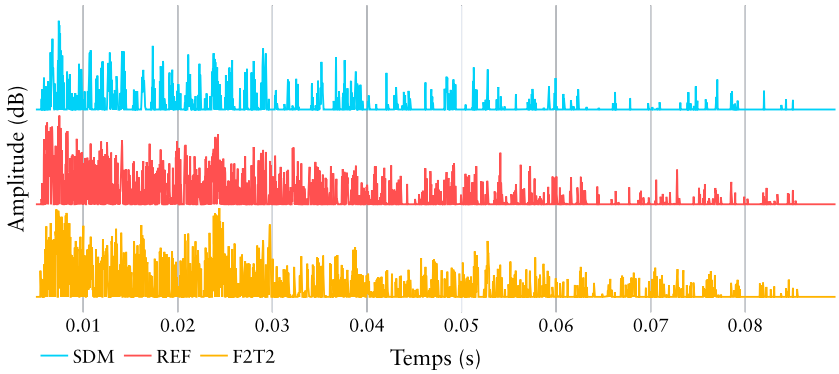


FIGURE 6.10 – Amplitude des premières réflexions en décibel pour trois méthodes de paramétrisation. Seuls les 20 premiers décibels sont affichés.

6.5.7. L'avantage de la paramétrisation par matrice de covariance par rapport à la SDM

La SDM permet un contrôle de la directivité des réflexions permettant de manipuler les caractéristiques spatiales de l'espace auralisé. Cependant la description du champ sonore représente une quantité importante de données : le signal omnidirectionnel associé à une matrice contenant les directions d'incidences pour chaque échantillon. Avec la paramétrisation proposée, un nombre plus faible de paramètres peut être utilisé selon la résolution spatiale, fréquentielle et temporelle employée. La

méthode de paramétrisation F1T2 semble la plus efficace des méthodes employées. En effet, la dissemblance avec la référence était significativement plus faible lorsque la résolution temporelle était la plus élevée. De plus, nous n'avons pas constaté de différences significatives entre les stimuli générés selon 4 ou 8 bandes de fréquence. A l'ordre 1, le nombre de paramètres générés sur une durée de 80 ms s'élève à 640 valeurs pour F1T2 et à 7680 valeurs pour la SDM, soit 12 fois plus de données. La réduction du nombre de paramètres permet d'avoir moins de valeurs à manipuler pour le contrôle perceptif de la réverbération et une réduction de l'utilisation de la mémoire. De plus, il est possible avec la méthode F1T2 de modifier simplement la directivité des réflexions par bande de fréquence. Néanmoins, les traitements a effectuer pour générer une SRIR paramétrée avec des matrices de covariance sont importants. Cette génération requiert la création de signaux décorrélés, un filtrage par bande de fréquence et une rotation des signaux résultants.

6.6. Conclusion

Dans le prolongement du chapitre précédent, nous avons étudié les résolutions temporelles et fréquentielles nécessaires à la restitution des premières réflexions de SRIRs. L'étude s'est placée dans le contexte d'un rendu binaural non-individualisé utilisant des SRIRs dont la partie réverbérée était encodée en ambisonique à l'ordre 1.

Pour reproduire le champ sonore selon différentes résolutions fréquentielles et temporelles, une paramétrisation basée sur le calcul de matrice de covariance a été proposée. Après avoir généré plusieurs SRIRs selon diverses méthodes de paramétrisation, un test perceptif a permis d'étudier l'influence des résolutions employées pour générer les SRIRs. Les stimuli issus de paramétrisations ont été comparés à des stimuli de référence ayant les plus grandes résolutions spatiales, temporelles et fréquentielles possibles. Des différences significatives entre les stimuli issus de paramétrisations et les stimuli de référence ont été obtenues pour toutes les méthodes de paramétrisation.

L'emploi d'une résolution temporelle élevée a significativement amélioré les résultats et aucune différence significative n'a été observée avec l'augmentation de la résolution fréquentielle. La méthode de paramétrisation employant une résolution temporelle de 5 ms dans quatre bandes de fréquence est apparue comme la plus efficace. Cette méthode n'était pas significativement différente de la SDM, une paramétrisation communément employée.

Néanmoins, la fidélité de reproduction des SRIRs dépend de la source et de l'espace considéré. La dégradation de la reproduction induite par la paramétrisation peut être importante dans certains cas. Les différences avec la référence étaient d'autant plus perceptibles que la proportion de signal modifié était importante. De plus, l'usage d'une source comprenant de nombreuses transitoires pouvait révéler des artefacts inhérents au processus de synthèse.

Il est possible que les paramétrisations testées ne respectaient pas suffisamment la structure temporelle des premières réflexions. En effet, même si la répartition de l'énergie des premières réflexions était correctement reproduite, il est possible que la directivité perçue était différente en raison de l'effet de précédence. Une identification fiable des réflexions spéculaires et une meilleure compréhension de leur interaction

permettrait d'ajouter de l'information nécessaire à une reproduction plus fidèle du champ sonore.

La paramétrisation proposée offre cependant la possibilité de contrôler le rendu sonore d'une auralisation en manipulant des SRIRs dans plusieurs bandes de fréquence. Le nombre de paramètres employés est moins important que celui de la SDM, ce qui facilite l'étude des effets sur les attributs perceptifs de modifications des caractéristiques spatiales d'une SRIR.

Bibliographie

[1] P. Coleman, A. Franck, D. Menzies et P. J. Jackson, "Object-based reverberation encoding from first-order ambisonic RIRs," dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[2] P. Stade, J. M. Arend et C. Pörschmann, "Perceptual evaluation of synthetic early binaural room impulse responses based on a parametric model," dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[3] D. R. Begault, B. U. McClain et M. R. Anderson, "Early reflection thresholds for virtual sound sources," dans *Proc. 2001 Int. Workshop on Spatial Media*, 2001.

[4] J. Merimaa et V. Pulkki, "Spatial impulse response rendering i : Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, n°. 12, p. 1115–1127, 2005.

[5] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger et M. Marschall, "Higher-order spatial impulse response rendering : Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *Journal of the Audio Engineering Society*, vol. 68, n°. 5, p. 338–354, 2020.

[6] S. Tervo, J. Pätynen, A. Kuusinen et T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, n°. 1/2, p. 17–28, 2013.

[7] J. Pätynen, S. Tervo et T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, n°. 2, p. 842–857, 2013.

[8] S. V. A. Garí, M. Kob et T. Lokki, "Investigations on stage acoustic preferences of solo trumpet players using virtual acoustics," dans *Proceedings of the 14th Sound and Music Computing Conference*, 2017, p. 8.

[9] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech et T. Lokki, "Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array," *Journal of the Audio Engineering Society*, vol. 63, n°. 11, p. 914–925, 2015.

[10] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot et S. H. Jensen, "Perceptual aspects of reproduced sound in car cabin acoustics," *The Journal of the Acoustical Society of America*, vol. 141, n°. 3, p. 1459–1469, 2017.

[11] S. V. A. Garí, W. O. Brimijoin, H. G. Hassager et P. W. Robinson, "Flexible binaural resynthesis of room impulse responses for augmented reality research," dans *EAA Spatial Audio Signal Processing Symposium*, Paris, France, 2019.

[12] G. Potard et I. Burnett, “Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays,” dans *Proc. Int. Conf. on Digital Audio Effects (DAFx’04)*, 2004.

[13] T. Pihlajamäki, O. Santala et V. Pulkki, “Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals,” *Journal of the Audio Engineering Society*, vol. 62, n°. 7/8, p. 467–484, 2014.

[14] J. Halton et G. Smith, “Radical inverse quasi-random point sequence, algorithm 247,” *Commun. ACM*, vol. 7, n°. 12, p. 701, 1964.

[15] A. Aguilera et R. Pérez-Aguila, “General n-dimensional rotations,” 2004.

[16] ITU-R BS.1534-3, “Method for the subjective assessment of intermediate quality level of audio systems.” International Telecommunication Union, Standard, 2015.

[17] P. Götz, K. Kowalczyk, A. Silzle et E. A. Habets, “Mixing time prediction using spherical microphone arrays,” *The Journal of the Acoustical Society of America*, vol. 137, n°. 2, p. EL206–EL212, 2015.

[18] MH Acoustics LLC, “Eigenmike em32 microphone array,” accessed 2020-08-22. [En ligne]. Disponible : https://mhacoustics.com/products

[19] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[20] C. Schörkhuber, M. Zaunschirm et R. Höldrich, “Binaural rendering of ambisonic signals via magnitude least squares,” dans *Proceedings of the DAGA*, vol. 44, 2018, p. 339–342.

[21] M. Zaunschirm, C. Schörkhuber et R. Höldrich, “Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint,” *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3616–3627, 2018.

[22] S. Zielinski, F. Rumsey et S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, n°. 6, p. 427–451, 2008.

[23] E. C. Poulton et S. Poulton, *Bias in quantifying judgements*. Taylor & Francis, 1989.

[24] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler et J. Herre, “webMUSHRA — a comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, vol. 6, n°. 1, 2018. [En ligne]. Disponible : https://github.com/audiolabs/webMUSHRA

[25] H. Wallach, “The role of head movements and vestibular and visual cues in sound localization.” *Journal of Experimental Psychology*, vol. 27, n°. 4, p. 339, 1940.

[26] D. R. Begault, E. M. Wenzel et M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *J. Audio Eng. Soc.*, vol. 49, p. 904–916, 2001.

[27] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz et C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, vol. 141, p. 3678–3688, 2017a, https ://doi.org/10.1121/1.4978612.

[28] W. O. Brimijoin, A. W. Boyd et M. A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PloS one*, vol. 8, n°. 12, p. e83068, 2013.

[29] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz et C. De Boishéraud, "Influence du «head-tracking» sur l'externalisation en écoute binaurale non-individualisée," dans *CFA'18 Le Havre, 14ème Congrès Français d'Acoustique*, 2018, p. 459–465.

[30] E. Kreyszig, "Advanced engineering mathematics, 10th eddition," 2009.

[31] H. Keselman, J. C. Rogan, J. L. Mendoza et L. J. Breen, "Testing the validity conditions of repeated measures f tests." *Psychological Bulletin*, vol. 87, n°. 3, p. 479, 1980.

[32] K. Weinfurt, "Repeated measures analysis : Anova, manova, and hlm," *Reading and Understanding More Multivariate Statistics*, 10 2000.

[33] S. B. Green et N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[34] J. Cohen, "A power primer." *Psychological bulletin*, vol. 112, n°. 1, p. 155, 1992.

[35] M. Morimoto, K. Nakagawa et K. Iida, "The relation between spatial impression and the law of the first wavefront," *Applied Acoustics*, vol. 69, n°. 2, p. 132–140, 2008.

[36] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Engineering Society*, vol. 20, n°. 2, p. 146–159, 1972.

III

Étude du contrôle perceptif des impressions spatiales

7

Les origines physiques des impressions spatiales

Dans la partie précédente, nous avons identifié parmi les échantillons composant une SRIR des paramètres permettant la reproduction de l'acoustique d'une salle avec un nombre de données limité. La résolution spatiale, temporelle et fréquentielle de SRIRs ont notamment été étudiées dans un contexte de rendu binaural non-individualisé. Nous souhaitons dans cette partie étudier la manière dont la modification des paramètres identifiés peut permettre de contrôler les impressions spatiales, à savoir la largeur apparente de source et l'enveloppement. Dans un premier temps, ce chapitre se concentre sur les origines physiques de ces deux attributs perceptifs en étudiant leurs liens avec des paramètres acoustiques extraits de réponses impulsionnelles de salle. Pour cela, nous utilisons une base de données fournie par l'Université technique de Berlin rassemblant des résultats issus d'une évaluation perceptive et des paramètres acoustiques associés aux stimuli évalués. Notre analyse des données confirme l'influence de la quantité d'énergie tardive, des réflexions latérales tardives et de la décorrélation interaurale sur l'enveloppement. Elle met également en évidence une influence de l'énergie tardive et du rapport entre champ direct et champ réverbéré sur la largeur apparente de source. Pourtant, ces deux propriétés ne sont pas prises en compte dans le calcul des paramètres acoustiques définis dans la littérature pour estimer cet attribut perceptif.

Le terme «impressions spatiales» désigne dans la littérature deux attributs perceptifs distincts liés à des aspects différents du champ sonore : la largeur apparente de source notée ASW (pour *Apparent Source Width*) et l'enveloppement noté LEV (pour *Listener Envelopment*) [1, 2]. Le premier correspond à l'étendue spatiale de la source dans le plan horizontal et le second correspond à la sensation d'être entouré d'un ensemble diffus d'images sonores qui ne sont pas associées à des emplacements de source particuliers [1]. Afin d'établir un contrôle perceptif de ces impressions spatiales, il est nécessaire d'identifier les propriétés du signal sonore liées à ces deux attributs. D'après

la littérature, la largeur apparente de source semble être principalement influencée par les propriétés spatiales des premières réflexions et l'enveloppement est généralement lié aux propriétés spatiales de la réverbération tardive.

Pour caractériser ces impressions spatiales, les études de Barron [3] et de Bradley et Soulodre [4, 5] ont permis d'établir des paramètres acoustiques présents dans la norme ISO 3382-1 [2] qui mesurent l'énergie latérale précoce (J_{LF}, J_{LFC}) et l'énergie latérale tardive (L_J). D'autres paramètres acoustiques tels que la décorrélation interaurale calculée sur la durée totale d'une réponse impulsionnelle binaurale de salle (IACC) ou sur la partie précoce seulement (BQI) permettent de caractériser les impressions spatiales.

Les paramètres acoustiques spécifiés dans la norme présentent cependant quelques défauts. D'une part, les paramètres acoustiques liés à la largeur apparente de source peuvent présenter de larges fluctuations entre deux mesures proches sans qu'une modification de cet attribut soit perceptible [6]. D'autre part, la sensation d'enveloppement est caractérisée dans la norme par l'énergie latérale tardive seulement, alors que d'autres directions d'incidence peuvent contribuer à cet attribut [7, 8]. Pour compléter ces mesures, d'autres paramètres acoustiques ont donc été proposés dans la littérature.

Dans ce chapitre, nous souhaitons vérifier la pertinence des paramètres définis dans la norme ISO 3382-1, étudier leur indépendance et tester d'autres paramètres acoustiques proposés dans la littérature. Pour être capable d'étudier ces paramètres pour plusieurs acoustiques et d'avoir une puissance statistique suffisante, il est nécessaire d'en calculer à partir d'un grand nombre de réponses impulsionnelles spatiales de salle (RIRs). Mesurer des RIRs dans de nombreuses salles étant un processus chronophage, une solution plus simple consiste à les calculer en simulant la propagation d'ondes sonores grâce à des modèles tri-dimensionnels de salles. La base de données GRAP [9], rassemble des réponses impulsionnelles simulées associés à une évaluation perceptive des impressions spatiales. Cette ressource a précisément été mise à disposition pour déterminer de nouveaux paramètres acoustiques au-delà de ceux définis dans la norme.

Nous présenterons dans un premier temps le contenu de la base de données GRAP. Afin de compléter ces données, des réponses impulsionnelles spatiales de salle (SRIR) ont été générées d'après des résultats de simulation présents dans la base. Le processus de génération des SRIRs sera présenté en section 7.2. Plusieurs paramètres acoustiques seront calculés d'après l'ensemble des réponses impulsionnelles à disposition. Les paramètres acoustiques identifiés comme pertinents pour caractériser la largeur apparente de source et l'impression d'enveloppement seront présentés dans les sections 7.4 et 7.5. Ces paramètres permettront de déterminer les propriétés physiques à modifier pour le contrôle perceptif des impressions spatiales étudié dans les chapitres 8 et 9.

7.1. La base de données GRAP

Ackermann *et al.* [9] ont développé la base de données GRAP (*Ground Truth on Room Acoustical Analysis and Perception*) afin de permettre le développement

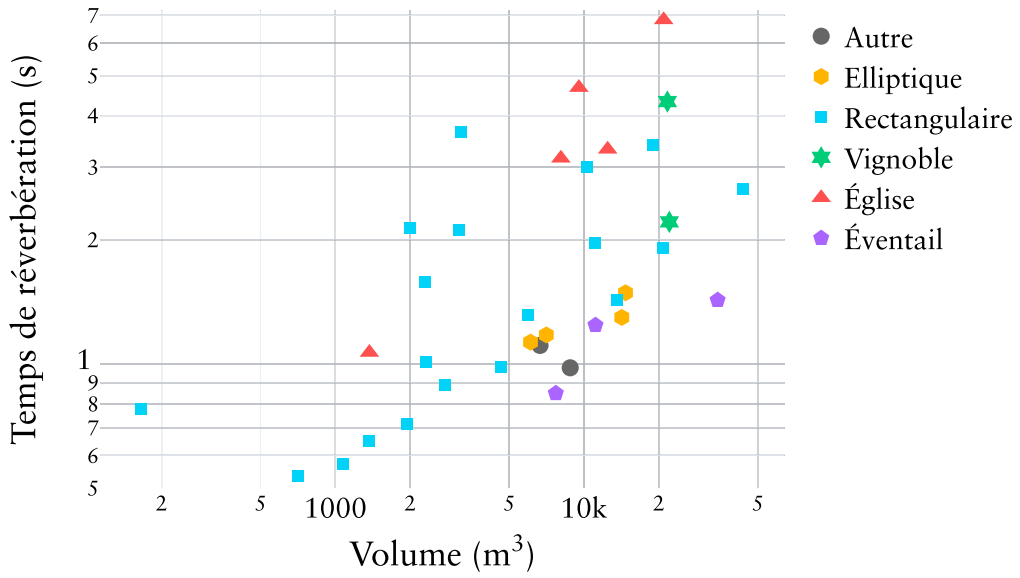


FIGURE 7.1 – Ensemble des 35 salles constituant la base de données GRAP représentées en fonction de leur volume, leur temps de réverbération et leur forme dans le plan horizontal. Une salle de concert est dite en «vignoble» lorsque le public entoure les musiciens en étant disposé sur différents balcons, à la manière de la philharmonie de Berlin et de Paris.

de paramètres acoustiques au-delà de ceux définis dans la norme ISO 3382-1 [2]. Elle rassemble des données physiques et perceptives liées à un ensemble de salles de concert, de théâtre, de séminaire ou d'amphithéâtres. Pour acquérir ces données, 35 modèles de salles virtuelles ont été créés d'après des salles existantes ou non dans le but de couvrir une grande variété de propriétés acoustiques. La figure 7.1 représente l'ensemble des salles incluses dans la base de données en fonction de leur volume, leur temps de réverbération et leur forme. Cet ensemble est constitué principalement de salles de grand volume dont la plupart sont de forme rectangulaire.

Pour chaque environnement acoustique, la base de données GRAP comporte un modèle tridimensionnel qui spécifie la géométrie et les propriétés acoustiques des surfaces de la salle ainsi que la position de la source et les positions de deux récepteurs. Une position de récepteur dans une salle sera par la suite nommée configuration acoustique. La base de données étant constituée de deux positions de récepteur pour 35 salles, elle contient donc 70 configurations acoustiques.

Pour chaque configuration acoustique, la base de données contient :

— des notes associées aux attributs perceptifs contenus dans le RAQI (cf. section 1.2.3) pour plusieurs sources sonores ;
— une réponse impulsionnelle omnidirectionnelle et 360 réponses impulsionnelles binaurales ;
— les paramètres acoustiques de la norme ISO 3382-1 calculés d'après les réponses impulsionnelles ;
— les résultats de simulation sous la forme d'une distribution d'énergie au cours du temps et d'une liste d'ondes planes associées à un temps d'arrivée, une direction et une amplitude par tiers d'octave.

7.1.1. Les notes du RAQI

Un test perceptif réalisé par Weinzierl *et al.* [10] a permis d'évaluer la perception de différentes sources pour chacune des 70 configurations acoustiques selon les 46 attributs perceptifs du RAQI (cf. section 1.2.3).

Pour ce faire, des réponses impulsionnelles binaurales (BRIRs) ont été simulées pour chaque configuration acoustique. Les réponses impulsionnelles relatives à la tête (HRIRs) utilisées pour générer les BRIRs étaient issues de la base de données FABIAN [11]. Dans le but de prendre en compte les mouvements de la tête de l'auditeur, 71 BRIRs ont été simulées pour des orientations de tête comprises entre -70° et +70° par pas de 2° dans le plan horizontal. Seule la partie précoce des BRIRs a été reproduite selon un rendu binaural dynamique. Le rendu binaural a été réalisé au moyen du SoundScape Renderer [12] en utilisant un casque extra-aural [1] et une compensation spectrale de la fonction de transfert du casque [13].

Les stimuli ont été obtenus par convolution avec les signaux de trois contenus sonores :

— Un solo de trompette - *Trumpet Voluntary* de J. Clarke - enregistré dans une chambre anéchoïque.
— La voix d'un homme déclamant un discours de Cicéron - enregistrée dans une chambre anéchoïque [14].
— Un orchestre symphonique dont chaque instrument a été enregistré séparément dans une chambre anéchoïque [15].

Seules 25 salles ayant la capacité de contenir un orchestre symphonique, l'évaluation perceptive de ce contenu n'a pas pu être réalisée pour 10 salles.

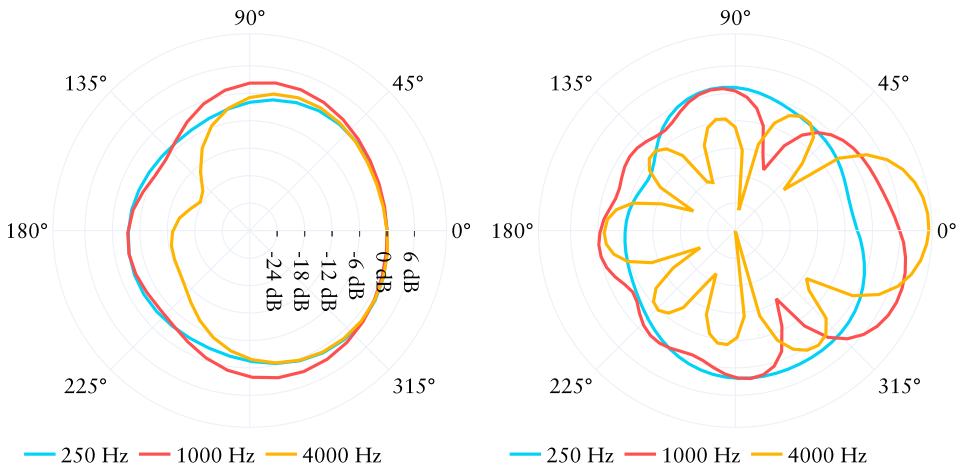
Chaque BRIR a été simulée pour chaque source sonore en prenant en compte leur directivité. La figure 7.2 représente les diagrammes de directivité des deux contenus monophoniques employés dans la génération des stimuli. On constate le caractère plus accidenté de la directivité de la trompette qui s'éloigne le plus d'une source omnidirectionnelle et implique une prédominance de certaines sources images. Ces directivités ont été mesurées en chambre anéchoïque avec une antenne sphérique de microphones [16].

190 sujets ont participé à l'évaluation des stimuli générés pour les 70 configurations acoustiques. Pour des questions de durée du test perceptif, chaque sujet a jugé une partie des stimuli seulement (14 stimuli) à la lumière des 46 attributs perceptifs. Pour chaque configuration acoustique, les notes des attributs pour chaque source sont incluses dans la base de données GRAP.

7.1.2. Les réponses impulsionnelles de salle

La base de donnée GRAP contient une réponse impulsionnelle omnidirectionnelle et un ensemble de BRIRs calculées tous les degrés dans le plan horizontal soit 360 BRIRs. Les procédés de simulation de propagation des ondes sonores décrits en annexe D ont été utilisés pour calculer les réponses impulsionnelles de GRAP. Les ré-

1. Un casque extra-aural consiste en une paire de haut-parleurs solidaires de l'arceau du casque et suspendus à une certaine distance des oreilles d'un auditeur.



(a) Voix (b) Trompette

FIGURE 7.2 – Diagrammes de directivité dans le plan horizontal des sources employées pour la simulation des BRIRs utilisées dans le test perceptif.

sultats sont issus d'une approche hybride qui emploie la méthode des sources images jusqu'au troisième ordre pour calculer les réflexions spéculaires et la méthode du lancé de particules (avec 200 000 particules) pour calculer une distribution temporelle de l'énergie des réflexions diffuses. Toutes les simulations ont été réalisées avec une résolution en fréquence en tiers d'octave de 20 Hz à 20 kHz, soit pour 31 bandes de fréquence à l'aide du logiciel de simulation RAVEN [17]. Comme recommandé par la norme ISO 3382-1, la source utilisée pour les simulations était une source omnidirectionnelle.

Dans la base de données GRAP, la médiane des fréquences de Schrœder est de 29.68 Hz et les valeurs minimales et maximales s'élèvent à 12.87 Hz et 136.92 Hz respectivement. Étant donné que seule l'approche géométrique a été employée pour la simulation des réponses impulsionnelles, la simulation n'est pas valide en dessous de ces valeurs fréquentielles. Néanmoins ces valeurs restent faibles pour la plupart des salles.

Les BRIRs fournies dans GRAP ont été obtenues d'après les HRIRs de la base de données FABIAN [11]. Le logiciel RAVEN calcule les BRIRs de la manière suivante :

— Pour la partie spéculaire, le signal résultant de chaque réflexion spéculaire est calculé d'après les HRIRs de la direction d'incidence, retardé par le temps de parcours et atténué par l'absorption des parois sur lesquelles la réflexion s'est réfléchie. L'absorption de l'air est également prise en compte en fonction de la distance du trajet effectué.

— La résolution de l'histogramme d'énergie résultant du lancé de particules étant inférieure à la résolution temporelle donnée par le taux d'échantillonnage, la structure temporelle des BRIRs est construite d'après un processus aléatoire basé sur une distribution de Poisson [2]. Une séquence temporelle est générée

2. la probabilité de collision d'une particule avec un détecteur vérifie une loi de Poisson [18]

pour chaque canal de diffusion afin de reproduire la partie non spéculaire. Ces séquences sont paramétrées de sorte que la densité d'écho théorique donnée par l'équation (1.1) soit respectée pour l'espace considéré.

— La décroissance de l'énergie est déterminée pour chaque bande de fréquence et plusieurs secteurs angulaires grâce aux résultats du lancé de particules. Par intervalle de temps régulier - de l'ordre d'une milliseconde - les bruits sont filtrés avec une paire de HRIRs correspondant à la direction d'incidence la plus probable. Cette probabilité est calculée d'après les courbes de décroissance obtenues par secteur angulaire, où une plus grande énergie mène à une plus grande probabilité d'incidence et inversement.

— Enfin, la décroissance des deux bruits ainsi générés (un par oreille) est ajustée par bande d'octave de manière à ce qu'elle correspondent à la décroissance calculée d'après l'histogramme d'énergie résultant du lancé de particules, tous secteur angulaire confondu.

Le processus est équivalent pour le calcul des réponses impulsionnelles omnidirectionnelles, sans considération de direction d'incidence.

7.2. Ajout de données complémentaires : la génération de SRIRs

Il nous a semblé intéressant de générer des réponses impulsionnelles de salle spatiales au format ambisonique avec une résolution spatiale comparable à celle obtenue avec une antenne sphérique de microphones communément employée telle que l'Eigenmike [19]. Ce choix est motivé par le souhait d'identifier des estimateurs d'impressions spatiales que l'on peut extraire de mesures de SRIRs, communément utilisées pour l'auralisation. En effet, l'utilisation d'une antenne sphérique de microphones est de plus en plus répandue pour la mesure de réponse impulsionnelle spatiale de salle.

De plus, dans le but d'établir d'autres paramètres acoustiques que ceux définis dans la norme, il est intéressant d'exploiter la résolution spatiale offerte par ce dispositif. Plusieurs études ont notamment proposé de mesurer des paramètres acoustiques avec une plus grande précision que ceux présents dans la norme. De Vries *et al.* [6] ont par exemple proposé le B_{LF}, le B_{LFC} pour pallier les variabilités de J_{LF}, J_{LFC}. Par ailleurs, pour calculer l'uniformité des directions d'incidence des réflexions, Hanyu et Kimura ont introduit le temps central spatialement équilibré SBT_s [8], qui utilise une analyse du temps central [3] dans plusieurs directions uniformément réparties dans le plan horizontal.

Pour chaque configuration acoustique, la base de données fournit les positions des réflexions spéculaires obtenues d'après la méthode source-image et les positions des réflexions diffuses détectées par le lancé de particules. Ces résultats de simulations ont été obtenus en considérant une source omnidirectionnelle comme recommandé dans la norme ISO 3382-1. Pour compléter la base de données, nous avons donc calculé une SRIR au format ambisonique pour chaque configuration acoustique grâce aux

3. Le temps central correspond au centre de gravité de l'énergie de la réponse impulsionnelle, mesuré en millisecondes

données disponibles dans la base. À la manière des opérations effectuées par RAVEN pour la génération des BRIRs, le calcul des SRIRs s'est effectué de la manière suivante :

— Pour générer la partie spéculaire des SRIRs, chaque réflexion spéculaire a été considérée comme une onde plane caractérisée par une direction d'incidence, une énergie dans 31 bandes de fréquence et un retard. Un filtre à phase linéaire de 4097 échantillons a été utilisé pour reproduire le spectre de chaque réflexion. Ce filtre fournit la meilleure approximation de la réponse en fréquence décrite par bande au sens des moindres carrés (l'intégrale de l'erreur quadratique moyenne dans les 31 bandes spécifiées est minimisée) [20].
— La structure temporelle des SRIRs a été construite d'après un processus aléatoire basé sur une distribution de Poisson. Des séquences temporelles sont générées pour 36 secteurs angulaires afin de reproduire la partie non spéculaire des SRIRs. Ces séquences sont paramétrées de sorte que la densité d'écho théorique donnée par l'équation (1.1) soit respectée pour l'espace considéré.
— La décroissance de l'énergie a été déterminée pour les 10 bande d'octaves et pour les 36 secteurs angulaires d'après l'histogramme d'énergie résultant du lancé de particules. La décroissance de l'énergie de chaque séquence temporelle générée (une par secteur angulaire) a été ajustée par bande d'octave pour satisfaire les décroissances d'énergie issues de la simulation.
— Les signaux correspondant à la partie spéculaire et non spéculaire des SRIRs ont été encodés en ambisonique à l'ordre 4.

Avec l'ensemble des données de la base GRAP complétée par les SRIRs ambisoniques ainsi générées, les sections suivantes mettent en relation des paramètres acoustiques associés aux 70 configurations acoustiques et les évaluations des impressions spatiales correspondantes.

7.3. Les paramètres acoustiques étudiés

Afin de déterminer les paramètres les plus à même de caractériser une acoustique selon la largeur apparente de source et l'enveloppement, 21 paramètres acoustiques ont été considérés. Ces paramètres sont répertoriés dans le tableau 7.1. On y trouve des durées de décroissance, des mesures de clarté, de force sonore, de fraction latérale d'énergie et des coefficients de décorrélation intéraurale. Aux paramètres acoustiques spécifiées par la norme ISO 3382-1 s'ajoutent le rapport champ direct sur champ réverbéré et ceux décrits dans la section 1.3.4. Néanmoins, le RAP_{ASW} et RAP_{LEV} ne se calculant pas d'après une réponse impulsionnelle binaurale, ils n'ont pas été pris en compte. L'ensemble des paramètres acoustiques ont été calculés pour 70 configurations acoustiques.

Bien que dans la littérature l'enveloppement soit généralement lié aux propriétés spatiales de la réverbération tardive, il semble que les réflexions précoces peuvent également contribuer à la sensation d'enveloppement. En effet, Dick *et al.* rapportent que des modifications apportées à la partie précoce d'une réponse impulsionnelle de salle peuvent influencer la sensation d'enveloppement tandis qu'au delà de 120 ms de telles modifications n'ont qu'un faible impact [21]. Par ailleurs, Klockgether *et al.* ont montré que la manipulation de la corrélation interaurale de la partie tardive

Tableau 7.1 – Paramètres acoustiques calculés pour chaque configuration acoustique et bandes d'octave sur lesquelles sont calculées les valeurs moyennes.

Nom	Sigle	Moyenne
Temps de réverbération∗	T_{30} (s)	500 - 1000 Hz
Durée de décroissance initiale∗	EDT (s)	500 - 1000 Hz
Temps central∗	T_s (ms)	500 - 1000 Hz
Temps central par filtrage angulaire†	SBT_s (ms)	500 - 1000 Hz
Clarté∗	C_{80} (dB)	500 - 1000 Hz
Définition∗	D_{50}	500 - 1000 Hz
Rapport champ direct sur champ réverbéré⋆	DRR (dB)	-
Force sonore∗	G (dB)	500 - 1000 Hz
Force sonore précoce⋆	G_E (dB)	500 - 1000 Hz
Force sonore précoce latérale†	G_{EL} (dB)	500 - 1000 Hz
Force sonore tardive⋆	G_L (dB)	500 - 1000 Hz
Force sonore latérale tardive†	L_J (dB)	125 - 1000 Hz
Fraction d'énergie latérale∗	J_{LF}	125 - 1000 Hz
Coefficient de fraction d'énergie latérale∗	J_{LFC}	125 - 1000 Hz
Fraction d'énergie latérale par filtrage angulaire†	B_{LFC}	125 - 1000 Hz
Coefficient de fraction d'énergie latérale par filtrage angulaire†	B_{LF}	125 - 1000 Hz
Fraction d'énergie latérale tardive†	LLF	125 - 1000 Hz
Rapport d'énergie avant-arrière†	FBR (dB)	500 - 1000 Hz
Binaural Quality Index⋆	BQI	500 - 2000 Hz
Coefficient de décorrélation interaurale tardive⋆	$1\text{-}IACC_L$	500 - 1000 Hz
Coefficient de décorrélation interaurale∗	1-IACC	500 - 1000 Hz

∗ Valeurs fournies dans la base de données.
⋆ Valeurs calculées d'après les réponses impulsionnelles fournies dans la base de données.
† Valeurs calculées d'après les SRIRs générées.

mais également précoce d'une BRIR affecte l'enveloppement perçu par l'auditeur. Ils ont également montré que la perception de la largeur de source peut être influencée par la manipulation de la corrélation interaurale de la partie tardive. Il nous a donc intéressant d'explorer aussi 1) les relations entre les paramètres acoustiques calculés sur la partie précoce des réponses impulsionnelles et les notes d'enveloppement 2) les relations entre les paramètres acoustiques calculés sur la partie tardive et les notes de largeur apparente de source. C'est pourquoi l'ensemble de ces 21 paramètres acoustiques a été considéré pour étudier chacune des deux impressions spatiales.

Le calcul de la fraction d'énergie latérale par filtrage angulaire

Le calcul de la fraction d'énergie latérale par filtrage angulaire a été adapté de l'étude réalisée par de Vries *et al.* [6]. En effet, dans leur étude le calcul de l'énergie

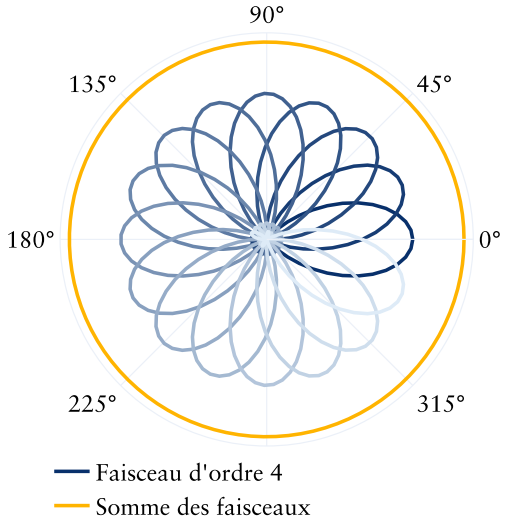


FIGURE 7.3 – Diagrammes de directivité des faisceaux utilisés pour calculer l'énergie de la SRIR dans 16 secteurs angulaires. L'analyse permet de couvrir l'ensemble du plan horizontal comme en atteste la somme constante des faisceaux dans le plan.

par secteur angulaire a employé une antenne linéaire de microphones. Pour l'effectuer d'après des signaux ambisoniques d'ordre 4, 16 secteurs angulaires dans le plan horizontal ont été considérés. Pour chaque secteur angulaire, un faisceau formé dans la direction correspondante a permis d'estimer la quantité d'énergie provenant de cette direction. Une pondération max-rE des signaux ambisoniques a été employée pour minimiser la taille du lobe arrière du faisceau [22]. La figure 7.3 représente le diagramme de directivité des faisceaux formés dans les secteurs angulaires. Les contributions des énergies de chaque secteur angulaire dans le dipôle d'ordre 1 latéral (c'est à dire perpendiculaire à la source) sont ensuite sommées pour former une estimation de l'énergie latérale. Sommer les énergies des secteurs angulaires permet de s'affranchir d'éventuelles interférences destructives ou constructives qui peuvent apparaitre en sommant les amplitudes. Ces interférences sont jugées responsables de la forte variabilité des mesures J_{LF} et J_{LFC} dans une salle [6]. La fraction d'énergie latérale par filtrage angulaire résulte du rapport entre l'énergie latérale ainsi calculée et l'énergie totale dans les 80 premières millisecondes de la SRIR.

Corrélations entre les paramètres acoustiques

La figure 7.4 représente sous forme matricielle la valeur absolue des coefficients de corrélation de Pearson, notés r, calculés entre les 70 valeurs des paramètres acoustiques. D'après les classifications établies par Evans pour interpréter la force de corrélations [23], les durées de décroissance et mesures de clarté présentent des corrélations élevées avec des coefficients r supérieurs à 0.72. De même, les mesures de force sonore sont fortement corrélées entre elles ($r > 0.78$) ainsi que les mesures de fraction latérale ($r > 0.82$).

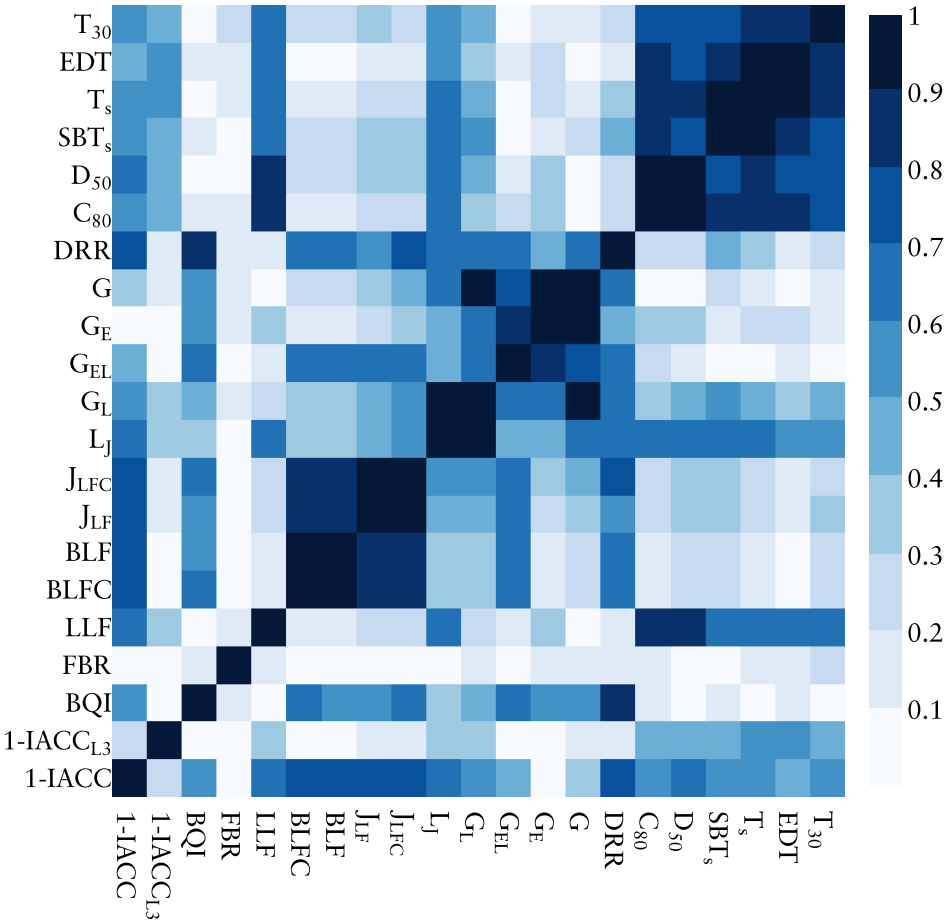


FIGURE 7.4 – Matrice des valeurs absolues des coefficients de corrélation de Pearson entre les différents paramètres acoustiques utilisés pour caractériser les impressions spatiales.

7.4. La prédiction de la largeur apparente de source

Le RAQI définit l'attribut perceptif *Width* comme « *the perceived spatial extent of a sound source in horizontal direction* ». Cet attribut sera par la suite nommé largeur apparente de source ou ASW. Dans cette section, nous souhaitons identifier les paramètres acoustiques calculés d'après les réponses impulsionnelles omnidirectionnelles, binaurales et spatiales permettant de prédire la largeur perçue d'une source.

7.4.1. Corrélations entre les notes d'ASW et les paramètres acoustiques

Le tableau 7.2 rassemble les coefficients de corrélation de Pearson entre les notes d'ASW associées aux source *Voix* et *Trompette* et les paramètres acoustiques considérés. Aucun paramètre acoustique n'apparait fortement corrélé aux notes de largeur apparente de source. Les valeurs maximales en valeur absolue du coefficient sont obtenues pour les paramètres G_L et L_J et s'élèvent respectivement à 0.574 et 0.624. Ce résultat est surprenant dans la mesure où ces paramètres caractérisent l'énergie tardive

Tableau 7.2 – Coefficients de corrélation de Pearson entre les paramètres acoustiques et les notes de largeur apparente de source.

Paramètres	Voix	Trompette	Paramètres	Voix	Trompette
T_{30}	0.413	0.488	L_J	0.624	0.620
EDT	0.392	0.473	J_{LF}	0.314	0.220
T_s	0.419	0.511	J_{LFC}	0.367	0.303
SBT_s	0.423	0.539	B_{LF}	0.229	0.246
C_{80}	-0.527	-0.523	B_{LFC}	0.226	0.290
D_{50}	-0.554	-0.552	LLF	0.382	0.349
DRR	-0.456	-0.547	FBR	0.049	0.014
G	0.365	0.404	BQI	0.268	0.258
G_E	0.163	0.196	$1\text{-}IACC_L$	0.550	0.341
G_{EL}	0.202	0.256	1-IACC	0.459	0.421
G_L	0.565	0.574			

et non les réflexions précoces qui sont jugées principalement responsables de l'ASW. De plus, il est intéressant de noter qu'une corrélation négative est obtenue avec les rapports d'énergie précoce et tardive (D_{50} et C_{80}) et le rapport entre champ direct et champ réverbéré (DRR).

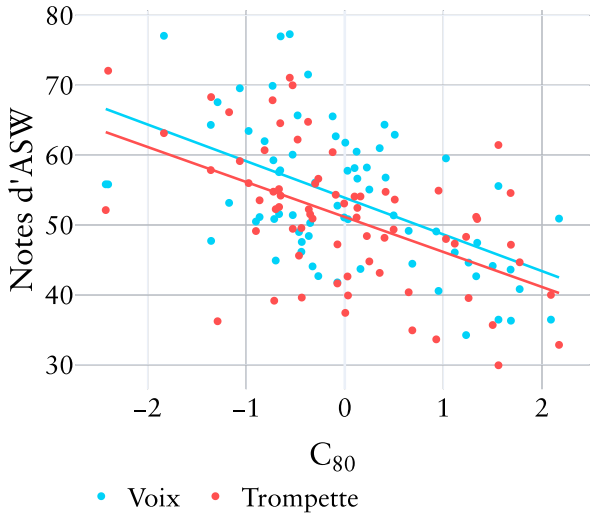


FIGURE 7.5 – Notes de largeur apparente de source en fonction des valeurs de C_{80}.

La figure 7.5 représente les notes d'ASW associées aux deux sources en fonction de la valeur du paramètre acoustique C_{80}. La relation apparait assez éloignée d'une relation linéaire entre les deux grandeurs.

Le fait que le paramètre acoustique C_{80} soit corrélé négativement avec l'ASW n'est pas en accord avec l'étude de Bradley *et al.* [1] qui rapportent qu'accroître l'énergie tardive tend à amoindrir la largeur apparente de source. Leur étude diffère avec la pré-

sente étude sur plusieurs points : 8 enceintes situées dans le plan horizontal ont permis la reproduction de 12 acoustiques (ne correspondant pas à des environnements réels de référence), les paramètres acoustiques C_{80}, J_{LF} et G_{EL}, L_J ont tous été moyennés sur les bandes de fréquence de 125 Hz à 1kHz et 6 à 11 sujets ont notés les impressions spatiales sur une échelle à 5 points. De plus, un faible nombre de réflexions a été employé et elles étaient reproduites sur une zone angulaire limitée, ce qui est moins représentatif du champ sonore réel vécu dans une salle. Dans l'évaluation perceptive réalisée par Weinzierl *et al.* [10] pour obtenir les notes d'ASW utilisées ici, l'emploi d'une méthode d'auralisation binaurale dynamique de réponses impulsionnelles de salles simulées d'après des environnements réels semble plus proche d'une situation d'écoute réelle (bien que limitée au cas des salles de grand volume). De plus, le grand nombre de salles considérées et de sujets impliqués dans l'évaluation perceptive des stimuli permet d'avoir une puissance statistique plus importante que dans l'étude de Bradley *et al.*.

7.4.2. Régression linéaire multidimensionnelle

Les corrélations observées ne sont pas suffisantes pour être en mesure d'expliquer les notes d'ASW en fonction des valeurs d'un seul paramètre acoustique. Néanmoins, les notes d'ASW peuvent être mises en relation avec des contributions conjointes de plusieurs paramètres acoustiques. C'est pourquoi nous avons réalisé une régression linéaire multidimensionnelle.

Nous cherchons à exprimer les $n = 70$ notes d'ASW associées aux deux sources sonores comme une combinaison linéaire des valeurs de p paramètres acoustiques. Soit $\boldsymbol{Y}$ la matrice contenant les vecteurs de notes $\boldsymbol{y}_{\text{voix}}$ et $\boldsymbol{y}_{\text{trompette}}$ associés aux deux contenus sonores *Voix* et *Trompette*. Soit $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$, la matrice des valeurs de paramètres acoustiques où $\boldsymbol{x}_i$ est un vecteur contenant n valeurs du ième paramètre acoustique considéré. La matrice $\hat{\boldsymbol{Y}}$ contenant le résultat de la combinaison linéaire s'écrit :

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W} + \mathbb{1}_{n\times 1}\boldsymbol{w}_0 \tag{7.1}$$

où $\boldsymbol{W} = [\boldsymbol{w}_{\text{voix}}, \boldsymbol{w}_{\text{trompette}}]$ est la matrice contenant les vecteurs de pondération permettant de prédire les notes associées à chaque source, $\boldsymbol{w}_0$ contient les deux ordonnées à l'origine (une pour chaque source) et $\mathbb{1}_{n\times 1}$ une matrice de 1 de dimension $n \times 1$. Pour calculer la matrice de pondération $\boldsymbol{W}$ et $\boldsymbol{w}_0$, nous souhaitons minimiser la quantité ϵ correspondant à la somme résiduelle des carrés entre les notes de largeur apparente de source $\boldsymbol{Y}$ et les notes prédites par régression linéaire $\hat{\boldsymbol{Y}}$. Cette quantité s'écrit :

$$\epsilon = \min_{\boldsymbol{W},\boldsymbol{w}_0} \|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_2^2 \tag{7.2}$$

Réduction du nombre de paramètres

Étant donné le nombre d'observations disponibles (les n notes d'ASW), il n'est pas judicieux d'utiliser l'ensemble des variables identifiées (les p paramètres acoustiques)

pour calculer un modèle prédictif. En effet, à mesure que le nombre de variables augmente, le nombre de données nécessaires pour généraliser le modèle prédictif de manière fiable augmente grandement. Pour la régression linéaire multidimensionnelle, certaines règles sur le nombre minimal d'observations nécessaires au calcul d'un modèle proposent une taille minimale d'observations basée sur le nombre de variables utilisées, par exemple 30 observations pour 1 variable [24] ou 10 observations pour 1 variable [25]. Tabachnick et Fidell ont proposé d'utiliser la formule $50 + 8m$ où m est le nombre de variables [26].

Afin de réduire le nombre de paramètres acoustiques utilisés, nous avons exclu de l'analyse les paramètres acoustiques présentant une très forte corrélation avec d'autres paramètres. Le seuil de corrélation a été fixé pour un coefficient de corrélation de Pearson de 0.9. Lorsque deux paramètres sont fortement corrélés, celui le plus fréquent dans la littérature a été retenu. Ainsi, les paramètres B_{LFC}, D_{50}, G_E, G_L, J_{LF}, T_s et SBT_s n'ont pas été utilisés dans l'analyse.

Avec 14 paramètres restants, le nombre de variables en jeu était encore trop important. Une méthode de sélection itérative des variables dite «descendante» (*Backward Feature Selection*) a été utilisée pour réduire ce nombre. Cette sélection consiste à :

1. Considérer l'ensemble des variables à disposition.
2. Calculer un modèle statistique d'après les variables sélectionnées. Dans notre cas, les coefficients de régression linéaire sont calculés pour chaque paramètre acoustique ainsi que les valeurs p de significativité associées. Cette valeur p permet de tester l'hypothèse nulle selon laquelle le coefficient est égal à zero (pas d'effet du paramètre).
3. Retirer une variable selon un critère de sélection, ici la valeur p de significativité. Une faible valeur p (< 0.05) indique que l'hypothèse nulle peut être rejetée. Au contraire, une valeur $p > 0.05$ suggère qu'un changement de valeur de la variable n'est pas associé à un changement de l'observation. Le paramètre acoustique lié à la plus grande valeur p est retiré du modèle.
4. Procéder aux étapes 2 et 3 de manière itérative jusqu'à ce que toutes les variables restantes du modèle possède une faible valeur p.

Suite à cette sélection, seuls les paramètres C_{80} et DRR ont été utilisés pour la régression linéaire. Il est intéressant de noter que ce sont des paramètres calculés d'après un signal omnidirectionnel et que les paramètres B_{LF}, J_{LF} et G_{EL}, qui caractérisent l'énergie précoce latérale, n'ont pas été retenus. Les valeurs associées ont été centrées et réduites avant de calculer les coefficients de pondération du modèle.

Résultats

Le tableau 7.3 rassemble les coefficients associés aux valeurs des paramètres C_{80} et DRR permettant de minimiser l'erreur quadratique entre les notes prédites $\hat{\boldsymbol{Y}}$ et les notes de la base de données $\boldsymbol{Y}$. Les coefficients de pondération respectifs des valeurs de C_{80} et DRR pour prédire les notes d'ASW sont comparables. Il semble que le DRR a eu une importance plus grande dans le jugement d'ASW pour *Trompette* bien que la différence de pondération avec les valeurs de C_{80} soit faible. Le coefficient de détermination R^2 de la régression linéaire s'élève à 0.436 : la variation des notes d'ASW

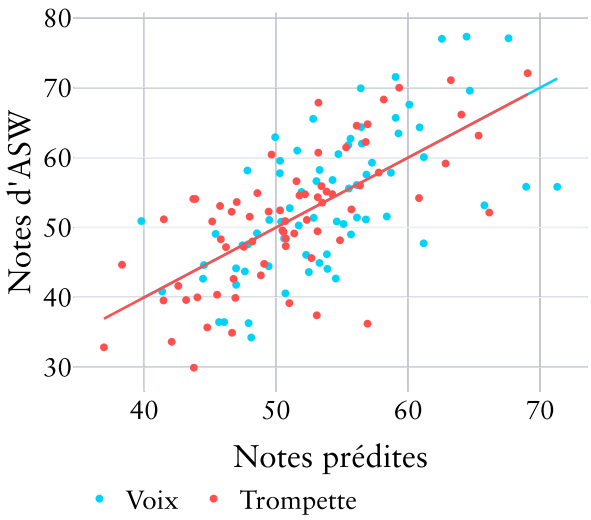


FIGURE 7.6 – Notes de largeur apparente de source $\boldsymbol{Y}$ en fonction des notes prédites $\hat{\boldsymbol{Y}}$ d'après les valeurs de C_{80} et DRR pour les deux sources.

	Voix	Trompette
C_{80} (dB)	-4.479	-4.074
DRR (dB)	-3.577	-4.368
Ordonnées à l'origine	53.897	51.154

Tableau 7.3 – Coefficients de la matrice de pondération $\boldsymbol{W}$ (valeurs $p < 0.001$) et du vecteur d'ordonnées à l'origine $\boldsymbol{w}_0$.

n'est prédite qu'à hauteur de 43.6% par les valeurs des deux paramètres acoustiques. La figure 7.6 représente les notes d'ASW en fonction des notes prédites. Bien qu'une tendance se dessine, l'établissement d'une relation linéaire entre les paramètres acoustiques et les notes d'ASW semble abusive.

Les résultats étant peu satisfaisants, une autre approche a été envisagée. Plutôt que de prédire des valeurs continues que sont les notes d'ASW, nous avons choisi d'analyser simplement deux catégories d'espaces associés à des valeurs d'ASW différentes grâce aux paramètres acoustiques.

7.4.3. Les paramètres acoustiques pertinents pour la classification

Le but de cette section est d'identifier les paramètres acoustiques permettant de catégoriser les configurations acoustiques selon les notes d'ASW obtenues. La figure 7.12 représente la répartition des notes de largeur apparente de source pour les deux sources sonores. Notons que les notes d'ASW attribuées aux deux sources sont faiblement corrélées ($r = 0.47$), l'influence de la source semble importante pour cet attribut. Les configurations acoustiques ont été classées selon quatre catégories :

— celles auxquelles ont été attribuées les notes d'ASW les plus élevées pour les deux sources (classe verte). Cette catégorie rassemble 25 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'ASW les plus faibles pour les deux sources (classe rouge). Cette catégorie rassemble 25 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'ASW les plus élevées pour *Trompette* et les plus faibles pour *Voix* (classe bleue). Cette catégorie rassemble 10 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'ASW les plus faibles pour *Trom-*

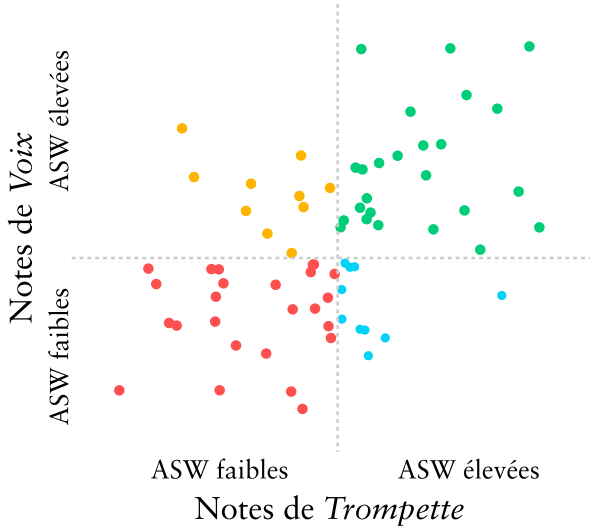


FIGURE 7.7 – Répartition des notes de largeur apparente de source attribuées aux 70 configurations acoustiques selon les deux sources sonores. Les notes les plus faibles pour les deux sources apparaissent en rouge et les notes les plus élevées pour les deux sources en vert. Les notes les plus faibles pour *Trompette* et les plus élevées pour *Voix* apparaissent jaune et les notes les plus élevées pour *Trompette* et les plus faibles pour *Voix* apparaissent en bleu.

pette et les plus élevées pour *Voix* (classe jaune). Cette catégorie rassemble 10 configurations acoustiques.

Une note est ici qualifiée d'élevée si elle est supérieure à la médiane et faible si elle y est inférieure.

Nous souhaitons étudier les configurations acoustiques ayant mené à des ASW élevées et faibles pour les deux sources sonores à la fois (classe verte et rouge respectivement). Ces deux catégories représentent 50 configurations acoustiques. La répartition de chaque paramètre acoustique a été analysée selon ces deux catégories de configurations acoustiques.

Dans l'idéal, un paramètre acoustique permettant de prédire une largeur apparente de source large ou étroite présenterait deux distributions distinctes, sans chevauchement. Néanmoins, pour certains paramètres, ces distributions sont non significativement différentes. Le test non-paramétrique de Wilcoxon-Mann-Whitney a été utilisé pour en attester. Ce test vérifie l'hypothèse nulle selon laquelle deux ensembles d'échantillons appartiennent à la même distribution. L'hypothèse alternative est que les ensembles d'échantillons sont issus de distributions dont les médianes sont différentes. Le tableau 7.4 recense les valeurs de significativité du test effectué pour chaque paramètre acoustique. L'hypothèse nulle a été rejetée pour l'ensemble des paramètres acoustiques sauf FBR qui correspond au rapport entre l'énergie frontale et arrière (proposé par Morimoto *et al.* [7] pour mesurer l'enveloppement) et G_E la force sonore de l'énergie précoce. La figure 7.8 représente les histogrammes de plusieurs paramètres ainsi que l'estimation de leur densité de probabilité [27]. On peut tirer des résultats du test et de l'inspection visuelle des histogrammes les remarques suivantes :

1. Le rapport champ direct sur champ réverbéré possède des valeurs significati-

Paramètre	Valeur p	Paramètre	Valeur p	Paramètre	Valeur p	Paramètre	Valeur p
T_{30}	*<0.001*	DRR	*<0.001*	J_{LF}	*0.004*	BQI	*0.010*
EDT	*<0.001*	G	*0.001*	J_{LFC}	*0.001*	1-$IACC_L$	*<0.001*
T_s	*<0.001*	G_E	0.204	B_{LF}	*0.005*	1-IACC	*<0.001*
SBT_s	*<0.001*	G_{EL}	*0.018*	B_{LFC}	*0.006*		
C_{80}	*<0.001*	G_L	*<0.001*	LLF	*0.006*		
D_{50}	*<0.001*	L_J	*<0.001*	FBR	0.567		

Tableau 7.4 – Valeurs de significativité p du test de Wilcoxon-Mann-Whitney. Hypothèse nulle : la médiane des paramètres acoustiques associés aux notes les plus élevées d'ASW est n'est pas significativement différente de celle des paramètres acoustiques associés aux notes les plus faibles. Les valeurs significatives apparaissent en bleu.

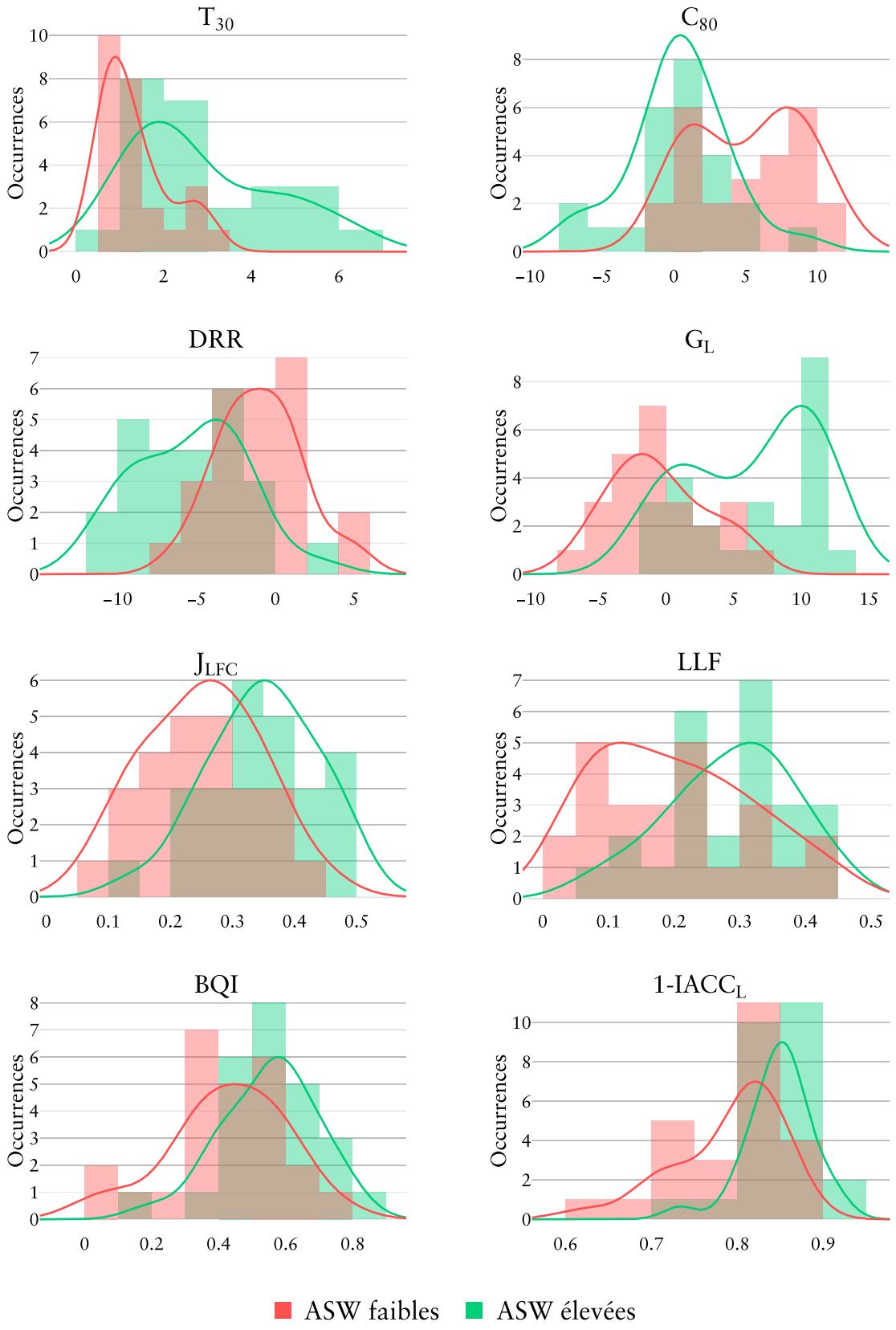


FIGURE 7.8 – Histogramme et densité de probabilité de paramètres acoustiques calculés pour les configurations acoustiques produisant les ASW les plus élevées et les plus faibles pour les deux sources sonores. Les distributions des paramètres acoustiques selon les deux catégories de configurations acoustiques sont significativement différentes.

vement différentes selon la catégorie de configurations acoustiques. La source sonore semble perçue moins large lorsque le son direct est prédominant (DRR positif).

2. La contribution des réflexions à l'amplification du son par l'acoustique, caractérisée par la force sonore G, semble liée à la perception d'une source large. Le calcul de la force sonore sur la partie précoce des réponses impulsionnelles de salle n'a pas permis de discriminer les deux groupes de configurations acoustiques. Au contraire, la force sonore sur la partie tardive est significativement plus élevée pour les configurations produisant une ASW élevée.
3. L'énergie latérale, caractérisée par les paramètres J_{LFC} et J_{LF}, semble avoir une influence significative sur la largeur apparente de la source. Ce résultat confirme l'importance des réflexions précoces latérales pour caractériser la largeur apparente de source. Les distributions d'énergie latérale tardive LLF sont également significativement différentes selon la condition d'ASW mais se chevauchent plus que celles de J_{LFC} et J_{LF}. Les paramètres acoustiques B_{LF} et B_{LFC} sont également significativement différents selon les deux catégories néanmoins les distributions n'apparaissent pas plus distinctes entre les deux catégories que celles de J_{LFC} et J_{LF}. Il est possible que le filtrage angulaire n'améliore pas la prédiction de cet attribut perceptif.
4. La décorrélation interaurale semble favoriser la perception d'une source large. Le calcul de la décorrélation sur la partie précoce des réponses impulsionnelles de salle, le BQI, est significativement différent selon les deux groupes de configurations acoustiques et il en va de même pour le coefficient calculé sur la partie tardive.
5. L'énergie tardive semble avoir une influence significative sur la largeur apparente de la source. Cette dernière est plus élevée lorsque l'acoustique est réverbérante (cf. T_{30}, EDT) et que l'énergie tardive est plus importante que l'énergie précoce (cf. C_{80}, D_{50}).

Détermination d'un hyperplan séparateur

Nous souhaitons déterminer les paramètres acoustiques capables de prédire à quelle catégorie (ASW élevée ou ASW faible) les configurations acoustiques considérées appartiennent. Soient $\boldsymbol{x}$, un vecteur contenant les valeurs de p paramètres acoustiques et $\boldsymbol{w}$ un vecteur de coefficients de pondération. Un hyperplan h dans un espace de dimension p est défini par l'équation suivante :

$$h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0 \tag{7.3}$$

Nous cherchons les valeurs de $\boldsymbol{w}$ et w_0 permettant de séparer l'espace de dimension p en deux régions distinctes, avec d'un coté les vecteurs associés à la classe d'ASW élevées et de l'autre d'ASW faibles. L'équation $h(\boldsymbol{x}) = 0$ correspond ainsi à une frontière de décision binaire. Par exemple, le vecteur $\boldsymbol{x}$ est associé à la classe d'ASW élevée si $h(\boldsymbol{x}) \leq 0$ et à la classe d'ASW faible sinon. Les coefficients de pondération sont calculés grâce à un ensemble d'apprentissage formé par un ensemble de vecteurs $\boldsymbol{x}_k$

chacun associé à une des deux classes. Les coefficients de pondérations sont calculés de manière à ce que la marge entre l'hyperplan h et les vecteurs d'apprentissage les plus proches de h soit maximale. Ces vecteurs sont appelés vecteurs de support. La figure 7.9 représente une séparation linéaire entre deux classes de données par un hyperplan.

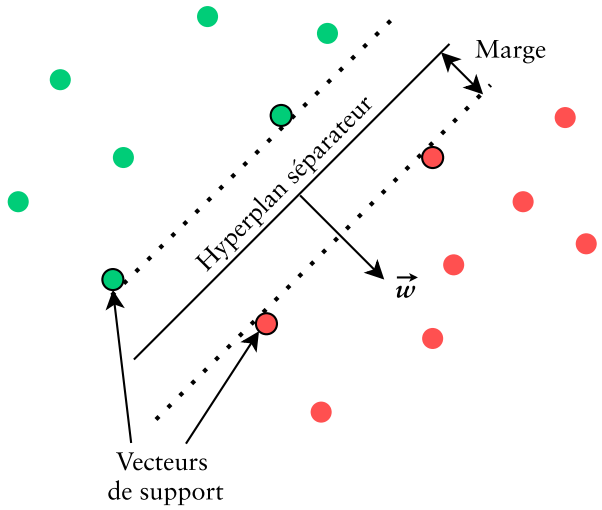


FIGURE 7.9 – Illustration du problème de séparation linéaire à deux classes. Les vecteurs de support se trouvent a une distance égale à la marge d'un côte ou de l'autre de l'hyperplan séparateur.

L'hyperplan est déterminé grâce à des vecteurs d'apprentissage $\boldsymbol{x}_k$ et la performance de la classification opérée par la séparation linéaire binaire peut être mesurée grâce à d'autres vecteurs $\boldsymbol{x}_t$, appelés vecteurs de test. La précision de la classification (*Accuracy*) est déterminée en termes de nombre de classes correctement prédites par rapport au nombre de classe de l'ensemble des vecteurs de test. Cette opération peut-être répétée plusieurs fois en choisissant les vecteurs d'apprentissage et les vecteurs de test de manière aléatoire. La mesure de performance rapportée par ce processus de validation croisée est alors la moyenne et l'écart-type des précisions calculées pour chaque séparateur.

Pour analyser les paramètres permettant de discriminer les configurations acoustiques en termes d'ASW nous avons procédé au calcul d'hyperplans séparateurs en considérant différents paramètres acoustiques. Pour identifier les paramètres acoustiques pertinents pour caractériser l'ASW, nous avons cherché les paramètres acoustiques permettant d'obtenir la meilleure performance de classification. Pour déterminer ces paramètres, l'ensemble des paramètres acoustiques ont été considérés à l'exception de FBR et G_E dont les valeurs ne sont pas significativement différentes d'une catégorie à l'autre et de ceux présentant une forte corrélation avec un autre paramètre acoustique (coefficient de Pearson > 0.9) : B_{LFC}, EDT, D_{50}, G_L, J_{LF}, T_s et SBT_s.

Nous avons considéré 30 vecteurs de paramètres acoustiques tirés aléatoirement pour le calcul de l'hyperplan séparateur et les tests étaient réalisés avec les 20 vecteurs restants. Pour un ensemble de paramètres acoustiques donné cette opération a été effectuée 20 fois. Le tableau 7.9 rassemble les moyennes et écart-types de la précision obtenue en fonction des paramètres considérés. Seuls les paramètres ayant permis

l'obtention d'une précision moyenne supérieure à 80% sont mentionnés. Au regard du nombre de vecteurs d'apprentissage, trois paramètres acoustiques maximums ont été considérés.

Paramètres acoustiques utilisés	Précision moyenne ± écart-type
1-$IACC_L$, DRR	80.3% ± 8.1%
C_{80}, DRR, T_{30}	80.5% ± 6.5%

Tableau 7.5 – Précision de classification des configurations acoustiques selon les notes d'ASW élevées (classe verte) ou faibles (classe rouge). Seuls les ensembles de deux paramètres acoustiques minimum et trois maximum ayant permis l'obtention d'une précision moyenne supérieure à 80% sont mentionnés.

Les valeurs de précision obtenues ne sont pas très élevées. Il est clair que le nombre de vecteurs d'apprentissage est trop faible pour établir un hyperplan séparateur suffisamment robuste et ayant un pouvoir prédictif satisfaisant. Néanmoins, il est intéressant de constater que l'un des meilleurs résultats de classification a été obtenu en considérant, entre autres, les paramètres C_{80} et DRR déjà identifiés comme expliquant la plus grande partie de la variance des notes d'ASW dans la section 7.4.1. Par ailleurs, la décorrélation interaurale de la partie tardive des réponses impulsionnelles, ($IACC_L$) semble permettre de discriminer les configurations acoustiques d'ASW élevée et faibles. Klockgether *et al.* [28] avaient déjà identifiés que la décorrélation intéraurale était lié à la perception d'une source large ou étroite. Cependant, l'$IACC_L$ n'est généralement pas utilisé pour caractériser cet attribut perceptif plus souvent associé au BQI (la décorrélation interaurale de l'énergie précoce). L'énergie tardive semble jouer un rôle significatif, comme en atteste également la présence du paramètre T_{30} parmi les paramètres acoustiques les plus discriminants entre les deux classes de configuration acoustiques.

7.5. Prédiction de l'enveloppement

Une analyse similaire a été effectuée d'après les notes de l'attribut perceptif *Envelopment by reverberation* du RAQI que l'on nommera simplement enveloppement ou LEV. Nous souhaitons identifier de la même manière les paramètres acoustiques calculés d'après les réponses impulsionnelles omnidirectionnelles, binaurales et spatiales qui permettent de prédire la sensation d'enveloppement.

7.5.1. Corrélations entre les notes d'enveloppement et les paramètres acoustiques

Le tableau 7.6 rassemble les coefficients de corrélation de Pearson entre les notes d'enveloppement associées aux deux sources pour l'ensemble des paramètres acoustiques considérés. Les valeurs maximales en valeur absolue du coefficient sont obtenues pour les paramètres D_{50} et C_{80} et s'élèvent respectivement à -0.852 et -0.857. Les corrélations avec l'enveloppement peuvent être considérées comme fortes pour les

Tableau 7.6 – Coefficients de corrélation de Pearson entre les paramètres acoustiques et les notes d'enveloppement.

Paramètres	Voix	Trompette	Paramètres	Voix	Trompette
T_{30}	0.780	0.783	L_J	0.696	0.636
EDT	0.804	0.839	J_{LF}	0.282	0.252
T_s	0.800	0.825	J_{LFC}	0.270	0.244
SBT_s	0.722	0.723	B_{LF}	0.161	0.230
D_{50}	-0.852	-0.805	B_{LFC}	0.152	0.229
C_{80}	-0.857	-0.836	LLF	0.725	0.627
DRR	-0.258	-0.303	FBR	0.068	0.088
G	0.157	0.149	BQI	0.086	0.109
G_E	-0.149	-0.148	$1\text{-}IACC_L$	0.575	0.507
G_{EL}	-0.047	-0.028	1-IACC	0.570	0.588
G_L	0.493	0.470			

mesures de clarté et les durées de décroissance ($r > 0.7$). Les paramètres acoustiques caractérisant la partie précoce des réponses impulsionnelles seulement (G_E, G_{EL}, J_{LF}, J_{LFC}, B_{LF}, B_{LFC}, BQI) présentent des corrélations très faibles avec les notes d'enveloppement ($r < 0.282$).

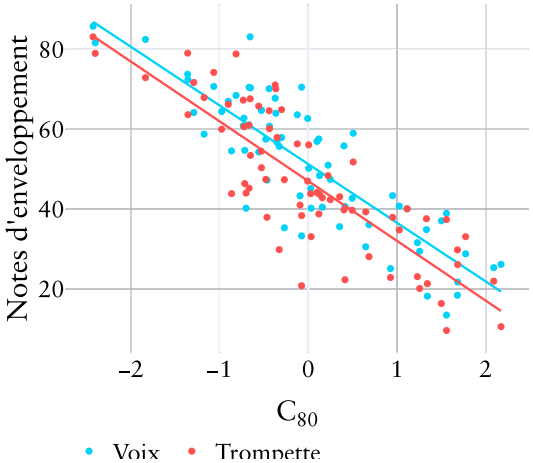


FIGURE 7.10 – Notes d'enveloppement en fonction des valeurs de C_{80}.

A titre d'exemple, la figure 7.10 représente les notes d'enveloppement associées aux deux sources en fonction de la valeur du paramètre acoustique C_{80}. Les deux grandeurs apparaissent bien mieux corrélées que dans le cas de l'ASW, bien que des écarts importants à la droite de régression linéaire apparaissent.

7.5.2. Régression linéaire multidimensionnelle

Une régression linéaire multidimensionnelle a été effectuée pour identifier la combinaison linéaire des paramètres acoustiques permettant de prédire au mieux l'enveloppement.

De la même manière que pour l'ASW, nous avons exclu de l'analyse les paramètres acoustiques présentant une très forte corrélation avec d'autres paramètres. En particulier, les valeurs de B_{LFC}, D_{50}, G_E, G_L, J_{LFC}, T_s et SBT_s n'ont pas été considérées dans l'analyse. Une sélection itérative descendante des paramètres acoustiques a également été effectuée. Suite à cette sélection, seuls les paramètres C_{80}, L_J et EDT ont été utilisés pour la régression linéaire. Les valeurs des paramètres restants ont été centrées et réduites avant de calculer les coefficients de pondération du modèle.

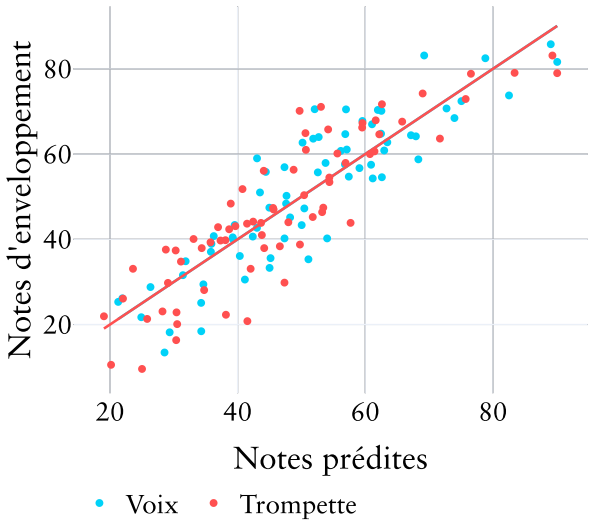


FIGURE 7.11 – Notes d'enveloppement en fonction des notes prédites d'après les valeurs de C_{80}, L_J et EDT pour les deux sources.

	Voix	Trompette
C_{80}	-7.843	-5.919
L_J	4.233	3.050
EDT	4.889	8.407
Ordonnées à l'origine	51.271	47.019

Tableau 7.7 – Coefficients de pondération des valeurs de paramètrs acoustiques (valeurs $p < 0.001$) et valeurs des ordonnées à l'origine.

Le tableau 7.7 rassemble les coefficients de pondération permettant de minimiser l'erreur quadratique entre les notes prédites et les notes de la base de données. Le coefficient de détermination R^2 de la régression linéaire s'élève à 0.785.

L'EDT semble être plus à même que le T_{30} d'expliquer la variance de l'enveloppement. Ces deux paramètres acoustiques sont liés à la réverbérance. Griesinger différencie deux types de réverbérance : la réverbérance courante (*running reverberance*) et la réverbérance «stoppée» (*stopped reverberance*) [29]. La première désigne la réverbération perçue lorsqu'une source sonore émet un signal et se mesure grâce à l'EDT. La seconde se fait entendre lorsque la source s'arrête et correspond à la queue de réverbération perçue entre deux événements sonores ponctuels. L'EDT est influencée par la distance à la source et la présence de réflexions spéculaires importantes. On peut émettre l'hypothèse que la sensation d'enveloppement puisse varier entre deux configurations acoustiques dans une même salle. L'importance de l'énergie tardive est également confirmée par la pondération négative du C_{80} : une proportion plus grande d'énergie tardive par rapport à l'énergie précoce mène à une sensation d'enveloppe-

ment plus importante. La force sonore latérale tardive semble également confirmer le rôle joué par les réflexions latérales dans la sensation d'enveloppement relevée par plusieurs études [4, 5, 30, 31].

La figure 7.11 représente les notes d'enveloppement en fonction des notes prédites. Une relation linéaire marquée se dessine, néanmoins une partie non négligeable de la variance n'est pas expliquée par le modèle de régression linéaire. Pour compléter cette analyse, la capacité des paramètres à discriminer les configurations acoustiques selon deux catégories d'enveloppement a également été étudiée.

7.5.3. Les paramètres acoustiques pertinents pour la classification

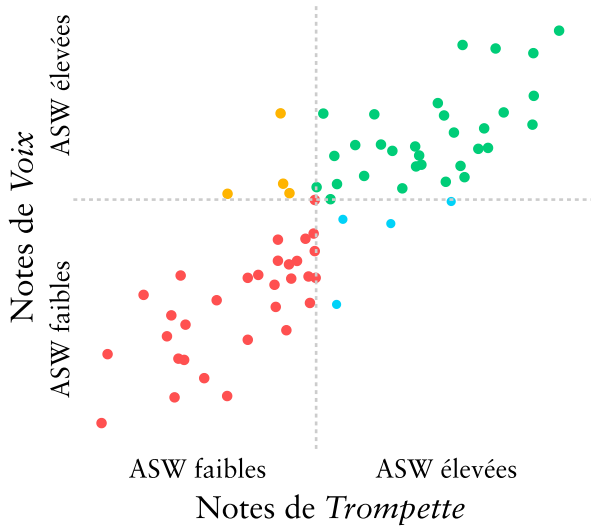


FIGURE 7.12 – Répartition des notes d'enveloppement attribuées aux 70 configurations acoustiques selon les deux sources sonores. Les notes les plus faibles pour les deux sources apparaissent en rouge et les notes les plus élevées pour les deux sources en vert. Les notes les plus faibles pour *Trompette* et les plus élevées pour *Voix* apparaissent jaune et les notes les plus élevées *Trompette* et les plus faibles pour *Voix* apparaissent en bleu.

Le but de cette section est d'identifier les paramètres acoustiques permettant de catégoriser les configurations acoustiques selon les notes d'enveloppement obtenues. La figure 7.12 représente la répartition des notes d'enveloppement pour les deux sources sonores. Les configurations acoustiques ont été classées selon quatre catégories :

— celles auxquelles ont été attribuées les notes d'enveloppement les plus élevées pour les deux sources (classe verte). Cette catégorie rassemble 31 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'enveloppement les plus faibles pour les deux sources (classe rouge). Cette catégorie rassemble 31 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'enveloppement les plus élevées pour *Trompette* et les plus faibles pour *Voix* (classe bleue). Cette catégorie rassemble 4 configurations acoustiques.
— celles auxquelles ont été attribuées les notes d'enveloppement les plus faibles

pour *Trompette* et les plus élevées pour *Voix* (classe jaune). Cette catégorie rassemble 4 configurations acoustiques.

Une note est ici qualifiée d'élevée si elle est supérieure à la médiane et faible si elle est inférieure.

Nous souhaitons étudier les configurations acoustiques ayant mené à des notes d'enveloppement élevées et faibles pour les deux sources sonores à la fois (classe verte et rouge respectivement). Ces deux catégories représentent 62 configurations acoustiques. La répartition de chaque paramètre acoustique a été analysée selon ces deux catégories de configurations acoustiques.

Paramètre	Valeur p	Paramètre	Valeur p	Paramètre	Valeur p	Paramètre	Valeur p
T_{30}	*<0.001*	DRR	0.166	J_{LF}	0.504	BQI	0.741
EDT	*<0.001*	G	0.301	J_{LFC}	0.657	1-$IACC_L$	*<0.001*
T_s	*<0.001*	G_E	0.328	B_{LF}	0.938	1-IACC	*<0.001*
SBT_s	*<0.001*	G_{EL}	0.251	B_{LFC}	0.871		
C_{80}	*<0.001*	G_L	*0.003*	LLF	*<0.001*		
D_{50}	*<0.001*	L_J	*<0.001*	FBR	0.269		

Tableau 7.8 – Valeurs de significativité p du test de Wilcoxon-Mann-Whitney. Hypothèse nulle : la médiane des paramètres acoustiques associés aux notes les plus élevées d'enveloppement est n'est pas significativement différente de celle des paramètres acoustiques associés aux notes les plus faibles. Les valeurs significatives apparaissent en bleu.

La répartition de chaque paramètre a été analysée selon le groupe des configurations acoustiques ayant les notes d'enveloppement les plus élevées d'une part et les plus faibles d'autre part. Le test non-paramétrique de Wilcoxon-Mann-Whitney a été utilisé pour vérifier si les paramètres acoustiques calculés pour chacune des catégories sont issus de la même distribution ou de distributions dont les médianes sont différentes. Le tableau 7.8 recense les valeurs de significativité du test effectué pour chaque paramètre acoustique. Les résultats du test montrent que les paramètres acoustiques caractérisant la partie précoce des réponses impulsionnelles seulement (G_E, G_{EL}, J_{LF}, J_{LFC}, B_{LF}, B_{LFC}, BQI) n'étaient pas significativement différents d'une catégorie à l'autre. Il en est de même pour G, DRR et FBR. La figure 7.13 représente les histogrammes de plusieurs paramètres acoustiques ainsi que l'estimation de leur densité de probabilité. On peut tirer des résultats du test et de l'inspection visuelle des histogrammes les remarques suivantes :

1. De la même manière que pour la largeur apparente de source, l'énergie tardive semble avoir une influence significative sur l'enveloppement. Celui-ci est plus élevé lorsque l'acoustique est réverbérante et que l'énergie tardive est plus importante que l'énergie précoce.
2. Alors que des caractéristiques de l'énergie tardive se sont montrées pertinentes pour la largeur apparente de source, les caractéristiques de l'énergie précoce ne semblent pas être liées à la sensation d'enveloppement. Aucun paramètre

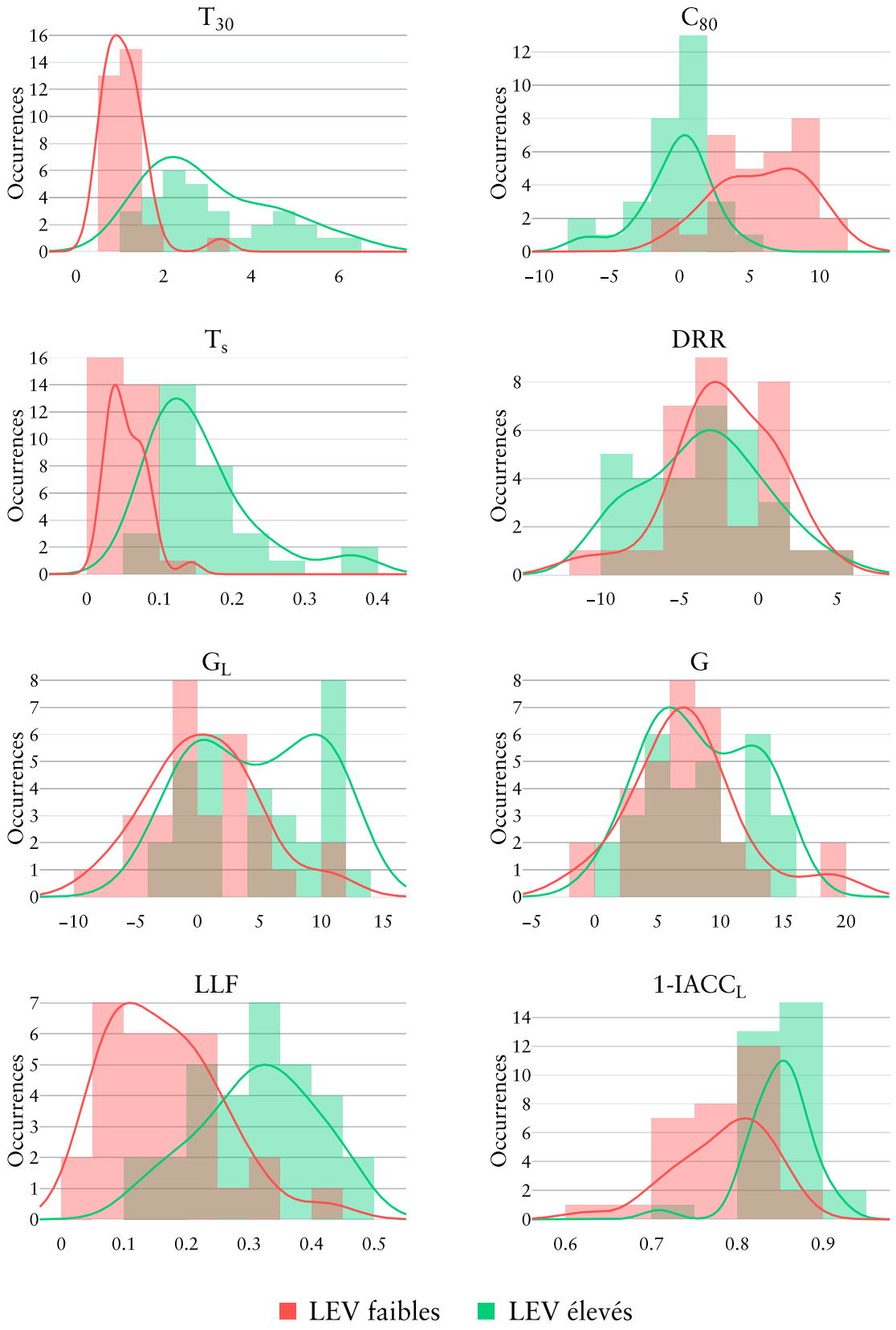


FIGURE 7.13 – Histogramme et densité de probabilité de paramètres acoustiques calculés pour les configurations acoustiques produisant les sensations d'enveloppement les plus élevées et les plus faibles pour les deux sources sonores. Les distributions des paramètres acoustiques selon les deux catégories de configurations acoustiques sont significativement différentes sauf pour G ($p = 0.301$) et DRR ($p = 0.166$).

propre à cette région temporelle n'est significativement différent selon la condition d'enveloppement.

3. L'énergie latérale, caractérisée dans les paramètres L_J et LLF, semble avoir une influence significative sur l'enveloppement ce qui confirme l'importance des réflexions tardives latérales pour caractériser cette impression.
4. L'emploi d'une analyse spatiale différente de celle spécifiée dans la norme n'a pas apporté d'amélioration des résultats. Le temps central pondéré dans l'espace SBT_S ne semble pas être plus pertinent que le temps central monodimensionnel. Le paramètre FBR correspondant au rapport entre l'énergie arrière et l'énergie frontale n'est pas différent selon le type de configuration acoustique considéré.
5. La décorrélation interaurale semble favoriser la sensation d'enveloppement. Le calcul de la décorrélation aussi bien sur la partie tardive que sur la durée totale des réponses impulsionnelles de salle mène à des résultats significativement différents selon les deux groupes de configurations acoustiques.

Détermination d'un hyperplan séparateur

Pour identifier les paramètres acoustiques permettant de discriminer les configurations acoustiques en termes d'enveloppement nous avons procédé au calcul d'hyperplans séparateurs en considérant différents paramètres acoustiques. L'ensemble des paramètres acoustiques ont été considérés à l'exception de G_E, G_{EL}, J_{LF}, J_{LFC}, B_{LF}, B_{LFC}, BQI, G, DRR et FBR dont les valeurs ne sont pas significativement différentes d'une catégorie à l'autre et de ceux présentant une forte corrélation avec un autre paramètre acoustique (coefficient de Pearson > 0.9) : D_{50}, G_L, T_s et SBT_s.

Nous avons utilisé 42 vecteurs de paramètres acoustiques tirés aléatoirement pour le calcul des hyperplans séparateurs et les tests étaient réalisés avec les 20 vecteurs restants. Pour un ensemble de paramètres acoustiques donné cette opération a été effectuée 20 fois. Le tableau 7.9 rassemble les moyennes et écart-types de la précision obtenue en fonction des paramètres considérés. Seuls les paramètres ayant permis l'obtention d'une précision moyenne supérieure à 90% sont mentionnés. Au regard du nombre de vecteurs d'apprentissage, trois paramètres acoustiques maximums ont été considérés.

Il est intéressant de constater que les meilleurs résultats de classification ont été obtenus en considérant, entre autres, les paramètres EDT, L_J et C_{80} déjà identifiés comme expliquant la plus grande partie de la variance des notes d'enveloppement dans la section 7.5.1. En plus de l'EDT, T_{30} semble également pertinent pour discriminer les configurations acoustiques selon la condition d'enveloppement. En plus de L_J, LLF (qui caractérise la proportion d'énergie latérale tardive) est également pertinent pour analyser les deux catégories. Par ailleurs, la décorrélation interaurale des réponses impulsionnelles, (1-IACC et $IACC_L$) semble permettre de discriminer les configurations acoustiques d'ASW élevée et faibles. Ainsi, il semble que l'analyse conjointe de trois grandeurs permette d'expliquer des notes d'enveloppement élevées ou faibles : la quantité d'énergie tardive, la quantité de réflexions latérales tardives et la décorélation interaurale.

Paramètres acoustiques utilisés	Précision moyenne ± écart-type
T_{30}, 1-$IACC_L$	90.7 ± 4.9
T_{30}, EDT, L_J	93.6 ± 4.3
EDT, L_J, C_{80}	92.8 ± 5.6
EDT, L_J, LLF	93.1 ± 5.5
EDT, 1-$IACC_L$, LLF	92.6 ± 4.8
EDT, 1-$IACC_L$, L_J	91.8 ± 3.8
EDT, 1-IACC, L_J	94.2 ± 3.6

Tableau 7.9 – Précision de classification des configurations acoustiques selon les notes d'enveloppement élevées (classe verte) ou faibles (classe rouge). Seuls les ensembles de deux paramètres acoustiques minimum et trois maximum ayant permis l'obtention d'une précision moyenne supérieure à 90% sont mentionnés.

7.6. Conclusion

La base de données GRAP rassemble des évaluations perceptives relatives à l'auralisation de 35 salles couvrant une grande variété de propriétés acoustiques. Plusieurs réponses impulsionnelles et résultats de simulations sont inclus dans la base et permettent ainsi de mettre en relation les paramètres acoustiques des salles avec les notes attribuées à des attributs perceptifs, ici la largeur apparente de source et l'enveloppement. Pour compléter les données à disposition, des SRIRs ont été générées d'après les résultats de simulation. Ces SRIRs ont été encodées en ambisonique à l'ordre 4, c'est-à-dire selon une résolution spatiale comparable à celle d'instruments de mesure communément employés tels que l'Eigenmike[19]. En plus des paramètres acoustiques définis dans la norme ISO 3382-1 et fournis dans la base de données GRAP, d'autres paramètres acoustiques proposés dans la littérature ont été calculés d'après ces SRIRs.

Afin d'établir des relations entre les paramètres acoustiques et la largeur apparente de source, des corrélations et une régression linéaire multidimensionnelle ont été calculés. La variance expliquée par le modèle linéaire est apparue trop faible pour être en mesure de prédire la largeur apparente de source de cette manière. L'identification des paramètres acoustiques permettant d'expliquer les différences entre les notes de largeur apparente de source les plus élevées et les plus faibles a été réalisée en utilisant une classification binaire. Les enseignements principaux de cette analyse sont que la prédominance du son direct apparait néfaste à la perception d'une source large et que celle-ci semble favorisée par une énergie tardive importante. Cette observation est en désaccord avec l'étude de Bradley *et al.* [1] selon laquelle l'accroissement de l'énergie tardive tend à amoindrir la largeur apparente de source. Les paramètres acoustiques propres à l'énergie précoce étaient significativement différents selon que la largeur de source était perçue large ou étroite, bien qu'ils n'aient pas permis d'obtenir les meilleures précisions de classification.

La largeur apparente de source semble difficile à caractériser avec les données à disposition. Les notes associées à cet attribut perceptif présentent une forte variabilité d'une source sonore à l'autre. Il serait utile d'obtenir des données liées à de multiples

sources sonores pour réduire l'influence de la source et ainsi mieux expliquer les résultats. Par ailleurs, comme en atteste la faible performance de la classification binaire, un jeu de données plus important paraît nécessaire pour analyser plus finement les liens entre les paramètres acoustiques et la largeur apparente de source.

Une étude similaire a été effectuée avec les notes d'enveloppement. La régression linéaire multidimensionnelle et la discrimination des notes d'enveloppement les plus élevées et les plus faibles d'après les paramètres acoustiques confirment que l'énergie tardive joue un rôle significatif dans la perception de l'enveloppement. La quantité d'énergie tardive, la quantité de réflexions latérales tardives et la décorrélation interaurale explique une partie importante des notes d'enveloppement et permettent de différencier celles les plus élevées des plus faibles. Alors que l'étude de Dick *et al.* rapporte un rôle significatif de l'énergie précoce sur la sensation d'enveloppement [21], aucun paramètre acoustique propre à cette région temporelle n'a permis d'expliquer les notes d'enveloppement.

D'autres études sont nécessaires pour mettre en lumière une possible influence de la région temporelle utilisée pour définir les réflexions précoces. La prise en compte du temps de mélange pourrait permettre d'améliorer les performances de certains paramètres tels que les fractions latérales d'énergie précoce ou de force sonore précoce. Comme proposé par Soulodre *et al.* [32], la région temporelle des premières réflexions utilisée pourrait dépendre de la fréquence.

Les paramètres acoustiques B_{LF}, B_{LFC} et SBT_s, calculés d'après une résolution angulaire élevée, ne se sont pas montrés plus pertinents que les paramètres J_{LF}, J_{LFC} et T_s spécifiés dans la norme. Aucun des paramètres acoustiques identifiés comme pertinents dans cette étude pour caractériser les impressions spatiales ne nécessite l'usage d'une antenne ambisonique d'ordre élevé.

Bibliographie

[1] J. Bradley, R. Reich et S. Norcross, "On the combined effects of early-and late-arriving sound on spatial impression in concert halls," *The Journal of the Acoustical Society of America*, vol. 108, n°. 2, p. 651–661, 2000.

[2] ISO 3382-1, "Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces," International Organization for Standardization, Geneva, CH, Standard, 2009.

[3] M. Barron et A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls : the derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, n°. 2, p. 211–232, 1981.

[4] J. S. Bradley et G. A. Soulodre, "Objective measures of listener envelopment," *The Journal of the Acoustical Society of America*, vol. 98, n°. 5, p. 2590–2597, 1995.

[5] J. S. Bradley et G. A. Soulodre, "The influence of late arriving energy on spatial impression," *The Journal of the Acoustical Society of America*, vol. 97, n°. 4, p. 2263–2271, 1995.

[6] D. de Vries, E. M. Hulsebos et J. Baan, "Spatial fluctuations in measures for spaciousness," *The journal of the Acoustical Society of America*, vol. 110, n°. 2, p. 947–954, 2001.

[7] M. Morimoto, K. Iida et K. Sakagami, “The role of reflections from behind the listener in spatial impression,” *Applied Acoustics*, vol. 62, n°. 2, p. 109–124, 2001.

[8] T. Hanyu et S. Kimura, “A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections,” *Applied Acoustics*, vol. 62, n°. 2, p. 155–184, 2001.

[9] D. Ackermann, M. Ilse, D. Grigoriev, S. Lepa, S. Pelzer, M. Vorländer et S. Weinzierl, “A ground truth on room acoustical analysis and perception (GRAP),” 2018.

[10] S. Weinzierl, S. Lepa et D. Ackermann, “A measuring instrument for the auditory perception of rooms : The room acoustical quality inventory (raqi),” *The Journal of the Acoustical Society of America*, vol. 144, n°. 3, p. 1245–1257, 2018.

[11] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, S. van de Par, M. Müller-Trapet, R. Opdam et M. Vorländer, “The FABIAN head-related transfer function data base,” 2017.

[12] M. Geier, J. Ahrens et S. Spors, “The soundscape renderer : A unified spatial audio reproduction framework for arbitrary rendering methods,” dans *In 124 th AES Conv*. Citeseer, 2008.

[13] V. Erbes, F. Schultz, A. Lindau et S. Weinzierl, “An extraaural headphone system for optimized binaural reproduction,” dans *Proc. 2012 German annual acoustic conference (DAGA)*, 2012, p. 313–314.

[14] C. Böhm, F. Fiedler, S. Weinzierl, E. Holter, S. Muth, U. Schaefer et S. Schwesinger, “An anechoic recording of cicero’s 3rd cataline oration : Italian, latin and german,” 2019.

[15] M. C. Vigeant, L. M. Wang, J. H. Rindel, C. L. Christensen et A. C. Gade, “Multi-channel orchestral anechoic recordings for auralizations,” 2010.

[16] S. Weinzierl, M. Vorländer, G. Behler, F. Brinkmann, H. v. Coler, E. Detzner, J. Krämer, A. Lindau, M. Pollow, F. Schulz et N. R. Shabtai, “A database of anechoic microphone array measurements of musical instruments,” 2017.

[17] D. Schröder et M. Vorländer, “RAVEN : A real-time framework for the auralization of interactive virtual environments,” dans *Forum Acusticum*. Aalborg Denmark, 2011, p. 1541–1546.

[18] M. Vorländer, *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[19] MH Acoustics LLC, “Eigenmike em32 microphone array,” accessed 2020-08-22. [En ligne]. Disponible : https://mhacoustics.com/products

[20] I. Selesnick, “Linear-phase fir filter design by least squares,” *Connexions*, 2005.

[21] D. A. Dick et M. C. Vigeant, “An investigation of listener envelopment utilizing a spherical microphone array and third-order ambisonics reproduction,” *The Journal of the Acoustical Society of America*, vol. 145, n°. 4, p. 2795–2809, 2019.

[22] J. Daniel, “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia,” Thèse de doctorat, University of Paris VI, 2000.

[23] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co, 1996.

[24] E. J. Pedhazur et L. P. Schmelkin, *Measurement, design, and analysis : An integrated approach*. psychology press, 2013.

[25] D. E. Miller et J. T. Kunce, "Prediction and statistical overkill revisited," *Measurement and evaluation in guidance*, vol. 6, n°. 3, p. 157–163, 1973.

[26] B. G. Tabachnick, L. S. Fidell et J. B. Ullman, *Using multivariate statistics*. Pearson Boston, MA, 2007, vol. 5.

[27] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, n°. 3, p. 1065–1076, 1962.

[28] S. Klockgether et S. van de Par, "A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception," *Acta Acustica united with Acustica*, vol. 100, n°. 5, p. 964–971, 2014.

[29] D. Griesinger, "Room impression, reverberance, and warmth in rooms and halls," dans *Audio Engineering Society Convention 93*. Audio Engineering Society, 1992.

[30] H. Furuya, K. Fujimoto, C. Y. Ji et N. Higa, "Arrival direction of late sound and listener envelopment," *Applied Acoustics*, vol. 62, n°. 2, p. 125–136, 2001.

[31] H. Furuya, K. Fujimoto, A. Wakuda et Y. Nakano, "The influence of total and directional energy of late sound on listener envelopment," *Acoustical science and technology*, vol. 26, n°. 2, p. 208–211, 2005.

[32] G. A. Soulodre, M. C. Lavoie et S. G. Norcross, "Objective measures of listener envelopment in multichannel surround systems," *Journal of the Audio Engineering Society*, vol. 51, n°. 9, p. 826–840, 2003.

8

L'influence de la spatialisation des premières réflexions sur la largeur apparente de source

Le chapitre précédent a mis en lumière le rôle prépondérant de l'énergie tardive et, dans une moindre mesure, de l'énergie précoce dans la perception de la largeur de source (ASW). Il en ressort également que la sensation d'enveloppement ne semble pas liée aux propriétés de l'énergie précoce. Il parait donc envisageable de contrôler la largeur apparente de source indépendamment de la sensation d'enveloppement en modifiant les propriétés de l'énergie précoce. De nombreuses études se sont concentrées exclusivement sur le lien entre la proportion d'énergie latérale précoce ou la décorrélation interaurale précoce de réponses impulsionnelles et la largeur apparente de source. Afin d'évaluer l'importance de ce lien, nous étudions ici dans quelle mesure la modification de ces deux propriétés acoustiques peut influencer l'ASW. L'étude s'inscrit dans le contexte d'un rendu binaural non-individualisé où le champ réverbéré est reproduit avec une faible résolution spatiale (ambisonique d'ordre 1). Les parties précoces de plusieurs SRIRs sont manipulées selon différentes transformations spatiales consistant en l'augmentation ou la réduction de l'énergie latérale et la décorrélation interaurale. Les résultats d'un test perceptif montrent que leur variation peut avoir une influence significative sur l'ASW. Néanmoins, la modification de l'énergie précoce de SRIRs ne présente pas d'influence significative sur l'ASW pour tous les espaces considérés. Il semble que l'efficacité des traitements appliqués pour accroître l'ASW soit dépendante de l'énergie des réflexions spéculaires proches du son direct.

8.1. Introduction

De nombreuses études mettent en évidence le rôle joué par l'énergie précoce de réponses impulsionnelles de salle sur la largeur apparente de source, couramment

désignée par ASW (pour *Apparent Source Width*) [1–6]. En raison de l'effet de précédence [7], les réflexions précoces en fusionnant perceptivement avec le son direct contribuent à flouter la localisation du son et ainsi élargir la source sonore. Barron [1] a également montré que, même lorsqu'une réflexion est perçue comme un écho distinct, elle peut contribuer à l'ASW. Cependant, cette contribution apparait lorsque la direction d'incidence provient d'une région particulière de l'espace. Pour Johnson *et al.* [6], deux réflexions contenues dans une région angulaire comprise entre 40° et 130° et dont le retard est inférieur à 30 ms par rapport à un son direct frontal produisent la même perception de largeur de source et cette largeur est maximale dans cette région spatio-temporelle. Ce résultat met en avant l'importance de l'énergie latérale précoce dans la perception d'une source large.

Plusieurs paramètres acoustiques sont définis dans la norme ISO 3382-1 [8] pour prédire l'ASW : la fraction latérale (J_{LF}), le coefficient de fraction latérale (J_{LFC}) et le coefficient de corrélation interaule (IACC). Beranek [5] a également proposé l'utilisation du Binaural Quality Index (BQI), qui correspond au coefficient de décorrélation interaurale calculé sur l'énergie précoce et moyenné sur les bandes de fréquences centrées sur 500 Hz, 1 kHz et 2 kHz. Dans le chapitre 7, l'étude des valeurs de BQI, J_{LF} et J_{LFC} pour de nombreux espaces sonores a confirmé la pertinence de ces paramètres pour caractériser l'ASW.

Plusieurs traitements de spatialisation sonore ont été proposés afin d'accroitre l'étendue spatiale d'une source anéchoïque ou d'une scène sonore [9–13]. Une méthode commune consiste à décomposer le signal à élargir en plusieurs sources sonores ponctuelles spatialement distinctes. Il faut cependant que ces sources soient décorrélées les unes des autres afin qu'elles soient perçues comme des événements sonores différents et non comme un unique événement sonore. La décorrélation est communément effectuée au moyen de filtres de décorrélation variants ou non dans le temps, ou d'après un procédé de *panning* d'amplitude (cf section 4.7.1). Cette décorrélation introduite dans la scène sonore permet une localisation moins précise de la source qui est comme floutée et l'étendue spatiale perçue de la source est alors plus importante.

En appliquant de tels traitements sur le son direct d'une SRIR, il serait possible d'accroitre l'ASW. Néanmoins, ces traitements ne permettent pas de réduire l'ASW. Nous souhaitons dans ce chapitre évaluer la possibilité d'augmenter et de réduire l'ASW dans un moteur de rendu binaural reproduisant un effet de réverbération encodé en ambisonique à l'ordre 1. Les premières réflexions de plusieurs SRIRs ont été traitées de manière à modifier l'énergie latérale précoce et la décorrélation interaurale résultante selon plusieurs méthodes. Les quatre méthodes de modifications spatiales employées seront d'abord exposées puis le test perceptif destiné à évaluer leur influence sur l'ASW sera présenté ainsi que les résultats obtenus.

8.2. Les modifications spatiales des premières réflexions

Dans cette section, plusieurs méthodes de spatialisation des premières réflexions d'une SRIR encodée en ambisonique sont présentées. Parmi ces méthodes, deux types de transformations visent à augmenter ou réduire l'énergie latérale précoce d'une SRIR. Une autre méthode permet l'accroissement de la décorrélation interaurale et

une dernière méthode offre la possibilité d'uniformiser la répartition de l'énergie dans le plan horizontal.

8.2.1. Amplification directionnelle

L'amplification directionnelle est un traitement permettant la modification de l'énergie sonore dans des directions spécifiques. Notons $\boldsymbol{b}(k)$ les composantes ambisoniques d'ordre L représentant les premières réflexions dans le domaine des harmoniques sphériques. Nous cherchons une matrice $\boldsymbol{T}$ permettant d'obtenir les signaux modifiés $\tilde{\boldsymbol{b}}(k)$ tels que :

$$\tilde{\boldsymbol{b}}(k) = \boldsymbol{T}\boldsymbol{b}(k) \tag{8.1}$$

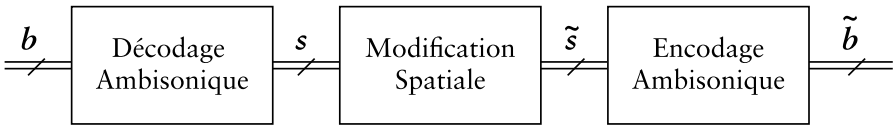


FIGURE 8.1 – Détails des opérations appliquées aux signaux ambisoniques des premières réflexions.

Les modifications spatiales peuvent être appliquées aux signaux $\boldsymbol{s}(k)$ extraits par décodage ambisonique selon un nombre suffisant de directions régulièrement réparties dans l'espace. Ces directions sont choisies de manière à être capable de reconstruire la représentation ambisonique initiale à l'ordre L. Soient $\boldsymbol{Y}_\Omega$ la matrice des harmoniques sphériques correspondant aux positions angulaires $\Omega_n = (\theta_n, \varphi_n)$ de N points régulièrement répartis dans l'espace :

$$\boldsymbol{Y}_\Omega = \begin{pmatrix} \boldsymbol{y}(\Omega_1) \\ \boldsymbol{y}(\Omega_2) \\ \vdots \\ \boldsymbol{y}(\Omega_N) \end{pmatrix} \tag{8.2}$$

avec

$$\boldsymbol{y}(\Omega_n) = [Y_{0,0}(\Omega_n)\ Y_{1,-1}(\Omega_n)\ \ldots\ Y_{L,L}(\Omega_n)]\ . \tag{8.3}$$

et

$$\Omega = [(\theta_1, \varphi_1)\ (\theta_2, \varphi_2)\ \ldots\ (\theta_N, \varphi_N)]\ . \tag{8.4}$$

Pour une implémentation simple nous choisissons des positions angulaires telles que :

$$\boldsymbol{I}_{(L+1)^2} = \frac{1}{N} Y_\Omega^\top Y_\Omega\ . \tag{8.5}$$

Les transformations sont appliquées aux signaux décodés $\boldsymbol{s}(k) = \frac{1}{N} Y_\Omega \boldsymbol{b}(k)$. Le processus de décodage, de modification spatiale des signaux et d'encodage permet de définir la matrice $\boldsymbol{T}$ comme illustré sur la figure 8.1.

De manière à amplifier ou réduire l'énergie latérale des premières réflexions une fonction de pondération doit être appliquée aux signaux décodés dans les N directions. Pour ce faire, le gain $\rho(\theta_n, \varphi_n)$ est appliqué au signal $s_n(k)$ issu du décodage ambisonique dans la direction (θ_n, φ_n). Dans le système de coordonnées cartésiennes défini en annexe B.1, ce gain directionnel est défini par l'équation suivante :

$$\rho(x_n, y_n, z_n) = \frac{\beta}{\sqrt{\beta^2 x_n^2 + y_n^2 + \beta^2 z_n^2}} \tag{8.6}$$

où β correspond au gain maximal à appliquer dans la direction latérale. De cette manière, le gain directionnel correspond au rayon des points d'une ellipsoïde dont les demi axes non latéraux sont unitaires et dont le demi axe latéral est égal à β. Un coefficient $\beta = 1$ signifie donc qu'un gain unitaire est appliqué dans toutes les directions. Ainsi la matrice de transformation $\boldsymbol{T}$ s'écrit :

$$\boldsymbol{T} = \frac{1}{N} Y_\Omega^\top \operatorname{diag}[\rho(\theta_n, \varphi_n)] \, Y_\Omega \tag{8.7}$$

La figure 8.2 représente le diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR encodée à l'ordre 1 et les modifications associées selon deux coefficients β.

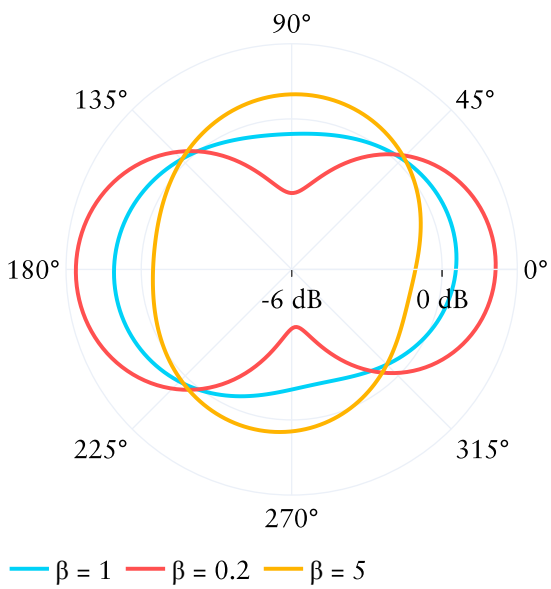


FIGURE 8.2 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR mesurée (*Toilettes*) et encodée en ambisonique à l'ordre 1 (en bleu). Celui résultant d'une amplification de l'énergie latérale est représenté en jaune et celui résultant de la réduction de l'énergie latérale est représenté en rouge.

8.2.2. Déformation angulaire

Formalisée par Pomberger et Zotter [14], le *warping* ou déformation angulaire, est une transformation spatiale qui permet de modifier la direction d'incidence des événements sonores contenus dans une scène ambisonique selon un coefficient de pondération fonction de leur azimut et/ou élévation. En d'autres termes, cette déformation est utilisée pour étirer une certaine région de l'espace tout en comprimant d'autres régions. Dans notre cas de figure, ce coefficient de pondération permet de concentrer les premières réflexions soit vers le plan médian, soit vers le plan interaural.

Pour ce faire, les signaux décodés $\boldsymbol{s}(k)$ sont ré-encodés en ambisonique selon des directions d'incidence modifiées. La largeur apparente de source correspondant à l'étendue spatiale de la source perçue dans le plan horizontal, nous avons choisi de modifier les azimuts θ_n des N directions d'encodage. Soient d la fonction de modification des azimuts et α le coefficient qui contrôle la direction et l'ampleur de la modification :

$$d(\theta) = \begin{cases} \frac{\pi}{2\alpha} \tan[(\frac{2\theta}{\pi} - 1) \arctan(\alpha)] + \frac{\pi}{2} & \text{si } \alpha > 0 \\ \frac{\pi}{2 \arctan \alpha} \arctan[(\frac{2\theta}{\pi} - 1)\alpha] + \frac{\pi}{2} & \text{si } \alpha < 0 \end{cases} \tag{8.8}$$

La figure 8.3 représente la fonction d pour différentes valeurs du coefficient α. On constate qu'en fonction du signe de α s'opère une concentration de l'azimut θ soit vers la direction frontale ($\theta = 0°$) soit vers la direction latérale ($\theta = 90°$).

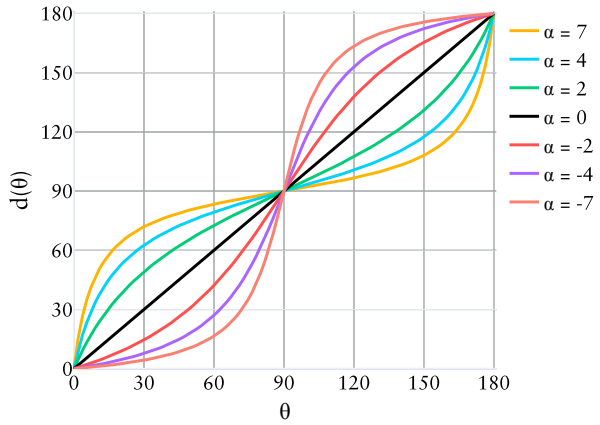


FIGURE 8.3 – Réseau de courbes représentant les valeurs de $d(\theta)$ pour différentes coefficients α.

Ainsi, la matrice de transformation $\boldsymbol{T}$ à appliquer aux signaux ambisoniques pour procéder à une déformation angulaire s'écrit :

$$T = \frac{1}{N} Y_{\tilde{\Omega}}^{\top} Y_{\Omega} \tag{8.9}$$

avec

$$\tilde{\Omega} = [(d(\theta_1), \varphi_1) \; (d(\theta_2), \varphi_2) \; \ldots \; (d(\theta_N), \varphi_N)] \, . \tag{8.10}$$

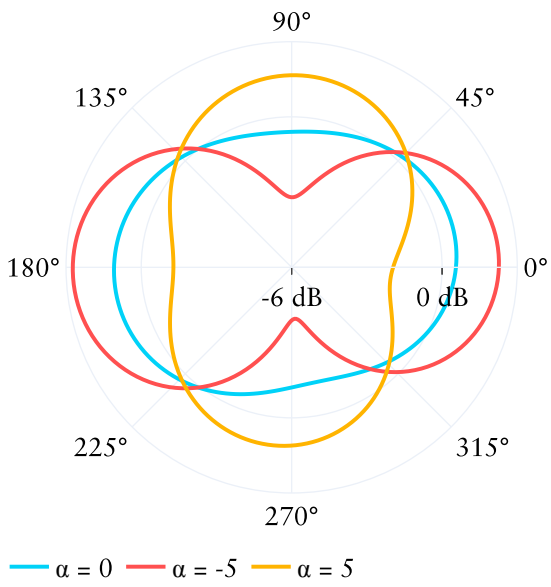


FIGURE 8.4 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR mesurée (*Toilettes*) et encodée en ambisonique à l'ordre 1 (en bleu). Celui résultant d'une déformation angulaire vers le plan interaural est représenté en jaune et celui résultant d'une déformation angulaire vers le plan médian e est représenté en rouge.

La figure 8.4 représente le diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR encodée à l'ordre 1 et les déformations angulaires associées selon deux coefficients α.

8

8.2.3. Élargissement

Zotter et Kronlachner [11] ont proposé une méthode permettant d'accroitre l'étendue spatiale d'une scène ambisonique en dispersant les directions d'arrivées des composantes fréquentielles des sources sonores constituant la scène. Autrement dit, une composante fréquentielle provenant de la direction ϕ_0 est encodée selon une direction d'incidence $\phi_0 + \phi(\omega)$, où ϕ est l'angle de modulation dépendant de la fréquence angulaire ω. En spatialisant différemment les composantes fréquentielles du signal, la localisation de la source est moins précise et l'énergie de la source est comme dispersée dans un région angulaire, ce qui permet d'accroitre la décorrélation interaurale.

Cette méthode peut s'appliquer après encodage ambisonique en effectuant une rotation de la scène ambisonique autour de l'axe z selon un angle dépendant de la fréquence. Les signaux ambisoniques d'ordre m de $\tilde{\boldsymbol{b}}$ issus de la rotation d'un angle ϕ autour de l'axe z d'une scène ambisonique décrite par les signaux $\boldsymbol{b}$ s'écrivent pour un nombre d'onde k :

$$\begin{pmatrix} \tilde{b}_m(k) \\ \tilde{b}_{-m}(k) \end{pmatrix} = \begin{pmatrix} \cos(m\phi) & -\sin(m\phi) \\ \sin(m\phi) & \cos(m\phi) \end{pmatrix} \begin{pmatrix} b_m(k) \\ b_{-m}(k) \end{pmatrix} \tag{8.11}$$

L'élargissement peut être mis en œuvre avec un angle de modulation

$$\phi = 2\pi\mu\cos(\omega\tau) \tag{8.12}$$

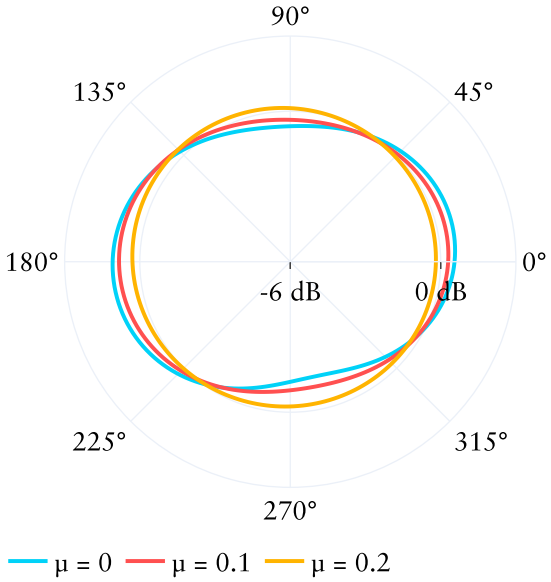


FIGURE 8.5 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR mesurée (*Toilettes*) et encodée en ambisonique à l'ordre 1 (en bleu) ainsi que les diagrammes de directivité du champ sonore modifié selon deux valeurs du coefficient d'élargissement μ de l'équation 8.12 ($\tau = 2.5$ ms dans les deux cas).

où μ est l'amplitude de modulation et τ est la période de modulation. L'expression de cette modulation dans le domaine temporel permet d'implémenter la méthode grâce à une convolution parcimonieuse, ce qui rend le traitement efficace en terme de calcul. La figure 8.5 représente des exemples d'élargissement effectuées sur la partie précoce d'une SRIR. On distingue un lissage de l'énergie dans le plan qui s'opère avec l'augmentation de l'amplitude de modulation.

8.2.4. Décorrélation des composantes sectorielles

Ce procédé consiste à modifier les corrélations entre signaux ambisoniques correspondants aux harmoniques sphériques sectorielles pour réduire la directionnalité du champ sonore dans le plan. Les composantes ambisoniques sectorielles $b_{l,m}$ d'ordre l et de degré m correspondent aux composantes ambisoniques vérifiant $|m| = l$. Pour ces composantes, quelque soit l'azimut, la phase ne change pas de signe selon l'élévation. Pour un ordre ambisonique L, $2L+1$ composantes sectorielles permettent de décrire une scène sonore dans le plan. Dans certain cas, la directionalité des premières réflexions est dominée par la direction frontale, celle du son direct et de la réflexion sur le sol. Cette méthode revient à réduire cette influence en répartissant spatialement l'énergie de manière plus équitable ou au contraire à exacerber cette directionalité.

La décorrélation entre des signaux ambisoniques peut être effectuée en modifiant les valeurs propres de la matrice de covariance associée. La matrice de covariance

$\boldsymbol{C}$ des valeurs fréquentielles $\boldsymbol{b}_s$ des signaux ambisoniques sectoriels est donnée par $\boldsymbol{C} = \mathbb{E}[\boldsymbol{b}_s \boldsymbol{b}_s{}^H]$ où ^{H} désigne l'opérateur adjoint. Soit $\boldsymbol{V}$ la matrice des vecteurs propres associés aux valeurs propres $(\lambda_i)_{1 \leq i \leq N}$ de $\boldsymbol{C}$. La décomposition en valeurs propres de la matrice $\boldsymbol{C}$ s'écrit :

$$\boldsymbol{C} = \boldsymbol{V}^H \operatorname{diag}(\Lambda) \boldsymbol{V} \tag{8.13}$$

avec le vecteur des valeurs propres $\Lambda = [\lambda_1 \ \lambda_2 \ \ldots \ \lambda_N]$.

Soit $\tilde{\boldsymbol{b}}_s = \boldsymbol{T}_s \boldsymbol{b}_s$. Nous souhaitons que les signaux $\tilde{\boldsymbol{b}}_s$ soient décorrélés et de même énergie pour ne favoriser aucune direction de l'espace. En d'autres termes, nous souhaitons que la matrice de covariance correspondante $\tilde{\boldsymbol{C}}$ soit proportionnelle à la matrice identité $\boldsymbol{I}_{2L+1}$:

$$\tilde{\boldsymbol{C}} = \mathbb{E}[\tilde{\boldsymbol{b}}_s \tilde{\boldsymbol{b}}_s{}^H] \propto \boldsymbol{I}_{2L+1} \tag{8.14}$$

$$\boldsymbol{T}_s \mathbb{E}[\boldsymbol{b}_s \boldsymbol{b}_s{}^H] \boldsymbol{T}_s^H \propto \boldsymbol{I}_{2L+1} \tag{8.15}$$

$$\boldsymbol{T}_s \boldsymbol{C} \boldsymbol{T}_s^H \propto \boldsymbol{I}_{2L+1} \tag{8.16}$$

$$\boldsymbol{T}_s \boldsymbol{T}_s^H \propto \boldsymbol{C}^{-1} \tag{8.17}$$

Il existe une infinité de solution à cette équation et une solution simple consiste à choisir l'expression suivante :

$$\boldsymbol{T}_s = \boldsymbol{V}^H \operatorname{diag}(\Lambda)^{-\frac{1}{2}} \boldsymbol{V} \tag{8.18}$$

Les signaux ambisoniques $\tilde{\boldsymbol{b}}$ sont ainsi décorrélés dans le plan. Il est possible de contrôler la décorrélation en appliquant un vecteur de pondération $\boldsymbol{g}$ aux valeurs propres. Cette pondération permet de réduire progressivement les différences entre les valeurs propres de la matrice de covariance $\tilde{\boldsymbol{C}}$ avec l'augmentation d'un coefficient γ jusqu'à l'obtention d'une matrice identité pour $\gamma = 1$. La matrice $\boldsymbol{T}_s$ s'écrit alors :

$$\boldsymbol{T}_s = \boldsymbol{V}^H \operatorname{diag}(\boldsymbol{g}_\gamma)^{\frac{1}{2}} \operatorname{diag}(\Lambda)^{-\frac{1}{2}} \boldsymbol{V} \tag{8.19}$$

où

$$\boldsymbol{g}_\gamma = [g_{\gamma,1} \ g_{\gamma,2} \ \ldots \ g_{\gamma,2L+1}] \ , \tag{8.20}$$

$$g_{\gamma,i} = \begin{cases} \max(p\lambda_i, \bar{\lambda}) & \text{si } \lambda_i > \bar{\lambda}_i \\ \min(p\lambda_i, \bar{\lambda}) & \text{si } \lambda_i < \bar{\lambda}_i \end{cases} \tag{8.21}$$

et

$$p = \begin{cases} (1-\gamma)(1-\bar{\lambda}/\lambda_{max}) + \bar{\lambda}/\lambda_{max} & \text{si } \lambda_i > \bar{\lambda}_i \\ (1-\gamma)(1-\bar{\lambda}/\lambda_{min}) + \bar{\lambda}/\lambda_{min} & \text{si } \lambda_i < \bar{\lambda}_i \end{cases} \tag{8.22}$$

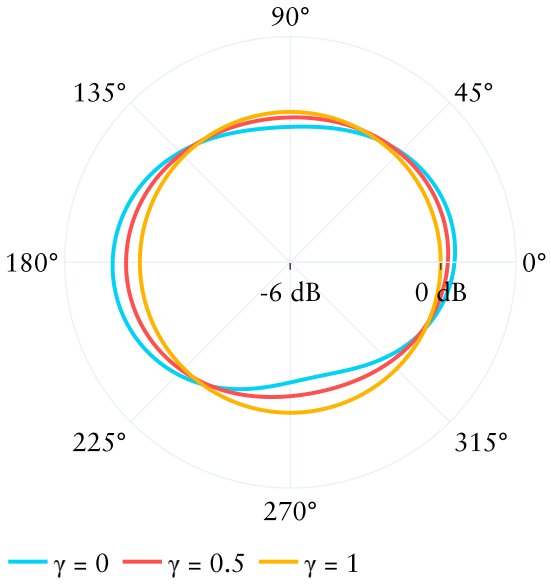


FIGURE 8.6 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions d'une SRIR mesurée (*Toilettes*) et encodée en ambisonique à l'ordre 1 (en bleu) ainsi que les diagrammes de directivité du champ sonore modifié en décorrélant les composantes ambisoniques sectorielles selon deux valeurs du coefficient γ.

avec $\bar{\lambda}$ la moyenne des valeurs propres, λ_{max} la valeur propre maximale et λ_{min} la valeur propre minimale.

La figure 8.6 représente les résultats de décorrélation des composantes ambisoniques dans le plan effectuées sur la partie précoce d'une SRIR. On distingue un lissage progressif de l'énergie dans le plan avec l'augmentation de la décorrélation entre les composantes ambisoniques.

8.3. Test Perceptif

Les transformations spatiales présentées dans la section précédente ont été appliquées à la partie précoce de plusieurs SRIRs dans le but de faire varier les paramètres acoustiques BQI, J_{LF} et J_{LFC} et un test perceptif a été réalisé afin d'évaluer la largeur apparente de source résultante. Le protocole de test était inspiré de la méthodologie MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*) [15]. Pour chaque espace auralisé, les stimuli sonores issus des SRIRs modifiées étaient comparés au stimulus de référence issu de la SRIR d'origine.

8.3.1. Réponses impulsionnelles spatiales de salle

Cinq espaces ont été utilisés pour créer les stimuli sonores : un amphithéâtre, une église, un hall d'université, un réfectoire et des sanitaires. Ces espaces seront désignés *Amphi*, *Église*, *Hall*, *Réfectoire* et *Toilettes* respectivement. Ils ont été choisis de façon à couvrir une large gamme de valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC}. Les différentes mesures présentées dans le tableau 8.1 illustrent cette variété. L'espace

Toilettes présente la plus grande proportion d'énergie précoce avec l'espace *Amphi* comme en attestent les mesures de clarté sonore C_{80}, D_{50} et T_s.

Paramètres	*Amphi*	*Église*	*Hall*	*Réfectoire*	*Toilettes*
T_{30} (s)	0.76	3.95	0.88	0.90	0.40
EDT (s)	0.71	3.25	0.60	0.71	0.37
T_s (ms)	16.3	54.2	34.0	31.7	22.4
C_{80} (dB)	11.80	7.53	8.78	8.57	13.85
D_{50} (%)	89.98	83.68	76.28	79.21	86.30
DRR (dB)	3.15	5.65	-1.70	2.09	-3.42
G_E (dB)	30.84	30.07	32.99	28.86	32.71
J_{LF}	0.119	0.059	0.296	0.609	0.684
J_{LFC}	0.114	0.070	0.270	0.429	0.474
BQI	0.257	0.097	0.490	0.618	0.800
1-IACC	0.297	0.264	0.548	0.665	0.810

Tableau 8.1 – Paramètres acoustiques des SRIRs utilisées dans le test perceptif. Les valeurs des paramètres correspondent à la moyenne des valeurs calculées dans les bandes de fréquence centrées sur 500 Hz et 1000 Hz sauf pour J_{LF} et J_{LFC} dont la moyenne comprend également les valeurs calculées dans les bandes de fréquence centrées sur 125 Hz et 250 Hz [8] et pour le BQI qui comprend également la valeur calculée dans la bande de fréquence centrées sur 2 kHz [5]. La force sonore des premières réflexions G_E a été calculée sur la région temporelle délimitée par le temps de mélange. Les valeurs minimales et maximales apparaissent en gras.

Les SRIRs ont été mesurées avec l'antenne sphérique de microphone Eigenmike EM32 et une enceinte Genelec 8040. L'enceinte était positionnée à 3 m du microphone pour tous les espaces sauf pour *Hall* où l'enceinte était positionnée à 4 m. Un signal à balayage sinusoïdal de 10 secondes a été utilisé durant la mesure. Les mesures de réponses impulsionnelles ont ensuite été débruitées selon la procédure décrite par Cabrera et al. [16]. Afin de compenser la réponse en fréquence de l'enceinte dans les mesures de SRIRs, un filtre à réponse impulsionnelle finie de 128 échantillons a été appliqué aux mesures. La correction fréquentielle a été appliquée entre 60 Hz et 16 kHz. Le filtre a été calculé d'après une mesure de la réponse impulsionnelle de l'enceinte effectuée dans une chambre anéchoïque avec un microphone omnidirectionnel situé dans l'axe de l'enceinte. Enfin, les réponses impulsionnelles issues des capsules du microphone sphérique ont été converties en ambisonique à l'ordre 4. Les procédés employés pour la mesure, le débruitage et l'encodage des SRIRs sont décrits en détail dans l'annexe A et B.

8.3.2. Sources sonores

Deux sources sonores ont été utilisées dans le test :

— Une voix masculine récitant *Fantaisie*, un poème de Gérard de Nerval, enregistrée dans une chambre anéchoïque ;

— Un instrument de percussion - un cajon - enregistré dans une chambre anéchoïque ;

Dans les sections suivantes, ces sources sonores seront désignées respectivement par *Voix* et *Cajon*. Les durées des signaux des sources sonores étaient respectivement de 9 s et 11 s. Chaque source sonore a été convoluée avec les SRIRs mesurées pour obtenir cinq scènes ambisoniques d'ordre 4 (une par espace sonore) pour créer les stimuli de référence.

8.3.3. Création des stimuli sonores

À la manière de la segmentation temporelle adoptée au chapitre 5, les cinq SRIRs ont été segmentées en trois régions temporelles (composantes initiale, précoce et tardive), d'après la méthode introduite par Götz *et al.* [17]. Le temps qui sépare temporellement les réflexions précoces de la réverbération tardive - nommé temps de mélange - figure dans le tableau 8.2 pour les cinq espaces utilisés.

Espace	*Amphi*	*Église*	*Hall*	*Réfectoire*	*Toilettes*
Temps de mélange (ms)	177	180	153	84	81

Tableau 8.2 – Temps de mélange des espaces sonores utilisés dans l'expérience.

Les parties précoces des cinq SRIRs considérées ont été modifiées selon les quatre méthodes de traitement spatial exposées précédemment. Bien que la méthode d'élargissement ne permette pas la réduction de l'ASW, elle a été employée à titre de comparaison pour évaluer son efficacité en termes d'accroissement de l'ASW. La méthode de décorrélation des composantes sectorielles a été utilisée pour garantir une répartition uniforme de l'énergie des premières réflexions dans le plan horizontal. Pour les stimuli résultants, le champ sonore formé par les premières réflexions ne possède pas de direction dominante.

Les paramétrages de ces traitements étaient les mêmes pour toutes les SRIRs et ont été réalisées de manière a réduire ou accroître les valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC} sans introduire d'artefact. Les coefficients de modification spatiale respectifs β, α, μ et γ ont été fixés de la manière suivante :

- Amplification directionnelle. Pour chaque SRIR, les valeurs des trois paramètres acoustiques étudiés sont représentées sur la figure 8.7 pour différentes valeurs du coefficient β. Deux valeurs du coefficient β ont été choisies pour paramétrer la méthode afin de générer les stimuli sonores. La valeur maximale correspondait à une valeur entière de β à partir de laquelle les paramètres acoustiques J_{LF} et J_{LFC} n'augmentaient pas de plus de 0.05, soit une JND (*Just Noticeable Difference*), en comparaison aux paramètres acoustiques obtenus avec la valeur entière supérieure de β. Ceci est vérifié pour toutes les SRIRs, lorsque $\beta = 5$. Par souci de symétrie, la valeur $\beta = 1/5$ a également été utilisée pour réduire les valeurs des paramètres acoustiques.

- Déformation angulaire. La figure 8.8 représente l'évolution des trois paramètres acoustiques étudiés pour chaque SRIR en fonction du coefficient de déformation angulaire α. De la même manière, deux valeurs de α ont été choisies pour paramétrer la

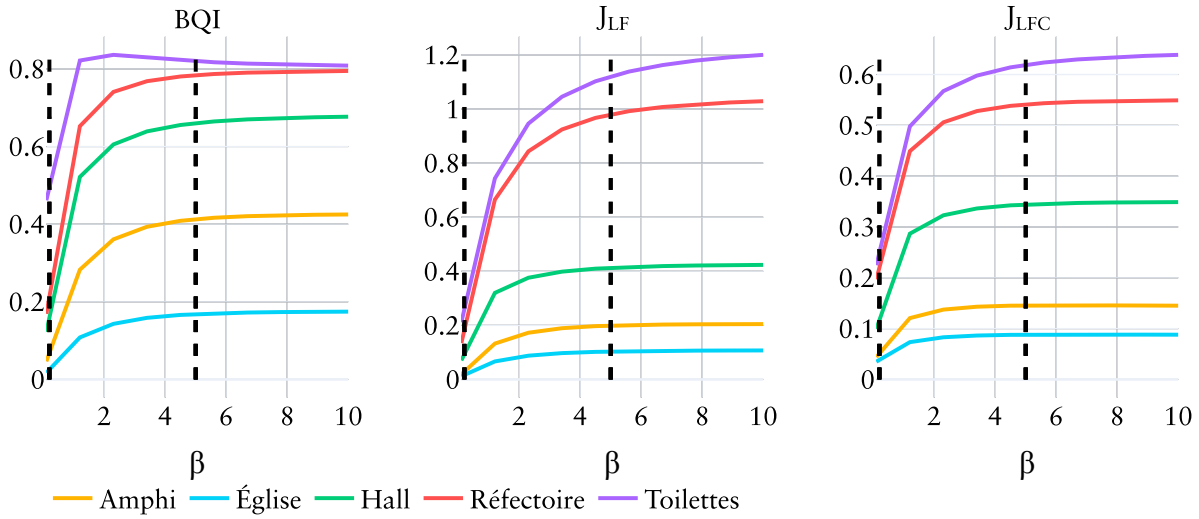


FIGURE 8.7 – Valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC} en fonction du paramètre d'amplification directionnel β pour les cinq SRIRs. Les seuils en pointillés indiquent les valeurs de β choisies pour générer les stimuli.

méthode afin de générer les stimuli sonores. En utilisant le même critère que pour la méthode précédente, des valeurs de α égales à 5 et -5 ont été employées.

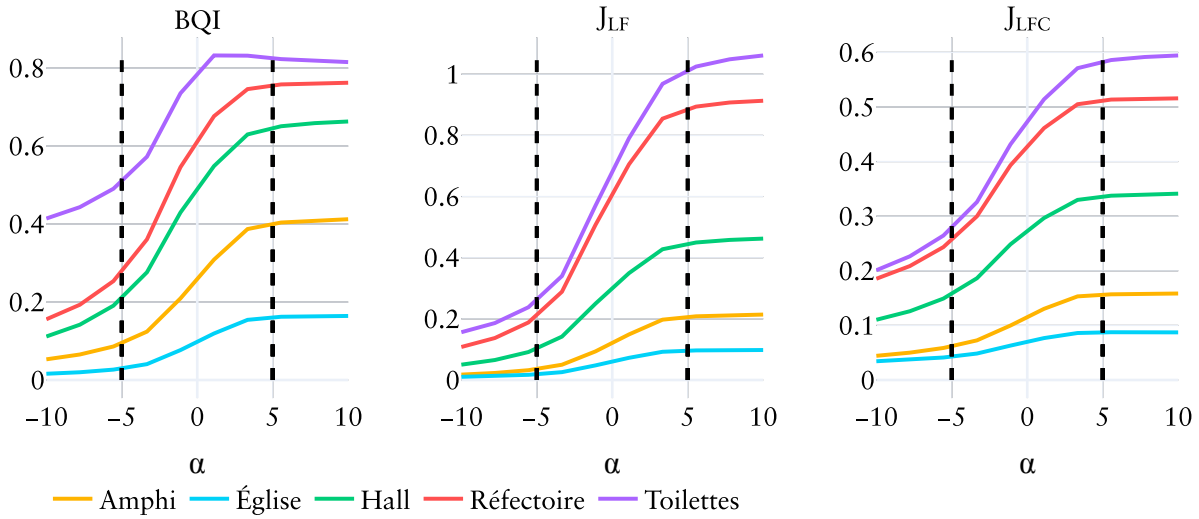


FIGURE 8.8 – Valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC} en fonction du paramètre de déformation angulaire α pour les cinq SRIRs. Les seuils en pointillés indiquent les valeurs de α choisies pour générer les stimuli.

- Élargissement. Les valeurs des trois paramètres acoustiques sont représentées sur la figure 8.9 en fonction du coefficient d'élargissement μ pour les cinq SRIRs. En dehors de ce coefficient, le paramétrage de la méthode fut celui de l'évaluation expérimentale réalisée par Zotter et Kronlachner dans leur étude [11]. La période de modulation τ a notamment été fixée à 2.5 ms. Les auteurs rapportent que pour des angles de modulation supérieurs à $\hat{\phi} = 80°$, la méthode génère un court effet

de réverbération perceptible. Ne souhaitant pas qu'un tel phénomène se produise, la valeur choisie du coefficient μ a été fixée à 0.2 soit $\hat{\phi} = 72°$.

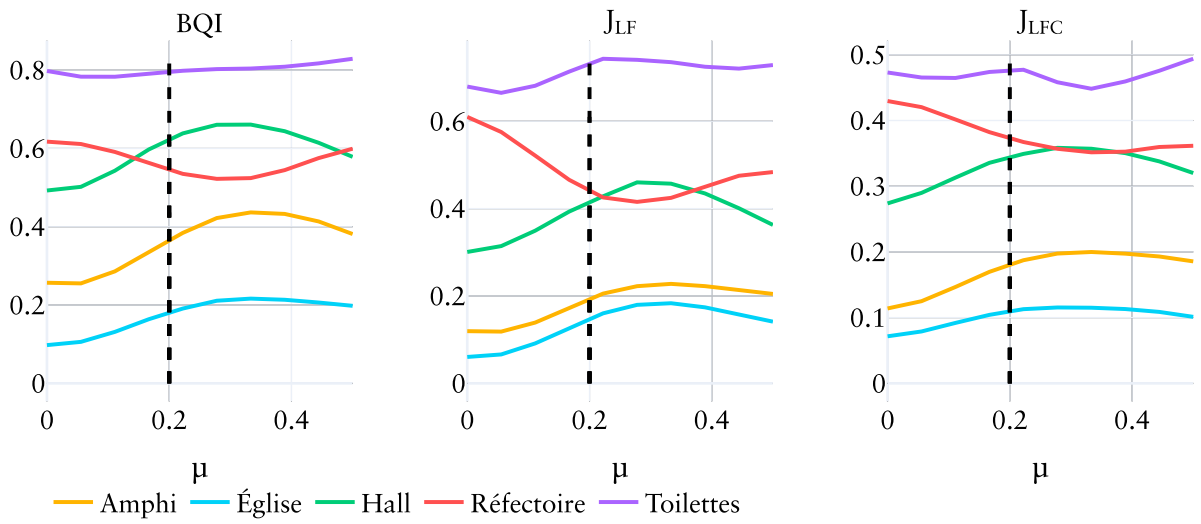


FIGURE 8.9 – Valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC} en fonction du coefficient d'élargissement μ pour les cinq SRIRs. Le seuil en pointillés indique la valeur de μ choisie pour générer les stimuli.

- Décorrélation des composantes sectorielles. Pour chaque SRIR, les valeurs des paramètres acoustiques sont représentées sur la figure 8.10 pour différentes valeurs du coefficient γ. Le coefficient $\gamma = 1$ produit une décorrélation maximale des composantes sectorielles dans le plan. Cette valeur a été choisie pour maximiser la variation des paramètres acoustiques avec cette méthode.

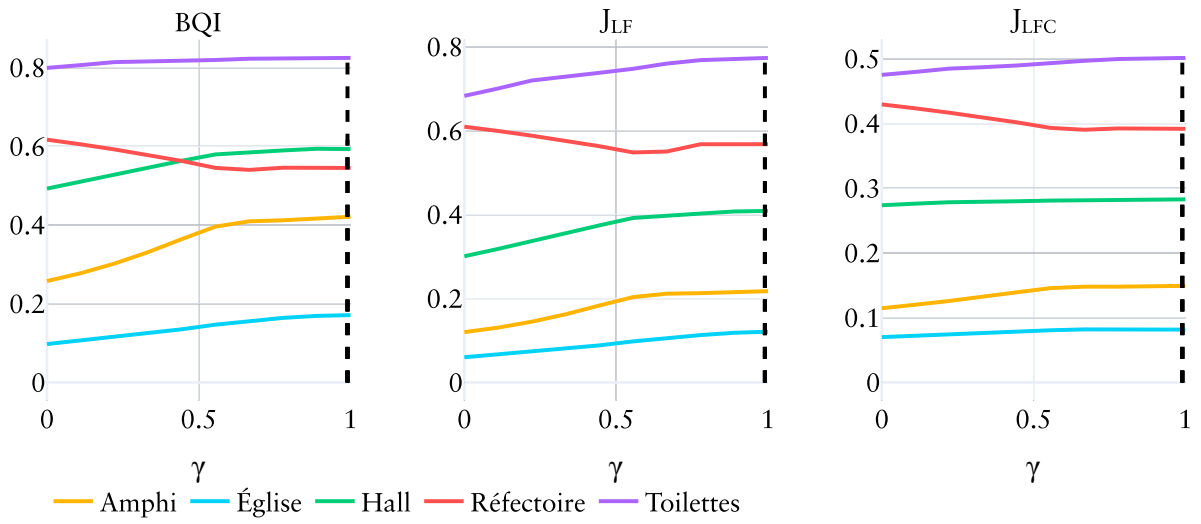


FIGURE 8.10 – Valeurs des paramètres acoustiques BQI, J_{LF} et J_{LFC} en fonction du coefficient de décorrélation γ pour les cinq SRIRs. Le seuil en pointillés indique la valeur de γ choisie pour générer les stimuli.

Pour toutes les méthodes, l'évolution des trois paramètres acoustiques est similaire. L'augmentation de l'énergie latérale augmente également la décorrélation car

elle exacerbe les différences dans l'axe interaural (à moins d'ajouter des ondes sonores en phase dans des directions opposées, ce qui ne correspond pas à une situation réelle). Inversement, avec les méthodes employées l'augmentation de la décorrélation se traduit par un «lissage» de la répartition de l'énergie sonore dans l'espace ce qui pour les SRIRs considérées participe à augmenter l'énergie latérale. Le tableau 8.3 répertorie les méthodes utilisées pour générer les stimuli sonores et les coefficients associés. Dans la suite du document, les méthodes d'amplification directionnelle seront désignées AD0 et AD1, les méthodes de déformation angulaire seront désignées DA0 et DA1, la méthode d'élargissement ELA et la méthode de décorrélation des composantes sectorielles DCS.

Abréviation	Méthode	Fonction	Coefficient
AD0	Amplification directionnelle	Réduction de l'énergie latérale	$\beta = 0.2$
AD1	Amplification directionnelle	Augmentation de l'énergie latérale	$\beta = 5$
DA0	Déformation angulaire	Réduction de l'énergie latérale	$\alpha = -5$
DA1	Déformation angulaire	Augmentation de l'énergie latérale	$\alpha = 5$
ELA	Élargissement	Augmentation de la décorrélation interaurale	$\mu = 0.2$
DCS	Décorrélation des composantes sectorielles	Équirépartition de l'énergie dans le plan	$\gamma = 1$

Tableau 8.3 – Récapitulatif des méthodes et coefficients associés utilisés pour générer les stimuli sonores du test perceptif.

Notons que les transformations appliquées aux SRIRs ne modifient pas seulement les paramètres acoustiques BQI, J_{LF} et J_{LFC} mais également le C_{80}, D_{50}, EDT, T_s ou encore G. Néanmoins, pour ces paramètres acoustiques, les changements de valeurs sont tous contenus dans les JND associées sauf dans les cas suivants :

— lorsque la méthode AD0 est appliquée à la SRIR de l'espace *Hall*. Une variation de -1.41 dB, -9.91% et -1.93 dB est obtenue pour C_{80}, D_{50} et G respectivement.
— lorsque la méthode AD0 est appliquée à la SRIR de l'espace *Toilettes*, le paramètre acoustique G diminue de 1.21 dB.
— lorsque la méthode DCS est appliquée à la SRIR de l'espace *Hall*, on observe une variation de -1.16 dB du paramètre acoustique G.

Les variations du DRR maximales étaient toutes obtenues lorsque la méthode DCS était utilisée avec des excursions s'élevant à 1.54 dB, 2.51 dB, 1.26 dB, 1.35 dB et 0.45 dB pour *Amphi*, *Église*, *Hall*, *Réfectoire*, *Toilettes* respectivement.

Création des fichiers binauraux

Les parties précoces des SRIRs modifiées ont été encodées en ambisonique à l'ordre 1 afin d'évaluer l'efficacité des méthodes employées à cet ordre. Le son direct était encodé à l'ordre 4 et la réverbération tardive à l'ordre 1. Les résultats du chapitre 5 nous informent que la réduction de l'ordre d'encodage de la réverbération peut avoir

un influence significative sur la perception de l'acoustique selon l'espace. Néanmoins, cette influence étant faible, nous avons choisi d'adopter un moteur de rendu de la réverbération à l'ordre 1 qui présente l'avantage de réduire considérablement les calculs nécessaires à l'auralisation.

Pour chaque espace considéré, six stimuli sonores ont été générés selon les six traitements figurant dans le tableau 8.3. En plus de ces stimuli, une référence cachée était jugée par les sujets. Avec cinq espaces et deux sources sonores, le test perceptif comprenait donc un total de 70 stimuli.

Les signaux ambisoniques ont été convertis en signaux binauraux en utilisant des filtres de décodage binaural dérivés de mesures d'une tête artificielle Neumann KU 100 [18] d'après les méthodes de rendu binaural de Schörkhuber *et al.* [19] et Zaunschirm *et al.* [20] dont il est question en annexe C.2. Comme les différentes régions temporelles des SRIRs étaient encodées avec des ordres différents (ordre 4 pour le son direct et ordre 1 pour les parties précoces et tardives), un décodeur binaural a été utilisé pour chaque région temporelle selon l'ordre correspondant. Les signaux décodés ont ensuite été additionnés pour créer les signaux binauraux.

8.3.4. Procédure

Le méthode d'évaluation perceptive utilisée pour ce test perceptif était inspirée de la méthode MUSHRA [15]. Pour chaque source et chaque espace - soit 10 configurations - 7 stimuli étaient présentés aux sujets en plus de la référence : une référence cachée et les 6 stimuli issus des modifications spatiales. La référence était décodée en binaural d'après une représentation ambisonique d'ordre 4 (pour le son direct ainsi que pour la partie précoce et tardive). Les stimuli étaient attribués de manière aléatoire à 6 boutons de lecture étiquetés « Son i » avec $i \in [1, 7]$. Les sujets avaient la possibilité d'écouter les stimuli de façon répétée et de passer de l'un à l'autre à leur gré. Un slider horizontal permettait de régler la boucle de lecture sur une région temporelle particulière.

Pour chacune des 10 configurations, les participants devaient évaluer la différence de largeur apparente de source entre la référence et les stimuli en déplaçant un curseur sur une échelle continue de -100 à 100 points dont les extrémités étaient étiquetées «Moins large» (-100) et «Plus large» (100). La valeur nulle correspondait à aucune différence perçue et était étiquettée «Identique». Aucune étiquette intermédiaire ni de graduation n'a été placée sur l'échelle pour éviter les biais dus à une distribution non linéaire des étiquettes et pour prévenir toute distorsion dans la distribution des données (avec des étiquettes intermédiaires, des groupes de notes peuvent en effet apparaître autour des valeurs où se trouvent les étiquettes [21, 22]). Les différentes configurations ont été présentées dans un ordre aléatoire.

Compte tenu du contexte sanitaire, le test perceptif a été réalisé sur internet à l'aide de l'application javascript *webMUSHRA* [23]. La figure 8.11 montre l'interface graphique utilisée pour le test. Pour des raisons pratiques, les mouvements de tête des auditeurs n'ont pas été pris en compte dans le test. Avant de débuter le test, les participants se sont familiarisés avec l'interface utilisateur en effectuant l'évaluation des stimuli associés à la source *Voix* dans l'espace «Amphi». Le test consistait en une

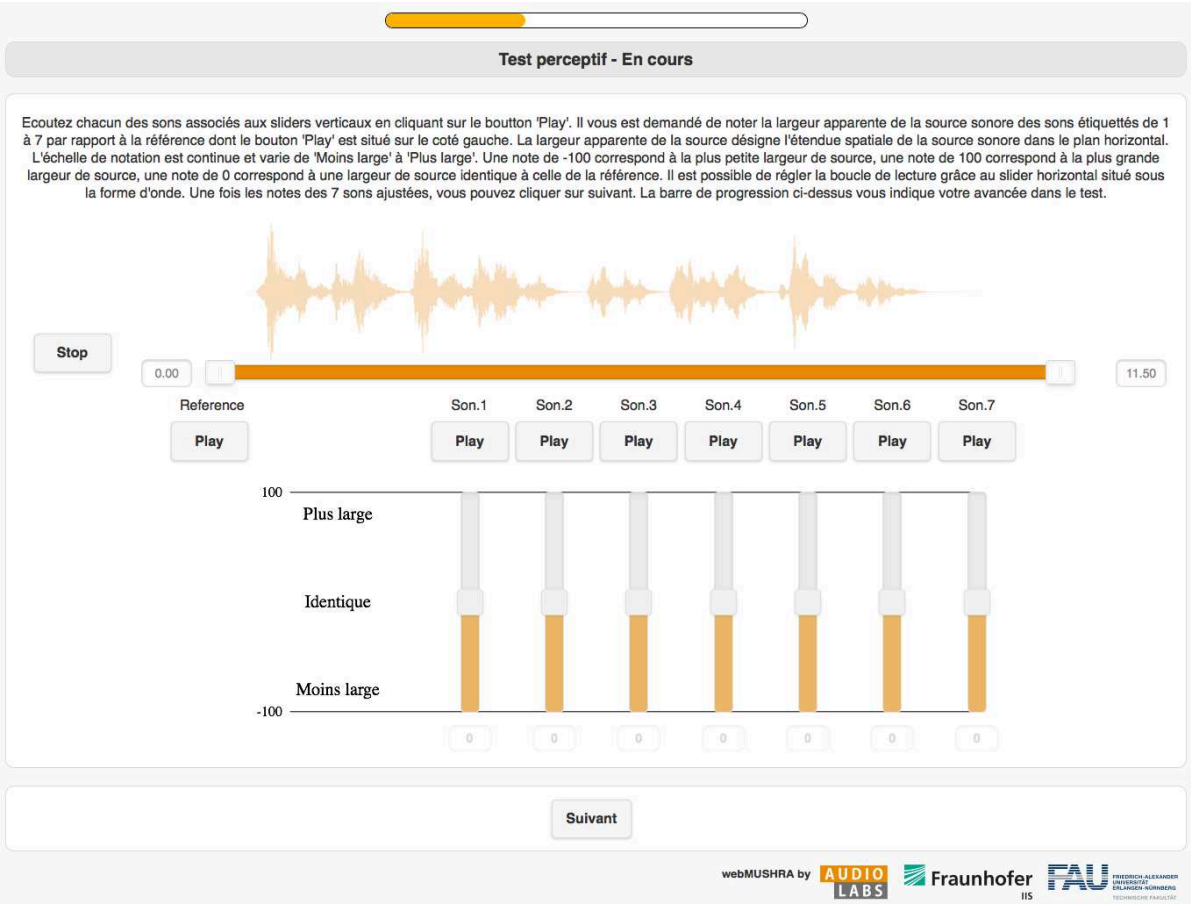


FIGURE 8.11 – Interface graphique du test perceptif.

seule session d'environ 40 minutes.

Réaliser un test sur internet réduit le contrôle expérimental du test perceptif. Des biais peuvent apparaître en raison de l'équipement utilisé et du volume sonore employé. En effet, ce n'est pas seulement l'appréciation des sujets qui varie mais également la restitution des stimuli en raison de la réponse fréquentielle des casques utilisés et de la variation de niveau. Afin de minimiser ces biais, il était demandé aux participants de réaliser le test en possession d'un casque professionnel et nous avons demandé de renseigner la marque du casque utilisé pour vérifier si les réponses fréquentielles étaient accidentées ou non. La coloration spectrale du casque peut notamment modifier l'externalisation. Néanmoins, cette influence peut être considérée comme négligeable en comparaison à la dégradation induite par l'usage d'HRTFs non-individualisées et de l'absence de dispositif de suivi des mouvements de tête. Par ailleurs, pour limiter les variations de niveau sonore, une étape de calibration décrite en section 6.3.5 a été réalisée par les sujets avant de procéder au test. D'autre biais sont liés au bruit environnant ou à l'implication des sujets. Le temps passé pour effectuer le test était mesuré pour juger de l'attention des sujets et il était demandé de réaliser le test dans un environnement calme.

8.3.5. Sujets

21 sujets âgés en moyenne de 24 ans ont pris part au test perceptif. En dehors de deux sujets - l'un ayant une formation musicale et l'autre étant ingénieur en traitement du signal audio - les participants étaient étudiants du Master Image & Son de l'Université de Brest ou de l'ENS Louis-Lumière. Ils étaient donc déjà formés à l'écoute critique en raison des enseignements reçus en prise de son, montage et mixage. Aucun d'entre eux n'a déclaré de perte auditive connue.

8.4. Résultats

Comme aucun point d'ancrage intermédiaire n'a été utilisé sur l'échelle de notation, les résultats ont été normalisés par rapport à la moyenne et à l'écart-type en utilisant la normalisation du score z [24].

8.4.1. Outliers

Les données ont été soumises à une ANOVA à mesures répétées avec les variables indépendantes suivantes : «source» (2) × «espace» (5) × «traitement» (7). L'analyse des observations appartenant à la même combinaison de niveaux de variables indépendantes a montré que les données récoltées contenaient des valeurs aberrantes (*outliers*) [25] : les résidus studentisés des résultats étaient supérieurs à 3 en valeur absolue dans certaines cellules pour 4 sujets. Parmi ces 4 sujets, deux ont réalisé le test en moins de 12 minutes (le temps moyen passé par les sujets était de 31 minutes), un participant avait une formation différente des sujets formés aux métiers du son et le dernier était également identifié comme *outlier* au chapitre 6. L'ANOVA étant sensible aux valeurs aberrantes nous avons fait le choix d'écarter ces sujets et d'effectuer l'analyse avec 17 sujets. Parmi les 17 sujets dont les réponses ont été utilisées pour l'analyse, 6 ont réajusté le volume sonore calculé à l'étape de calibration.

8.4.2. Analyse de la variance

Pour pouvoir légitimement réaliser une ANOVA à mesures répétées, l'hypothèse de normalité doit être remplie : les résidus des observations appartenant à la même combinaison de niveaux de variables indépendantes doivent être normalement distribués [26]. Un test de Shapiro-Wilk de normalité sur les résidus studentisés a été effectué sur chaque cellule avec un niveau de significativité de 5% et a rejeté l'hypothèse nulle d'une distribution normale pour 10 cellules sur 70. La plupart des études statistiques indiquent que l'ANOVA peut être robuste à ce genre de violation [27], surtout si la taille de l'échantillon est supérieure à 15 observations par cellule [28]. Avec 17 observations par cellule, nous avons considéré que l'utilisation de l'ANOVA pour l'analyse statistique était toujours légitime.

Le test de sphéricité de Mauchly a indiqué que l'hypothèse de sphéricité était remplie pour chaque variable indépendante ainsi que pour leurs interactions. Les résultats

Facteur	df	SS	MS	F	p
SO	1	8550	8550	5.233	0.036
SP	4	8679	1633	2.503	0.051
TM	6	265768	44295	22.659	<0.001
SO * SP	4	1465	366	0.383	0.820
SO * TM	6	31001	5166	6.120	<0.001
SP * TM	24	60141	2505	3.107	<0.001

Tableau 8.4 – Résultats de l'ANOVA. SO : source, SP : espace, TM : traitement, df : degré de liberté, SS : somme des carrés de type III, MS : moyenne des carrés, F : valeur f, p : valeur p.

sont présentés dans le tableau 8.4. L'analyse a révélé une influence significative de la variable indépendante «source» [F(1, 16) = 5.233 ; p = 0.036] et la variable indépendante «traitement» [F(6, 96) = 22.659 ; p = <0.001] ainsi que pour les interactions «source» × «traitement» [F(6, 96) = 6.120 ; p = <0.001] et «espace» × «traitement» [F(24, 384) = 3.107 ; p = <0.001]. La variable indépendante «espace» est à la limite du seuil de significativité mais l'étude de cet effet simple - comme celui de «source» - n'apportant pas d'information sur les traitements effectués seules leurs interactions avec «traitement» seront analysées.

8.4.3. Comparaisons des niveaux du facteur «traitement»

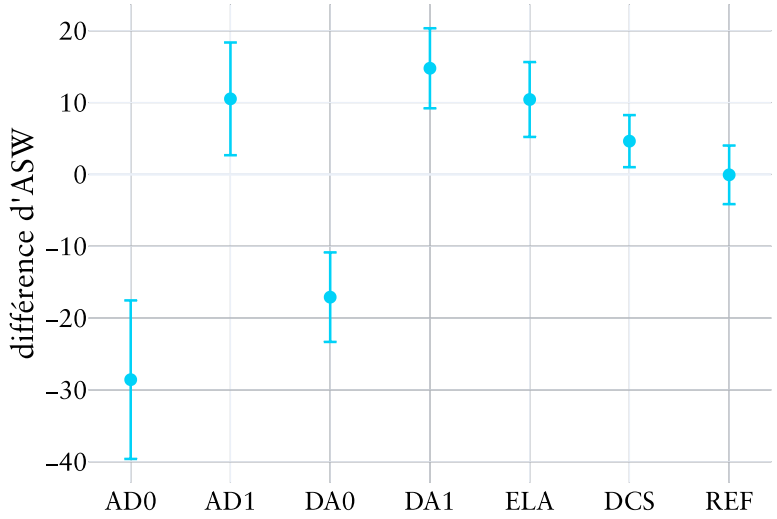


FIGURE 8.12 – Notes moyennes des différences de largeur apparente de source et intervalles de confiance à 95% associés pour les 7 traitements. REF désigne la référence cachée.

La figure 8.12 représente les notes moyennes des différences de largeur apparente de source et l'intervalle de confiance à 95% associé pour les différents niveaux de la variable indépendante «traitement». Des tests LSD de Fisher avec correction de Bonferroni ont été effectués pour déterminer les différences significatives entre les

Traitement	AD0	AD1	DA0	DA1	ELA	DCS	REF
AD0	-	0.005	0.263	<0.001	0.001	0.001	0.003
AD1		-	0.008	1.000	1.000	1.000	0.811
DA0			-	<0.001	0.001	0.001	0.006
DA1				-	1.000	0.071	0.016
ELA					-	1.000	0.069
DCS						-	0.579

Tableau 8.5 – Valeurs p des tests LSD de Fisher avec correction de Bonferroni. Les valeurs significatives sont reportées en gras. REF désigne la référence cachée.

différents traitements. Aucune différence significative n'a été obtenue entre les notes associées à la référence cachée (qui n'a subi aucun traitement) et aux méthodes AD1, ELA et DCS ($p = 0.811$, $p = 0.069$ et $p = 0.579$ respectivement). La méthode DA1 était la seule méthode visant à accroître l'ASW qui soit significativement différente de la référence. Les notes des méthodes DA0 et AD0 qui permettaient la réduction de l'ASW n'étaient pas significativement différentes entre elles ($p = 0.263$) mais significativement différentes de toutes les autres méthodes y compris de la référence.

Les traitements visant à réduire l'énergie latérale précoce pour réduire l'ASW semblent avoir eu une plus grande influence que les traitements visant à accroître l'ASW. En effet, les moyennes des notes associées aux méthodes AD0 et DA0 s'élèvent à -28.57 et -17.9 respectivement alors que la moyenne maximale des notes des autres méthodes s'élève à 14.77 pour DA1.

8

8.4.4. Résultats des interactions avec le facteur «traitement»

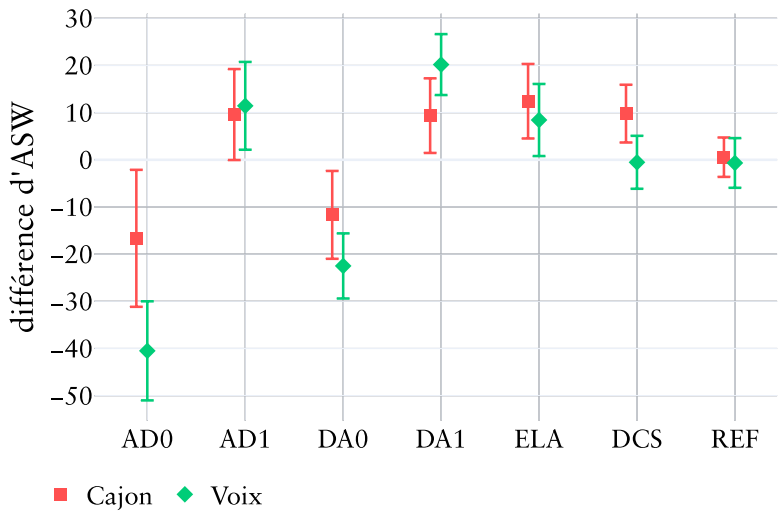


FIGURE 8.13 – Notes moyennes des différences de largeur apparente de source et intervalles de confiance à 95% associés pour les 7 traitements selon les deux sources sonores.

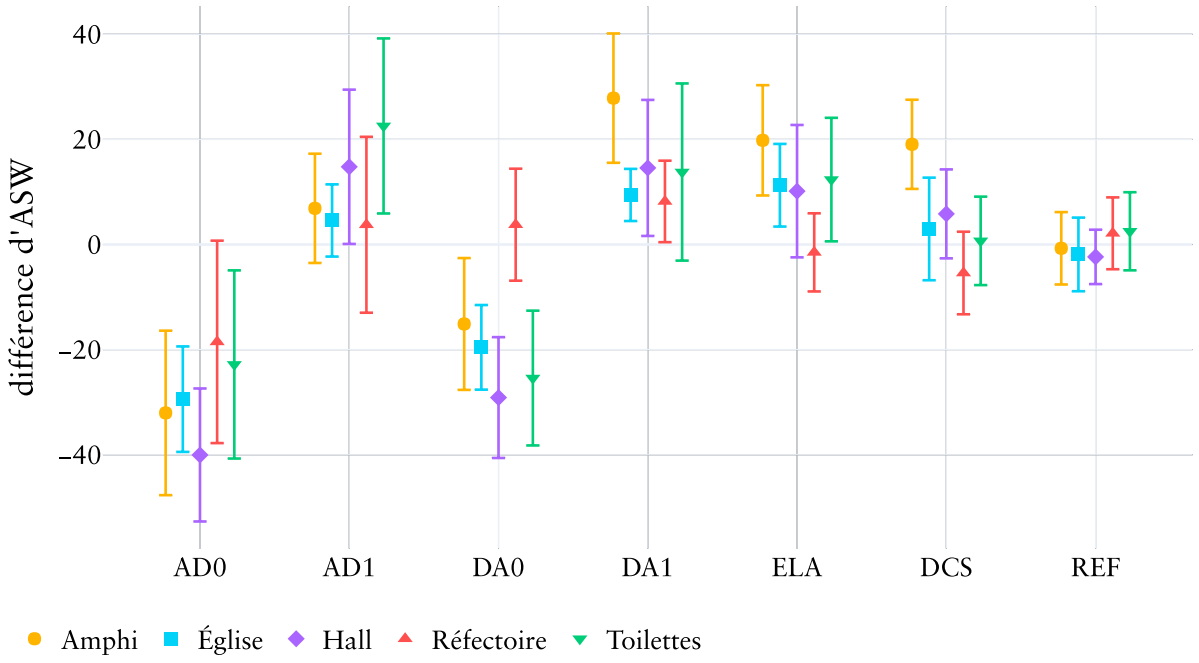


FIGURE 8.14 – Notes moyennes des différences de largeur apparente de source et intervalles de confiance à 95% associés pour les 7 traitements selon les 5 espaces sonores.

La figure 8.13 affiche les notes moyennes de dissemblance et l'intervalle de confiance à 95% pour chaque traitement selon la source sonore. Les différences perçues en termes d'ASW apparaissent plus importantes lorsque la source *Voix* était utilisée avec les méthodes visant à accroître l'énergie latérale.

La figure 8.14 affiche les notes moyennes de dissemblance et l'intervalle de confiance à 95% pour chaque traitement selon l'espace sonore. Les notes moyennes des différences perçues en termes d'ASW apparaissent plus faibles pour l'espace *Réfectoire*. Les intervalles de confiance associés sont assez larges comme pour l'espace *Toilettes*. Des test t de Student ont permis d'analyser plus en détails cette interaction.

Un test *t* de Student a été réalisé pour déterminer si la moyenne des notes attribuées à la référence cachée était significativement différente de 0 pour chaque condition «espace» (c'est-à-dire si elle a été perçue significativement différente de la référence). Pour tous les espaces, la moyenne des notes n'était pas significativement différente de 0 ($p > 0.271$) ; ce qui atteste d'une absence de différence significative avec la référence.

Des tests LSD de Fisher avec correction de Bonferroni ont été réalisés afin de déterminer si les notes moyennes des différents traitements étaient significativement différentes de la référence cachée pour chaque espace. Le tableau 8.6 rassemble les valeurs p obtenues. On constate que pour tous les espaces, le traitement AD1 n'a mené à aucun résultat significativement différent de la référence cachée. Les traitements DA1, ELA et DCS étaient significativement différents de la référence cachée pour l'espace *Amphi* seulement. Pour l'espace *Réfectoire*, aucun traitement n'a mené à un résultat significativement différent de la référence cachée.

Méthode	Amphi	Église	Hall	Réfectoire	Toilettes
AD0	0.035	0.001	0.001		
AD1					
DA0		0.031	0.025		0.040
DA1	0.022				
ELA	0.023				
DCS	0.014				

Tableau 8.6 – Valeurs p des tests LSD de Fisher avec correction de Bonferroni appliqués aux notes des traitements pour déterminer des différences significatives avec la référence cachée. Seules les valeurs significatives sont reportées.

8.5. Discussion

8.5.1. Le niveau de diffusion

Six des 21 participants ont modifié le niveau de diffusion calculé à l'étape de calibration avant de réaliser le test perceptif. Il était précisé que ce niveau n'était pas destiné à être modifié et qu'il ne devait être ajusté qu'en cas d'inconfort. Malheureusement, les modifications du niveau appliquées par les sujets n'ont pas été enregistrées. Nous ne pouvons donc pas rendre compte de l'écart entre le niveau calculé et le niveau jugé par les sujets comme étant confortable.

Plutôt que de permettre la modification du niveau de diffusion, une nouvelle étape de calibration ou une mise en relation avec l'expérimentateur à ce stade auraient été des options plus judicieuses. Le biais expérimental lié au niveau sonore de diffusion à sans doute été limité grâce à la méthode de calibration mais n'a pas pu être éliminé. Il vient s'ajouter aux biais liés aux équipements audio employés, au bruit environnant et à l'implication des sujets.

8.5.2. L'efficacité des traitements employés

Les résultats de l'ANOVA ont mis en lumière une influence significative du traitement spatial des premières réflexions sur l'ASW. Les variations des paramètres acoustiques sont en cohérence avec les tendances observées sur cet attribut perceptif : les notes moyennes des différences d'ASW par rapport à la référence sont positives lorsque les paramètres acoustiques ont été augmentés avec les méthodes AD1, DA1, ELA, DCS (> 4.62) et négatives lorsqu'ils ont été diminués avec les méthodes AD0 et DA0 (< -17.09). Ces résultats viennent confirmer l'influence de l'énergie latérale précoce et de la décorrélation sur l'ASW.

Les traitements visant à réduire l'ASW semblent avoir eu une plus grande influence que les traitements visant à l'accroître. Le paramétrage de la méthode d'amplification directionnelle et de la déformation angulaire présentés en section 8.3.3 peut expliquer ce phénomène. Leurs paramètres ont été fixés de manière à effectuer la réduction

et l'augmentation de l'énergie latérale dans les mêmes proportions : en choisissant des paramètres inverses pour AD0 et AD1 ou opposés pour DA0 et DA1. Pourtant, l'influence du paramétrage sur les valeurs résultantes des paramètres acoustiques est moindre avec les méthodes AD1 et DA1. Cette augmentation limitée des paramètres acoustiques se reflète dans les accroissements limités d'ASW.

8.5.3. Les variations d'énergie des premières réflexions

Des variations d'énergie des premières réflexions par rapport à la référence ont été constatées et sont recensées dans le tableau 8.7. Ces variations de niveau peuvent expliquer des différences observées entre les méthodes. On constate notamment que le traitement AD1 appliqué aux SRIRs de «Amphi» et «Église» a mené à des réductions de l'énergie des premières réflexions aux alentours de -1.5 dB. Il est possible que ceci explique la non significativité de cette méthode pour ces deux espaces. La normalisation de l'énergie des premières réflexions a été effectuée à l'ordre 4 avant de réduire l'ordre d'encodage ambisonique à l'ordre 1. Une normalisation à l'ordre 1 de l'énergie aurait permis de comparer simplement la spatialisation des réflexions au delà du changement de niveau induit.

	AD0	AD1	DA0	DA1	ELA	DCS
Amphi	0.78	-1.52	-0.79	-0.18	-0.25	-0.94
Église	0.62	-1.49	-1.02	-0.40	-0.22	-1.25
Hall	-0.62	-0.58	-0.20	-0.53	-0.03	-1.04
Réfectoire	-1.11	-0.09	-0.29	-0.66	-0.12	-1.20
Toilettes	-0.98	-0.48	-0.24	-0.48	-0.28	-0.74

Tableau 8.7 – Variations en décibel de l'énergie des premières réflexions par rapport à la référence pour les 6 traitements.

8.5.4. L'emploi d'un autre indicateur d'ASW

Plutôt que d'utiliser à la fois les paramètres acoustiques J_{LF} et J_{LFC} qui présentent une forte corrélation (le coefficient de Pearson s'élève à 0.97 pour l'ensemble des SRIRs), un autre indicateur que nous nommerons B_{LF} permet d'interpréter les résultats. Cet indicateur représente le rapport entre l'énergie d'une SRIR calculée dans les directions latérales et l'énergie de la SRIR de référence calculée dans le plan horizontal pour chaque espace. Pour calculer B_{LF} d'après des signaux ambisoniques d'ordre 1, quatre secteurs angulaires ont été considérés dans la direction frontale, arrière, gauche et droite. Pour chaque secteur angulaire, un faisceau formé dans la direction correspondante a permis d'estimer la quantité d'énergie provenant de cette direction. Les contributions des énergies des secteurs latéraux sont ensuite sommées pour former une estimation de l'énergie latérale. L'indicateur B_{LF} s'écrit :

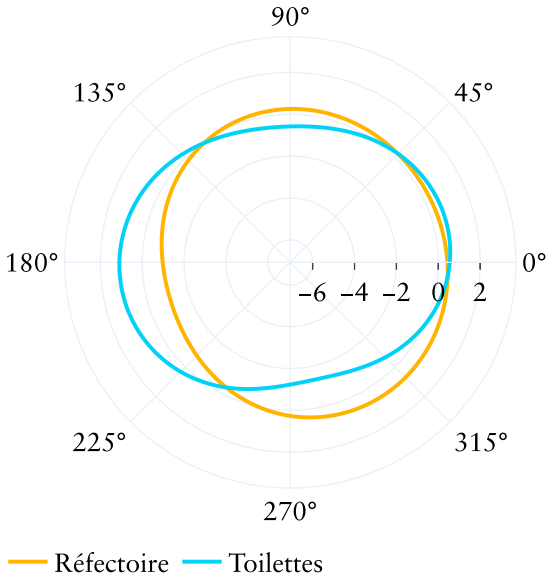


FIGURE 8.15 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions de la SRIR associée à *Réfectoire* et à *Toilettes*.

$$\mathrm{B_{LF}} = \frac{\sum_{n=0}^{K-1} e_{\mathrm{G}}[n] + e_{\mathrm{D}}[n]}{\sum_{n=0}^{K-1} e_{\mathrm{D}}^{\mathrm{ref}}[n] + e_{\mathrm{G}}^{\mathrm{ref}}[n] + e_{\mathrm{Av}}^{\mathrm{ref}}[n] + e_{\mathrm{Ar}}^{\mathrm{ref}}[n]} \tag{8.23}$$

où $e_{\mathrm{G}}^{\mathrm{ref}}$, $e_{\mathrm{D}}^{\mathrm{ref}}$, $e_{\mathrm{Av}}^{\mathrm{ref}}$, $e_{\mathrm{Ar}}^{\mathrm{ref}}$ représente l'énergie du secteur angulaire gauche, droite, avant et arrière de la SRIR de référence respectivement et e_{G}, e_{D} représente l'énergie du secteur angulaire gauche et droite de la SRIR étudiée. L'énergie dans un secteur angulaire s se calcule de la manière suivante :

$$\mathrm{e_s}[n] = (\boldsymbol{y}^{\top}(\theta_s, \varphi_s)\, \boldsymbol{b}[n])^2 \tag{8.24}$$

où $\boldsymbol{b}$ est la composante ambisonique de la SRIR à l'instant n et $\boldsymbol{y}$ est le vecteur des harmoniques sphériques calculées pour la direction (θ_s, φ_s).

Sommer les énergies des secteurs angulaires permet de s'affranchir d'éventuelles interférences destructives ou constructives qui peuvent apparaitre en sommant les amplitudes. Ces interférences sont jugées responsables de la forte variabilité des mesures $\mathrm{J_{LF}}$ et $\mathrm{J_{LFC}}$ dans une salle [29]. Comme pour ces deux paramètres acoustiques, les valeurs de $\mathrm{B_{LF}}$ résultent de la moyenne des mesures calculées dans les bandes de fréquence centrées sur 125, 250, 500 et 1000 Hz. Nous avons choisi de calculer ces valeurs sur la partie des SRIRs modifiée, c'est-à-dire sur l'ensemble des premières réflexions et non sur les 80 ms seulement et de ne pas inclure le son direct. Dans notre cas de figure, cette mesure s'avère plus pertinente que $\mathrm{J_{LF}}$. La figure 8.15 représente le diagramme de directivité dans le plan horizontal pour les espaces *Toilettes* et *Réfectoire*. Malgré une énergie latérale plus importante pour *Réfectoire*, le paramètre

acoustique J_{LF} est pourtant plus faible (0.684 contre 0.609). Au contraire, les valeurs de B_{LF} sont plus cohérentes car elles s'élèvent à 0.387 et 0.556 respectivement. Le tableau 8.8 répertorie les valeurs de B_{LF} pour toutes les SRIRs de référence.

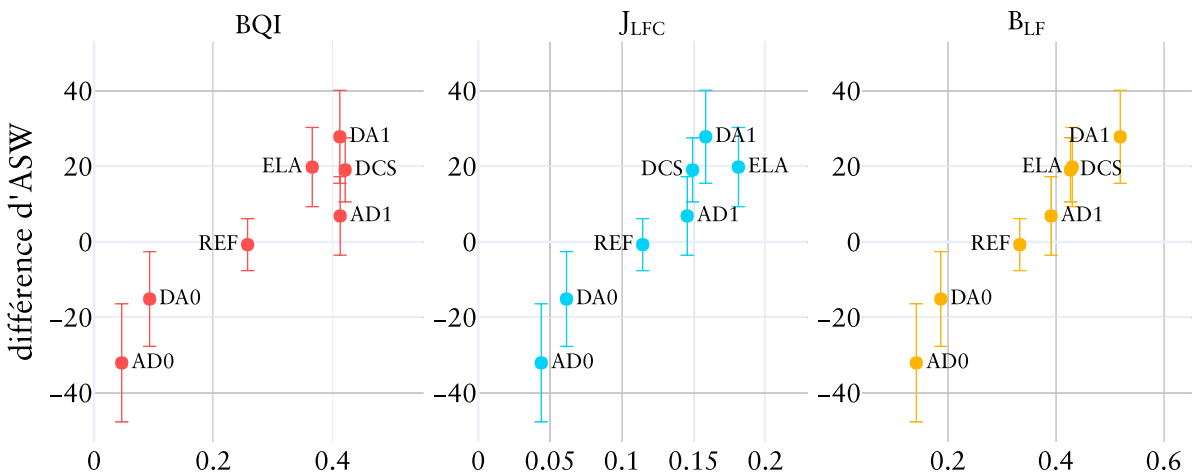
Amphi	Église	Hall	Réfectoire	Toilettes
0.333	0.291	0.406	0.556	0.387

Tableau 8.8 – Valeurs de l'indicateur B_{LF} pour les SRIRs de référence des espaces sonores utilisés dans l'expérience.

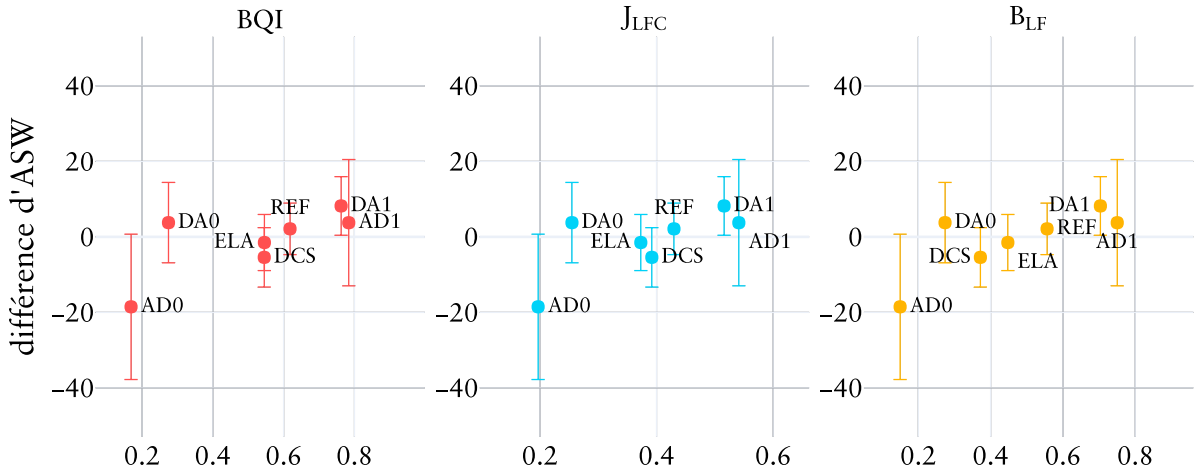
8.5.5. L'influence de la source

L'analyse statistique a révélé que la perception des traitements appliqués dépendait de la source sonore employée. Ces traitements semblent avoir eu une plus grande influence avec la source *Voix*. Plusieurs études rapportent que l'ASW est un attribut perceptif lié à la perception des médiums et basses fréquences [3, 30], c'est pourquoi les paramètres acoustiques associés J_{LF} et J_{LFC} sont calculés entre 125 Hz et 1000 Hz. Il est possible que la source *Voix* ait fournit plus d'indices dans cette région fréquentielle car l'énergie dans cette bande fréquence était plus importante de 2.72 dB par rapport à *Cajon*. De plus, les sujets sont susceptibles d'avoir une certaine connaissance à long terme des caractéristiques de sources sonores naturelles, et cette connaissance a priori peut influencer la façon dont ils jugent ces sons quotidiens, notamment pour juger de la distance de sources sonores [31]. Il est possible que ce soit également le cas pour juger de la largeur de source : une connaissance a priori de la voix a pu aider les sujets à juger sa largeur avec précision, ce qui n'était peut-être pas le cas avec la source *Cajon*.

8.5.6. L'influence de l'espace



(a) Amphi



(b) Réfectoire

FIGURE 8.16 – Notes moyennes des différences d'ASW en fonction des paramètres acoustiques BQI, J_{LF} et B_{LF} selon les 7 traitements spatiaux pour deux espaces sonores.

L'efficacité des traitements dépend fortement des espaces employés. La figure 8.16 représente les moyennes des notes d'ASW associées aux intervalles de confiance à 95% en fonction des paramètres acoustiques BQI, J_{LF} et B_{LF} pour *Amphi* et *Réfectoire*. Ces espaces représentent deux cas extrêmes : pour le premier, des méthodes de réduction et d'accroissement de l'ASW ont mené à des notes significativement différents de celles de la référence cachée et pour le second aucune méthode n'a mené à de différences significatives.

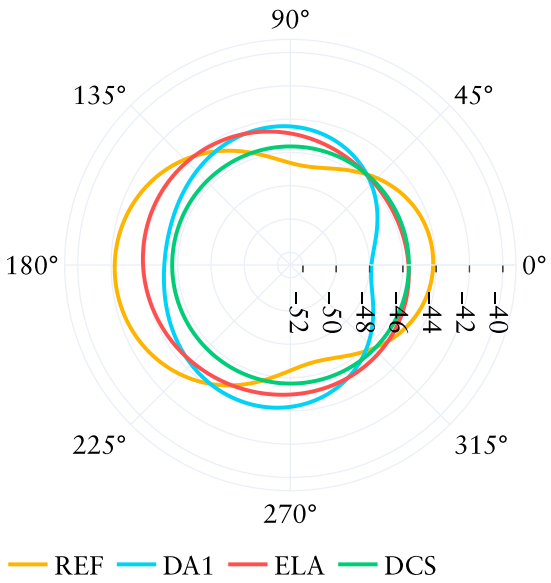


FIGURE 8.17 – Diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions de la SRIR de référence associée à *Amphi* et de celles issues des traitements ayant menés à des effets significatifs.

Pour l'espace *Amphi*, la méthode DA1 a mené à des notes d'ASW significativement

différentes de la référence cachée et la moyenne associée est la plus élevée de toutes les méthodes. Ce résultat est en cohérence avec l'indicateur B_{LF} qui est le plus élevé pour la méthode DA1. La figure 8.17 représente le diagramme de directivité dans le plan horizontal du champ sonore formé par les premières réflexions de la SRIR de référence et celles issues des traitements DA1, ELA et DCS. Bien que les variations d'énergie soient faibles entre les méthodes, le champ résultant de DA1 possède moins d'énergie frontale et plus d'énergie dans l'axe interaural que dans le cas des autres méthodes. Il est intéressant de constater que malgré une réduction de l'énergie moyenne dans le plan horizontal de 0.93 dB avec la méthode DCS, le stimulus sonore résultant à été perçu comme significativement plus large que la référence. Ce résultat peut être dû à la réduction de l'énergie frontale et arrière au profit des directions latérales. Ainsi la décroissance de l'énergie des premières réflexions ne mène pas nécessairement à une réduction de l'ASW si l'énergie latérale reste importante. On observe également que les traitements ont mené à des augmentations de l'énergie latérale dans des régions angulaires différentes comprises entre 45° et 130° sans que ces différences ne mènent à des ASW significativement différentes. La zone définie entre 40° et 130° identifiée par Johnson *et al.* [6] pour laquelle les réflexions participent de la même manière à la largeur apparente de source semble valide.

L'espace *Amphi* est le seul pour lequel les méthodes visant l'accroissement de l'ASW ont mené a des notes moyennes significativement différentes de celle de la référence cachée. Cet espace possède la valeur de D_{50} la plus grande et le temps central le plus court. Le D_{50}, nommé définition, correspond au rapport entre l'énergie contenue dans les 50 premières millisecondes et l'énergie totale de la réponse impulsionnelle de salle. Le temps central T_s représente le moment où l'énergie contenue dans la réponse impulsionnelle en amont est égale à l'énergie contenue en aval. La réponse impulsionnelle de cette salle contient donc une grande proportion d'énergie dans les premières millisecondes, c'est-à-dire dans une région temporelle où les réflexions précoces sont susceptibles de fusionner perceptivement avec le son direct. Il semble que la présence de réflexions ayant une énergie importante et étant proches du son direct soit une condition nécessaire pour être en mesure d'accroître efficacement l'ASW.

La figure 8.16b correspondant aux résultats de l'espace *Réfectoire* ne suggère pas de relation particulière entre les paramètres acoustiques et l'ASW. Contrairement aux autre espaces, une réduction des valeurs de paramètres acoustiques ne s'est pas traduit par une diminution significative de l'ASW. La proportion d'énergie précoce était faible pour cet espace par rapport aux autres ($C_{80} = 8.57$ dB) et le rapport champ direct sur champ réverbéré était positif (DRR = 2.09 dB). L'espace *Église* présente également ces deux caractéristiques, néanmoins la modification de la spatialisation des premières réflexions a été effectuée sur 180 ms pour *Église* et 84 ms pour *Réfectoire*. De plus, la force sonore des premières réflexions associée à *Refectoire* était la plus faible de tous les espaces ($G_E = 28.86$ dB) . Il est donc possible que le son direct ait masqué la modification des premières réflexions dont l'énergie était faible et le support temporel court.

Application à des premières réflexions paramétrées

L'amplification directionnelle et la déformation angulaire, qui se sont avérées efficaces pour certains espaces, peuvent être appliquées efficacement à l'ordre 1. Les

signaux modifiés ont été obtenus en appliquant une matrice $\boldsymbol{T}$ aux signaux ambisoniques de référence. Dans le cadre de la paramétrisation des premières réflexions adoptée au chapitre 6, il est possible d'appliquer de tels traitements directement sur les paramètres : les matrices de covariance $\boldsymbol{C}$. En effet, la matrice de covariance des signaux traités $\tilde{\boldsymbol{C}}$ s'écrit : $\tilde{\boldsymbol{C}} = \boldsymbol{T}\boldsymbol{C}\boldsymbol{T}^H$.

8.6. Conclusion

Dans ce chapitre, les parties précoces de SRIRs ont été manipulées dans le domaine ambisonique pour étudier la perception de la largeur apparente de source. L'étude s'est placée dans le contexte d'un rendu binaural non-individualisé de contenus ambisoniques d'ordre 4 où l'effet de réverbération est reproduit à l'ordre 1. Quatre méthodes de manipulation du champ sonore ont été employées de manière à faire varier l'énergie latérale et la décorrélation interaurale de cinq SRIRs de référence.

Un test perceptif a montré que la variation de l'énergie latérale et de la décorrélation interaurale peut avoir une influence significative sur l'ASW. Cependant, l'effet de ces traitements dépendait fortement de l'espace considéré. D'une part, une diminution de l'ASW a pu être observée pour tous les espaces considérés sauf pour l'un d'entre eux. Pour cet espace, il semble que les premières réflexions n'étaient pas suffisamment importantes en comparaison du son direct pour que l'effet de leur modification puisse être significatif. D'autre part, une augmentation significative de l'ASW n'a pu être observée que pour un seul espace. Les premières réflexions de la SRIR associée qui étaient proches du son direct possédaient une énergie importante. Cette observation est en accord avec l'hypothèse selon laquelle les réflexions précoces fusionnant avec le son direct contribuent à augmenter l'ASW. Dès lors, la région temporelle fixée à 80 ms et utilisée pour calculer les paramètres acoustiques censés caractériser l'ASW (J_{LF}, J_{LFC} et BQI) ne semble pas judicieuse. Une région temporelle plus courte devrait être définie pour être en mesure de mieux prédire l'ASW.

Les résultats de cette étude mettent en évidence les limites de l'accroissement de la proportion d'énergie latérale pour augmenter l'ASW. Dans cette optique, d'autres travaux pourraient étudier l'efficacité de l'augmentation de l'énergie précoce. Néanmoins, des tests informels suggèrent que d'autres approches basées sur le traitement spatial du son direct semblent plus efficaces que la modification des premières réflexions. La réduction de l'énergie latérale des premières réflexions reste cependant une option efficace pour réduire l'ASW.

Par ailleurs, Un autre indicateur que J_{LF} et J_{LFC} s'est montré plus à même d'expliquer les résultats : B_{LF}. Cet indicateur quantifie la proportion d'énergie latérale des premières réflexions par rapport à leur énergie dans le plan horizontal. Son calcul repose sur des formations de faisceaux pour quantifier l'énergie par secteur angulaire et ainsi s'affranchir d'éventuelles interférences entre les réflexions qui peuvent se produire lorsque les contributions de plusieurs secteurs sont mélangées.

Bibliographie

[1] M. Barron, "The subjective effects of first reflections in concert halls—the need for lateral reflections," *Journal of sound and vibration*, vol. 15, n°. 4, p. 475–494, 1971.

[2] M. Barron et A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls : the derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, n°. 2, p. 211–232, 1981.

[3] J. Bradley, R. Reich et S. Norcross, "On the combined effects of early-and late-arriving sound on spatial impression in concert halls," *The Journal of the Acoustical Society of America*, vol. 108, n°. 2, p. 651–661, 2000.

[4] M. Morimoto, K. Nakagawa et K. Iida, "The relation between spatial impression and the law of the first wavefront," *Applied Acoustics*, vol. 69, n°. 2, p. 132–140, 2008.

[5] L. Beranek, *Concert halls and opera houses : music, acoustics, and architecture*. Springer Science & Business Media, 2012.

[6] D. Johnson et H. Lee, "Just noticeable difference in apparent source width depending on the direction of a single reflection," dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[7] R. Y. Litovsky, H. S. Colburn, W. A. Yost et S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, n°. 4, p. 1633–1654, 1999.

[8] ISO 3382-1, "Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces," International Organization for Standardization, Geneva, CH, Standard, 2009.

[9] G. Potard et I. Burnett, "Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays," dans *Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.

[10] T. Pihlajamäki, O. Santala et V. Pulkki, "Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals," *Journal of the Audio Engineering Society*, vol. 62, n°. 7/8, p. 467–484, 2014.

[11] F. Zotter, M. Frank, M. Kronlachner et J.-W. Choi, "Efficient phantom source widening and diffuseness in ambisonics," *10.14279/depositonce-4103*, 2014.

[12] T. Carpentier, "Ambisonic spatial blur," dans *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[13] T. Schmele et U. Sayin, "Controlling the apparent sourcesize in ambisonics using decorrelation filters," dans *Audio Engineering Society Conference : 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society, 2018.

[14] H. Pomberger et F. Zotter, "Warping of 3d ambisonic recordings," dans *Proc. of the 3rd Int. Symp. on Ambisonics & Spherical Acoustics*, 2011.

[15] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems." International Telecommunication Union, Standard, 2015.

[16] D. Cabrera, D. Lee, M. Yadav et W. L. Martens, “Decay envelope manipulation of room impulse responses : Techniques for auralization and sonification,” dans *Proceedings of Acoustics*, 2011.

[17] P. Götz, K. Kowalczyk, A. Silzle et E. A. Habets, “Mixing time prediction using spherical microphone arrays,” *The Journal of the Acoustical Society of America*, vol. 137, n°. 2, p. EL206–EL212, 2015.

[18] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

[19] C. Schörkhuber, M. Zaunschirm et R. Höldrich, “Binaural rendering of ambisonic signals via magnitude least squares,” dans *Proceedings of the DAGA*, vol. 44, 2018, p. 339–342.

[20] M. Zaunschirm, C. Schörkhuber et R. Höldrich, “Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint,” *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3616–3627, 2018.

[21] S. Zielinski, F. Rumsey et S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, n°. 6, p. 427–451, 2008.

[22] E. C. Poulton et S. Poulton, *Bias in quantifying judgements*. Taylor & Francis, 1989.

[23] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler et J. Herre, “webMUSHRA — a comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, vol. 6, n°. 1, 2018. [En ligne]. Disponible : https://github.com/audiolabs/webMUSHRA

[24] E. Kreyszig, “Advanced engineering mathematics, 10th eddition,” 2009.

[25] R. D. Cook et S. Weisberg, *Residuals and influence in regression*. New York : Chapman and Hall, 1982.

[26] H. Keselman, J. C. Rogan, J. L. Mendoza et L. J. Breen, “Testing the validity conditions of repeated measures f tests.” *Psychological Bulletin*, vol. 87, n°. 3, p. 479, 1980.

[27] K. Weinfurt, “Repeated measures analysis : Anova, manova, and hlm,” *Reading and Understanding More Multivariate Statistics*, 10 2000.

[28] S. B. Green et N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[29] D. de Vries, E. M. Hulsebos et J. Baan, “Spatial fluctuations in measures for spaciousness,” *The journal of the Acoustical Society of America*, vol. 110, n°. 2, p. 947–954, 2001.

[30] T. Okano, L. L. Beranek et T. Hidaka, “Relations among interaural cross-correlation coefficient (iacc e), lateral fraction (lf e), and apparent source width (asw) in concert halls,” *The Journal of the Acoustical Society of America*, vol. 104, n°. 1, p. 255–265, 1998.

[31] P. Zahorik, D. S. Brungart et A. W. Bronkhorst, “Auditory distance perception in humans : A summary of past and present research,” *ACTA Acustica united with Acustica*, vol. 91, n°. 3, p. 409–420, 2005.

9

Étude du contrôle anisotrope de la réverbération tardive

Dans le chapitre 7, les origines physiques de la sensation d'enveloppement ont été étudiées en établissant des liens entre les résultats de l'évaluation de cet attribut perceptif et des paramètres acoustiques extraits de réponses impulsionnelles. Les données à disposition ont permis de confirmer l'influence de la quantité d'énergie tardive, des réflexions latérales tardives et de la décorrélation interaurale sur la sensation d'enveloppement. Cette analyse nous conforte dans l'idée que les paramètres acoustiques associés sont bien représentatifs de cet attribut perceptif et qu'il est nécessaire de reproduire convenablement les propriétés spatiales qu'ils caractérisent pour reproduire la sensation d'enveloppement. Dans ce chapitre, un réverbérateur artificiel basé sur un réseau récursif de lignes à retard (FDN) est utilisé pour générer la partie tardive de SRIRs encodées en ambisonique à l'ordre 1. L'architecture employée permet la reproduction des caractéristiques spectrales, spatiales et temporelle d'une SRIR tout en fournissant un contrôle du temps de réverbération par bande de fréquence et par secteur angulaire. La capacité du réverbérateur artificiel à reproduire les paramètres acoustiques identifiés comme pertinents au chapitre 7 pour caractériser la sensation d'enveloppement est étudiée. Les erreurs commises sur ces paramètres acoustiques étant dans certains cas importantes, des pistes d'amélioration du réverbérateur artificiel sont proposées.

Parmi les méthodes de traitement du signal permettant d'appliquer un effet de réverbération à une source sonore, deux grandes approches sont couramment utilisées : la convolution avec une réponse impulsionnelle de salle et l'utilisation d'un réseau récursif de lignes à retard ou FDN (pour *Feedback Delay Network*).

Auraliser un espace sonore en convoluant une source sonore avec une réponse impulsionnelle de salle permet d'obtenir un rendu similaire à l'enregistrement de la source dans l'espace auralisé. Des techniques de convolution basées sur la décomposition de la réponse impulsionnelle en blocs de petites tailles permettent d'implémenter

le filtrage efficacement en termes de calcul et avec une latence faible [1, 2]. Néanmoins, le coût de calcul étant dépendant de la longueur de la réponse impulsionnelle, il peut être très élevé lorsque le temps de réverbération est important. Cette approche demande d'autant plus de calculs lorsque l'on souhaite traiter des réponses impulsionnelles spatiales de salles pour lesquelles le filtrage est nécessairement multicanal.

Le FDN permet de reproduire les propriétés statistiques d'une réverbération tardive avec une faible complexité de calcul tout en produisant un effet de réverbération jugé naturel [3, 4]. Il permet également une modification aisée du temps de décroissance et de la réponse en fréquence de la réverbération tardive [5]. Néanmoins, il peut produire des artefacts audibles dans la partie précoce de la réverbération générée (cas le plus fréquent) mais également dans la partie tardive si ses paramètres ne sont pas convenablement sélectionnés. En effet, dans cette partie la densité d'écho est généralement trop faible en raison de la nature récursive du traitement qui implique un temps nécessaire à la génération d'un nombre d'échos suffisants. De plus, les transitoires francs produits par les échos successifs produisent des colorations spectrales et artefacts temporels peu naturels.

Carpentier *et al.* [6] ont donc proposé une implémentation hybride d'un effet de réverbération permettant un traitement temps-réel d'un contenu sonore. Cette implémentation hybride est à la fois composée d'un étage de convolution pour générer la partie précoce de la réverbération et d'un FDN pour générer la partie tardive. De cette manière, le coût de calcul induit par la convolution est limité au traitement des premières réflexions dont la zone temporelle s'étend généralement jusqu'à 50 à 200 ms. Parallèlement, la réverbération tardive - qui peut s'étendre sur plusieurs secondes - est générée par le FDN avec une qualité sonore comparable à celle de la réponse impulsionnelle de salle de référence. La méthode proposée estime d'abord le temps de réverbération dans différentes bandes de fréquences à partir de la réponse impulsionnelle afin de paramétrer le FDN. Un filtre est ensuite appliqué pour assurer une transition entre le rendu issu de la convolution et celui issu du FDN. Cette approche apparait comme un compromis avantageux entre coût de calcul et qualité de reproduction.

L'approche hybride pourrait être employée pour reproduire fidèlement les caractéristiques spatiales d'une SRIR. À cette fin, l'architecture du FDN employé pour générer la partie tardive de la réverbération doit permettre, pour plusieurs raisons, une reproduction anisotrope de la partie tardive : une distribution non-uniforme de l'énergie moyenne dans l'espace. D'une part, il est commun que des espaces sonores présentent des propriétés de réverbération anisotropes dans des situations réelles du fait des irrégularités de formes qui constituent un espace, des propriétés d'absorption des matériaux, ou de l'emplacement de l'auditeur. Ces caractéristiques sont perceptibles même pour une faible variation d'énergie moyenne dans l'espace. Romblom et Guastavino [7] ont montré qu'une réverbération tardive anisotrope peut être perçue comme significativement différente de son équivalent isotrope pour des variations d'énergies inférieures à 3 dB. D'autre part, l'étude réalisée au chapitre 7 atteste de l'importance de la proportion d'énergie latérale pour reproduire une impression d'enveloppement. Le contrôle perceptif des impressions spatiales doit donc permettre une modification de la répartition d'énergie tardive dans l'espace.

Plusieurs architectures de FDN ont été proposées pour contrôler la pondération

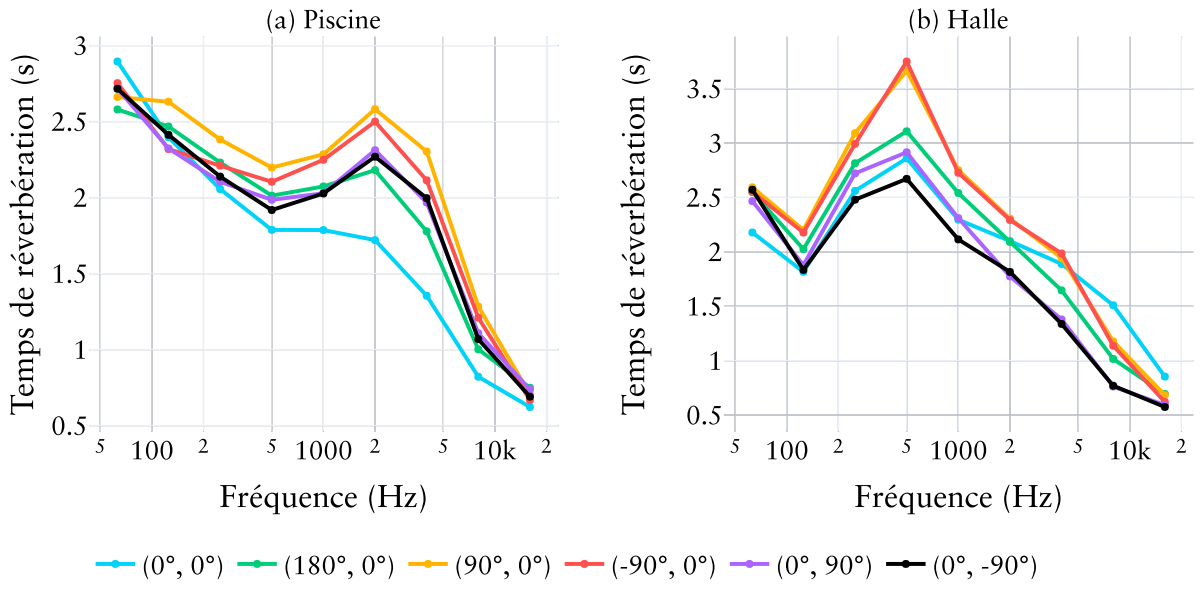


FIGURE 9.1 – Exemples de temps de réverbération calculés par bande de fréquence dans 6 directions de l'espace (θ, ϕ). Les temps de réverbération ont été calculés d'après les signaux issus de formations de faisceau appliqués à des SRIRs ambisoniques d'ordre 4.

de l'énergie dans différentes directions d'incidences [8, 9] en appliquant des gains directionnels sur les signaux de sortie de FDNs avant encodage ambisonique. Néanmoins, avec ces méthodes les caractéristiques spatiales de la partie tardive d'une SRIR ne sont pas entièrement respectées car la décroissance de l'énergie reste la même dans toutes les directions : elle reste isotrope. La figure 9.1 illustre des variations du temps de réverbération dans plusieurs bandes de fréquence et dans différentes directions pour deux réponses impulsionnelles spatiales de salle. On constate aisément que les variations du temps de réverbération dans une salle peuvent être importantes.

Récemment, Alary et Politis [10] ont proposé une architecture de FDN - nommé FDN directionnel - permettant de contrôler par bande de fréquence la décroissance de l'énergie sonore dans l'espace. Dans cette architecture, les lignes à retard sont multicanales et chaque canal correspond au signal issu d'une direction d'incidence particulière. Les directions d'incidence sont les mêmes d'une ligne à l'autre et sont uniformément réparties sur la sphère. Cette méthode permet de satisfaire une décroissance d'énergie sonore donnée avec une résolution angulaire dépendante du nombre de directions considérées. Les directions sont choisies de manière à représenter la SRIR dans le domaine des harmoniques sphériques à un ordre ambisonique donné.

Nous verrons dans ce chapitre comment reproduire avec un FDN des temps de réverbération cibles par bande de fréquence et les procédés permettant de minimiser des artefacts spectraux et temporels courants. Puis, en employant une architecture similaire à celle proposée par Alary et Politis [10], nous étudierons la capacité d'un FDN directionnel d'ordre 1 à reproduire les paramètres acoustiques liés à l'impression d'enveloppement.

9.1. Le réseau récursif de lignes à retard

Le réseau récursif de ligne à retard ou FDN est un système simple et peu coûteux en calculs qui est utilisé pour générer un effet de réverbération. Ce système, parfois nommé réverbérateur artificiel, consiste à mélanger des signaux retardés entre eux de manière récursive. La figure 9.2 illustre l'architecture de ce type de traitement dans le cas d'un signal de sortie monodimensionnel. Un FDN est constitué de N lignes à retard ou la ième ligne est retardée de m_i échantillons. Les signaux retardés $\boldsymbol{s}_i$, $0 < i \leq N$, sont mélangés par la matrice $\boldsymbol{A}$ avant d'être réinjectés dans les lignes à retard. Dans le domaine temporel, ce traitement est régi par le système d'équations suivant :

$$\begin{cases} y[n] = \boldsymbol{c}^\top \boldsymbol{s}[n] + dx[n] \\ \boldsymbol{s}[n + \boldsymbol{m}] = \boldsymbol{A}\boldsymbol{s}[n] + \boldsymbol{b}x[n] * \boldsymbol{h} \end{cases} \tag{9.1}$$

où x et y sont les signaux d'entrée et de sortie respectivement, d est un gain scalaire, $\boldsymbol{c}$ et $\boldsymbol{b}$ sont des vecteurs de gains, $\boldsymbol{h}$ est un vecteur de filtres et $\boldsymbol{m}$ est le vecteur des retards m_i.

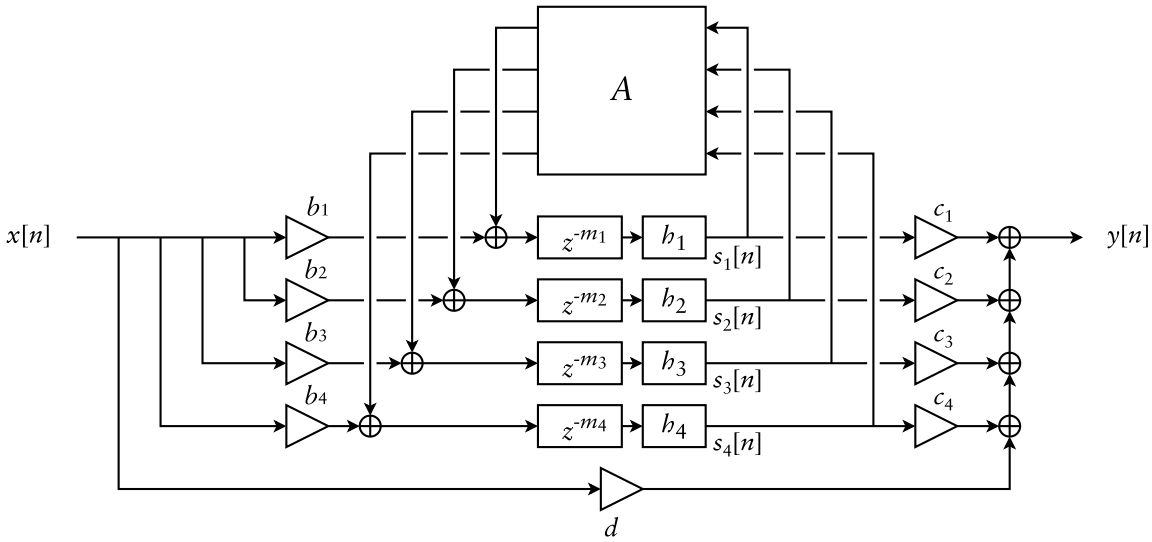


FIGURE 9.2 – Architecture d'un réseau récursif de 4 lignes à retard

L'un des principaux objectifs de la conception d'un réverbérateur artificiel est d'obtenir un système dont la réponse impulsionnelle possède un temps de réverbération dépendant de la fréquence. La convolution par le filtre $\boldsymbol{h}_i$ vise à satisfaire un motif de décroissance par bande de fréquence en atténuant l'énergie dans la boucle de rétroaction. Pour maitriser cette atténuation, il est nécessaire qu'en dehors de ce filtrage le système soit sans perte quels que soient les retards employés. Jot et Chaigne ont montré que les matrices de mélange unitaires permettent de satisfaire cette condition [5]. Ces matrices vérifient $\boldsymbol{A}\boldsymbol{A}^H = \boldsymbol{I}_N$ où $\boldsymbol{I}_N$ est la matrice identité de taille N et $()^H$ l'opérateur adjoint. De cette manière l'énergie à l'entrée de la matrice de mélange $\boldsymbol{A}$ est

préservée en sortie de cette matrice et la seule source d'atténuation du signal y est le filtrage réalisé dans chaque boucle de rétroaction.

9.1.1. Le calcul des filtres d'atténuation

Réponse en fréquence théorique

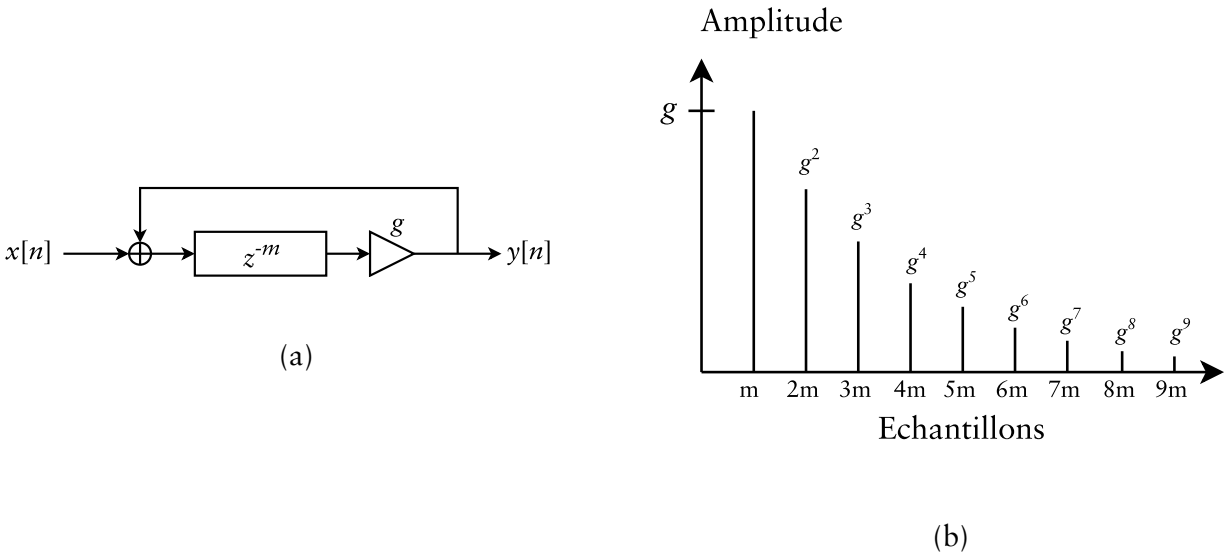


FIGURE 9.3 – Architecture (a) et réponse impulsionelle (b) d'un filtre en peigne de retard m qui peut être considéré comme un cas particulier de FDN à une ligne à retard.

Considérons dans un premier temps, une ligne à retard sur laquelle un gain d'atténuation indépendant de la fréquence est appliqué. La figure 9.3 représente un filtre en peigne de retard m, que l'on peut considérer comme un cas particulier de FDN à une ligne de retard. La réponse impulsionnelle r d'un tel système s'écrit :

$$r[n] = \begin{cases} g^k & \text{si } n = km \text{ où } k \text{ est un entier positif} \\ 0 & \text{sinon.} \end{cases} \tag{9.2}$$

où k est un entier strictement positif. Pour que la réponse impulsionnelle du système possède un temps de réverbération T_{60}, l'expression r_{dB} de r en décibel doit vérifier l'équation suivante :

$$r_{\mathrm{dB}}[f_s \mathrm{T}_{60}] = -60 \tag{9.3}$$

où f_s est la fréquence d'échantillonnage. Le gain en décibel g_{dB} à appliquer dans la boucle de rétroaction est donc :

$$g_{\mathrm{dB}} = \frac{-60m}{f_s \mathrm{T}_{60}} \tag{9.4}$$

Pour satisfaire un temps de réverbération dépendant de la fréquence, plutôt que d'utiliser un gain scalaire, un filtre h est inséré dans la boucle de rétroaction. Il découle

de l'équation précédente que la magnitude h_{dB} de ce filtre s'écrit en fonction de la fréquence de pulsation ω :

$$h_{\mathrm{dB}}(\omega) = \frac{-60m}{f_s \mathrm{T}_{60}(\omega)} \tag{9.5}$$

Implémentation du filtre

Afin de réaliser le filtrage pour chaque ligne d'un FDN, Prawda *et al.* [11] proposent d'utiliser un banc de filtres permettant d'appliquer une atténuation dans plusieurs bandes de fréquence. Ce filtres est constitué de Q filtres à réponse impulsionnelle infinie d'ordre 2 ou filtres biquadratiques. La réponse en fréquence de ces filtres est dite «en cloche» (*peak-notch filters*). La transformée en z du filtre h s'exprime comme le produit des transformées en z des filtres biquadratiques :

$$h(z) = \prod_{q=1}^{Q} \frac{b_{q,0} + b_{q,1}z^{-1} + b_{q,2}z^{-2}}{1 + a_{q,1}z^{-1} + a_{q,2}z^{-2}} \tag{9.6}$$

où $b_{q,i}$ et $a_{q,i}$ ($i = \{1, 2\}$) sont les coefficients du filtre biquadratique centré sur la qième bande de fréquence. La figure 9.4 représente la réponse en fréquence de 9 filtres en cloches repartis en bande d'octaves. Cet ensemble de filtres sera considéré dans la suite de ce chapitre.

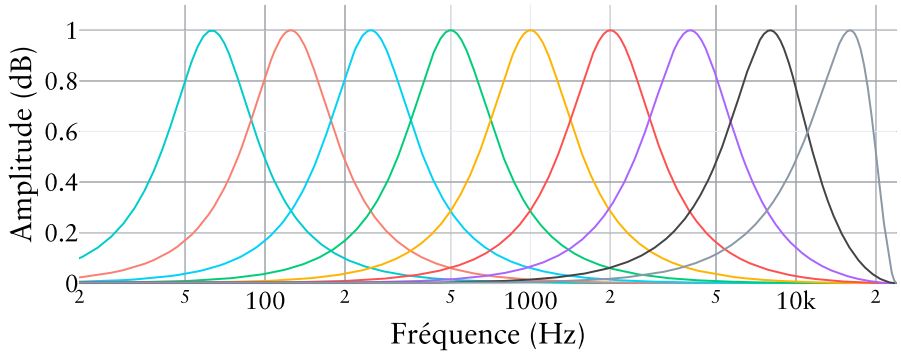


FIGURE 9.4 – Réponse en fréquence normalisée à 1 dB des filtres biquadratiques constituant le filtre h.

Pour reproduire la réponse en fréquence théorique $\boldsymbol{h}_{\mathrm{dB}}$ de l'équation (9.5) au moyen de ce banc de filtre, il est nécessaire de pondérer la réponse de chaque filtre biquadratique par un gain g_q. Cette pondération doit être réalisée en prenant en compte le recouvrement spectral visible sur la figure 9.4 : chaque filtre contribue au delà de la bande d'octave sur laquelle il est centré. Soit $\boldsymbol{B}$ la matrice de dimension $N \times Q$ contenant les réponses en fréquence de taille N des Q filtres biquadratiques et $\boldsymbol{g}$ le vecteur des gains g_q appliqués aux réponses en fréquence des filtres. La réponse en fréquence du banc de filtre, notée $\boldsymbol{h}_{\mathrm{IIR,\,dB}}$, s'écrit alors :

$$\boldsymbol{h}_{\mathrm{IIR,\,dB}} = \boldsymbol{B}\boldsymbol{g} \tag{9.7}$$

Il est nécessaire de déterminer le vecteur $\boldsymbol{g}$ permettant que la réponse $\boldsymbol{h}_{\mathrm{IIR,\,dB}}$ du banc de filtre s'approche autant que possible de la réponse en fréquence théorique de

taille N désignée par le vecteur $\boldsymbol{h}_{\text{dB}}$. Pour ce faire, il est possible de minimiser l'erreur quadratique entre les deux réponses. Il parait cependant plus judicieux de minimiser l'erreur entre les inverses des réponses en fréquence. En effet, l'amplitude de la réponse en fréquence théorique est inversement proportionnelle au temps de réverbération. La figure 9.5 représente le temps de réverbération en fonction de l'atténuation en décibel avec deux intervalles d'erreur. Selon l'erreur commise sur l'atténuation, le temps de réverbération correspondant peut varier grandement. Par exemple, pour une erreur de $\pm$ 0.05 dB commise sur un gain de -0.2 dB, le temps de réverbération peut varier entre 5 et 8.1 s.

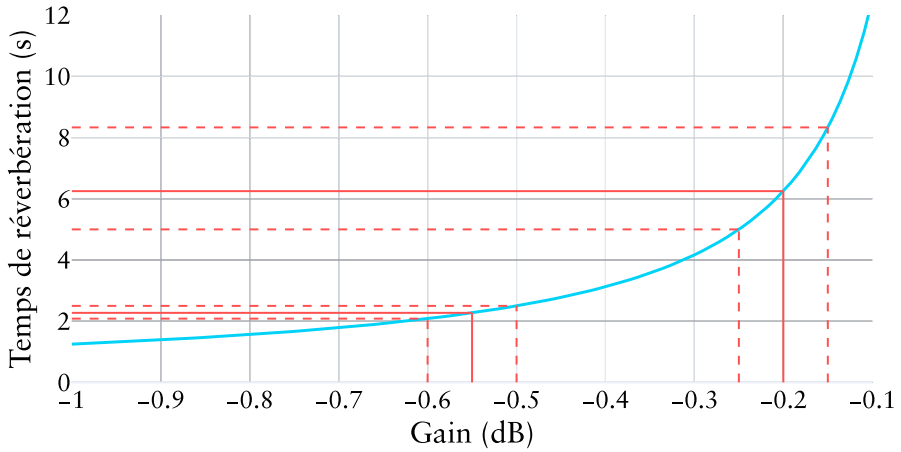


FIGURE 9.5 – Temps de réverbération obtenu avec un filtre en peigne de retard m = 48 ms et un gain d'atténuation scalaire. Les marges d'erreur sur le temps de réverbération sont affichées en pointillés pour une erreur de $\pm$ 0.05 dB sur l'atténuation.

Schlecht et Habets proposent donc de minimiser une erreur e proportionnelle au temps de réverbération [12]. Dans ce cas, l'erreur quadratique à minimiser peut s'écrire [11] :

$$e = \| \frac{1}{\boldsymbol{h}_{\text{IIR, dB}}} - \frac{1}{\boldsymbol{h}_{\text{dB}}} \|_2^2. \tag{9.8}$$

On peut également choisir de minimiser l'erreur $\tilde{e}$ suivante :

$$\tilde{e} = \| \mathbb{1}_{N\times 1} - \frac{\boldsymbol{h}_{\text{IIR, dB}}}{\boldsymbol{h}_{\text{dB}}} \|_2^2 \tag{9.9}$$

soit

$$\tilde{e} = \| \mathbb{1}_{N\times 1} - \boldsymbol{W}\boldsymbol{B}\boldsymbol{g} \|_2^2 \tag{9.10}$$

avec

$$\boldsymbol{W} = \text{diag}\left(\frac{1}{\boldsymbol{h}_{\text{dB}}}\right) \tag{9.11}$$

et $\mathbb{1}_{N\times 1}$ un vecteur de 1 de dimension $N \times 1$. Le vecteur de gains $\boldsymbol{g}$ permettant de minimiser cette erreur est alors donnée par l'équation suivante :

$$\boldsymbol{g} = (\boldsymbol{W}\boldsymbol{B})^{\dagger} \mathbb{1}_{N\times 1} \tag{9.12}$$

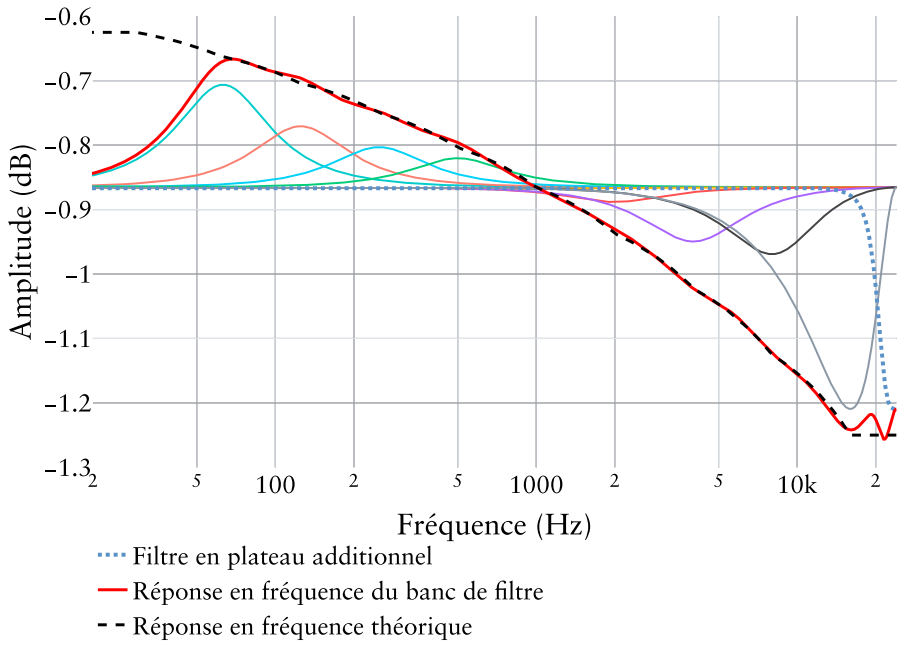


FIGURE 9.6 – Réponse en fréquence du filtre d'atténuation $\boldsymbol{h}_{\mathrm{IIR,\,dB}}$ (en rouge) résultant des contributions des Q filtres biquadratiques pondérés par le vecteur de gains $\boldsymbol{g}$. La réponse en fréquence théorique $\boldsymbol{h}_{\mathrm{dB}}$, calculée pour un retard m = 48 ms, est matérialisée par les pointillés noirs.

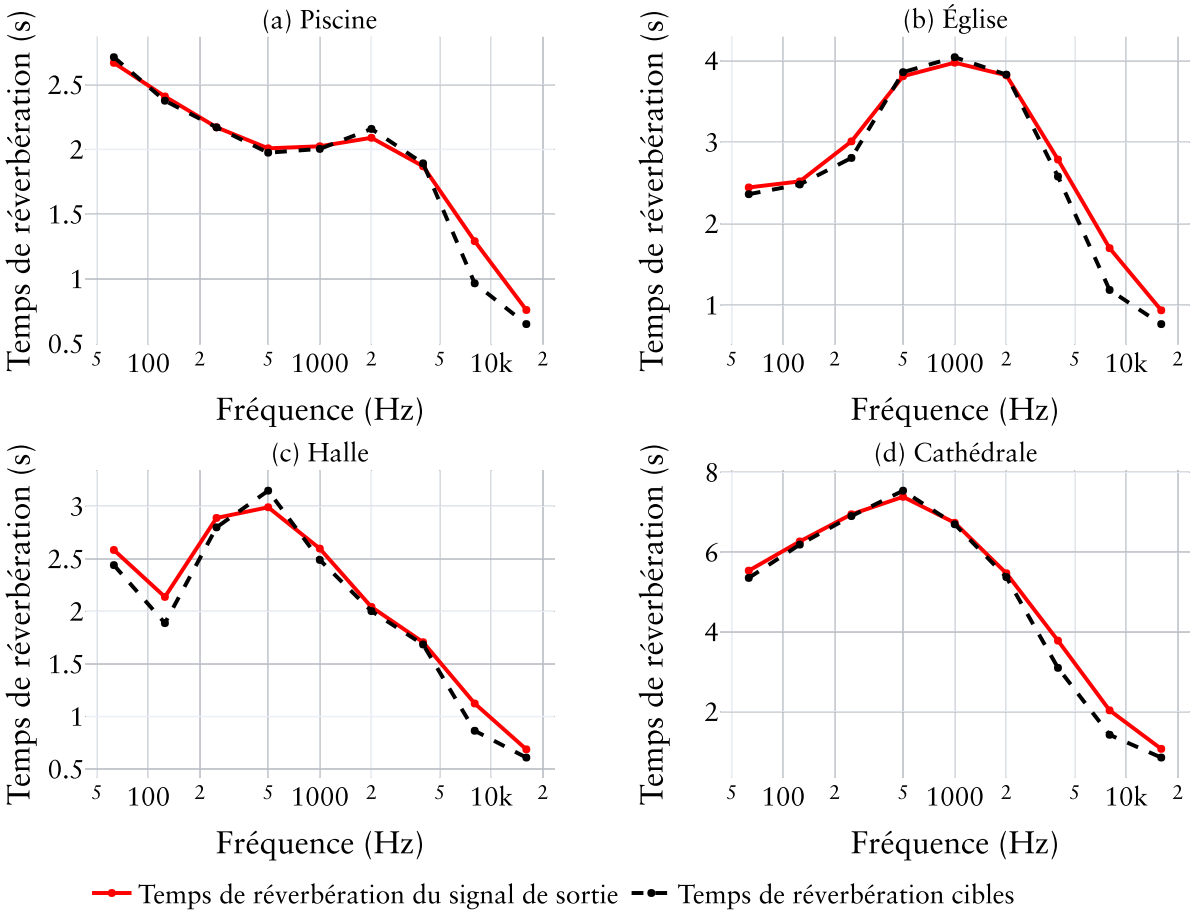


FIGURE 9.7 – Temps de réverbération de réponses impulsionnelles spatiales de salle et temps de réverbération des réponses impulsionnelles de sortie d'un FDN de 8 lignes constitué d'un filtre d'atténuation sur chaque ligne à retard.

où $(\boldsymbol{WB})^\dagger$ est la matrice pseudo inverse de $\boldsymbol{WB}$. Les coefficients des filtres biquadratiques sont calculés d'après une fréquence centrale et un facteur de qualité. Or, lorsque l'on applique un gain à la réponse en fréquence des filtres, le facteur de qualité correspondant est modifié. Un processus itératif permet d'ajuster la valeur des gains à appliquer afin de respecter les facteurs de qualité des filtres [13].

En utilisant les 9 filtres de la figure 9.4, les gains sont définis pour chaque octave dont la fréquence centrale est comprise entre 63 Hz et 16 kHz. Cependant, en dehors de cette région fréquentielle, la réponse en fréquence du banc de filtre peut s'approcher d'une atténuation nulle et ainsi produire des temps de réverbération très élevés. Pour éviter ce cas de figure, Prawda *et al.* [11] proposent d'appliquer une atténuation sur l'ensemble du spectre de la réponse en fréquence du banc de filtre. De cette manière, une atténuation non nulle est réalisée en dehors de la zone spectrale couverte par les filtres biquadratiques.

La figure 9.6 représente la somme pondérée des réponses en fréquence des filtres biquadratiques d'après le vecteur de gains $\boldsymbol{g}$. Un filtre en plateau (*shelf filter*) est appliqué en haute fréquence pour éviter une augmentation du temps de réverbération au dessus de 16 kHz [11]. Le gain de ce filtre correspond à celui du dernier filtre biquadratique du banc de filtres.

La figure 9.7 représente les temps de réverbération de réponses impulsionnelles spatiales de salle d'après lesquelles des filtres d'atténuation ont été calculés. On observe également, les temps de réverbération des réponses impulsionnelles issues de FDN de 8 lignes intégrant les filtres d'atténuation correspondants. Le profil des temps de réverbération est respecté. Les erreurs maximales obtenues s'élèvent cependant à 33.7%, 43.2%, 30.0% et 43.0% - toutes dans la bande de fréquence centrée sur 8 kHz - pour les espaces *Piscine*, *Église*, *Halle* et *Cathédrale* respectivement.

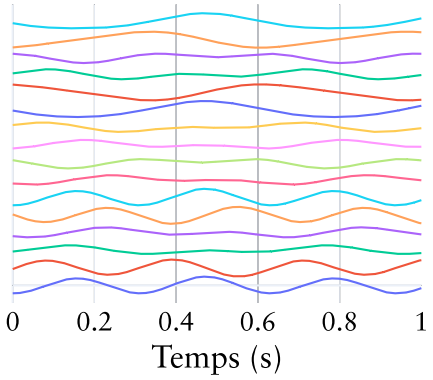
9.1.2. La modulation de la matrice de mélange

Les valeurs des retards m_i et des coefficients de la matrice de mélange du système récursif étant fixes, des motifs périodiques peuvent apparaitre en sortie de FDN et se traduire par une coloration notable du signal de sortie. Afin d'éviter de tels artefacts, Schlecht et Habets ont proposé de faire varier les coefficients de la matrice de mélange au cours du temps [14, 15].

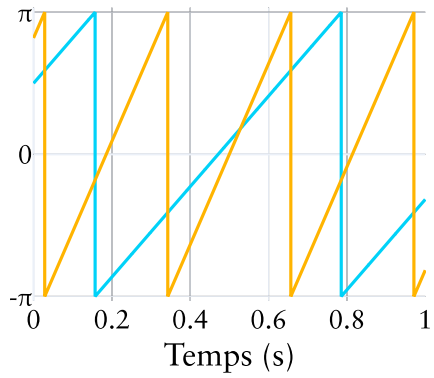
Afin d'assurer que le FDN reste sans perte, c'est à dire que les signaux ne soient pas atténués dans la boucle de rétroaction en dehors du filtrage, il est nécessaire que la matrice de mélange $\boldsymbol{A}$ soit unitaire pour tout échantillon n. Soient $\boldsymbol{U}$ la matrice des vecteurs propres associés aux valeurs propres $(\lambda_i)_{1\leq i\leq N}$ de $\boldsymbol{A}$, la décomposition en valeurs propres de la matrice de mélange s'écrit à l'instant n :

$$\boldsymbol{A}[n] = \boldsymbol{U}[n]^H \Lambda[n] \boldsymbol{U}[n] \tag{9.13}$$

où Λ est une matrice diagonale dont les coefficients diagonaux sont les valeurs propres $(\lambda_i)_{1\leq i\leq N}$. $\boldsymbol{A}$ étant unitaire, le module de ses valeurs propres est égal à 1 et la matrice $\boldsymbol{U}$ est unitaire. En fixant la matrice $\boldsymbol{U}$ au cours du temps, la matrice $\boldsymbol{A}$ sera unitaire également à l'instant $n+1$ si les valeurs propres $(\lambda_i)_{1\leq i\leq N}$ sont de norme 1. Une



(a) Évolution des coefficients de la matrice. Les valeurs associées à chaque coefficient sont décalées sur l'axe des ordonnées pour une meilleure visibilité.



(b) Évolution de l'angle des valeurs propres de la matrice. Seules les angles de deux des valeurs propres sont représentées car les deux autres sont leur antisymétriques (la matrice de mélange étant réelle).

FIGURE 9.8 – Variation d'une matrice de mélange de dimension 4×4 au cours du temps.

solution simple pour faire varier la matrice $\boldsymbol{A}$ de sorte qu'elle reste unitaire consiste donc à modifier les valeurs propres selon l'équation suivante :

$$\lambda_i[n+1] = \lambda_i[n]\, e^{j\phi_i} \text{ avec } \quad \phi_i = \frac{2\pi}{f_s \mu_i} \tag{9.14}$$

où f_s correspond à la fréquence d'échantillonnage et μ_i est nommée période de modulation. De cette manière, la mise à jour de la matrice à chaque pas de temps consiste à appliquer la transformation suivante :

$$\mathrm{A}[n+1] = \mathrm{A}[n]\mathrm{R} \tag{9.15}$$

avec

$$\boldsymbol{R} = \boldsymbol{U}^H \operatorname{diag}\left(e^{j\phi_1}, \ldots, e^{j\phi_N}\right) \boldsymbol{U}. \tag{9.16}$$

La figure 9.8 représente la variation des coefficients et de l'angle des valeurs propres d'une matrice de mélange au cours du temps. Bien que la modification de la matrice de mélange soit simple, l'évolution de ses coefficients apparait plus complexe, ce qui permet d'éviter de courts motifs périodiques dans le signal de sortie du FDN. Les résultats d'un test perceptif mené par Schlecht et Habets [14], ont montré que la variation de la matrice de mélange permet d'améliorer significativement la qualité de la réverbération. En particulier, les jugements de qualité de signaux issus d'un FDN de 8 lignes avec une matrice de mélange modulée n'étaient pas significativement différents des jugements de qualité de signaux issus d'un FDN de 16 lignes avec une matrice de mélange non modulée.

9.1.3. Le choix des délais et des périodes de modulation

Il est essentiel de choisir convenablement les retards appliqués aux signaux dans la boucle de rétroaction d'un FDN car ils permettent à la fois d'obtenir un signal de sortie ayant une densité d'écho précise à un instant donné [16] et de décorréler ces signaux entre eux de manière à ce que le signal de sortie s'apparente à celui issu d'un processus aléatoire [17]. Il est en effet souhaitable que la réponse en fréquence d'un FDN soit la plus plate possible pour éviter toute coloration indésirable. Il est courant que des phénomènes de résonances apparaissent en raison du choix des retards qui accentuent distinctement certaines fréquences. La présence de motifs périodiques dans le domaine temporel et fréquentiel provoque également des colorations peu naturelles.

Bien que les réverbérateurs artificiels basés sur un FDN soient couramment utilisés, il n'existe pas de règle claire sur la façon de choisir les retards et ils sont le plus souvent déterminés de manière heuristique. Néanmoins, une règle répandue consiste à utiliser des retards premiers entre eux [17]. Si tous les retards partagent un facteur premier q, alors tous les instants où apparait un écho sont des multiples entiers de q et aucun écho n'apparait pour les instants non multiples entiers de q. En utilisant des délais qui ne sont pas premiers entre eux, les échos issus de chaque ligne à retard apparaissent à des instants différents, ce qui permet d'accroitre la densité d'écho au cours du temps sans que trop d'échos ne s'accumulent à certains moments.

Il est également souhaitable qu'une partie des retards ne soient pas concentrée autour d'une certaine valeur et de ses multiples. Si tel est le cas, des concentrations d'échos apparaissent dans le signal de sortie et ce qui se traduit par une forte fluctuation de la densité d'écho. Pour éviter une concentration de valeurs de retards, Schlecht et Habets [17] proposent d'utiliser des retards dont l'écart-type géométrique $\sigma_{\boldsymbol{m}}$ vérifie :

$$\sigma_{\boldsymbol{m}} \geq 1.2 \quad \text{avec} \quad \sigma_{\boldsymbol{m}} = \exp\sqrt{\frac{\sum_{i=1}^{N}\left(\ln\frac{m_i}{\bar{m}}\right)^2}{N}} \tag{9.17}$$

où m_i est le retard de la ième ligne du FDN, $\bar{m}$ est la moyenne géométrique des retards et N le nombre de lignes du FDN.

Schlecht et Habets [17] mettent également en lumière les effets liés aux dépendances entre les retards, c'est à dire aux combinaisons linéaires des retards qui coïncident avec d'autres combinaisons linéaires des retards. A titre d'exemple, si un FDN de 3 lignes contient des retards de 49, 51 et 100 échantillons, un écho issu de la dernière ligne sera coïncident avec un écho successivement retardé par la première et la deuxième ligne. La densité d'écho résultante correspond à celle d'un FDN à deux lignes ayant pour retards 49 et 51 échantillons. Il paraît donc nécessaire d'éviter de telles dépendances.

Concernant les périodes de modulation des valeurs propres de la matrice de mélange, il est nécessaire d'utiliser des valeurs qui soient suffisamment courtes pour que la variation réduise les périodicités du signal de sortie sans toutefois introduire d'artefact audible. D'une manière générale, il est souhaitable que les retards et périodes de modulation choisies permettent d'obtenir un signal de sortie ayant un spectre le plus constant possible si aucun filtre d'atténuation n'est appliqué dans la boucle de

rétroaction. De cette manière, la coloration du signal de sortie sera seulement due à l'atténuation conçue pour satisfaire les temps de réverbération cible par bande de fréquence. Le test de Ljung-Box [18] est employé pour quantifier la similarité d'un signal sonore avec un bruit blanc gaussien. En calculant le coefficient statistique Q du test de Ljung-Box, Agus *et al.* [19] ont étudié le seuil à partir duquel une différence est perceptible entre un signal sonore $\boldsymbol{x}$ et un bruit blanc gaussien. Le coefficient statistique Q correspond à l'autocorrélation normalisée moyenne du signal pour différentes valeurs de décalage temporel :

$$Q(\boldsymbol{x}) = K(K+2)\sum_{k=1}^{L}\frac{r_k^2}{(K-k)} \tag{9.18}$$

avec

$$r_k = \frac{\sum_{n=k+1}^{K} x[n]x[n-k]}{\sum_{n=1}^{K} x[n]^2} \tag{9.19}$$

où K est la longueur du signal $\boldsymbol{x}$ en échantillon et L le décalage temporel maximal. Une valeur du coefficient Q élevée traduit une forte variation spectrale.

Ce coefficient statistique Q suit une loi du χ^2 et pour des raisons de comparaison à d'autre indicateurs, Agus *et al.* ont appliqué la transformation de Wilson-Hilferty [20] de manière à obtenir un coefficient $\hat{Q}$ suivant une loi normale :

$$\hat{Q}(\boldsymbol{x}) = \frac{\frac{Q(\boldsymbol{x})^{1/3}}{L} - \mu}{\sqrt{\sigma}} \tag{9.20}$$

avec $\mu = 1 - \frac{2}{9L}$ et $\sigma = \frac{2}{9L}$.

Un test perceptif basé sur des comparaisons par paire entre un bruit blanc gaussien et un bruit coloré a montré que la moitié des 49 participants n'a pas pu correctement distinguer les stimuli sonores lorsque le bruit coloré avait une valeur $\hat{Q}(\boldsymbol{x}) \leq 30.35$. Cette valeur peut constituer un critère pour sélectionner des retards et périodes de modulation permettant d'obtenir un signal sonore en sortie de FDN ayant une faible coloration spectrale lorsqu'aucune atténuation n'est appliquée dans la boucle de rétroaction.

9.2. Le FDN directionnel

Dans cette section, nous employons une architecture de réverbérateur artificiel proche de celle proposée par Alary et Politis [10] destinée à reproduire fidèlement la partie tardive d'une SRIR. A la lumière des résultats obtenus au chapitre 5, nous avons fait le choix de générer des SRIRs d'ordre 1. La figure 9.9 représente les différents éléments de l'architecture proposée.

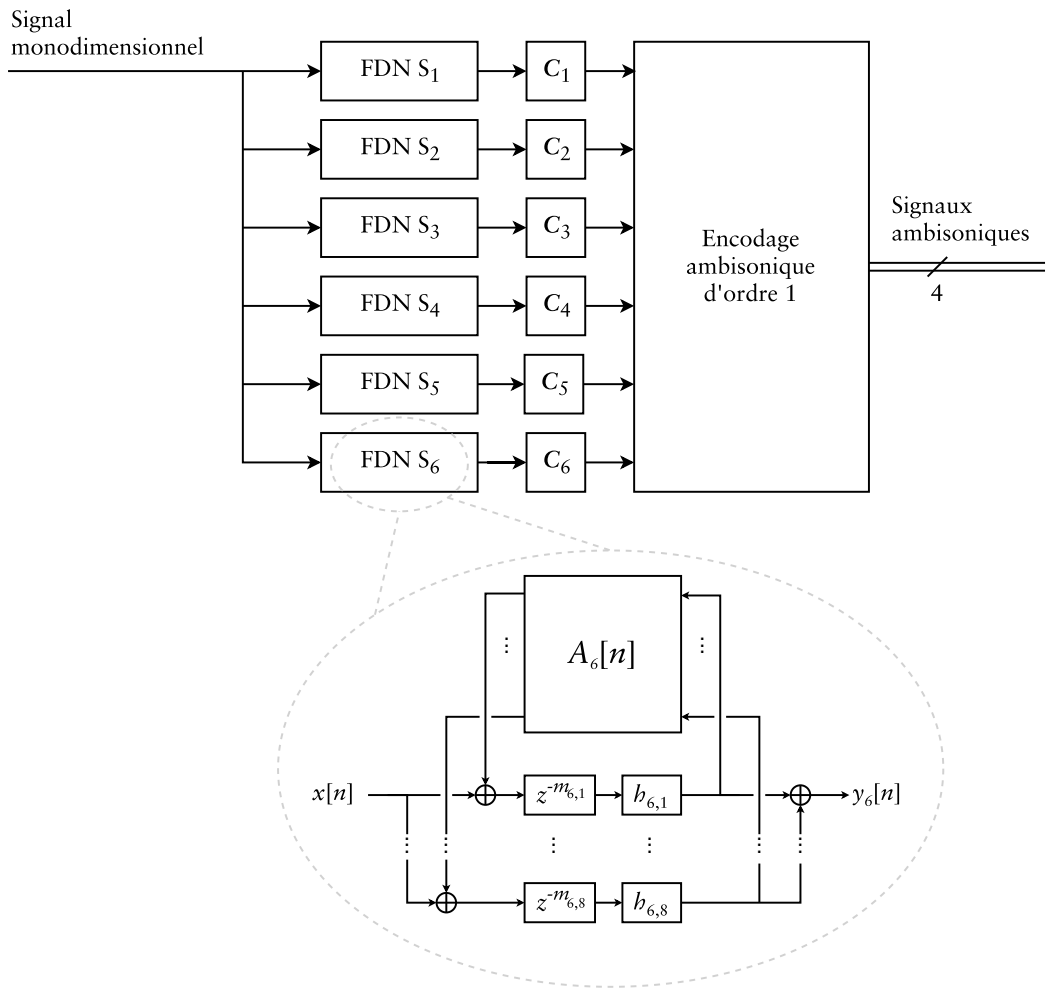


FIGURE 9.9 – Architecture du FDN directionnel employé. Ce réverbérateur artificiel génère 6 signaux dans 6 directions uniformément réparties dans l'espace. Chaque FDN est constitué de 8 lignes à retards dans lesquelles un banc de filtre d'atténuation $\boldsymbol{h}_{p,i}$ est appliqué. Un filtre de correction spectral $\boldsymbol{C}_p$ est appliqué avant d'encoder les signaux en ambisonique.

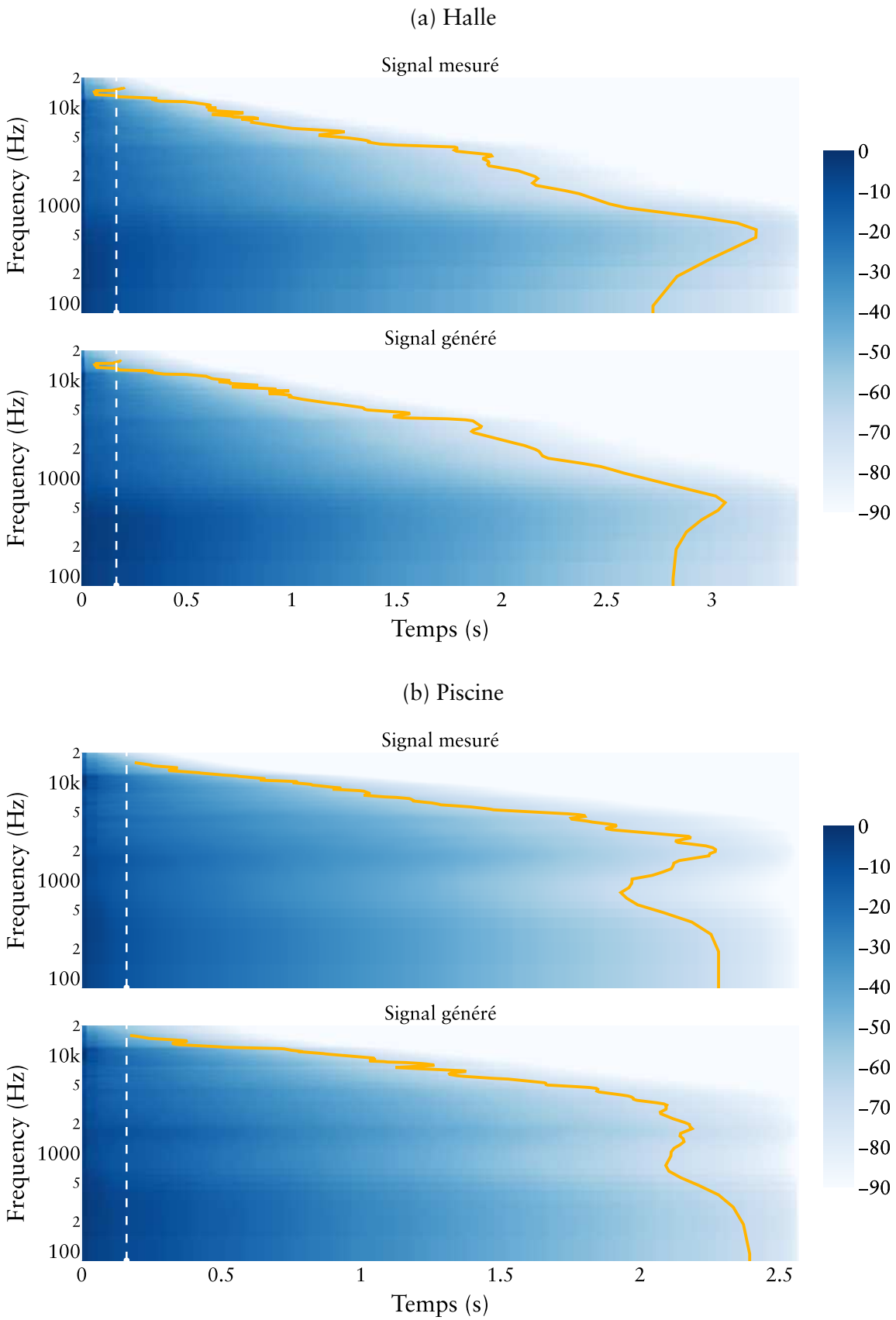


FIGURE 9.10 – EDRs de la SRIR mesurée et de la SRIR générée pour deux salles. Le temps de réverbération en fonction de la fréquence apparait en jaune. Le temps de mélange est matérialisé par la ligne en pointillés noirs. Les parties en amont du temps de mélange sont les mêmes entre un signal mesuré et le signal généré associé. Les signaux sont issus d'une formation de faisceau appliquée aux SRIRs d'ordre 1 dans la direction $(\theta, \varphi) = (90°, 0°)$.

Six FDNs de 8 lignes dont les matrices de mélange sont variables sont employés pour générer 6 signaux encodés en ambisonique selon des directions notées $S_p, 1 \leq p \leq 6$. Ces directions ont été choisies de manière 1) à ce qu'elles soient suffisamment nombreuses pour représenter la SRIR générée en ambisonique à l'ordre 1, 2) à ce qu'elles soient réparties uniformément dans l'espace, 3) à pouvoir contrôler l'énergie tardive dans les directions latérales. Chaque FDN génère un signal dans la direction associée afin de satisfaire les temps de réverbération cibles par bande de fréquence dans cette direction.

Une caractéristique importante d'une réponse impulsionnelle de salle est la densité d'écho. Comme les signaux des différentes directions se somment aux oreilles de l'auditeur, le FDN directionnel exige une densité d'écho moindre des signaux par rapport à un FDN monophonique et le nombre de lignes à retard peut donc être réduit. Dans notre version du FDN directionnel, nous avons fait le choix d'utiliser un faible nombre de lignes à retard ($N = 8$).

Dans le but de reproduire l'énergie tardive d'une SRIR de référence, un filtre de correction spectrale est ajouté en sortie de FDN de manière à respecter la réponse en fréquence cible. Pour chaque direction, ce filtre est calculé d'après le rapport entre le spectre issu de la formation de faisceau appliquée à la SRIR de référence dans la direction correspondante et le spectre issu du FDN associé, après le temps de mélange.

La figure 9.10 représente les densités spectrales d'énergie subsistante à chaque instant ou EDR (pour *Early Decay Relief*) des signaux générés et des signaux de référence dans une direction latérale pour deux salles. L'EDR est une évaluation locale de la courbe de Shrœder, qui permet une représentation temps-fréquence d'une réponse impulsionnelle [21]. L'inspection visuelle confirme un bon accord entre les EDRs. Le profil du temps de réverbération est respecté bien que des erreurs manifestes soient commises : le temps de réverbération est légèrement sous-estimé avec le FDN directionnel. Par ailleurs, on constate que certaines fréquences sont plus accentuées dans les signaux générés par le FDN directionnel, en particulier pour la salle *Piscine* à 2 kHz.

Dans l'optique d'implémenter une réverbération hybride, il est possible de soustraire les signaux du FDN directionnel à la partie précoce de la SRIR destinée à la convolution jusqu'au temps de mélange [22]. Il également possible de soustraire aux signaux du FDN directionnel les signaux issus d'un autre FDN directionnel dont la décroissance de l'énergie par bande de fréquence est plus importante [23]. Ainsi, il est possible de restituer efficacement l'acoustique d'une salle tout en ayant la capacité de contrôler les temps de réverbération par bande de fréquence, l'amplitude et le spectre de la réverbération tardive dans 6 directions de l'espace.

9.3. Évaluation de la méthode

Afin d'évaluer la capacité de la méthode à reproduire la sensation d'enveloppement, nous souhaitons étudier les paramètres acoustiques identifiés comme pertinents à la section 7.5.3 pour caractériser cet attribut. Pour ce faire, les parties tardives des 70 SRIRs issues de la base de donnée GRAP [24] (cf. section 7.1.2) ont été reproduites avec le FDN directionnel de la section précédente.

9.3.1. Les paramètres utilisés

Pour reproduire la partie tardive des SRIRs de référence, plusieurs paramètres du FDN directionnel ont été fixés :

1. Le choix des retards et des périodes de modulation. 40 000 jeux de 48 retards et périodes de modulation maximales et minimales ont été générés. Ces retards étaient premiers entre eux et le ratio entre le retard maximum et minimum était égal à 2.5. Les périodes de modulation étaient inférieures à 3.77 secondes. Des signaux de 5 secondes ont été générés par le FDN directionnel avec ces paramètres et sans filtres d'atténuation. Le coefficient statistique $\hat{Q}$ du test de Ljung-Box a été calculé sur la somme des sorties des 6 FDNs à partir de la seconde moitié des signaux. Le jeu de retards ayant mené au coefficient le plus faible ($\hat{Q} = 3.43$) a été sélectionné pour la suite de l'étude. Une faible valeur du coefficient atteste que la sortie du FDN directionnel sans filtre d'atténuation est proche d'un bruit blanc gaussien.
2. Le filtre d'atténuation. 9 filtres en cloche répartis en bande d'octave entre 63 et 16 000 Hz ont été utilisés ainsi qu'un filtre en plateau haute fréquence dont la fréquence de coupure est fixée à 20 200 Hz et dont le gain correspond à celui du filtre centré sur 16 000 Hz.
3. Le filtre de correction spectrale. 6 filtres à phase linéaire de 513 échantillons ont été utilisés de manière à ce que le spectre du signal généré par le FDN directionnel soit similaire à celui du signal de référence dans les 6 directions.

9.3.2. L'estimation du temps de mélange

La méthode introduite par Abel *et al.* [25] a été utilisée pour déterminer le temps de mélange. Une valeur η, nommé profil de densité d'écho, est calculée à chaque instant d'après l'écart-type de l'amplitude de la réponse impulsionnelle dans une fenêtre temporelle glissante. Il correspond à la proportion d'échantillons dont l'amplitude est supérieure à l'écart-type de l'amplitude dans la fenêtre considérée par rapport au cas d'une distribution gaussienne :

$$\eta[n] = \frac{1}{\epsilon} \sum_{\tau=n-\delta}^{n+\delta} w[\tau] \mathbf{1}\{|r[\tau]| > \sigma\} \tag{9.21}$$

où r correspond à la réponse impulsionnelle, w est la fenêtre temporelle de taille $2\delta+1$ vérifiant $\sum_\tau w[\tau] = 1$, $\mathbf{1}\{.\}$ est une fonction renvoyant 1 lorsque son argument est vrai et zéro sinon et ϵ est le nombre estimé d'échantillons se situant en dehors d'un écart-type de la moyenne pour une distribution gaussienne. L'écart-type σ est défini tel que :

$$\sigma = \sqrt{\sum_{\tau=n-\delta}^{n+\delta} w[\tau] r^2[\tau]} \tag{9.22}$$

Lorsque $\eta = 1$, le nombre d'échantillons dont l'amplitude est supérieure à l'écart-type de l'amplitude dans la fenêtre considérée correspond à celui obtenu avec un processus gaussien. On peut alors considérer que si cette condition est remplie, la portion du signal sur laquelle est centrée la fenêtre glissante est similaire à un bruit gaussien, c'est à dire que cette fenêtre est située dans la région de la réverbération tardive.

Cette méthode a été appliquée sur le signal omnidirectionnel des 70 SRIRs. La moyenne des temps de mélange estimés était de 187.2 ms avec un écart-type de 72.1 ms.

9.3.3. Résultats

Paramètre	JND	Moyenne	Écart-type	Minimum	Maximum	Score
T_{30}	5%	3.36%	4.46%	-9.1%	18.0%	72.8%
EDT	5%	***10.15%***	5.88%	-31.1%	23.2%	18.5%
T_s	10 ms	***10.7*** ms	24.8 ms	-36.5 ms	159.6 ms	77.1%
D_{50}	0.05	0.042	0.040	-0.168	0.128	71.4%
C_{80}	1 dB	0.93 dB	0.87 dB	-3.80 dB	4.25 dB	67.1%
G_L	1 dB	0.85 dB	0.78 dB	-2.38 dB	3.83 dB	70.0%
L_J	-	1.52 dB	1.27 dB	-4.31 dB	5.25 dB	-
LLF	0.05	0.043	0.045	-0.038	0.207	68.5%
1-$IACC_L$	0.075	***0.128***	0.071	-0.136	0.315	25.7%
1-IACC	0.075	0.040	0.041	-0.056	0.249	85.7%

Tableau 9.1 – Erreurs obtenues entre les paramètres acoustiques calculés d'après les SRIRs générées avec le FDN directionnel et ceux calculés d'après les 70 SRIRs de références. Les valeurs des paramètres correspondent à la moyenne des valeurs calculées dans les bandes de fréquence centrées sur 500 Hz et 1000 Hz (sauf pour L_J et LLF dont la moyenne comprend également les valeurs calculées dans les bandes de fréquence centrées sur 125 Hz et 250 Hz). Les valeurs de JND sont issus de la norme ISO 3382-1 [26]. Les valeurs supérieures aux JNDs sont indiquées en gras. Le score correspond à la proportion d'espaces pour lesquelles la différence entre le paramètre acoustique et celui de la référence est inférieure à la JND.

Le tableau 9.1 recense les différences obtenues entre les paramètres acoustiques calculés d'après les SRIRs générées avec le FDN directionnel et ceux calculés d'après les 70 SRIRs de référence. Pour le calcul des coefficients de décorrélation interaurale, les SRIRs ambisoniques ont été décodées en binaural avec des filtres de décodage calculés d'après les méthodes de rendu binaural proposées par Schörkhuber *et al.* [27] et Zaunschirm *et al.* [28]. Ces filtres ont été obtenus à partir d'HRTFs mesurées avec une tête artificielle Neumann KU 100 [29].

L'erreur moyenne excède les seuils d'audibilité ou JND (pour *Just Noticeable Differences*) pour deux paramètres acoustiques : EDT et 1-$IACC_L$. Pour tous les paramètres acoustiques, les plus grandes erreurs commises excèdent toutes les valeurs de JNDs. De plus, aucun espace ne respecte les JNDs pour l'ensemble des paramètres acoustiques à la fois.

Le score affiché dans le tableau correspond à la proportion d'espaces dont la différence entre le paramètre acoustique concerné et celui de la référence est inférieure à la JND correspondante. Les scores les plus faibles sont obtenus pour l'EDT et 1-IACC$_L$ que seule une faible proportion de SRIRs générées parvient à reproduire convenablement (18.5% et 25.7% respectivement). Le coefficient de décorrélation interaural 1-IACC calculé sur l'ensemble de la réponse impuslionnelle binaurale obtient le meilleur score avec des valeurs similaires à celles des SRIRs de référence pour 85.7% des SRIRs générées.

9.3.4. Discussion

La différence importante constatée pour l'EDT peut s'expliquer par la présence de réflexions spéculaires au-delà du temps de mélange et qui ne sont donc pas présentes dans les SRIRs générées. Le temps d'apparition moyen des réflexions spéculaires sur l'ensemble des 70 SRIRs de référence s'élève à 222 ms et l'écart-type est égal à 113 ms. Ces réflexions sont issues d'un calcul des sources-images de troisième ordre. En raison de l'amplitude importante de ces réflexions par rapport à la composante diffuse, des paliers dans la courbe de Schrœder apparaissent et réduisent ainsi la décroissance initiale de l'énergie sonore calculée. Lorsqu'elles ne sont pas prises en compte, la décroissance est plus importante et la durée de décroissance initiale plus faible. Ceci peut expliquer la moyenne négative de l'erreur commise sur l'EDT.

D'autre part, les retards du FDN directionnel étant fixes, la densité d'écho reste la même pour tous les signaux issus du FDN directionnel et est différente de celles des SRIRs de référence. Selon la valeur du temps de mélange estimé il est possible que la densité d'écho des signaux générés ne soit pas adaptée à celle d'origine. Le nombre d'échos introduits après le temps de mélange peut dans certains cas être trop faible et introduire moins d'énergie dans la portion du signal généré qui n'a pas encore atteint une densité d'écho suffisante. Pour pallier ce problème, la sélection des retards peut être effectuée en cohérence avec la densité d'écho de la SRIR de référence [16].

Les valeurs du coefficient 1-IACC$_L$ sont en moyenne plus élevées pour les BRIRs générées avec le FDN directionnel. La moyenne du coefficient de décorrélation interaurale tardive des signaux binauraux issus du FDN directionnel s'élève à 0.600 contre 0.479 pour les signaux binauraux de références. Les signaux de sortie du FDN directionnel sont décorrélés en raison des différents retards et mélanges appliqués au signal d'entrée. La partie tardive des SRIRs de référence, résulte quant à elle de l'encodage de 36 signaux synthétisés d'après un processus de Poisson et régulièrement répartis dans l'espace. La nature de ces différents signaux utilisés pour créer la réverbération tardive semble avoir une influence importante sur la décorrélation interaurale. On peut également faire l'hypothèse que la présence de réflexions spéculaires non prises en compte dans les SRIRs générées produit des différences importantes en termes de décorrélation. Afin de faire correspondre les décorrélations interaurales de la partie tardive, il serait intéressant de réaliser des mélanges entre les signaux des 6 FDNs pour être en mesure de contrôler la décorrélation interaurale résultante.

Les décorrélations interaurales calculées sur l'ensemble des BRIRs (1-IACC) sont beaucoup plus proches. La prise en compte des premières réflexions réduit signifi-

cativement la décorrélation en raison de la forte directionnalité introduite dans les réponses impulsionnelles binaurales. Les valeurs de décorrélation résultantes sont dominées par la présence de ces forts indices directionnels.

L'architecture du FDN directionnel utilisée permet un contrôle du temps de réverbération dépendant de la fréquence et de la direction. Néanmoins, les écarts observés sur le temps de réverbération calculé aux fréquences moyennes sont importants (de -9.1% à 18.0%). L'erreur commise dans la boucle de rétroaction sur l'atténuation par bande de fréquence peut expliquer les différences observées sur les temps de réverbération. Le calcul des gains d'atténuation issus de l'équation (9.12) ne fournit pas de gains satisfaisants en toute circonstance. Une piste d'amélioration possible serait d'étudier les liens entre les temps de réverbération obtenus et les gains d'atténuation utilisés afin d'établir une fonction de correspondance plus précise.

9.4. Conclusion

En raison d'un coût de calcul plus faible que celui induit par une convolution pour reproduire des temps de réverbération longs, le FDN est un algorithme couramment utilisé pour créer un effet de réverbération. Nous avons exposé dans ce chapitre comment paramétrer un FDN pour permettre la reproduction de temps de réverbération dans plusieurs bandes de fréquence, tout en réduisant l'introduction d'artefacts spectraux et temporels. Un traitement utilisant plusieurs FDNs a ensuite été présenté dans le but de reproduire les caractéristiques spectrales, spatiales et temporelles de la partie tardive de SRIRs tout en permettant le contrôle du temps de réverbération en fonction de la fréquence et de la direction.

L'examen des erreurs produites par le FDN directionnel sur les paramètres acoustiques pertinents pour caractériser l'impression d'enveloppement permet de conclure que ce traitement n'est pas assez précis pour reproduire de manière imperceptible des SRIRs simulées de référence. Les erreurs commises sur ces paramètres étant parfois importantes, on peut faire l'hypothèse que cette impression spatiale n'est pas fidèlement reproduite avec ce système. Un test perceptif pourrait être mis en œuvre pour confirmer cette hypothèse. Des pistes d'amélioration consistent à reproduire convenablement les réflexions spéculaires tardives, à mieux maîtriser le choix des retards pour satisfaire la densité d'écho de l'espace à reproduire et à parfaire le calcul des gains d'atténuation.

Néanmoins, cette architecture permet d'auraliser un espace à moindre coût tout en fournissant un rendu sonore crédible et cohérent par rapport à un enregistrement réalisé *in situ* grâce à la convolution des premières réflexions qui permet de préserver le naturel et la signature spectrale de la salle [6]. De plus, elle permet de réaliser un contrôle des caractéristiques spatiales, spectrales et temporelles de l'effet de réverbération généré. Même lorsque le gain en coût de calcul est faible en comparaison à celui de la convolution, le FDN directionnel permet une modification plus flexible d'un effet de réverbération dans la mesure où, dans le cas d'une convolution, une modification du temps de réverbération nécessite une mise à jour de l'ensemble de la réponse impulsionnelle spatiale. Il serait intéressant d'étudier le comportement du FDN directionnel lorsque ces paramètres sont modifiés en temps-réel afin d'identifier

de possibles artefacts.

Bibliographie

[1] W. G. Gardner, “Efficient convolution without input/output delay,” dans *Audio Engineering Society Convention 97*. Audio Engineering Society, 1994.

[2] G. Garcia, “Optimal filter partition for efficient convolution with short input/output delay,” dans *Audio Engineering Society Convention 113*. Audio Engineering Society, 2002.

[3] M. A. Gerzon, “Synthetic stereo reverberation : Part one,” *Studio Sound*, vol. 13, p. 632–635, 1971.

[4] J. Stautner et M. Puckette, “Designing multi-channel reverberators,” *Computer Music Journal*, vol. 6, n°. 1, p. 52–65, 1982.

[5] J.-M. Jot et A. Chaigne, “Digital delay networks for designing artificial reverberators,” dans *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[6] T. Carpentier, M. Noisternig et O. Warusfel, “Hybrid reverberation processor with perceptual control,” dans *17th International Conference on Digital Audio Effects-DAFx-14*, 2014, p. 93–100.

[7] D. Romblom, C. Guastavino et P. Depalle, “Perceptual thresholds for non-ideal diffuse field reverberation,” *The Journal of the Acoustical Society of America*, vol. 140, n°. 5, p. 3908–3916, 2016.

[8] J. Anderson et S. Costello, “Adapting artificial reverberation architectures for b-format signal processing,” dans *Ambisonics Symposium*, 2009, p. 2–6.

[9] B. Wiggins et M. Dring, “Ambifreeverb 2—development of a 3d ambisonic reverb with spatial warping and variable scattering,” dans *Audio Engineering Society Conference : 2016 AES International Conference on Sound Field Control*. Audio Engineering Society, 2016.

[10] B. Alary et A. Politis, “Frequency-dependent directional feedback delay network,” dans *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, p. 176–180.

[11] K. Prawda, V. Välimäki, S. J. Schlecht *et al.*, “Improved reverberation time control for feedback delay networks,” dans *Proc. 22th Int. Conf. Digital Audio Effects (DAFx-19)*, 2019.

[12] S. J. Schlecht et E. A. Habets, “Accurate reverberation time control in feedback delay networks,” *Proc. Digital Audio Effects (DAFx-17), Edinburgh, UK*, p. 337–344, 2017.

[13] V. Välimäki et J. Liski, “Accurate cascade graphic equalizer,” *IEEE Signal Processing Letters*, vol. 24, n°. 2, p. 176–180, 2016.

[14] S. J. Schlecht et E. A. Habets, “Time-varying feedback matrices in feedback delay networks and their application in artificial reverberation,” *The Journal of the Acoustical Society of America*, vol. 138, n°. 3, p. 1389–1398, 2015.

[15] S. J. Schlecht et E. A. Habets, “Practical considerations of time-varying feedback delay networks,” dans *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[16] S. J. Schlecht et E. A. Habets, “Feedback delay networks : Echo density and mixing time,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, n°. 2, p. 374–383, 2016.

[17] M. R. Schroeder et B. F. Logan, “«Colorless» artificial reverberation,” *IRE Transactions on Audio*, n°. 6, p. 209–214, 1961.

[18] G. M. Ljung et G. E. Box, “On a measure of lack of fit in time series models,” *Biometrika*, vol. 65, n°. 2, p. 297–303, 1978.

[19] N. Agus, H. Anderson, J.-M. Chen, S. Lui et D. Herremans, “Perceptual evaluation of measures of spectral variance,” *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3300–3311, 2018.

[20] E. B. Wilson et M. M. Hilferty, “The distribution of chi-square,” *proceedings of the National Academy of Sciences of the United States of America*, vol. 17, n°. 12, p. 684, 1931.

[21] J.-M. Jot, L. Cerveau et O. Warusfel, “Analysis and synthesis of room reverberation based on a statistical time-frequency model,” dans *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[22] A. B. Greenblatt, J. S. Abel et D. P. Berners, “A hybrid reverberation crossfading technique,” dans *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, p. 429–432.

[23] N. Meyer-Kahlen, S. J. Schlecht et T. Lokki, “Fade-in control for feedback delay networks,” dans *Proceedings of the 23rdInternational Conference on Digital Audio Effects (DAFx2020)*, Vienna, Austria, 2020, p. 227–233.

[24] D. Ackermann, M. Ilse, D. Grigoriev, S. Lepa, S. Pelzer, M. Vorländer et S. Weinzierl, “A ground truth on room acoustical analysis and perception (GRAP),” 2018.

[25] J. S. Abel et P. Huang, “A simple, robust measure of reverberation echo density,” dans *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[26] ISO 3382-1, “Acoustics-measurement of room acoustic parameters, part 1 : Performance spaces,” International Organization for Standardization, Geneva, CH, Standard, 2009.

[27] C. Schörkhuber, M. Zaunschirm et R. Höldrich, “Binaural rendering of ambisonic signals via magnitude least squares,” dans *Proceedings of the DAGA*, vol. 44, 2018, p. 339–342.

[28] M. Zaunschirm, C. Schörkhuber et R. Höldrich, “Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint,” *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3616–3627, 2018.

[29] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” dans *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. AIA/DAGA, 2013, p. 29.

Conclusion

Les études présentées dans cette thèse participent à l'élaboration d'outils de création de contenus sonores immersifs. De tels outils permettraient de modifier conjointement les propriétés spatiales, fréquentielles et temporelles d'un effet de réverbération selon différents attributs perceptifs dont notamment les impressions spatiales. L'ensemble des études a été mené dans un contexte d'écoute communément employé pour la reproduction de contenus immersif, c'est à dire en utilisant un rendu binaural dynamique non-individualisé de scènes sonores encodées en ambisonique.

Une partie conséquente de cette thèse s'est portée sur la décomposition de réponses impulsionnelles spatiales de salle (SRIRs) qui caractérisent le trajet acoustique entre la source sonore et l'auditeur. La représentation spatiale d'une SRIR peut constituer une grande quantité de données et la réduction de cette quantité à des éléments pertinents d'un point de vue perceptif pour reproduire l'acoustique d'une salle permet notamment de réduire le coût de calcul et de rendre le contrôle plus simple et compréhensible. Une fois la quantité de données utilisées pour décrire une SRIR réduite à quelques éléments, nous avons étudié comment les modifier pour contrôler deux attributs perceptifs : la largeur apparente de source et l'enveloppement.

Avant de se concentrer sur le contrôle d'un effet de réverbération, nous avons choisi d'étudier préalablement deux facteurs ayant un rôle important dans la perception de l'espace : l'environnement visuel et la source sonore. Un premier test perceptif a montré que la perception visuelle d'une salle n'affectait pas les différences perçues entre les espaces sonores. En analysant seulement des jugements de dissemblance sous différentes conditions visuelles, cette étude n'a présumé ni du nombre ni de la nature des dimensions perceptives impliquées dans la perception de l'espace sonore. Les résultats soutiennent les quelques études faisant état d'une absence d'influence de l'environnement visuel sur la perception de l'acoustique d'une salle. L'attention portée aux informations sonores n'étant pas moindre en présence d'informations visuelles, il n'est pas envisageable de réduire la quantité d'informations sonores en raison de la présence d'images. N'ayant pas observé d'influence significative de l'image sur la perception de l'acoustique, nous avons fait le choix de poursuivre nos travaux sans prendre en compte l'environnement visuel des espaces étudiés.

Nous nous sommes ensuite demandés dans quelle mesure des sources sonores usuelles peuvent influencer notre appréciation de l'acoustique des salles. Un test perceptif a montré que la source sonore a une forte influence sur la perception de l'acoustique d'une salle. En particulier, les principales dimensions perceptives utilisées pour différencier plusieurs espaces sonores étaient différentes d'une source à l'autre. Seule la réverbérance était un critère commun pour différencier les espaces sonores. Ainsi, les sources sonores usuelles révélant différentes caractéristiques perceptives d'un espace sonore, un outil de contrôle perceptif doit être capable de modifier un grand

nombre de propriétés acoustiques pour être adapté à une pluralité de sources sonores. Dans la suite de la thèse, nous avons toutefois fait le choix de ne pas multiplier les sources sonores pour limiter la durée des tests. Toute extrapolation des résultats issus de ces tests à d'autres types de sources sonores doit donc être considérée avec précautions.

Dans la seconde partie de cette thèse nous avons étudié les éléments pertinents d'une SRIR à manipuler pour contrôler un effet de réverbération. Plusieurs méthodes de paramétrisation de réponses impulsionnelles spatiales ont été détaillées. Deux types de paramétrisation ont été identifiés : les méthodes analysant les SRIRs dans des fenêtres temporelles ou temps-fréquence restreintes (de l'ordre de quelques millisecondes et quelques dizaines de hertz) et les méthodes d'analyse qui estiment les éléments de la modélisation d'une réponse impulsionnelle de salle : des réflexions spéculaires associées à des pentes de décroissance de l'énergie sonore. Les méthodes présentées ne sont pas exemptes de biais dans l'analyse des paramètres et produisent parfois des artefacts audibles, notamment en raison des procédés de décorrélation utilisés pour la reproduction.

Deux tests perceptifs ont été réalisés dans l'optique de réduire le nombre d'éléments à manipuler d'une SRIR pour faciliter le contrôle du rendu sonore. En particulier, une première étude a été effectuée afin de déterminer la résolution spatiale (en termes d'ordre ambisonique et de dimensionnalité) avec laquelle il est nécessaire de reproduire une SRIR. L'analyse des résultats a révélé que lorsque le son direct est restitué avec une résolution spatiale importante, le nombre de canaux ambisoniques décrivant la partie réverbérée peut être grandement réduit sans entraîner de différences perceptives importantes voire même significatives. Ce résultat dépend de l'espace considéré, la présence de réflexions précoces importantes ne permettant pas la réduction de la résolution spatiale de manière imperceptible (même si la différence perçue reste faible). L'intérêt d'une représentation ambisonique d'ordre élevé pour décrire une SRIR parait donc limité dans le contexte d'un rendu binaural non-individualisé. Par la suite nous avons fait le choix d'utiliser une représentation spatiale de l'acoustique selon un ordre ambisonique faible étant donnée la réduction importante de données qui en résulte.

Une seconde étude s'est portée sur les résolutions fréquentielles et temporelles nécessaires à la reproduction des premières réflexions avec une résolution spatiale réduite. Pour permettre cette analyse, une méthode de paramétrisation basée sur le calcul de matrices de covariance a été proposée. Cette méthode a permis de reproduire la répartition de l'énergie sonore de réflexions précoces dans plusieurs régions temps-fréquence. Un test perceptif a révélé que, parmi les méthodes évaluées, la méthode de paramétrisation employant une résolution temporelle de 5 ms dans quatre bandes de fréquence était la plus efficace. En effet, elle a produit des stimuli jugés parmi les plus proches des scènes sonores de référence tout en limitant le nombre de données utilisées. L'accroissement de la résolution fréquentielle à huit bandes de fréquence n'a pas permis d'améliorer les résultats. Les résultats obtenus avec la méthode la plus efficace n'étaient pas significativement différents de ceux obtenus avec une autre méthode de paramétrisation communément employée - la SDM - tout en

réduisant le nombre de paramètres nécessaires. Néanmoins, la paramétrisation proposée introduit des artefacts audibles, particulièrement lorsque la source comprend de nombreux transitoires. De plus, la quantité de calcul nécessaire à la reproduction de la partie précoce d'une SRIR avec cette méthode est également importante. Au vu de ces observations, l'intérêt de cette paramétrisation paraît limité. L'enseignement principal de cette étude est que la reproduction de la répartition spatiale de l'énergie sonore d'une SRIR doit être précise dans le temps mais peut être restreinte en fréquence.

Dans la dernière partie de cette thèse, nous nous sommes concentrés sur le contrôle de deux attributs perceptifs : la largeur apparente de source et l'enveloppement. Nous avons fait le choix d'étudier des attributs perceptifs ayant un lien spécifique au dispositif de restitution employé : un rendu binaural non-individualisé de scènes sonores encodées en ambisonique. En effet, ces attributs perceptifs sont liés à la directivité du champ sonore et le rendu sonore utilisé présente l'intérêt de pouvoir reproduire et manipuler cette directivité dans les trois dimensions de l'espace.

Une première étude a tenté d'identifier les origines physiques de ces deux attributs perceptifs en analysant la base de données GRAP fournie par l'Université technique de Berlin. Nous avons mis en lien les résultats issus de leur évaluation perceptive avec des paramètres acoustiques. L'analyse des données a mis en évidence que la prédominance du son direct est néfaste à la perception d'une source large et que cette perception semble favorisée par une énergie tardive importante. Cette influence de l'énergie tardive doit être corroborée par d'autres études car elle contredit des observations antérieures. Dans une moindre mesure, l'analyse a également mis en lumière le rôle de l'énergie précoce dans la perception de la largeur de source. Cependant, les données à disposition n'ont pas permis d'établir des modèles prédictifs suffisamment performants. D'autres données seraient nécessaires pour affiner la caractérisation de la largeur apparente de source.

La sensation d'enveloppement n'ayant pas présenté de lien avec les propriétés de l'énergie précoce, nous avons choisi d'étudier le contrôle de la largeur apparente de source en modifiant les premières réflexions seulement. De cette manière, la modification de cet attribut perceptif peut être réalisée sans altérer la sensation d'enveloppement. De nombreuses études se sont concentrées sur le lien entre la proportion d'énergie latérale précoce ou la décorrélation interaurale précoce de réponses impulsionnelles et la largeur apparente de source. Pour étudier ce lien, nous avons appliqué différentes méthodes de transformation spatiale à l'énergie précoce de SRIRs et analysé les modifications induites sur la largeur apparente de source. Ces transformations consistaient en l'augmentation ou la réduction de l'énergie latérale et de la décorrélation intéraurale. Un test perceptif a révélé que la variation de ces deux propriétés spatiales permet en effet d'accroître ou de réduire significativement la largeur apparente de source. Néanmoins, la modification de la spatialisation des premières réflexions n'a pas eu d'effet significatif sur la largeur apparente de source pour tous les espaces considérés. Il semble que l'énergie des réflexions proches du son direct doit être suffisamment importante pour permettre une modification perceptible de la largeur apparente de source. Suite à ces études, il apparaît que les paramètres acous-

tiques définis pour prédire la largeur apparente de source doivent être calculés d'après une région temporelle différente de celle fixée aux 80 premières millisecondes dans la norme ISO 3382-1.

En cohérence avec les observations présentes dans la littérature, notre analyse statistique a permis de mettre en lumière l'influence de la quantité d'énergie tardive, des réflexions latérales tardives et de la décorrélation interaurale sur la sensation d'enveloppement. La paramétrisation de la partie tardive d'une SRIR consiste le plus souvent à calculer les pentes de décroissance de l'énergie sonore par bande de fréquence. Étant donné le rôle joué par les réflexions latérales dans la sensation d'enveloppement, il nous a semblé judicieux de modifier ces paramètres dans différents secteurs angulaires. Pour cela, un réverbérateur artificiel basé sur plusieurs réseaux récursifs de lignes à retard (FDN) a été utilisé. Le FDN est couramment utilisé pour créer un effet de réverbération car il nécessite un coût de calcul plus faible que celui induit par une convolution pour reproduire des temps de réverbération longs. Il permet également de modifier aisément les pentes de décroissance par bande de fréquence. Une étude a évalué la capacité de ce réverbérateur artificiel à reproduire les paramètres acoustiques identifiés comme pertinents pour caractériser la sensation d'enveloppement. Les erreurs commises sur ces paramètres acoustiques étant dans certains cas importantes, des pistes d'amélioration du réverbérateur artificiel ont été proposées. Notamment, de la même manière que pour la reproduction des premières réflexions, la restitution de réflexions spéculaires semble nécessaire pour respecter certains des paramètres étudiés.

Il nous semble intéressant de poursuivre l'étude des impressions spatiales en examinant plus précisément les interactions entre la largeur apparente de source et enveloppement. La modification des propriétés spatiales, fréquentielles et temporelles de l'énergie tardive que permet l'architecture du réverbérateur artificiel employé aura vraisemblablement des conséquences sur la perception de la largeur apparente de source. Bien qu'une influence significative des premières réflexions sur cet attribut soit avérée, l'influence de la réverbération tardive semble avoir été négligée dans la littérature. Ces régions temporelles sont définies arbitrairement dans la norme ISO 3382-1 et une étude plus précise de cette segmentation pourrait aider à la compréhension des origines physiques de la largeur apparente de source.

Suite aux travaux effectués, il semble judicieux de créer un effet de réverbération avec un rendu des premières réflexions par convolution et un rendu de la réverbération tardive selon plusieurs réseaux récursifs de ligne à retard. L'emploi de cette approche n'est pas seulement motivé par une économie de calcul. De cette manière, la sensation d'enveloppement peut être contrôlée simplement en modifiant les énergies et les pentes de décroissance définies dans plusieurs bandes de fréquence et secteurs angulaires. Il serait d'ailleurs intéressant d'étudier le nombre de bandes de fréquences auxquelles il faut attribuer une pente de décroissance. Pour la reproduction des premières réflexions, nous avons vu qu'au delà de quatre bandes de fréquence l'accroissement de la résolution fréquentielle n'était pas significativement avantageuse pour restituer la répartition énergétique du champ sonore. Afin de contrôler la largeur apparente de source, nous suggérons d'accroitre cet attribut perceptif en modifiant les propriétés

spatiales du son direct et de le réduire en modifiant les premières réflexions selon une déformation angulaire ou une amplification directionnelle.

Dans un contexte qui implique des contraintes fortes en termes de calcul et de mémoire, l'emploi d'une restitution hybride à un ordre ambisonique faible semble raisonnable étant donné les faibles différences perceptives observées avec la réduction de la résolution spatiale. Nous n'avons pas identifié de paramétrisation permettant une réduction des données imperceptible. Il semble que la prise en compte des réflexions spéculaires serait un ajout substantiel permettant une plus grande fidélité dans la perception de la directivité du champ sonore et un plus grand respect de certains paramètres acoustiques tels que l'EDT ou la décorrélation interaurale. Néanmoins, la méthode de sélection de réflexions spéculaires pertinentes n'est pas encore établie. Une meilleure modélisation de l'effet de précédence semble nécessaire.

La quantité de calcul à effectuer pour l'auralisation d'un espace sonore selon le procédé de rendu proposé reste importante. Dans les études perceptives réalisées au cours de cette thèse, le critère utilisé pour juger de la réduction des données à manipuler était la similarité avec des scènes sonores de référence. Ce critère privilégie une contrainte forte : la fidélité de restitution. Nous avons choisi cette contrainte car la visée de cette thèse est de fournir aux ingénieurs du son des outils de qualité pour la manipulation d'un espace sonore dans un contexte de production de contenus immersifs. Selon l'application, d'autres études privilégieront la qualité d'expérience, la plausibilité ou encore la préférence. Il est nécessaire d'arbitrer entre la qualité du rendu et le coût de calcul selon l'application. Avec l'avènement de technologies immersives visant la navigation dans un environnement virtuel ou se déployant sur des systèmes embarqués, les contraintes en termes de traitements temps-réel sont fortes et les possibilités de contrôle seront *a fortiori* différentes.

Annexes

A

La mesure d'une réponse impulsionnelle de salle

Une salle peut être considérée comme un système de transmission acoustique linéaire et invariant dans le temps entre une source et un récepteur situés dans la salle. La réponse impulsionnelle de la salle décrit les modifications appliquées au signal émis par la source lorsqu'il se propage d'un point à un autre de la salle. Cette réponse impulsionnelle peut être mesurée. C'est une mesure couramment utilisée pour extraire des paramètres acoustiques (cf section 1.3) qui caractérisent certains aspects de la perception d'une salle. Elle peut également être employée pour la simulation d'espaces sonores en la convoluant avec une source sonore anéchoïque. Une telle approche s'apparente à un processus d'auralisation. L'auralisation désigne «*la technique permettant de créer des fichiers sonores audibles à partir de données numériques (simulées, mesurées ou synthétisées)*» [1].

La réponse impulsionnelle d'un système correspond à la sortie du système après son excitation par une impulsion de Dirac. En pratique, la reproduction d'une impulsion de Dirac est complexe. L'utilisation de dispositifs impulsifs (pistolet, ballon) présente des variabilités importantes dans les résultats et l'emploi d'une enceinte introduit des distorsions non-linéaires en raison de la trop grande quantité d'énergie délivrée sur une très courte durée par rapport aux propriétés électro-mécaniques du transducteur [2]. D'autres signaux d'entrée peuvent être utilisés pour accéder à la réponse impulsionnelle de salle dans la mesure où ils contiennent suffisamment d'énergie dans la gamme de fréquence du spectre audible.

Soit Y le résultat de l'excitation à la fréquence de pulsation ω d'un signal d'entrée X reproduit à travers une enceinte dont la fonction de transfert est notée G :

$$Y(\omega) = H(\omega) \cdot G(\omega) \cdot X(\omega) + N(\omega) \tag{A.1}$$

avec H la fonction de transfert complexe de la salle et N la réponse en fréquence d'un bruit blanc gaussien décorrélé du signal d'entrée qui modélise le bruit ambiant.

Le signal Y contient à la fois la réponse de la salle et de l'enceinte utilisée. La coloration de l'excitation par la réponse en fréquence de l'enceinte G est indésirable à des fins d'auralisation car impacterait spectralement la réponse impulsionnelle obtenue. Compenser cette coloration *a posteriori* en corrigeant la réponse impulsionnelle, pourrait amplifier le bruit ambiant dans certaines régions fréquentielles. C'est pourquoi on utilise un signal d'excitation $\widetilde{X}$ résultat du pré-traitement du signal d'entrée par la réponse en fréquence inverse de l'enceinte :

$$\widetilde{X}(\omega) = \frac{X(\omega)}{G(\omega)}. \tag{A.2}$$

De cette manière on obtient un signal $\widetilde{Y}$ indépendant de la réponse en fréquence de l'enceinte :

$$\widetilde{Y}(\omega) = H(\omega) \cdot G(\omega) \cdot \widetilde{X}(\omega) + N(\omega) = H(\omega) \cdot X(\omega) + N(\omega) \tag{A.3}$$

En pratique cette compensation est réalisée dans un plage de fréquence donnée pour laquelle un niveau d'amplification reste raisonnable.

Notons que si la réponse est utilisée pour extraire des paramètres acoustiques ou si elle est utilisée pour une auralisation, les directivités requises de la source et du récepteur peuvent différer ; l'utilisation d'une source sonore omnidirectionnelle étant privilégiée pour l'extraction de paramètres acoustiques.

A.1. Balayage sinusoïdal

Les signaux d'entrée couramment utilisés fournissent un énergie constante par fréquence, soit un spectre «plat». La réponse en fréquence est alors de la forme :

$$X(\omega) = e^{i\phi(\omega)} \tag{A.4}$$

avec ϕ le spectre de phase de X. Il en résulte que $X(\omega)^{-1} = X^*(\omega)$ ou $*$ désigne le complexe conjugué. Ainsi

$$\widetilde{H}(\omega) = Y(\omega) \cdot X^*(\omega) \tag{A.5}$$

où $\widetilde{H}$ est une estimation de la fonction de transfert du système acoustique. La transformée de Fourier inverse de $X^*(\omega)$ étant $x(-t)$, la déconvolution peut s'effectuer dans le domaine temporel en utilisant le retournement temporel du signal d'entrée.

Il est souhaitable d'utiliser des signaux d'excitation avec une énergie importante afin d'obtenir un rapport signal-bruit suffisant dans toute la gamme de fréquences concernée. Les balayages sinusoïdaux se sont révélés être un bon choix pour les mesures acoustiques car ils permettent d'alimenter l'enceinte fréquence par fréquence avec une énergie considérablement plus importante que lorsqu'une impulsion est utilisée, sans introduire d'artefacts dans la réponse impulsionnelle acquise [3, 4]. Ils correspondent à des signaux sinusoïdaux dont la fréquence varie en continu :

$$x(t) = \sin(\phi(t)). \tag{A.6}$$

Pour un balayage sinusoïdal linéaire, le retard de groupe, c'est à dire la variation de la phase en fonction du temps, à partir d'une fréquence de pulsation ω_1 jusqu'à ω_2 sur un temps T est donné par :

$$\frac{\mathrm{d}\phi(t)}{\mathrm{dt}} = \omega_1 + \frac{\omega_2 - \omega_1}{\mathrm{T}} \cdot t \tag{A.7}$$

Le fait de tronquer la variation du balayage sinusoïdal à des fréquences finies introduit une certaine ondulation aux extrémités du spectre d'amplitude qui peut être minimisée en appliquant un court fenêtrage en début et fin du balayage.

Le balayage sinusoïdal logarithmique (également nommé exponentiel) est couramment utilisé. La variation de la phase en fonction du temps d'une fréquence de pulsation ω_1 jusqu'à ω_2 sur un temps T est alors de la forme :

$$\frac{\mathrm{d}\phi(t)}{\mathrm{dt}} = \frac{K}{L} \cdot e^{\frac{t}{L}} \tag{A.8}$$

avec

$$K = \frac{\omega_1 \cdot T}{\ln(\frac{\omega_2}{\omega_1})} \quad \text{et} \quad L = \frac{T}{\ln(\frac{\omega_2}{\omega_1})} \tag{A.9}$$

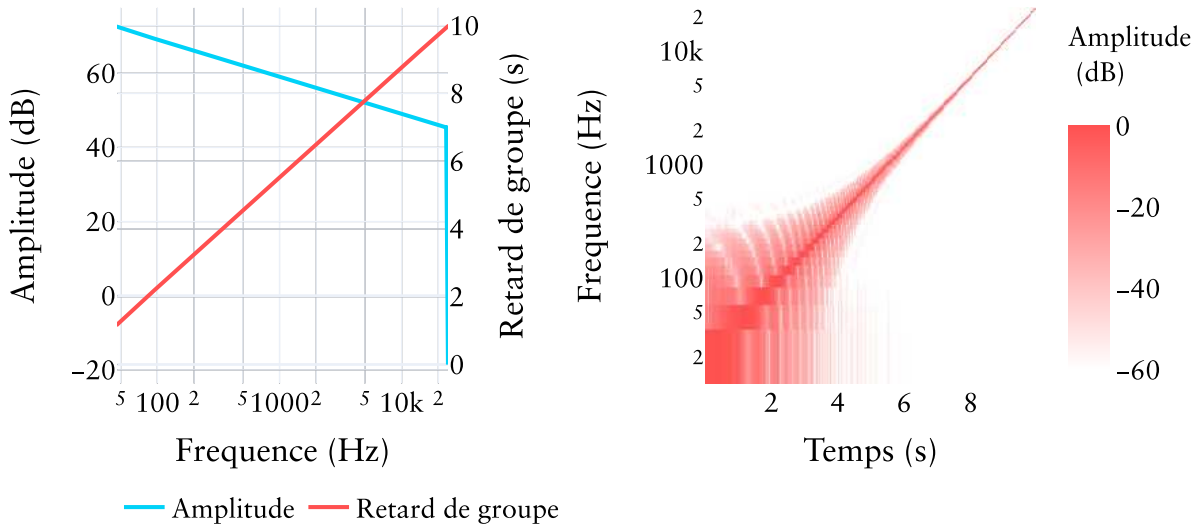


FIGURE A.1 – Réponse en fréquence, retard de groupe et spectrogramme d'un exemple de balayage sinusoïdal logarithmique.

La figure A.1 représente le spectre, le retard de groupe et le spectrogramme d'un tel signal. Les balayages logarithmiques ont un spectre dont l'amplitude diminue de 3 dB par octave si bien que chaque octave contient la même énergie. Lors de la déconvolution, une modulation d'amplitude est alors ajoutée pour compenser cette décroissance

en énergie. L'un des atouts de ce signal est que la précision temporelle en basses fréquences est plus grande, il correspond donc plus étroitement à la perception humaine du spectre des sons.

L'avantage prépondérant d'une mesure à balayage sinusoïdal est le fait que les composantes de distorsions harmoniques peuvent être entièrement isolées de la réponse impulsionnelle mesurée [3, 4]. Après déconvolution du signal mesuré par l'excitation d'entrée, ces distorsions apparaissent avant le son direct et peuvent ainsi être complètement effacées. Afin d'illustrer la propriété de rejet de la distorsion Muller et Massarini [4] prennent l'exemple d'un balayage qui contient une fréquence de 100 Hz après 2 s et atteint 200 Hz après 3 s. Le signal inverse utilisé pour la déconvolution va reconstituer le signal d'excitation en une impulsion de Dirac. Le spectre de ce signal contient un retard de groupe correspondant de -2 s à 100 Hz et de -3 s à 200 Hz. Lorsque la fréquence instantanée est de 100 Hz et que l'enceinte produit des harmoniques, une composante de 200 Hz avec le même retard que la fondamentale de 100 Hz sera présente dans la mesure. Cette composante de 200 Hz sera alors traitée lors de la déconvolution avec le retard de groupe de -3 s du spectre de référence à 200 Hz et apparaîtra donc à -1 s après le processus de déconvolution. De même, des harmoniques d'ordre supérieur apparaîtront encore plus en amont du son direct.

L'utilisation d'un balayage un peu plus long que la réverbération à mesurer permet d'exclure tous les produits de distorsions harmoniques, ne laissant pratiquement que le bruit de fond comme limite pour mesurer la réverbération sur une plage dynamique suffisante. Le rapport signal sur bruit est en effet très bon avec cette excitation car l'énergie est étirée sur une longue période, puis ramenée à une réponse courte avec la déconvolution.

La mesure obtenue dépend de la position de source et du récepteur, de leur réponse en fréquence - si elle n'est pas compensée - et de leur directivité. Elle dépend également de l'instant de mesure car deux mesures successives d'une réponse impulsionnelle sont toujours différentes en raison de la variabilité du milieu de propagation.

Au delà de la variation des conditions de mesure, des évènements sonores non-stationnaires de courte durée peuvent apparaitre. Afin de réduire leur influence, Massé et al. [5] proposent d'analyser les écarts d'énergie contenus dans plusieurs spectrogrammes issus de différentes répétitions de mesures par balayage sinusoïdal. Les déviations significatives par rapport au spectrogramme moyen sont alors remplacées dans la région temps-fréquence considérée par la magnitude moyenne calculée d'après les répétitions non-affectées.

A.2. Extension de la décroissance de l'intensité sonore

Malgré l'usage d'un signal d'entrée permettant d'obtenir un bon rapport signal sur bruit, les mesures de réponses impulsionnelles de salle contiennent inévitablement un bruit de fond. Lorsque les réponses impulsionnelles sont utilisées pour extraire des paramètres acoustiques ou dans le cadre d'une auralisation, ce bruit est préjudiciable dans la mesure où il peut biaiser les calculs de paramètres ou créer des artefacts temporels audibles. Il est alors nécessaire de s'affranchir de ce «plancher» de bruit en extrapolant la décroissance de l'intensité sonore en deçà du niveau du bruit. Étant

donné que le rapport signal sur bruit peut varier selon la fréquence, ces extrapolations sont effectuées par bande de fréquences.

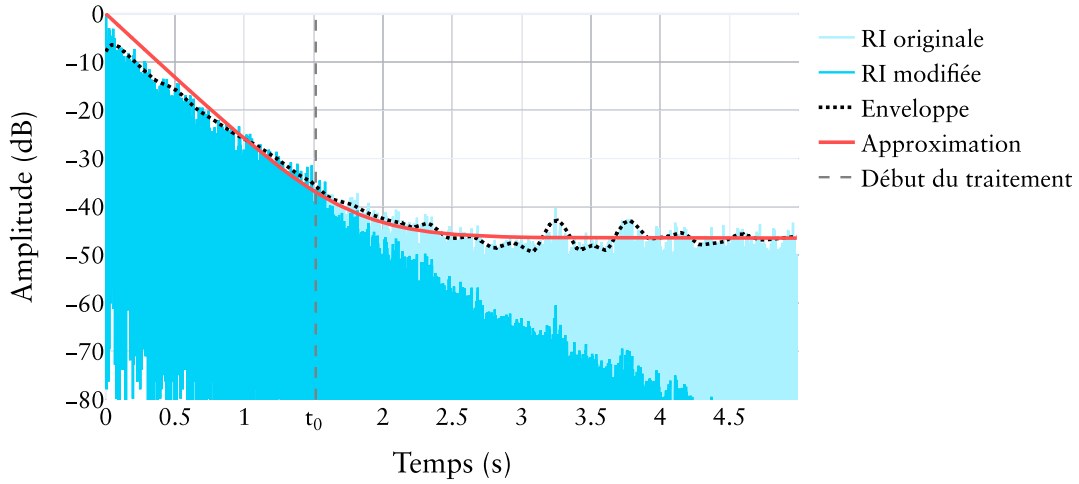


FIGURE A.2 – Exemple de traitement d'une réponse impulsionnelle mesurée pour la suppression du plancher de bruit. Les réponses impulsionnelles originale et modifiée sont identique jusqu'à t_0, la valeur temporelle pour laquelle le niveau de l'enveloppe approchée est supérieur à 10 dB du niveau de bruit.

Dans cette optique, Cabrera et al. [6] proposent une approche simple qui consiste à modéliser l'enveloppe de décroissance de la réponse imulsionnelle mesurée comme la somme d'une exponentielle décroissante et d'une constante de bruit. Afin d'obtenir l'enveloppe de décroissance, l'énergie de la réponse impulsionnelle dans la bande de fréquences considérée est lissée selon un filtre passe-bas dont la fréquence de coupure est fixée à 4 Hz. Une courbe c de la forme suivante est ensuite ajustée à l'enveloppe :

$$c(t) = 10 \log_{10} \left(10^{\frac{at}{10}} + b \right) \tag{A.10}$$

Le paramètre a contrôle la pente de décroissance et le paramètre b contrôle la constante de bruit dont le niveau asymptotique est $10 * \log_{10}(b)$. Une fois les paramètres a et b de la courbe identifiés, une fonction de compensation de gain peut être définie :

$$g(t) = \sqrt{\frac{10^{\frac{at}{10}}}{10^{\frac{at}{10}} + b}} \tag{A.11}$$

La fonction g est définie comme le rapport entre le niveau de l'enveloppe théorique de la décroissance exponentielle et le niveau de l'enveloppe modélisée d'après la mesure. Cette fonction de compensation est appliquée à l'enveloppe à partir de l'instant pour lequel le niveau d'énergie est 10 dB au dessus du niveau de bruit de fond. La figure A.2 illustre le résultat de cette méthode sur une réponse impulsionnelle mesurée.

La modélisation de l'enveloppe peut être problématique si le rapport signal sur bruit est trop faible, si le niveau de bruit n'est pas constant ou si la décroissance de l'intensité sonore ne correspond pas à celle d'une fonction exponentielle décroissante en raison, par exemple, de la présence d'un espace couplé lors de la mesure. De plus,

étant donné que la partie du signal constituant le bruit de fond est utilisée telle quelle pour le prolongement de l'enveloppe de décroissance, il est nécessaire que celui-ci ne soit pas spectralement coloré. Dans ce cas, l'usage d'un bruit de synthèse suivant la décroissance exponentielle a du modèle semble plus judicieux. Massé et al. [5] proposent notamment d'identifier le temps de mélange au moyen d'un estimateur de diffusion [7] afin de remplacer le plancher de bruit par une extension exponentiellement décroissante de la queue de réverbération. Cette synthèse est effectuée en employant un bruit gaussien généré dans le respect des caractéristiques spectrales et spatiales de la réverbération à traiter.

* * *

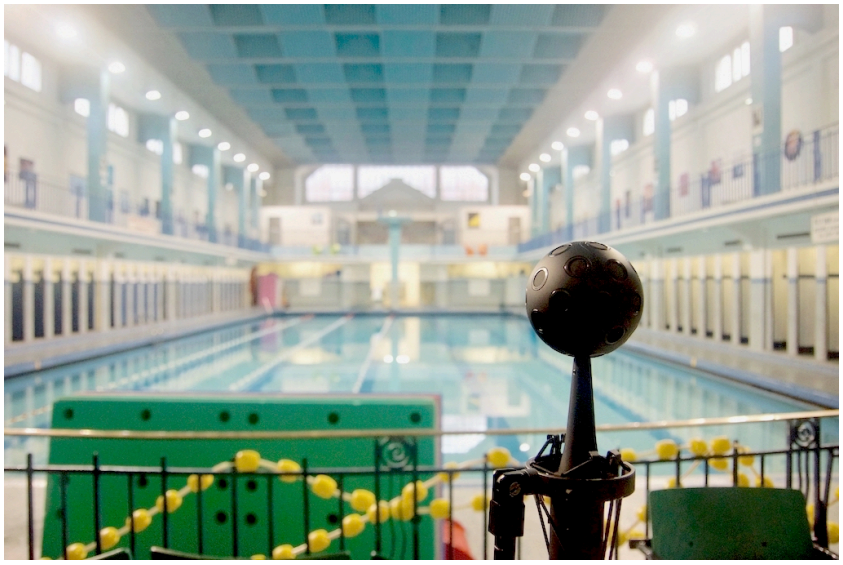


FIGURE A.3 – Le microphone sphérique Eigenmike EM32 dans la piscine Saint-Georges de Rennes. Avec l'aimable autorisation de Nicolas Épain.

En pratique, toutes les mesures de réponse impulsionnelle mentionnées dans ce document ont été mesurées avec le microphone sphérique Eigenmike EM32 [8] ($r_s = 4.2$ cm) et une enceinte Genelec 8040. Un signal à balayage sinusoïdal logarithmique de 10 secondes a été utilisé lors des mesures. Afin de compenser la réponse en fréquence de l'enceinte, un filtre à réponse impulsionnelle finie de 128 échantillons a été employé pour une correction fréquentielle comprise entre 60 Hz et 16 kHz. Ce filtre a été calculé d'après une mesure de la réponse impulsionnelle de l'enceinte effectuée dans une chambre anéchoïque avec un microphone omnidirectionnel situé dans l'axe de l'enceinte.

Bibliographie

[1] M. Vorländer, *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[2] B. A. Blesser, "An interdisciplinary synthesis of reverberation viewpoints," *Journal of the Audio Engineering Society*, vol. 49, n°. 10, p. 867–903, 2001.

[3] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” dans *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[4] S. Müller et P. Massarani, “Transfer-function measurement with sweeps,” *Journal of the Audio Engineering Society*, vol. 49, n°. 6, p. 443–471, 2001.

[5] P. Massé, T. Carpentier, O. Warusfel et M. Noisternig, “A robust denoising process for spatial room impulse responses with diffuse reverberation tails,” *The Journal of the Acoustical Society of America*, vol. 147, n°. 4, p. 2250–2260, 2020.

[6] D. Cabrera, D. Lee, M. Yadav et W. L. Martens, “Decay envelope manipulation of room impulse responses : Techniques for auralization and sonification,” dans *Proceedings of Acoustics*, 2011.

[7] N. Epain et C. T. Jin, “Spherical harmonic signal covariance and sound field diffuseness,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, n°. 10, p. 1796–1807, 2016.

[8] MH Acoustics LLC, “Eigenmike em32 microphone array,” accessed 2020-08-22. [En ligne]. Disponible : https://mhacoustics.com/products

B

La représentation spatiale d'une réponse impulsionnelle multicanale

En employant la méthode de mesure présentée dans la section précédente, il est possible de caractériser le contenu temporel et fréquentiel de la réverbération en un point de l'espace. Cependant pour être capable d'analyser les directions d'incidences des multiples réflexions ou de les spatialiser pour reproduire la réverbération sur un dispositif admettant plus d'une enceinte, plusieurs mesures doivent être effectuées selon différents secteurs angulaires. Un ensemble de réponses impulsionnelles mesurées en direction de plusieurs secteurs angulaires fournissent une représentation spatiale du champ réverbéré que nous nommerons SRIR (*Spatial Room Impulse Response*) dans la suite du document. Selon le nombre de mesures effectuées, il est possible d'avoir une description plus ou moins précise de la spatialisation des réflexions qui constituent la réverbération. Nous introduisons dans cette section la représentation de la réverbération au format ambisonique, qui présente l'avantage de décrire la composante spatiale du champ sonore indépendamment du système d'acquisition et de restitution. De plus, l'usage de ce format permet la reproduction d'un champ sonore à la position de la tête d'un auditeur [1, 2] et de prendre en compte efficacement les mouvements de sa tête [3] ce qui présente un avantage en terme de réalisme.

B.1. Système de coordonnées spatiales adopté

Dans ce document, un point de l'espace $P(x, y, z)$ est décrit en coordonnées sphériques selon un rayon $r \in [0, \infty[$, un azimut $\theta \in [0, 2\pi[$ et une élévation $\varphi \in [-\pi/2, \pi/2]$ d'après le système de coordonnées suivant illustré en figure B.1 :

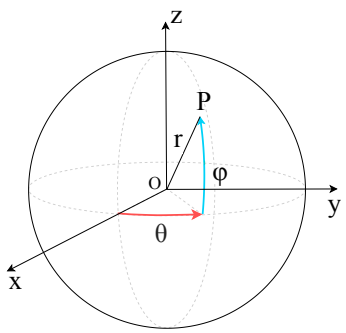


FIGURE B.1 – Système de coordonnées sphérique. Un point P de l'espace est représenté par son rayon r, son azimut θ et son élévation φ.

$$\begin{cases} x = r\cos(\theta)\cos(\varphi) \\ y = r\sin(\theta)\cos(\varphi) \\ z = r\sin(\varphi) \end{cases} \tag{B.1}$$

B.2. Décomposition en harmoniques sphériques

Dans le système de coordonnées sphériques ainsi défini, tout champ sonore constitué d'ondes sonores incidentes peut être décomposé sur une base orthogonale de $\mathbb{L}^2$ [1] formée par des fonctions nommées harmoniques sphériques. La pression sonore au point de mesure (r, θ, φ) peut être entièrement décrite dans le domaine fréquentiel par une série d'harmoniques sphériques [4] :

$$p(k, r, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} p_{l,m}(k, r) Y_{l,m}(\theta, \varphi), \tag{B.2}$$

où

$$p_{l,m}(k, r) = i^l j_l(kr) b_{l,m}(k), \tag{B.3}$$

avec k le nombre d'onde, $j_l(kr)$ la fonction de Bessel sphérique d'ordre l et $Y_{l,m}(\theta, \varphi)$ l'harmonique sphérique à valeur réelle d'ordre l et de degré m. Les coefficients $b_{l,m}(k)$ sont nommés composantes ambisoniques. Les coefficients $p_{l,m}$ représentent les coefficients de la transformée de Fourier sphérique issus de la projection de la pression sonore sur la base formée par les harmoniques sphériques. Ces fonctions sont définies par [4] :

$$Y_{l,m}(\theta, \varphi) = A_{l,m} P_{l,|m|}(\sin(\varphi)) \begin{cases} \cos(|m|\theta) & \text{si } m \geqslant 0 \\ \sin(|m|\theta) & \text{si } m < 0 \end{cases} \tag{B.4}$$

1. $\mathbb{L}^2$ désigne l'espace hilbertien des fonctions à valeurs réelles de carré intégrable, c'est à dire dont le carré du module est sommable, sur la sphère unité $\Omega = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$.

où

$$A_{l,m} = \sqrt{(2l+1)\epsilon_m \frac{(l-|m|)!}{(l+|m|)!}}, \tag{B.5}$$

$P_{l|m|}$ sont les polynômes associés de Legendre d'ordre l et de degré $|m|$ avec $m \leq |l|$, $(l, m) \in (\mathbb{N}, \mathbb{Z})$ et $\epsilon_m = 1$ si $m = 0$, $\epsilon_m = 2$ si $|m| > 0$. La normalisation des harmoniques sphériques contenue dans l'équation (B.5) correspond à une normalisation dite N3D [2], adoptée dans la suite de ce document. Les harmoniques sphériques ainsi définies sont de norme unitaire et forment donc une base orthonormée.

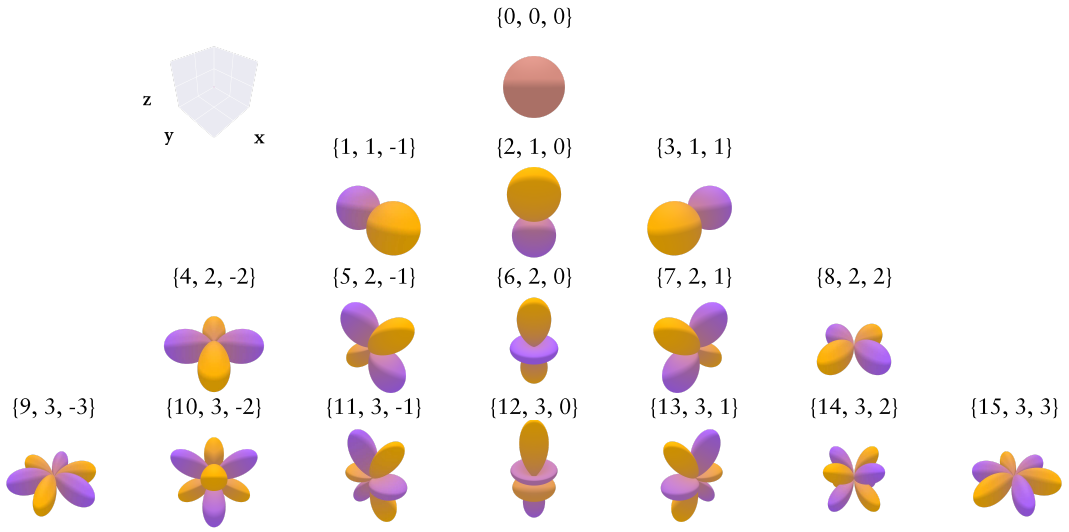


FIGURE B.2 – Diagramme de directivité des harmoniques sphériques jusqu'à l'ordre $L = 3$. Les valeurs positives et négatives sont respectivement représentées par les couleurs orange et bleue. L'indice ACN i, l'ordre l et le degré m des harmoniques sphériques sont indiqués par les ensembles $\{i, l, m\}$.

Les harmoniques sphériques de degré $|m| = l$ sont nommées harmoniques sphériques sectorielles. Pour ces harmoniques, quelque soit l'azimut, la phase ne change pas de signe selon l'élévation. Les harmoniques sphériques de degré 0 sont dites zonales (leur valeur ne varie pas selon l'azimut). Les autres sont nommées harmoniques sphériques tessérales.

En pratique, la décomposition en harmoniques sphériques d'un champ acoustique en un point de l'espace est tronquée. Elle est obtenue en conservant les composantes ambisoniques $b_{l,m}(k)$ jusqu'à un ordre l maximal noté L et appelé par la suite ordre ambisonique. Le nombre N de composantes ambisoniques impliquées dans la représentation du champ sonore dans toutes les directions de l'espace dépend de l'ordre ambisonique :

$$N = (L+1)^2. \tag{B.6}$$

Les composantes ambisoniques qui résultent de la décomposition temporelle du champ sonore fournissent une représentation au format dit ambisonique - également

nommé ambisonique d'ordre élevé (*Higher Order Ambisonics*) pour les ordres supérieurs à 1. Elles forment un vecteur $\boldsymbol{b} = [b_0, ..., b_i, ..., b_{N-1}]^t$ où b_i représente la i^{me} composante ambisonique et ^{t} la transposée. La concaténation des vecteurs de composantes ambisoniques par pas de temps forme une matrice de signaux ambisoniques. L'ordonnancement des composantes adopté dans ce document est le *Ambisonic Channel Order* (ACN) dont la notation permet d'identifier une harmonique sphérique selon un unique indice i fonction de l'ordre et du degré correspondants : $i = l^2+l+m$ [5]. La figure B.2 représente le diagramme de directivité des harmoniques sphériques jusqu'à l'ordre 3 selon cet ordonnancement.

B.3. Erreur de troncature

La précision physique de la représentation du champ sonore dépend de l'ordre ambisonique utilisé pour la décomposition. Plus l'ordre est élevé, plus la zone dans laquelle cette description est précise est grande. La troncature de la représentation du champ sonore mène à une erreur d'approximation ϵ_L entre le champ de pression théorique p et le champ de pression tronquée à l'ordre L, p_L :

$$\epsilon_L = \frac{\displaystyle\int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |p(k,r,\theta,\varphi) - p_L(k,r,\theta,\varphi)|^2 \, d\varphi d\theta}{\displaystyle\int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |p(k,r,\theta,\varphi)|^2 d\varphi d\theta} \tag{B.7}$$

D'après le calcul de cette erreur, Ward et Abhayapala [6] et Poletti [7] - dans le cas des ondes planes et sphériques respectivement - ont établit que l'ordre minimal fournissant une représentation du champ sonore avec une erreur ϵ_L de 4% dans une sphere de rayon r est estimé par la formule suivante :

$$L = \lceil kr \rceil \tag{B.8}$$

où $\lceil . \rceil$ représente l'arrondi à l'entier supérieur. Ainsi, pour un rayon r donné, plus l'ordre ambisonique augmente, plus la fréquence jusqu'à laquelle la représentation du champ sonore est correcte augmente. Par exemple, la représentation du champ sonore dans une sphère de rayon $r = 8.5$ cm (le rayon approximatif d'une tête humaine) et selon un ordre $L = 4$, est jugée satisfaisante jusqu'à une fréquence $f = 2561$ Hz environ. Pour un ordre $L = 3$, la représentation du champ sonore dans cette sphère est jugée satisfaisante jusqu'à une fréquence $f = 1921$ Hz environ. Pour un rayon r de représentation et une erreur moyenne fixes, chaque ordre de la représentation possède donc une bande de fréquence utile : la contribution des différents ordres harmoniques n'est pas égale à toutes les fréquences. En particulier, les ordres élevés sont moins utiles en basse fréquence qu'en haute fréquence.

B.4. Captation et encodage d'un champ sonore

Il est possible d'estimer les composantes ambisoniques d'un champ sonore au moyen d'une antenne sphérique de microphones. En effet, la connaissance de la pression sonore à la surface d'une sphère de rayon r_s en un nombre fini de positions permet de calculer les composantes ambisoniques pour un ordre ambisonique donné.

Soit un ensemble de Q microphones situés sur un surface sphérique aux positions angulaires (θ_q, φ_q), $q \in [1; Q]$. Le développement en harmonique sphérique de la pression sonore $p\left(kr_s, \theta_q, \varphi_q\right)$ captée par le q^{ime} microphone s'écrit :

$$p\left(k, r_s, \theta_q, \varphi_q\right) = \sum_{l=0}^{L} \sum_{m=-l}^{l} W_l(kr_s) b_{l,m}(k) Y_{l,m}\left(\theta_q, \varphi_q\right) \tag{B.9}$$

où $W_l(kr_s)$ est une fonction de pondération dont l'expression dépend de la directivité des microphones employés, de la géométrie et des matériaux constituant l'antenne.

Le nombre de microphones utilisés ainsi que leur position sur la sphère dépend du nombre de composantes ambisoniques à estimer. Pour réaliser une estimation correcte de ces composantes, il est nécessaire que le nombre de capteurs soit supérieur ou égal au nombre de composantes à estimer et préférable qu'ils soient uniformément répartis sur la sphère pour échantillonner au mieux le champ sonore dans l'espace.

B.4.1. L'encodage des signaux captés

Pour réaliser un encodage les signaux $\boldsymbol{p}$ mesurés par les capteurs, une pratique courante consiste à résoudre le système d'équations linéaires suivant :

$$\boldsymbol{p}(k) = \boldsymbol{T}\boldsymbol{b}(k) \tag{B.10}$$

avec

$$\boldsymbol{T} = \boldsymbol{Y}\,\mathrm{diag}\left[W_l(kr_s)\right], \tag{B.11}$$

$$\boldsymbol{b}(k) = \left[b_{0,0}(k)\ b_{1,-1}(k)\ \ldots\ b_{L,L}(k)\right]^t, \tag{B.12}$$

et

$$\boldsymbol{p}(k) = \left[p\left(k, r_s, \theta_0, \varphi_0\right)\ p\left(k, r_s, \theta_1, \varphi_1\right)\ \ldots\ p\left(k, r_s, \theta_Q, \varphi_Q\right)\right]^t \tag{B.13}$$

où k est le nombre d'onde et $\boldsymbol{Y}$ correspond à la matrice des harmoniques sphériques pour les positions angulaires (θ_q, φ_q), $q \in [1; Q]$ des microphones :

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}(\theta_1, \varphi_1) \\ \boldsymbol{y}(\theta_2, \varphi_2) \\ \vdots \\ \boldsymbol{y}(\theta_Q, \varphi_Q) \end{pmatrix} \tag{B.14}$$

avec

$$\boldsymbol{y}(\theta_q, \varphi_q) = [Y_{0,0}(\theta_q, \varphi_q)\ Y_{1,-1}(\theta_q, \varphi_q)\ \ldots\ Y_{L,L}(\theta_q, \varphi_q)]\ . \tag{B.15}$$

Les composantes ambisoniques $\tilde{\boldsymbol{b}}$ à estimer grâce à l'antenne de microphones correspondent à la solution minimisant le résidu quadratique $\|\boldsymbol{p} - \boldsymbol{T} \cdot \boldsymbol{b}\|_2^2$. La résolution de ce problème d'optimisation aux moindres carrés admet pour solution :

$$\tilde{\boldsymbol{b}}(k) = \mathrm{diag}\,[EQ_l(kr_s)]\,\boldsymbol{Y}^{\dagger}\boldsymbol{p}(k) \tag{B.16}$$

où $\boldsymbol{Y}^{\dagger}$ est la pseudo-inverse de $\boldsymbol{Y}$ et le filtre d'égalisation $EQ_l = 1/W_l$. Le calcul de ces composantes correspond à l'encodage ambisonique.

La figure B.3 représente en pointillés les gains d'amplification des filtres d'égalisation EQ_l ainsi définis selon plusieurs ordres. Avec cette solution les gains d'amplification sont très importants pour les ordres strictement positifs et d'autant plus importants que la fréquence est basse et le rayon du microphone petit. Ce phénomène est présent en basses fréquences en raison de la taille de l'antenne sphérique trop petite par rapport à leur longueur d'onde. En effet, dans le but d'obtenir des informations précises sur la direction des ondes incidentes à ces fréquences, il est nécessaire d'appliquer une forte amplification pour percevoir des différences entre les signaux captés par les microphones. Or plus les microphones sont proches, plus l'information de propagation est difficile à obtenir. Des amplifications excessives peuvent mener à de fortes instabilités car le bruit de mesure devient considérablement amplifié et vient perturber l'estimation des composantes ambisoniques.

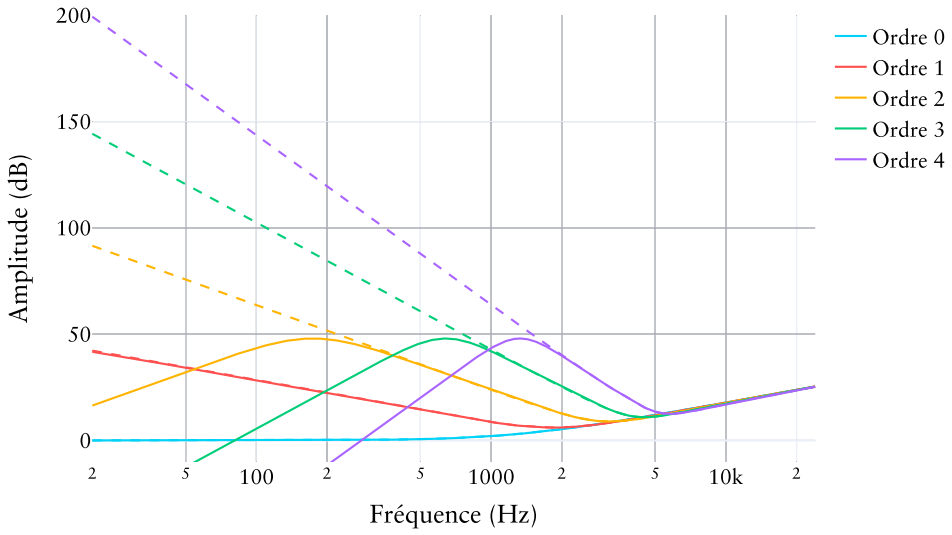


FIGURE B.3 – Gain des filtres d'égalisation pour une sphère rigide de rayon r_s = 4.2 cm aux ordres 0 à 4. Les gains théoriques EQ_l apparaissent en pointillés et les gains des filtres régularisés $EQ_{l,\lambda}$ apparaissent en traits pleins avec un paramètre de régularisation $\lambda = 2\mathrm{e}{-3}$.

Moreau et al. [8] proposent alors d'estimer les composantes ambisoniques grâce à une régularisation de Thikhonov. Les filtres d'égalisation régularisés sont donnés par l'équation suivante :

$$E_{l,\lambda}(kr_s) = \frac{(1/E_l(kr_s))^*}{|1/E_l(kr_s)|^2 + \lambda^2} \tag{B.17}$$

où λ est un paramètre de régularisation compris entre 0 et 1. Ce paramètre peut être fixé en fonction de l'amplification maximale souhaitée a :

$$\lambda = \frac{1 - \sqrt{1 - 1/a^2}}{1 + \sqrt{1 - 1/a^2}}, \tag{B.18}$$

avec $a \geq 1$. On peut faire en sorte que le gain d'amplification du filtre régularisé pour un ordre donné corresponde au gain d'amplification du filtre théorique à partir d'une certaine fréquence. Cette fréquence peut être fixée en cohérence avec le concept de bande fréquentielle utile présentée en annexe B.3.

B.4.2. Fréquence de repliement spatial

L'usage d'un nombre fini de microphone à la surface de la sphère entraîne un autre inconvénient : un phénomène de repliement spatial. En effet, la représentation spatiale du champ sonore est réalisable jusqu'à une certaine fréquence en raison de la taille de l'antenne sphérique. Cette fréquence maximale dépend de la distance entre un capteur et ses plus proches voisins : l'intervalle d'échantillonnage spatial. Selon le critère de Shannon, cette distance doit être inférieure à la moitié de la plus petite longueur d'onde à estimer. Au delà de la fréquence maximale, appelée fréquence de repliement, l'estimation des composantes ambisoniques est erronée. Cette fréquence est donnée par la formule suivante :

$$f_{\text{re}} = \frac{c}{2r_s\gamma} \tag{B.19}$$

où c est la célérité du son dans l'air et γ l'angle maximal entre deux capsules. Par exemple pour un rayon $r_s = 4.2$ cm et un angle $\gamma = 37.6°$, la fréquence de repliement spatial est $f_{re} = 6168$ Hz. Il est possible d'augmenter cette fréquence en diminuant les distances entre les microphones, soit en augmentant leur nombre, soit en réduisant le rayon de l'antenne sphérique (au risque de détériorer les performances de l'antenne à basse fréquence).

L'emploi d'une antenne sphérique de microphones permet de mesurer des réponses impulsionnelles dans différentes directions de l'espace selon la procédure présentée en annexe A. L'ensemble des ces réponses impulsionnelles forme la réponse impulsionnelle spatiale d'une salle : la SRIR. Afin d'acquérir la représentation ambisonique de la SRIR, l'encodage ambisonique s'applique aux réponses impulsionnelles mesurées par chacun des microphones de l'antenne. Notons qu'en raison de la géométrie et du nombre de microphones constituant l'antenne, la représentation ambisonique est limitée à un ordre fini et n'est valable que dans une plage de fréquence donnée autour de la tête d'un auditeur.

Bibliographie

[1] M. A. Gerzon, “Ambisonics in multichannel broadcasting and video,” *Journal of the Audio Engineering Society*, vol. 33, n°. 11, p. 859–871, 1985.

[2] J. Daniel, “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia,” Thèse de doctorat, University of Paris VI, 2000.

[3] J.-M. Jot, V. Larcher et J.-M. Pernaux, “A comparative study of 3-d audio encoding and rendering techniques,” dans *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[4] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[5] C. Nachbar, F. Zotter, E. Deleflie et A. Sontacchi, “Ambix-a suggested ambisonics format,” dans *Ambisonics Symposium, Lexington*, 2011, p. 11.

[6] D. B. Ward et T. D. Abhayapala, “Reproduction of a plane-wave sound field using an array of loudspeakers,” *IEEE Transactions on speech and audio processing*, vol. 9, n°. 6, p. 697–707, 2001.

[7] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *Journal of the Audio Engineering Society*, vol. 53, n°. 11, p. 1004–1025, 2005.

[8] S. Moreau, J. Daniel et S. Bertet, “3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone,” dans *120th Convention of the AES*, 2006, p. 20–23.

C

L'écoute binaurale dynamique d'une scène sonore réverbérée

Une réponse impulsionnelle spatiale de salle (SRIR) encodée en ambisonique peut être utilisée pour créer une scène sonore réverbérée contenant une source, en réalisant le filtrage du signal anéchoïque de cette source par la SRIR. La scène sonore résultante peut ensuite être décodée sur différents systèmes de restitution. Dans ce document, le seul système de restitution utilisé est le casque d'écoute.

C.1. Fonctions de transfert relatives à la tête

La localisation d'une source sonore dans l'espace repose sur la perception de plusieurs indices qui résultent des déformations subies par le son depuis la source jusqu'aux tympans lorsqu'il rencontre le torse, la tête et les oreilles de l'auditeur. Parmi ces indices figurent la différence interaurale de temps notée ITD (pour *Interaural Time Difference*) et la différence interaurale de niveau notée ILD (pour *Interaural Level Difference*). Rayleigh [1] a déterminé que l'ITD était un indice prépondérant dans les basses fréquences et l'ILD dans les fréquences plus élevées ; bien que les gammes de fréquences correspondantes se chevauchent. Les seules informations d'ILD et d'ITD ne suffisent cependant pas à identifier la direction d'incidence d'une source. Les régions de l'espace produisant des valeurs similaires à la fois d'ITD et d'ILD forment des «cônes de confusion» illustré par la figure C.1. Ces informations n'étant pas suffisantes, d'autres indices tels que les indices spectraux permettent de résoudre ces ambiguïtés liées à la perception de la direction d'incidence [2].

Pour une direction donnée, la propagation du son depuis une source jusqu'au tympan peut être modélisée par un système linéaire et invariant dans le temps. On peut donc caractériser cette propagation dans le domaine temporel par une réponse impulsionnelle relative à la tête notée HRIR (pour *Head Related Impulse Response*) ou dans le domaine fréquentiel par une fonction de transfert relative à la tête notée

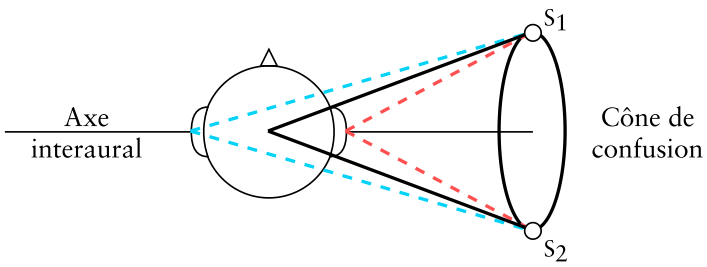


FIGURE C.1 – Illustration d'un cône de confusion. Les sources S_1 et S_2 fournissent les mêmes valeurs d'ITD et d'ILD.

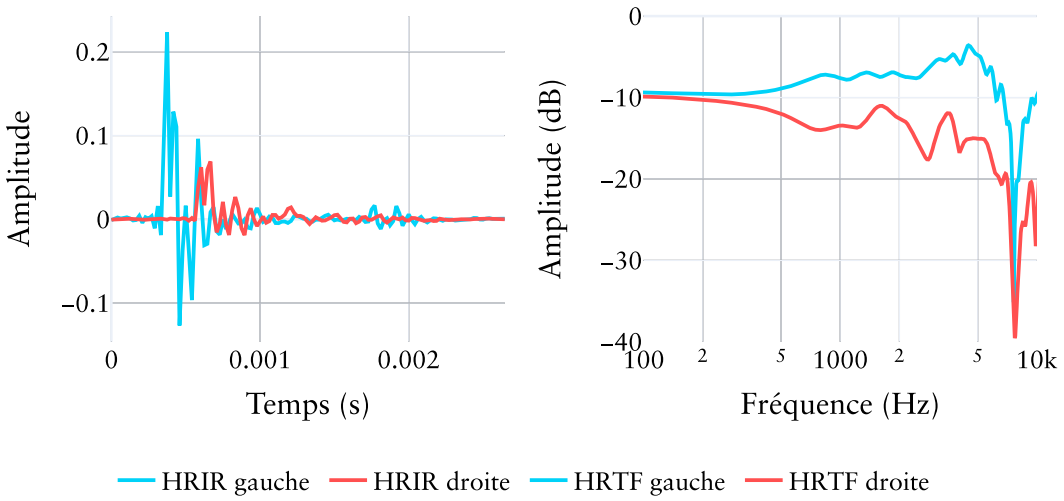


FIGURE C.2 – HRIRs et HRTFs gauche et droite mesurées pour une onde incidente d'azimut $\theta = 30°$ et d'élévation $\varphi = 0°$. On distingue la différence de niveau ainsi que la différence de temps d'arrivée de l'onde sonore entre l'oreille gauche plus proche (ipsilatérale) et l'oreille droite plus éloignée (controlatérale). La représentation fréquentielle fait figurer les différentes régions spectrales atténuées lors de la propagation de l'onde jusqu'aux deux oreilles.

HRTF (pour *Head Related Transfer Function*). Ces filtres comprennent *a fortiori* les différences interaurales de niveau, de temps et les indices spectraux. Les HRTFs sont employées par paire : pour une position de source donnée, la propagation du son est caractérisée pour chaque oreille. Une exemple de paire d'HRTFs est représenté dans la figure C.2. Ces HRTFs sont propres à chaque individu en raison de morphologies différentes, sources de déformations, déphasages et diffractions diverses du son. Ainsi un rendu sonore binaural est généralement obtenu par convolution entre un signal monophonique et des HRIRs mesurées ou modélisées.

Idéalement, les filtres de binauralisation devraient être personnalisés pour chaque auditeur, cependant mesurer les HRTFs est un processus fastidieux et complexe [3]. Une solution communément utilisée est d'écouter «à travers» les oreilles d'un autre auditeur, dont les HRTFs sont disponibles, ou d'une tête artificielle. L'utilisation d'HRTFs non-individualisées peut accroître les occurrences de confusions avant-arrière [4] ou produire des localisations intra-crâniennes [5, 6]. Néanmoins, ces problèmes sont efficacement atténués par la présence de réverbération dans les signaux

sonores [7, 8] et grâce à un dispositif de suivi des mouvements de tête [9, 10] (cf annexe C.3).

C.2. Calcul des filtres de décodage binaural

Le rendu binaural de signaux ambisoniques est couramment réalisé 1) en décodant les signaux vers des haut-parleurs virtuels, 2) en convoluant les signaux des haut-parleurs virtuels avec les HRIR des directions correspondantes, 3) en sommant les signaux convolués [11, 12]. Les résultats obtenus avec cette approche dépendent cependant du type de décodage employé pour construire les signaux des haut-parleurs virtuels et de la configuration spatiale de ces haut-parleurs. De récents travaux ont alors proposé d'effectuer le décodage binaural sans l'étape intermédiaire du décodage vers une configuration de haut-parleurs virtuels [13–16].

La formulation du problème peut être décrite comme suit. Soit une onde plane dont la forme d'onde mesurée au centre du repère est s et dont la direction d'incidence est (θ_p, φ_p), la pression générée par cette onde donnée au point $\mathrm{P}(r, \theta, \varphi)$ par :

$$p(k, r, \theta, \varphi) = s(k) e^{-ikr\cos(\gamma)} \tag{C.1}$$

où γ correspond à l'angle entre les directions (θ_p, φ_p) et (θ, φ). La projection de la pression sonore sur la base d'harmonique sphérique telle que définie à l'équation (B.2) fournit les composantes ambisoniques d'ordre L, $\boldsymbol{b}(k)$, correspondantes [17] :

$$\boldsymbol{b}(k) = s(k)\boldsymbol{y}(\theta_p, \varphi_p) \ , \tag{C.2}$$

où $\boldsymbol{y}(\theta_p, \varphi_p)$ est le vecteur contenant les valeurs des harmoniques sphériques pour la direction (θ_p, φ_p) jusqu'à l'ordre L. Soit $h(k, \theta_p, \varphi_p)$ la fonction de transfert en champ lointain relative à la tête. Le signal binaural cible est donné par :

$$x(k) = s(k)h(k, \theta_p, \varphi_p) \ . \tag{C.3}$$

Le but est de trouver les filtres de décodage $\boldsymbol{w}(k)$ qui conduisent au signal binaural $\hat{x}(k)$ qui est le plus proche possible de $x(k)$ pour toute direction :

$$\hat{x}(k) = \boldsymbol{w}^T(k)\, \boldsymbol{b}(k) \tag{C.4}$$

La fonction de coût qui modélise la dissemblance perçue entre les signaux x et $\hat{x}$ doit alors être minimisée pour un ensemble dense de directions. Cependant, Bernschutz et al. [13] ont montré que le calcul de ces filtres par la minimisation de l'erreur quadratique entre x et $\hat{x}$ engendre une altération des différences interaurales et d'importants artefacts spectraux qui dépendent de la direction d'incidence de la source, pour les ordres ambisoniques usuels des composantes $\boldsymbol{b}$.

Diverses méthodes ont alors été explorées afin d'améliorer la qualité spectrale des filtres de décodage en utilisant des filtres d'égalisation [14, 18] ou en ré-échantillonnant l'ensemble des directions pour lesquelles l'optimisation est effectuée [13]. Une méthode récente permettant de réduire de manière significative les artefacts spectraux

consiste à pré-traiter les HRIRs en vue de l'étape d'optimisation [16]. Dans un premier temps, l'alignement temporel en fonction de la fréquence est appliqué de manière à ce que les HRIRs soient alignées temporellement en hautes fréquences tandis que les différences interaurales de temps sont maintenues aux basses fréquences. Pour différents types de signaux, il a été démontré que cette modification de phase à haute fréquence peut être effectuée au-dessus d'une fréquence de coupure $f_c = 2$ kHz [16]. Dans un second temps, l'optimisation est effectuée en minimisant l'erreur quadratique entre x et $\hat{x}$ pour un ensemble dense de directions en basse fréquence tandis que seule l'erreur des magnitudes est minimisée à haute fréquence pour cet ensemble de directions. Les filtres de décodage optimisés, $\hat{\boldsymbol{w}}(k)$, doivent satisfaire autant que possible la solution au moindre carré du système d'équations suivant :

$$\hat{\boldsymbol{w}}(k) = \begin{cases} \underset{\boldsymbol{w}}{\operatorname{argmin}} \| \boldsymbol{Y}.\boldsymbol{w}(k) - \boldsymbol{h}(k)\|_2^2 & \text{si } k < k_c \\ \underset{\boldsymbol{w}}{\operatorname{argmin}} \| \, |\boldsymbol{Y}.\boldsymbol{w}(k)| - |\boldsymbol{h}(k)|\|_2^2 & \text{sinon,} \end{cases} \tag{C.5}$$

où k_c est le nombre d'onde à f_c, $\boldsymbol{Y}$ est la matrice des valeurs des harmoniques sphériques pour un ensemble de M directions :

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}(\theta_1, \varphi_1) \\ \boldsymbol{y}(\theta_2, \varphi_2) \\ \vdots \\ \boldsymbol{y}(\theta_M, \varphi_M) \end{pmatrix} \tag{C.6}$$

et $\boldsymbol{h}(k)$ est le vecteur des valeurs d'HRTFs correspondantes :

$$\boldsymbol{h}(k) = [h(k, \theta_1, \varphi_1) \; h(k, \theta_2, \varphi_2) \; \ldots \; h(k, \theta_M, \varphi_M)]^T \; . \tag{C.7}$$

Cependant, bien que l'alignement temporel des HRTFs à haute fréquence améliore la reconstruction de la magnitude des HRTF, il tend également à produire des signaux binauraux qui sont anormalement corrélés entre eux. En particulier, les filtres de décodage binaural d'ordre faible fournissent des signaux binauraux trop corrélés au dessus de la fréquence de coupure f_c. Afin de contourner ce problème, Zaunschirm *et al.* [15] ont proposé d'ajouter une étape finale à la méthode décrite ci-dessus, dans laquelle les filtres de décodage binaural sont ajustés de manière à ce que la corrélation interaurale des signaux binauraux corresponde à celle des HRTFs d'origine. Cette méthode applique une contrainte sur la corrélation interaurale des filtres de décodage binaural qui au dessus de la fréquence de coupure f_c doit correspondre à celle calculée d'après les HRTFs mesurées dans une configuration en champ diffus. Cette correction améliore la décorrélation des signaux binauraux pour les ordres 1 et 2 essentiellement.

La figure C.3 représente les diagrammes de directivité des filtres de décodage binaural calculés pour une oreille et permet de rendre compte de la précision spatiale

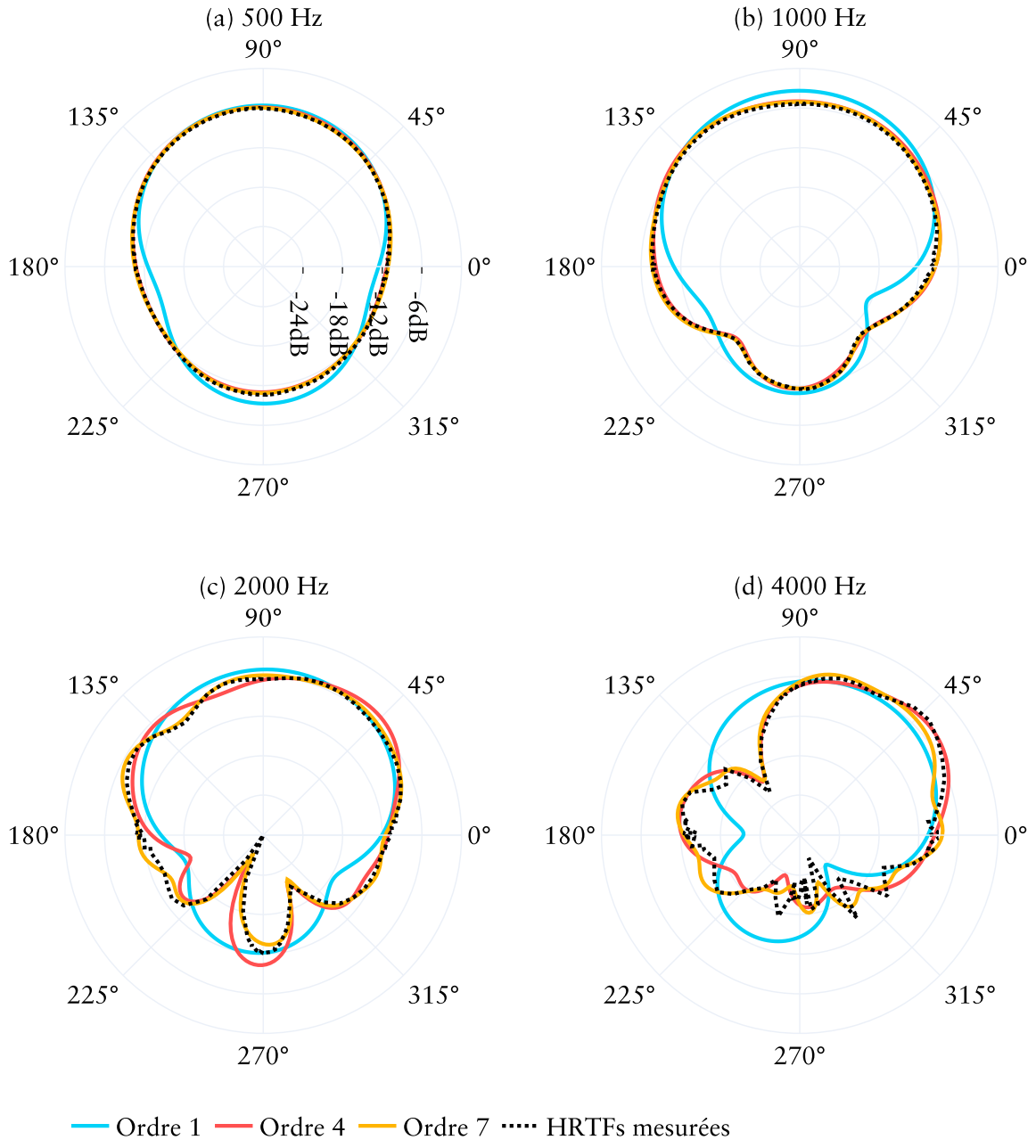


FIGURE C.3 – Diagrammes de directivité des filtres de décodage binaural aux ordres 1, 4 et 7 pour des fréquences de (a) 500 Hz, (b) 1000 Hz, (c) 2000 Hz, (d) 4000 Hz,.

des filtres ainsi calculés à différentes fréquences en comparaison aux HRTFs mesurées. La figure C.4 représente le spectre et la décorrélation en fonction de la fréquence des signaux binauraux calculés d'après les HRTFs mesurées et d'après les filtres de décodage aux ordres 1, 4 et 7, après correction. On constate que les variations spectrales hautes fréquences sont contenues grâce à l'opération de l'équation (C.5) qui permet de respecter le spectre de référence et que la décorrélation suit le même profil de décroissance en fonction de la fréquence que les HRTFs mesurées grâce à la correction proposée par Zaunschirm *et al.* [15].

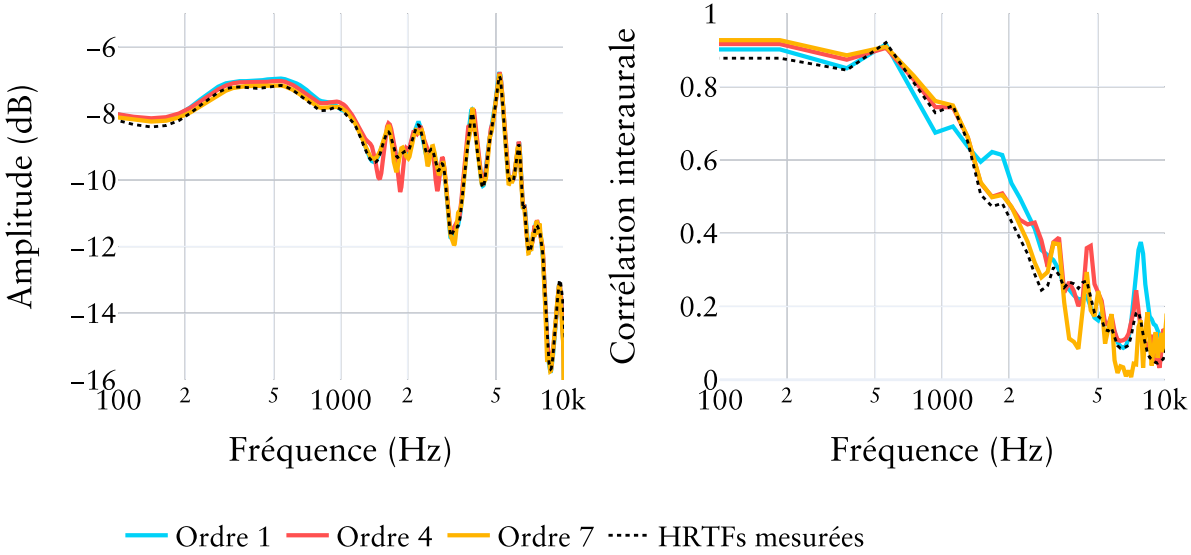


FIGURE C.4 – Spectres moyennés sur les composantes ambisonique et corrélation croisée des filtres de décodage binaural calculés aux ordres 1, 4 et 7.

C.3. Prise en compte des mouvements de la tête

Le rendu binaural dynamique désigne une restitution au casque pour laquelle la rotation de la tête de l'auditeur est prise en compte. L'utilisation d'un dispositif de suivi des mouvements de la tête permet de s'approcher d'une situation d'écoute naturelle et présente l'avantage de réduire à la fois les ambiguïtés de localisation et d'améliorer l'externalisation - c'est à dire la capacité à percevoir des événement auditifs en dehors de la tête en écoute binaurale [7, 9, 19].

La prise en compte des mouvements de la tête se traduit par une compensation des mouvement en appliquant une rotation inverse au champ sonore. Pour un rendu binaural dynamique utilisant directement les HRIRs pour spatialiser un signal monophonique (approche dite objet), la rotation implique d'effectuer un fondu ou commutation entre les filtres pour éviter l'apparition d'artefacts. Dans le domaine ambisonique, la prise en compte des mouvements de tête est plus efficace car la rotation de la scène sonore s'effectue par un simple produit matriciel entre les signaux ambisoniques et une matrice de rotation [12, 20].

La rotation d'un harmonique sphérique d'ordre l résulte en une combinaison linéaire d'harmoniques sphériques de même ordre l. La matrice de rotation des signaux ambisoniques prend donc la forme d'une matrice diagonale par bloc :

$$\boldsymbol{R} = \begin{bmatrix} 1 & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{0} & \boldsymbol{R}_1 & \boldsymbol{0} & \dots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}_2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{C.8}$$

où chaque bloc $\boldsymbol{R}_l$ est de dimension $(2m+1) \times (2m+1)$. Le calcul de la matrice $\boldsymbol{R}$ pour une rotation quelconque selon les axes x, y, z et un ordre ambisonique souhaité n'est cependant pas trivial. Il est néanmoins possible de calculer cette matrice par récurrence [21]. Zotter propose également de décomposer la rotation $\boldsymbol{R}$ en plusieurs matrices de rotations fixes de $\frac{\pi}{2}$ autour de l'axe y et de rotations variables autour de l'axe z pour lesquelles les expressions analytiques sont simples [22].

Il est également possible d'exprimer $\boldsymbol{R}$ comme la combinaison de deux opérations :

1. une décomposition en ondes planes d'ordre L vers un ensemble de $N \geq (L+1)^2$ directions ;
2. un encodage des signaux résultants selon ce même ensemble de directions auquel a été appliquée une matrice de rotation.

Soient la matrice Θ des coordonnées cartésiennes des N directions et $\boldsymbol{Y}_\Theta$ la matrice des valeurs des harmoniques sphériques associées. Notons $\Theta_\mathcal{R}$ la matrice des coordonnées cartésiennes de ces N directions à laquelle a été appliquée une matrice de rotation $\mathcal{R}(\chi, \eta, \xi)$ d'angles χ autour de l'axe x, η autour de l'axe y et ξ autour de l'axe z :

$$\Theta_\mathcal{R} = \mathcal{R}(\chi, \eta, \xi)\Theta. \tag{C.9}$$

Soit $\boldsymbol{Y}_{\Theta_\mathcal{R}}$ la matrice des valeurs des harmoniques sphériques associées aux directions obtenues après rotation. La matrice de rotation des signaux ambisoniques $\boldsymbol{R}$ s'écrit alors :

$$\boldsymbol{R} = \boldsymbol{Y}^\dagger_{\Theta_\mathcal{R}} \boldsymbol{Y}_\Theta \tag{C.10}$$

où $\boldsymbol{Y}^\dagger_{\Theta_\mathcal{R}}$ désigne la pseudo-inverse de $\boldsymbol{Y}_{\Theta_\mathcal{R}}$, c'est à dire $\boldsymbol{Y}^\dagger_{\Theta_\mathcal{R}} \boldsymbol{Y}_{\Theta_\mathcal{R}} = \boldsymbol{I}_{(L+1)^2}$.

* * *

Ainsi il est possible de restituer en binaural dynamique le champ sonore d'une scène réverbérée en utilisant les procédés de traitement du signal exposés dans les annexes de ce document. La figure C.5 illustre le processus d'auralisation d'une scène sonore convoluée par la réponse impulsionnelle spatiale d'une salle. Ce processus est celui employé dans les tests perceptifs présentés dans ce document.

Bibliographie

[1] L. Rayleigh, "XII. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, n°. 74, p. 214–232, 1907.

[2] S. Carlile, R. Martin et K. McAnally, "Spectral information in sound localization," *International review of neurobiology*, vol. 70, p. 399–434, 2005.

[3] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira et J. A. Santos, "On the improvement of localization accuracy with non-individualized HRTF-based sounds," *J. Audio Eng. Soc.*, vol. 60, p. 821–830, 2012.

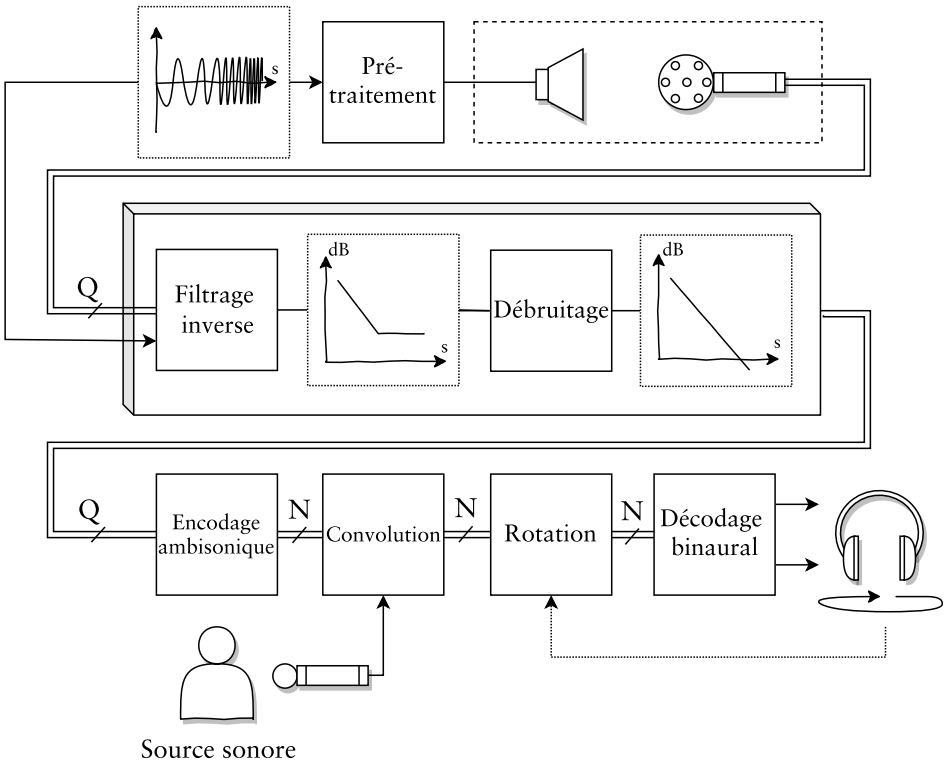


FIGURE C.5 – Processus d'auralisation d'une scène sonore réverbérée au moyen d'une réponse impulsionnelle spatiale de salle.

[4] E. M. Wenzel, M. Arruda, D. J. Kistler et F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, p. 111–123, 1993, https ://doi.org/10.1121/1.407089.

[5] S. M. Kim et W. Choi, "On the externalization of virtual sound images in headphone reproduction : A Wiener filter approach," *J. Acoust. Soc. Am.*, vol. 117, p. 3657–3665, 2005, https ://doi.org/10.1121/1.1921548.

[6] D. R. Begault et E. M. Wenzel, "Headphone localization of speech," *Hum. Fac. Erg. Soc.*, vol. 35, p. 361–376, 1993, https ://doi.org/10.1177/001872089303500210.

[7] D. R. Begault, E. M. Wenzel et M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, p. 904–916, 2001.

[8] D. R. Begault, "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc.*, vol. 40, p. 895–904, 1992.

[9] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz et C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, vol. 141, p. 3678–3688, 2017a, https ://doi.org/10.1121/1.4978612.

[10] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz et C. de Boishéraud, "Improvement of externalization by listener and source movement using a "binau-

ralized" microphone array," *J. Audio Eng. Soc.*, vol. 65, p. 589–599, 2017b, https ://-doi.org/10.17743/jaes.2017.0018.

[11] J.-M. Jot, S. Wardle et V. Larcher, "Approaches to binaural synthesis," dans *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

[12] M. Noisternig, A. Sontacchi, T. Musil et R. Holdrich, "A 3d ambisonic based binaural sound reproduction system," dans *Audio Engineering Society Conference : 24th International Conference : Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[13] B. Bernschütz, A. V. Giner, C. Pörschmann et J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, n°. 5, p. 972–983, 2014.

[14] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl et B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, n°. 6, p. 4087–4096, 2017.

[15] M. Zaunschirm, C. Schörkhuber et R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, n°. 6, p. 3616–3627, 2018.

[16] C. Schörkhuber, M. Zaunschirm et R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," dans *Proceedings of the DAGA*, vol. 44, 2018, p. 339–342.

[17] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[18] J. Sheaffer, S. Villeval et B. Rafaely, "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction," *10.14279/depositonce-4103*, 2014.

[19] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, vol. 27, n°. 4, p. 339, 1940.

[20] J.-M. Jot, V. Larcher et J.-M. Pernaux, "A comparative study of 3-d audio encoding and rendering techniques," dans *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[21] J. Ivanic et K. Ruedenberg, "Rotation matrices for real spherical harmonics. direct determination by recursion," *The Journal of Physical Chemistry*, vol. 100, n°. 15, p. 6342–6347, 1996.

[22] F. Zotter, *Analysis and synthesis of sound-radiation with spherical arrays*. Citeseer, 2009.

D

La simulation de réponses impulsionnelles de salle

Dans le but d'obtenir la réponse impulsionnelle d'une salle, une simulation numérique peut être mise en œuvre. La propagation du son dans un milieu homogène et isotrope est décrite par l'équation d'onde. Cette équation relie le comportement spatial de la pression sonore p à son comportement temporel selon une équation linéaire aux dérivées partielles de second ordre :

$$\Delta p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{D.1}$$

où c est la vitesse du son et Δ l'opérateur laplacien scalaire de dérivation spatiale d'ordre 2. Cette équation doit être résolue pour simuler une réponse impulsionnelle de salle mais, en dehors de configurations simples, il n'existe pas de solutions analytiques. Deux approches permettent d'obtenir une approximation de ces solutions : les méthodes de résolution numérique et les méthodes d'acoustique géométrique [1].

Les méthodes de résolutions numériques, telles que la méthode des éléments finis [2] ou la méthode des éléments finis de frontière [3], permettent d'approcher une solution de l'équation d'onde par discrétisation du volume délimité par l'environnement modélisé ou des parois délimitant cet environnement. La solution au problème est calculée sur les nœuds d'un maillage de l'environnement, ce qui nécessite une interpolation des résultats pour décrire la pression sonore en tout point. Plus les points sont distants les uns des autres, plus l'approximation liée à l'interpolation risque de s'écarter des phénomènes physiques étudiés. Cependant, le maillage de l'environnement dépend de la longueur d'onde : les éléments du maillage doivent avoir des dimensions environ six fois plus petites que la longueur d'onde de la fréquence à laquelle la pression est calculée [4]. Par conséquent, ces approches demandent beaucoup de ressources en calcul et ne sont pas appliquées en hautes fréquences pour des environnements à grand volume. Ces méthodes sont généralement employées pour le

calcul de la partie basse fréquence de la réponse impulsionnelle de salle jusqu'à la fréquence de Schrœder. Au-delà de cette fréquence, les modes de la salle se chevauchent et peuvent être décrits par des propriétés statistiques. D'autres approches fondées sur l'acoustique géométrique, plus efficaces en terme de calcul, sont alors utilisées dans cette région fréquentielle. On compte notamment parmi ces approches, la méthode des sources images [5, 6] et du lancé de particules [7].

D.1. La méthode des sources images

Cette méthode simple permet de calculer les réflexions spéculaires générées par une source S et reçues par un récepteur dans une salle. Elle consiste à calculer la position des points symétriques de S par rapport à toutes les parois de la salle. Ces points, nommés sources images, sont considérés comme des sources secondaires produisant une onde sonore résultant de la réflexion de S avec la paroi associée. Les sources images symétriques de S par rapport à la source principale sont des sources images d'ordre 1. Les symétriques de ces sources images sont à leur tour déterminées de la même façon et sont qualifiées d'ordre 2. L'opération est réalisée N fois où N désigne l'ordre maximal des sources images. Il est ensuite nécessaire de vérifier si il existe un trajet reliant chacune des sources images au récepteur pour sélectionner les sources images audibles en ce point.

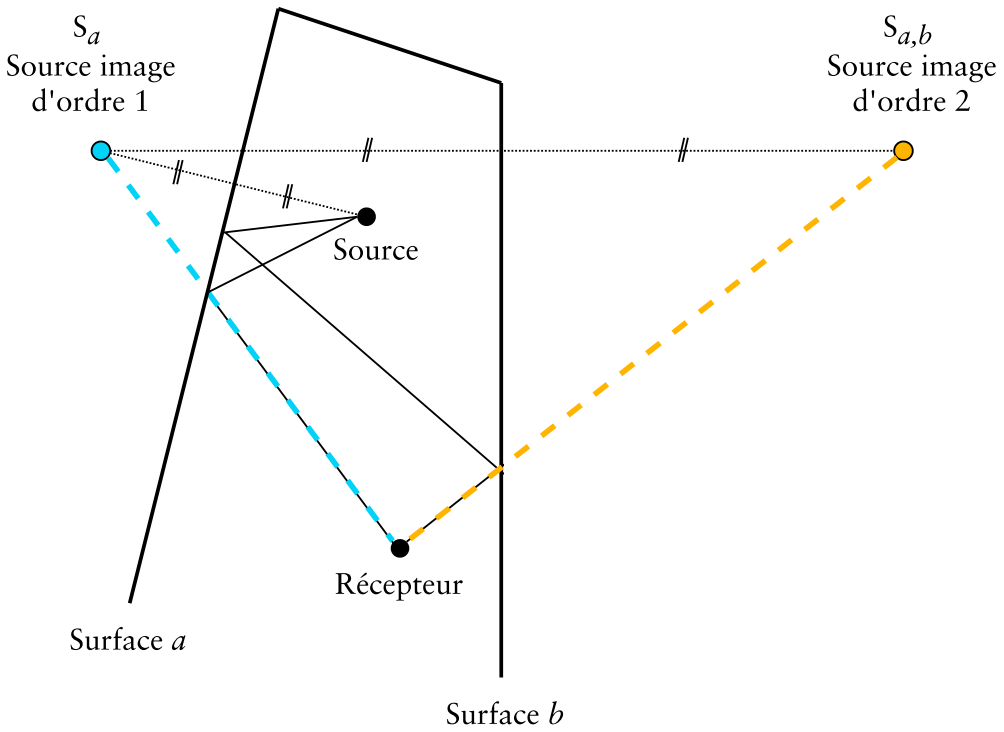


FIGURE D.1 – Illustration de la méthode des sources images.

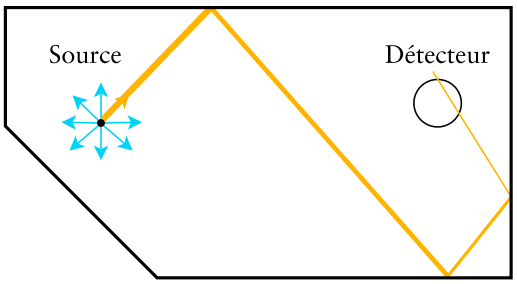
La figure D.1 illustre cette méthode. La source-image S_a est l'image de S par rapport à la paroi a. Le trajet entre S_a et le récepteur ne rencontrant pas d'autres obstacles que la paroi a, ce trajet est retenu et associé à la réflexion sur la paroi a d'une onde

émise par la source. L'image source d'ordre 2, $S_{a,b}$, est ensuite obtenue en calculant le symétrique de S_a par rapport à la paroi b.

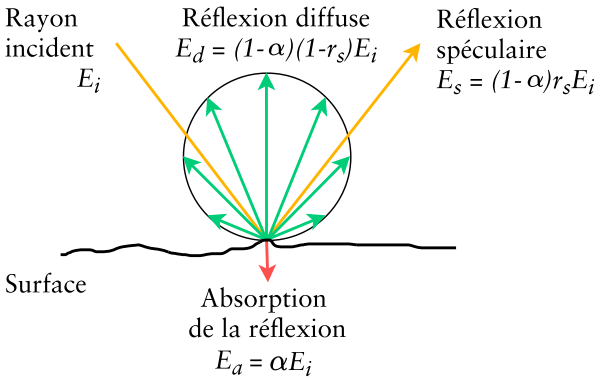
La méthode des sources images n'est pas en mesure de reproduire la diffusion qui apparait lors de la réflexion des ondes sonores sur les parois de l'environnement. Elle permet néanmoins d'estimer correctement la position des premières réflexions d'ordre faible. Une approche hybride combinant cette méthode avec le lancé de particules est souvent adoptée pour être en mesure de générer la partie tardive de la réponse impulsionnelle.

D.2. Le lancé de particules

Le lancé de particule simule la propagation d'une impulsion sonore provenant de la source en émettant des particules d'énergie qui se propagent à la vitesse du son dans un nombre fini de directions réparties autour de la source. Chaque particule transporte une quantité d'énergie qui dépend de la directionnalité de la source. Ces particules perdent de l'énergie en se propageant du fait des frottements dans l'air et en raison des phénomènes d'absorption et de diffusion lorsqu'elles se réfléchissent sur les parois de l'environnement.



(a) Propagation d'une particule générée par une source omnidirectionnelle. La largeur du rayon symbolise la quantité d'énergie portée par la particule.



(b) Illustration de la réflexion d'une particule incidente sur une surface de manière spéculaire ou diffuse. Les énergies de la particule incidente, de sa réflexion diffuse, spéculaire et de son absorption sont notées E_i, E_d, E_s et E_a respectivement.

FIGURE D.2 – Illustration de la méthode du lancé de particules.

Lorsqu'une onde se réfléchit sur une paroi, une proportion de son énergie réfléchie (notée ici r_s) se propage selon un angle d'incidence égal à l'angle de réflexion et l'autre partie (notée $r_d = 1 - r_s$) se propage de manière diffuse dans les autres directions. Lors du lancé de particules, la réflexion d'une particule est soit considérée comme étant spéculaire - l'angle d'incidence est égal à l'angle de réflexion - soit diffuse - l'angle de réflexion est déterminé selon une loi de probabilité définie par la diffusivité de la paroi. La nature de la réflexion est déterminée de manière aléatoire en prenant

en compte le pouvoir diffuseur du matériaux. Si une particule est réfléchie de manière diffuse, l'énergie portée par cette particule correspond à la proportion de l'énergie réfléchie r_d. Si une particule est réfléchie de manière spéculaire, l'énergie portée par cette particule correspond à la proportion de l'énergie réfléchie r_s. Les deux types de réflexions considérés dans la simulation sont illustrés par la figure D.2a. Ainsi, contrairement à la méthode des sources images, le lancé de rayon présente l'avantage de prendre en compte la diffusion des matériaux.

Les particules sont captées par un récepteur caractérisé par un volume ou une surface car la probabilité d'atteindre un récepteur ponctuel dans un espace avec un nombre fini de particules est quasi nulle. Ce récepteur non ponctuel, nommé détecteur, est illustré sur la figure D.2a. Chaque fois qu'une particule frappe le détecteur, l'énergie, l'angle d'incidence et le temps de trajet de la particule sont enregistrés.

Le lancé de particules est effectué par bande de fréquence car la directionnalité de la source, l'absorption et la diffusion des matériaux et l'absorption par l'air dépendent fortement de la fréquence. La propagation de ces particules est calculée jusqu'à un temps maximum ou un seuil minimum d'énergie donné. En raison de la nature stochastique de cette méthode, les résultats de simulation peuvent varier d'une simulation à l'autre. Les valeurs enregistrées permettent néanmoins d'établir une estimation de la distribution d'énergie temporelle tridimensionnelle des réflexions en fonction de la fréquence.

D.3. La prise en compte de la diffraction

Ces méthodes ne couvrent cependant pas les phénomènes de diffraction qui peuvent se produire aux bords des parois et en présence d'objets dans l'environnement qui font obstacle à la propagation du son à une fréquence donnée [8]. Ce phénomène apparait pour des fréquences dont la longueur d'onde est du même ordre de grandeur que les dimensions de l'obstacle.

Sans prise en compte de la diffraction, les méthodes d'acoustique géométrique ne peuvent donc pas correctement simuler la propagation d'ondes sonores provenant d'une source située derrière un obstacle. Bien que le calcul nécessaire pour simuler des phénomènes de diffraction soit important, certains cas de figure peuvent être couverts pas des méthodes simples.

Selon la théorie de la diffraction uniforme (UTD) [9], les bords d'un objet faisant obstacle à la propagation du son deviennent des sources secondaires de rayons diffractés qui à leur tour peuvent se propager dans l'environnement. La diffraction d'une onde sur un bord est alors représentée par un rayon dont l'énergie est atténuée par un coefficient de diffraction. L'UTD est une approximation haute fréquence et s'applique en théorie à des bords infinis, lorsque la source et le récepteur sont éloignés (en termes de longueur d'onde) des surfaces diffractantes.

Pour adapter la méthode de calcul des sources images au calcul de la diffraction, Schröder [10] propose donc d'ajouter un autre type de sources secondaires appelées sources de diffraction. Ces sources sont positionnées à chaque arête d'un obstacle et sont prises en compte si celui-ci se situe entre la source émettrice ou une source-image et le récepteur. Comme pour la génération de sources images, les sources symétriques

d'une source de diffraction sont calculées par rapport à toutes les surfaces de la salle jusqu'à un ordre de réflexion prédéfini.

Pour inclure le calcul de la diffraction dans la méthode du lancé de particules, Schröder [10] propose également d'utiliser des détecteurs cylindriques situés sur les arêtes des obstacles, nommés détecteur de déviation. L'axe du détecteur correspond à l'axe de l'arête et le rayon du cylindre est proportionnel à la longueur d'onde de la fréquence considérée. Si un détecteur de déviation est frappé par une particule d'énergie, l'angle de déviation de la particule peut être obtenu d'après une fonction de densité de probabilité proposée par Stephenson [11]. Ce processus est illustré par la figure D.3.

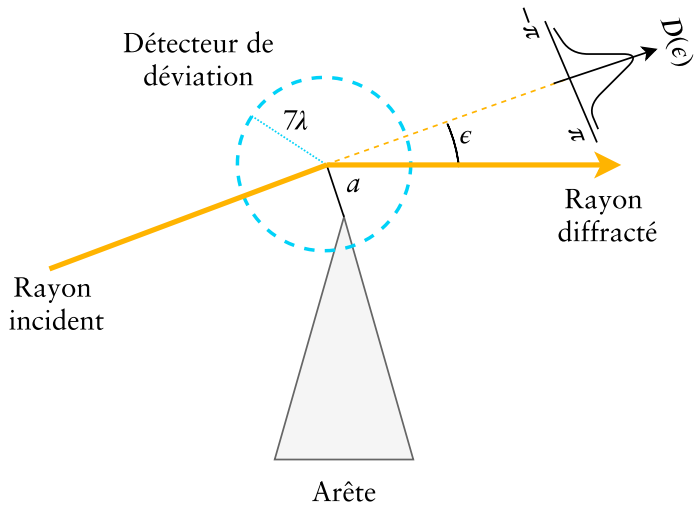


FIGURE D.3 – Illustration de la prise en compte du phénomène de diffraction dans la méthode du lancé de particules où α désigne la distance entre la particule et l'arête, λ la longueur d'onde, ϵ l'angle de déviation et $D(\epsilon)$ la densité de probabilité de l'angle de déviation dont l'expression est fonction de la distance α [11].

Bibliographie

[1] S. Siltanen, T. Lokki et L. Savioja, "Rays or waves ? understanding the strengths and weaknesses of computational room acoustics modeling techniques," dans *Proc. Int. Symposium on Room Acoustics*, 2010.

[2] F. Ihlenburg, *Finite element analysis of acoustic scattering*. Springer Science & Business Media, 2006, vol. 132.

[3] R. D. Ciskowski et C. A. Brebbia, *Boundary element methods in acoustics*. Springer, 1991.

[4] S. Marburg, "Six boundary elements per wavelength : Is that enough ?" *Journal of computational acoustics*, vol. 10, n°. 01, p. 25–51, 2002.

[5] J. B. Allen et D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, n°. 4, p. 943–950, 1979.

[6] J. Borish, “Extension of the image model to arbitrary polyhedra,” *The Journal of the Acoustical Society of America*, vol. 75, n°. 6, p. 1827–1836, 1984.

[7] M. R. Schroeder, “Digital simulation of sound transmission in reverberant spaces,” *The Journal of the acoustical society of america*, vol. 47, n°. 2A, p. 424–431, 1970.

[8] M. Vorländer, *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[9] R. G. Kouyoumjian et P. H. Pathak, “A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface,” *Proceedings of the IEEE*, vol. 62, n°. 11, p. 1448–1461, 1974.

[10] D. Schröder et A. Pohl, “Real-time hybrid simulation method including edge diffraction,” 01 2009.

[11] U. M. Stephenson et U. P. Svensson, “An improved energetic approach to diffraction based on the uncertainty principle,” *19th Int. Cong. on Acoustics (ICA)*, 2007.

E

Publications liées à la thèse

Revues

Article accepté

— Salmon, F., Hendrickx, É., Épain, N., & Paquier, M. (2020). «The Influence of Vision on Perceived Differences Between Sound Spaces». *Journal of the Audio Engineering Society*, 68(7/8), 522-531.

Article soumis

— Salmon, F., Hendrickx, É., Épain, N., & Paquier, M. (2020). «The influence of Spatial Room Impulse Response Resolution on the Perception of Sound Spaces When Using Non-Individualized HRTFs». *Applied Acoustics* [Under Review].

Conférence

— Salmon, F., Hendrickx, É., Épain, N., George, Q., & Paquier, M. (2019). «The Influence of the Sound Source on Perceived Differences Between Binaurally Rendered Sound Spaces». *Proceedings of the AES International Conference on Headphone Technology (San Francisco, August 2019)*.

Brevet

— Salmon, F. & Épain, N. (2020). «Procédé de conversion d'un premier ensemble de signaux représentatifs d'un champ sonore en un second ensemble de signaux et dispositif électronique associé». *Déclaration d'invention auprès de l'INPI*.

F

The Influence of Vision on Perceived Differences Between Sound Spaces

The influence of vision on perceived differences between sound spaces

François Salmon[1,2] Étienne Hendrickx[2] Nicolas Épain[1] Mathieu Paquier[1,2]

[1]b<>com Institute of Research and Technology,
1219 Avenue des Champs Blancs, 35510 Cesson-Sévigné, France
[2]University of Brest, CNRS, Lab-STICC UMR 6285,
6 avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France

Abstract

Few studies have investigated the influence of visual cues on sound space perception, beyond the influence of visual cues on sound source position. Previous studies suggest that the perception of late reflections is not affected by the visual impression of a room, however only a limited number of spatial sound attributes were investigated. In the present paper, audiovisual interactions were examined without making assumptions on the number and the nature of perceptual dimensions involved in the perception of sound space. In a virtual environment that employed an Head Mounted Display and dynamic binaural playback, subjects were asked to judge the perceived dissimilarity between sound spaces while watching the same visual stimulus. Pairwise comparisons were repeated using multiple visual conditions, including an audio-only condition. One sound source, a male voice reciting a poem, was considered in the listening test. It appeared that the visual modality did not impact the perceived differences between sound spaces.

1 Introduction

Various studies have defined numerous perceptual attributes related to spatial qualities of sound in the field of spatial quality assessment, spatial audio reproduction or concert hall acoustics [1–3]. Some work employed real existing rooms and performed studies directly in concert halls or in laboratories using dummy head recordings. However, with laboratory test setups, it can be argued that there is a lack of experimental control concerning the stimuli presented: the visual impression of the sound space under study was usually not considered [4]. In this paper, we address the influence of vision on the perception of sound space in order to assess whether the lack of experimental control regarding the visual environment was detrimental in such studies. To this end, we used state-of-the-art auralization and display technologies that give us the opportunity to have a better experimental control on the audio-visual stimuli under study.

Several studies on audiovisual cross modality show an interaction between auditory and visual cues related to the perception of space. Some

studies have especially examined the influence of vision on the assessment of sound source localization [5–7], auditory distance perception [8–12], externalization [13–17] or spatial impression [18–20].

1.1 Sound source localization

It seems that vision has a strong impact on the perception of audiovisual source localization. Following the early work of Jack and Thurlow [5], numerous experiments were performed to determine whether subjects experienced an auditory stimulus perceptually unified with a visual object of a different location (the so-called ventriloquism effect) [21–25]. All came to the conclusion that a strong visual capture occurs and that the effect increases with decreasing angular difference between the positions of the sound and visual stimuli. Hendrickx et al. [6] also reported that the phenomenon had a greater importance in elevation than in azimuth. Bishop et al. [7] studied another cross-modal interaction regarding source localization: the impact of visual cues on the precedence effect. According to the precedence effect, it is considered that when the delay to the reflection is short, the reflection is not perceived as a separate event. In this case, the perceived source localization is dominated by the location of the leading sound. It was shown that the strength of the precedence effect can be enhanced when visual information spatially and temporally coincides with the leading wave. Conversely, the precedence effect is lessened when vision coincides with the reflection.

1.2 Auditory distance perception

Visual capture also occurs in auditory distance perceptions even in the presence of multiple auditory cues [8, 9]. In particular, Hládek et al. [10] measured distance localization performance in a dark reverberant room using noise bursts that were spatially congruent or incongruent with visual stimuli (LEDs). They reported a ventriloquism effect in the distance dimension: a shift towards the visual stimuli was perceived when they were presented closer to or farther away from the auditory stimuli. Further, Calcagno et al. [11] reported that, even when the sound source was visually occluded, auditory distance judgements were more accurate when visual information of the whole scene was available. They hypothesized that visual information other than the perceived distance to the source, such as room size, can be used by listeners to perform auditory distance judgements.

1.3 Externalization

Several investigations show that visual aspects also have an influence on externalization [13–17]. This capacity to hear out-of-head auditory events in binaural playback decreases if the listening room and the synthesized room are incongruent, *i.e.* when there is a mismatch between the virtual sound reverberation and the visual impression of the listening room.

In a study conducted by Udesen et al. [16], externalization judgements were performed in two rooms: one congruent and the other incongruent with the virtual sound space of binaural stimuli. Although the same binaural sound sample was used - only visual cues differed - results showed significant differences between test environments: externalization ratings were lower

in the incongruent condition. Authors hypothesized that when expectations of a realistic sound environment are not met, particularly when a divergence between vision and auditory cues occurs, the realism of the sound scene is affected and leads to internalized perception.

Thus it seems that auditory perception is significantly affected by visual impressions of the listening room and listener expectations - which can be altered through training [17]. Nevertheless, results of an externalization test conducted by Gil-Carvajal et al. [15] showed that the auditory modality had a greater impact on externalization than the visual modality. In their experiment, Binaural Room Impulse Response (BRIR) were measured in a reference room to create individualized sound stimuli for 18 subjects. Stimuli were played-back in the reference room and in two other rooms: 1) a smaller and more reverberant room, 2) a larger and anechoic room. The incongruence between the reference room and the test rooms differed depending on volume (visual cues) and reverberation time (auditory cues). Three conditions were tested: an auditory-only condition (test performed in the dark and noise bursts were used as additional auditory cues), a visual-only condition (subjects could see the room and no additional auditory cues were available), visual and auditory cues (subjects could see the room and noise bursts were used as additional auditory cues). They reported that the highest degree of externalization was obtained when both audio and visual information were congruent. Externalization ratings were significantly reduced when subjects received additional auditory cues from the playback room that did not match those from the virtual sound space. However, externalization ratings remained unaffected when subjects could see a room that differed from the one they heard through the headphone reproduction. Hence, it can be hypothesized that prior knowledge of acoustical features of the listening environment may have a greater impact on externalization than acoustical expectations based on room-related visual cues. Further studies are needed to resolve this ambiguity. It is clear that vision can bias audition in multiple non-trivial ways and that the connections between visual and auditory perception are not fully understood.

1.4 Spatial impressions

Still photographs have been employed as visual cues in numerous studies on audiovisual cross modality. However, Larsson et al. [18] showed that the degree of visual realism affected auditory room quality assessment. Particularly, their study indicated that sound sources were considered significantly wider when the subject was in the real room or used a Head-Mounted Display (HMD) in comparison to conditions with no visual cues or still photographs. Consequently, Postma & Katz [20, 26] used a test setup with a high degree of visual realism - a CAVE-light system - to investigate the influence of vision on room acoustics perception. For the room under study, they observed a significant influence of visual cues on distance perception while no significant influence regarding the apparent source width or envelopment was found. It can be argued that the different test setups or sound spaces employed in these studies can explain the contradictory results obtained regarding the apparent source width. Further studies are needed to clarify such ambiguity and determine whether visual cues impact our perception of a sound space. While the perception of sound source distance, localization and externalization seem to be impacted by visual cues, the influence of vi-

sion on other spatial sound attributes remains unclear.

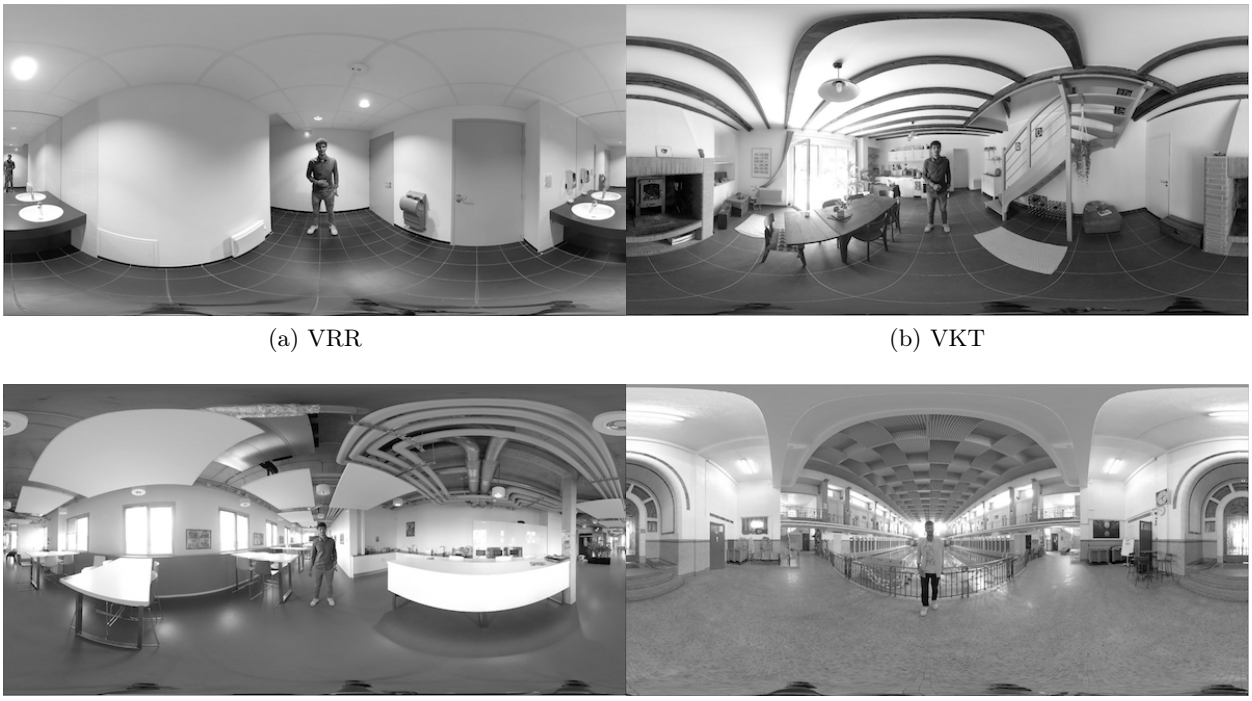
Recently, Schutte et al [27] have investigated the influence of vision on the perceived degree of reverberation in audiovisual virtual environments. Three conditions were tested : 1) congruent auditory and visual environments, 2) incongruent auditory and visual environments, and 3) audio-only condition. They reported no influence of vision on the subjects' answers, whether the environment were congruent or incongruent. However, this study focused the attention on the perceived degree of reverberation only and subjects may thus have based their judgment on limited features, such as reverberation time or direct to reverberant energy ratio. The perception of sound space is generally considered multidimensional and involves other perceptual attributes such as clarity, envelopment, depth or width [28]. Therefore, a test protocol that does not steer the attention towards a particular sound feature would be more ecologically valid, and might highlight factors other than reverberance that do interact with the visual modality.

1.5 Aim of the present study

We examined the influence of multiple visual conditions on dissimilarity ratings related to a set of sound stimuli. Subjects had to rate the perceived differences between sound stimuli while looking at the same visual stimulus and the pairwise comparisons were repeated with multiple visual stimuli (including an audio-only condition). Hence, we did not make assumptions on the number and the nature of perceptual dimensions involved in the perception of sound space. Additionally, Multidimensional Scaling (MDS) was employed to assess that the perceptual structure underlying the perception of the sound spaces was not uni-dimensional.

As mentioned earlier, when there are spatially-related conflicts between visual and auditory cues, vision usually dominates and biases audition [29]. For example, the perception of externalization can be significantly lessened if there is a mismatch between the visual impressions of the listening room and the acoustical features of the virtual sound. Additionally, biases of auditory localization towards the visual stimulus are usually observed when subjects are presented with a spatially discordant auditory-visual stimulus with respect to direction or distance. This could be explained by the fact that the visual system has a greater spatial precision than the auditory system and is therefore more reliable [6, 30, 31]. Likewise, it can be hypothesized that vision offers more information about a given room (such as room-size [12]), which may bias sound space perception towards visual expectations, especially if the visual and auditory environments are incongruent. In the present experiment, it may lead to subjects perceiving fewer differences when they compare sound spaces while watching the same visual space than when the comparisons are conducted without any visual cues.

Since auditory room quality assessment seems to be affected by the degree of visual realism [18], Schutte et al. employed a Head Mounted Display (HMD) to provide a high visual fidelity [27]. Likewise, we employed 360° videos displayed in a HMD along with a dynamic binaural sound reproduction to ensure a high simulation quality. Moreover, the sound spaces involved were measured with a dense spherical microphone array to allow an accurate reproduction of the corresponding sound field.



(a) VRR (b) VKT

(c) VRF (d) VSW

Figure 1: Equirectangular projection of the different visual stimuli. VRR: *Restroom*, VKT: *Kitchen*, VRF: *Refectory*, VSW: *Swimming pool*. The volume of the visual spaces are respectively 23 m^3, 81 m^3, 360 m^3 and 7448 m^3.

2 Perceptual Test

2.1 Visual stimuli

The visual stimuli consisted of an actor reciting *Fantaisie*, a poem by Gérard de Nerval, in four different spaces: a restroom, a kitchen, a refectory, and a swimming pool. The visual stimuli will be designated VRR, VKT, VRF, and VSW, respectively, in the following sections. A fifth visual condition, being no image at all, was added. This condition will be thereafter referred to as V0.

The visual stimuli were 360° videos, recorded in 4K (3840 × 1920 pixels) at 30 frames per seconds using the Insta360 camera. The videos covered the entire space in both azimuth and elevation. Screenshots of the visual stimuli are displayed in Figure 1.

In order to ensure minimum differences of actor performance between stimuli, a loudspeaker was playing an anechoic recording of the voice and the actor had to lip sync. Thus, no sound was recorded while capturing the visual stimuli. A post-synchronization step was then required to synchronize sound and video. An informal test conducted by the authors and informal discus-

Table 1: Abbreviations, early decay times (EDT), direct-to-reverberant ratios (DRR), early-to-late index (C80) and reverberation times (RT) of the sound spaces used in the experiment. Measurements were computed at mid-frequencies.

Spaces	Abbreviation	EDT(s)	DRR(dB)	C80(dB)	D50(%)	RT(s)
Small box	SBX	0.31	-3.73	16.07	85.87	0.38
Restroom	TLT	0.33	-2.13	15.14	82.62	0.41
Meeting room	MTR	0.35	0.98	15.22	89.90	0.43
Small concert hall	SCH	0.31	6.08	16.6	96.19	0.46
Classroom	CLS	0.40	3.67	15.53	93.17	0.55
Kitchen	KTC	0.51	-2.56	10.19	81.28	0.60
Refectory	RFC	0.65	2.61	11.85	87.40	0.83
Medium concert hall	MCH	0.27	8.16	18.37	96.88	0.85
Swimming pool	SWP	0.72	3.74	11.95	90.13	1.91
Hall	HLL	0.55	9.11	16.36	96.57	2.58
Church	CHR	2.17	6.55	9.84	88.75	3.56
Cathedral	CTH	2.48	6.85	12.72	92.78	6.55

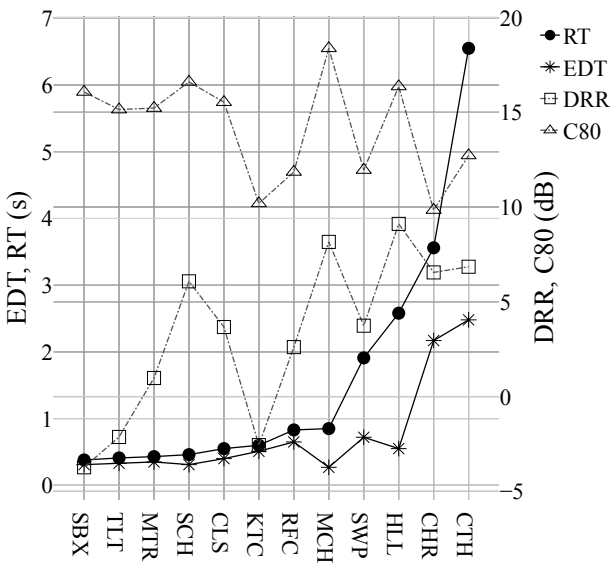


Figure 2: Acoustic parameters of the sound spaces used in the experiment: reverberation times (RT), early decay times (EDT), direct-to-reverberant ratios (DRR) and early-to-late index (C80).

sions with the subjects suggested that there was no perceivable desynchronization.

2.2 Spatial room impulse responses

Twelve different spaces were used for creating the audio stimuli (cf. Table 1), including the four spaces used for the visual stimuli. Sound spaces where chosen on the basis of an informal test conducted by the authors so as to cover a wide span of the perceptual range of sound spaces. The sound spaces were not limited to concert halls or large spaces, that are favoured by a part of the literature in room acoustics, but also included everyday places. The various measures displayed in Table 1 and Fig. 2 are an example of this heterogeneity. The considered rooms covered a wide range of decay times (EDT, RT) and presented various energy ratios (DRR, C80) for similar decay times.

The spatial room impulse responses (SRIRs) were measured using the EigenMike EM32 microphone array from mh acoustics LLC and a Genelec 8040 loudspeaker. The loudspeaker was

positioned at the same distance from the microphone array (150 cm) in every room, in order to compare rooms only and not distances. A 10 second long exponential sweep-sine signal was used for the measurements. Since measured impulse responses are typically corrupted by measurement noise, the noise floor was used to extend the reverberation decay [32]. Moreover, the influence of the loudspeaker was compensated in the SRIRs using equalization filters that were derived from measurements done in an anechoic chamber. Lastly, the measured microphonic SRIRs were converted to order-4 ambisonic SRIRs.

2.3 Sound stimuli

One sound source was considered in this test: a male voice reciting *Fantaisie*, recorded in an anechoic room. The recorded signal was convolved with the 12 SRIRs corresponding to the aformentionned sound spaces. Hence, we obtained 12 order-4 ambisonic stimuli. A loudness equalization was done subjectively by the experimenters prior to the perceptual test (as recommended in [33]), so that the perceived loudness remained the same when switching between spaces. The total duration of the stimuli was 20 s (sound source duration) + 6.5 s (reverberation time of the longest SRIRs) = 26.5 s. Stimuli were sampled at a rate of 48000 Hz with a 24 bit resolution.

2.4 Binauralization and head-tracking

The ambisonic signals were converted into binaural headphone signals using the open-source *BinauralDecoder* VST plug-in from the IEM plug-in suite with Head Related Transfer Functions (HRTFs) derived from measurements of a Neumann KU 100 dummy head. For further information about these measurements and the binaural renderer, please refer to [34, 35].

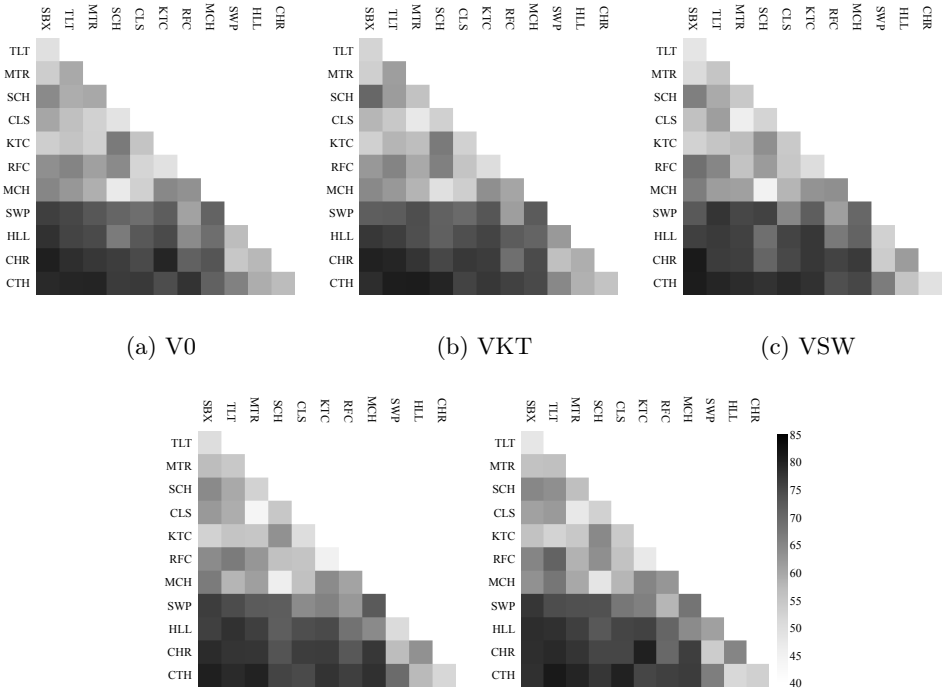
Ideally, binauralization filters should be personalized for each listener, however measuring individualized HRTFs is a complex and expensive process. A common solution is to listen "through the ears" of another listener, whose HRTFs are already available, or a dummy head, as in the present experiment. The use of such non-individualized HRTFs can increase the occurrence of front-back confusions or result in "in-head" localization, but these issues are effectively alleviated by head tracking [36, 37] and reverberation [38, 39]. Furthemore, Begault et al. reported that individualized HRTFs offered no advantage in localization accuracy and externalization for the binaural synthesis of speech stimuli reproduced in the horizontal plane [38].

Head-tracking was performed using the HMD along with the *ambix_rotator_o7* plugin from Matthias Kronlachner [40].

2.5 Procedure

The perceptual test consisted of successive trials based on pairwise dissimilarity ratings involving the 12 sound stimuli. A trial consisted of rating the dissimilarity between two sound stimuli while watching a video clip synchronized with the sound. Note that the same visual stimulus was used when comparing two sound stimuli. Dissimilarity ratings for all pairwise comparisons of sound stimuli were performed under the five different visual conditions. A total of $\binom{12}{2} = 66$ comparisons for each visual condition were thus rated by each subject.

For each pair, subjects were instructed to judge to what extent the sound stimuli were either similar or different by moving a slider along a continuous 100-point scale displayed in the HMD, whose extremities were labeled identical



(a) V0 (b) VKT (c) VSW

(d) VRF (e) VRR

Figure 3: Dissimilarity matrices for the five visual conditions of the experiment. Each square represents the dissimilarity rating between two spaces. The grey scale refers to the dissimilarity scores averaged across subjects. Darker means greater dissimilarity.

Table 2: Results from ANOVA. V: visual conditions, P: pairs, SS: sum of squares, MS: mean of squares, F: f-value, p: p-values, η^2: percentage of variance accounted for.

Factor	df	SS	MS	F	p-value	Eta squared (%)
V	4	334	84	1.6	0.193	0.06
P	65	335374	5160	41.1	< 0.001	60.57
V * P	260	13762	53	1.14	0.067	2.48

(0) and very different (100). There were no predefined intermediate labels nor gradations on the scale to avoid any undesirable bias [41]. Subjects were specifically instructed not to close their eyes during the whole experiment and were allowed to listen to the two stimuli repeatedly and to switch between them at will. Playback, commands, and data capture were controlled using both Unity and Max. The experiment was carried out in virtual reality using a HTC Vive HMD and a pair of Sennheiser HD 650 headphones.

The experiment consisted of two one-hour long

sessions on different days. Before each session, participants were presented with the sound source reverberated in the 12 sound spaces, in order to apprehend the diversity of stimuli included in the test.

The different pairs, along with the associated visual stimulus, were presented in a randomized order. The order of the sound stimuli within each pair was also randomized. These randomizations were different for each subject.

Overall, subjects had to rate 330 pairs of sound stimuli in different visual conditions. Note that the tested pairs of stimuli either included sound spaces that were both incongruent with the visual space or where only one of the sound spaces was congruent with the visual space.

2.6 Subjects

Eleven subjects (1 woman and 10 men, aged 21 to 25 years old) took part in the perceptual test. They were all Masters degree students from the Image & Sound course at the University of Brest. None reported any known hearing loss, and none had experience with Virtual Reality nor laboratory listening tests. The experience of subjects with binaural content was globally low : some of them had already listened to static natural binaural recordings (i.e. real sound sources captured with microphones placed in the ears of a dummy head or of a listener), yet none of them had ever experienced dynamic binaural synthesis.

3 Results

3.1 Dissimilarity matrices

The pair-wise dissimilarity ratings resulted in a dissimilarity matrix of dimension 12×12 for each subject and for each visual condition. Since the order of sound stimuli within each pair was not distinguished, the dissimilarity matrices are symmetric. Matrices averaged over subjects are plotted in Figure 3 for the five visual conditions. This figures appeared to be very similar, which was confirmed by large Pearson correlation coefficients between each other (> 0.94, $p < 0.001$). This seemed to imply that the vision cues had a relatively small influence on the perceived differences between spaces.

The data were submitted to a repeated-measures ANOVA with the following factors: visual conditions V (5) × sound spaces pairs P (66). Results are presented in Table 2: only the effect of sound space pairs had a significant influence on the dissimilarity scores. The p-value associated to the interaction between visual conditions and sound space pairs was not sufficiently low to achieve significance [$F(260, 2600) = 1.142$, $p = 0.067$], and the percentage of variance accounted for by this interaction was only 2.48%.

3.2 Multidimensional scaling

Since no influence of the visual cues was found in the experiment, the corresponding dissimilarity matrices were averaged over the different visual conditions. Multidimensional Scaling (MDS) was used for the investigation of the perceptual structure underlying judgments of stimuli dissimilarities. MDS searches for the perceptual dimensionality assumed to underlie the perception of a set of stimuli and provides coordinates for these stimuli on each perceptual dimension [42]. In particular, INDSCAL analysis [43] finds a group of perceptual dimensions that are common to all subjects and provides the corresponding coordinates for each stimuli and subjects.

One of the challenges in MDS analysis is to find

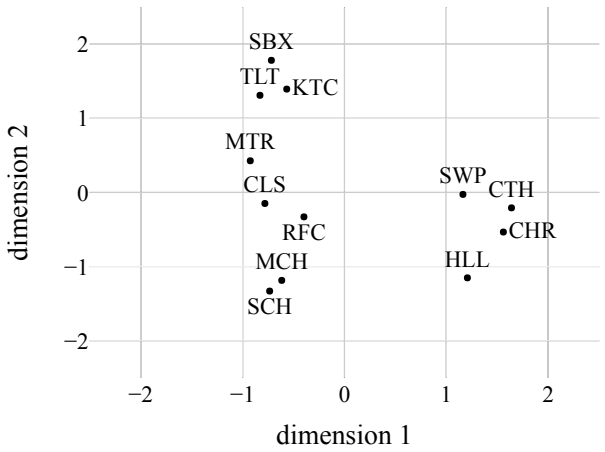


Figure 4: Two dimensional perceptual space resulting from INDSCAL analysis. Variances of the overall ratings related to dimension 1 and 2 are 0.757 and 0.096 respectively.

the right number of perceptual dimensions, *i.e.* which dimensions are worth interpreting. The analysis of the explained variance as a function of dimension gives a hint on the number of dimensions to consider: it is of interest to maximize the overall explained variance of the solution until it increases by less than 0.05 per added dimension [44]. The analysis revealed that the associated perceptual space was a two-dimensional one. The first dimension explains most of the perceived differences between sound spaces with 75.7 % of the variance explained, while the second dimension explains almost 10 % of the variance. The representation of the 12 sound stimuli in this perceptual space is displayed in Figure 4.

Additional statistical analyses were carried out to find correlations between the obtained perceptual dimensions and acoustic measurements. Twenty-two room acoustic parameters were derived from the 12 SRIRs, defined in the ISO 3382-1 standard [45]. These parameters were computed at mid-frequencies (average between values from 250Hz to 2kHz) and on broadband SRIRs. The first dimension was found to be well correlated to the logarithm of reverberation times at mid-frequencies with a Pearson correlation coefficient of 0.966. It can be postulated that the corresponding perceptual attribute is reverberance, *i.e.* the perceived amount of reverberation which was studied by Schutte et al. [27]. Measurement D_{50} seem to explain the second perceptual dimensions with a Pearson correlation coefficient of 0.926. D_{50} measures the early to total energy ratio of the room impulse response. This metric is a measure of clarity which is particularly suited to characterize speech intelligibility [45].

4 Discussion

The analysis of variance revealed that visual conditions had no significant influence on the perceived differences between sound space. Thus, whatever the divergence between the visual and auditory conditions, may there be visual content or not, the visual modality did not change the perceived differences between sound spaces.

The results of the present experiment are in agreement with the findings of Schutte et al. [27] that vision has no significant impact on the perception of reverberance. As the present experiment did not make any assumptions on the dimensions involved in the perception of sound space, it suggests that vision not only has a weak impact on reverberance, but also more globally on the overall perception of sound spaces.

A multidimensional analysis was performed to investigate whether other dimensions than reverberance were involved to distinguish between

sound spaces. It revealed that the structure underlying the perception of the sound spaces was multidimensional. In addition to reverberance, the listeners seem to have assessed features related to clarity. However, we can hypothesize that the second perceptual dimension is related to the sound source employed and that the use of other sound sources might have revealed different perceptual features of the sound spaces [46]. Moreover, it is possible that the range of auditory spatial attributes covered by the considered rooms was not wide enough. In other words, subjects may not have been able to assess other attributes such as envelopment or apparent source width, because not enough differences were present in the sound stimuli regarding these attributes. Further, most of the sound spaces under consideration had a positive DRR. (this is due to the fact that the sound source was located at 1.5 m away from the subject) and for SRIRs with the highest DRR, the reverberant part may not have been loud enough in comparison with the direct sound to highlight differences in the perception of some attributes. Various distances could be employed to investigate whether high DRR impacted dissimilarity ratings, *i.e.* whether there is a threshold above which reverberation differences are masked by the direct sound. Hence, further studies should be performed using other sound sources and sound spaces to confirm our results.

The limited number of perceptual dimensions involved in this analysis is due to the fact that only 12 sound spaces were investigated, as pairwise comparisons are very time-consuming. Other protocols such as free classification tasks would have enabled to use a much larger amount of sound stimuli and thus potentially uncover other dimensions. However, the goal of the study was rather to examine the influence of vision on the perceived differences between sound spaces, rather than to determine the perceptual dimensions involved, and to this end pair-wise comparisons seemed a better option, as it is known to be more accurate than free classification tasks [47].

5 Conclusion

The present work investigated the effect of vision on the perception of sound spaces using pairwise comparisons of sound stimuli under multiple visual conditions. Results showed that the visual impression of a room did not affect the perceived differences between sound spaces. This study was performed using dissimilarity ratings without presuming of the number and nature of the perceptual dimensions involved: additional multidimensional scaling analysis revealed that two dimensions - reverberance and clarity - were considered by subjects to distinguish between sound spaces. Following the studies of Schutte [27] and Postma [26], the present study therefore adds new evidence for a lack of visual influence on the auditory perception of space. Further experiments must be undertaken to investigate the influence of visual impression on room acoustic perception, particularly using more sound sources and a greater diversity of sound spaces.

6 Acknowledgments

This work was partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02). The authors would like to thank Charles Verron, Yaël Laouar, Jérôme Daniel, Cathy Colomes and Louis Anglionin for their help in recording visual and sound stimuli. The authors would also like to thank all the subjects who took part in the

subjective experiments.

References

[1] T. Lokki, J. Pätynen, A. Kuusinen, S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3148–3161 (2012), URL https://doi.org/10.1121/1.4756826.

[2] N. Kaplanis, S. Bech, S. H. Jensen, T. van Waterschoot, "Perception of reverberation in small rooms: a literature study," presented at the *Audio Engineering Society Conference: 55th International Conference: Spatial Audio* (2014), URL https://doi.org/10.1121/1.5135582.

[3] N. Zacharov, C. Pike, F. Melchior, T. Worch, "Next generation audio system assessment using the multiple stimulus ideal profile method," presented at the *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pp. 1–6 (2016), URL https://doi.org/10.1109/QoMEX.2016.7498966.

[4] S. Weinzierl, M. Vorländer, "Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know?" *Acoustics Australia*, vol. 43, no. 1, pp. 41–48 (2015), doi:https://doi.org/10.1007/s40857-015-0007-6.

[5] C. E. Jack, W. R. Thurlow, "Effects of degree of visual association and angle of displacement on the ventriloquism effect," *Perceptual and motor skills*, vol. 37, no. 3, pp. 967–979 (1973), URL https://doi.org/10.1177/003151257303700360.

[6] E. Hendrickx, M. Paquier, V. Koehl, J. Palacino, "Ventriloquism effect with sound stimuli varying in both azimuth and elevation," *The Journal of the Acoustical Society of America*, vol. 138, no. 6, pp. 3686–3697 (2015), URL https://doi.org/10.1121/1.4937758.

[7] C. W. Bishop, S. London, L. M. Miller, "Visual influences on echo suppression," *Current Biology*, vol. 21, no. 3, pp. 221–225 (2011), URL https://doi.org/10.1016/j.cub.2010.12.051.

[8] M. B. Gardner, "Proximity image effect in sound localization," *The Journal of the Acoustical Society of America*, vol. 43, no. 1, pp. 163–163 (1968), URL https://doi.org/10.1121/1.1910747.

[9] D. H. Mershon, D. H. Desaulniers, T. L. Amerson, S. A. Kiefer, "Visual capture in auditory distance perception: Proximity image effect reconsidered." *Journal of Auditory Research* (1980).

[10] L. Hládek, C. C. Le Dantec, N. Kopčo, A. Seitz, "Ventriloquism effect and aftereffect in the distance dimension," presented at the *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, p. 050042 (2013), URL https://doi.org/10.1121/1.4799881].

[11] E. R. Calcagno, E. L. Abregu, M. C. Eguía, R. Vergara, "The role of vision in auditory distance perception," *Perception*, vol. 41, no. 2, pp. 175–192 (2012), URL https://doi.org/10.1068/p7153.

[12] H.-J. Maempel, M. Jentsch, "Audio-visual interaction of size and distance perception in concert halls-a preliminary study," presented at the *Proc. International Symposium on Room Acoustics (ISRA) Toronto* (2013).

[13] G. Plenge, "On the problem of in head localization," *Acta Acustica united with Acustica*, vol. 26, no. 5, pp. 241–252 (1972).

[14] A. Neidhardt, N. Knoop, "Investigating the room divergence effect in binaural playback," presented at the *3rd International Conference on Spatial Audio (ICSA)* (2015).

[15] J. C. Gil-Carvajal, J. Cubick, S. Santurette, T. Dau, "Spatial hearing with incongruent visual or auditory room cues," *Nature Scientific Reports*, vol. 6, p. 37342 (2016), URL `https://doi.org/10.1038/srep37342`.

[16] J. Udesen, T. Piechowiak, F. Gran, "The effect of vision on psychoacoustic testing with headphone-based virtual sound," *Journal of the Audio Engineering Society*, vol. 63, no. 7/8, pp. 552–561 (2015), URL `https://doi.org/10.17743/jaes.2015.0061`.

[17] S. Werner, F. Klein, T. Mayenfels, K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," presented at the *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pp. 1–6 (2016), URL `https://doi.org/10.1109/QoMEX.2016.7498973`.

[18] P. Larsson, D. Västfjäll, M. Kleiner, *et al.*, "Auditory-visual interaction in real and virtual rooms," presented at the *Proceedings of the Forum Acusticum, 3rd EAA European Congress on Acoustics, Sevilla, Spain* (2002).

[19] D. Cabrera, A. Nguyen, Y. Choi, "Auditory versus visual spatial impression: A study of two auditoria," presented at the *ICAD 04-Tenth Meeting of the International Conference on Auditory Display* (2004).

[20] B. N. Postma, B. F. Katz, "The influence of visual distance on the room-acoustic experience of auralizations," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. 3035–3046 (2017), URL `https://doi.org/10.1121/1.5009554`.

[21] P. Bertelson, M. Radeau, "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," *Perception & psychophysics*, vol. 29, no. 6, pp. 578–584 (1981), URL `https://doi.org/10.3758/BF03207374`.

[22] D. H. Warren, R. B. Welch, T. J. McCarthy, "The role of visual-auditory compellingness in the ventriloquism effect: Implications for transitivity among the spatial senses," *Perception & Psychophysics*, vol. 30, no. 6, pp. 557–564 (1981), URL `https://doi.org/10.3758/bf03202010`.

[23] J. Lewald, R. Guski, "Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli," *Cognitive brain research*, vol. 16, no. 3, pp. 468–478 (2003), URL `https://doi.org/10.1016/S0926-6410(03)00074-0`.

[24] M. T. Wallace, G. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan,

J. A. Schirillo, "Unifying multisensory signals across time and space," *Experimental Brain Research*, vol. 158, no. 2, pp. 252–258 (2004), URL https://doi.org/10.1007/s00221-004-1899-9.

[25] C. R. André, E. Corteel, J.-J. Embrechts, J. G. Verly, B. F. Katz, "Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and wave field synthesis," *International journal of human-computer studies*, vol. 72, no. 1, pp. 23–32 (2014), URL https://doi.org/10.1016/j.ijhcs.2013.09.004.

[26] B. N. Postma, B. F. Katz, "Influence of visual rendering on the acoustic judgements of a theater auralization," presented at the *Proceedings of Meetings on Acoustics 173EAA*, vol. 30, p. 015008 (2017), URL https://doi.org/10.1121/2.0000575.

[27] M. Schutte, S. D. Ewert, L. Wiegrebe, "The percept of reverberation is not affected by visual room impression in virtual environments," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL229–EL235 (2019), URL https://doi.org/10.1121/1.5093642.

[28] N. Zacharov, T. Pedersen, C. Pike, "A common lexicon for spatial sound quality assessment-latest developments," presented at the *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pp. 1–6 (2016), URL https://doi.org/10.1109/QoMEX.2016.7498967.

[29] S. E. Guttman, L. A. Gilroy, R. Blake, "Hearing what the eyes see: Auditory encoding of visual temporal sequences," *Psychological science*, vol. 16, no. 3, pp. 228–235 (2005), URL https://doi.org/10.1111/j.0956-7976.2005.00808.x.

[30] D. Alais, D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current biology*, vol. 14, no. 3, pp. 257–262 (2004), URL https://doi.org/10.1016/j.cub.2004.01.029.

[31] P. W. Anderson, P. Zahorik, "Auditory/visual distance estimation: accuracy and variability," *Frontiers in psychology*, vol. 5, p. 1097 (2014), URL https://doi.org/10.3389/fpsyg.2014.01097.

[32] D. Cabrera, D. Lee, M. Yadav, W. L. Martens, "Decay envelope manipulation of room impulse responses: Techniques for auralization and sonification," presented at the *Proceedings of ACOUSTICS* (2011).

[33] A. E. Society, "AES Recommended Practice for Professional Audio: Subjective Evaluation of Loudspeakers," Tech. rep., Audio Engineering Society (2008).

[34] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," presented at the *40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, p. 29 (2013).

[35] M. Zaunschirm, C. Schörkhuber, R. Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627 (2018), URL https://doi.org/10.1121/1.5040489.

[36] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, vol. 141, pp. 3678–3688 (2017a), URL `https://doi.org/10.1121/1.4978612`.

[37] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, C. de Boishéraud, "Improvement of Externalization by Listener and Source Movement Using a "Binauralized" Microphone Array," *J. Audio Eng. Soc.*, vol. 65, pp. 589–599 (2017b), URL `https://doi.org/10.17743/jaes.2017.0018`.

[38] D. R. Begault, E. M. Wenzel, M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, pp. 904–916 (2001).

[39] D. R. Begault, "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc.*, vol. 40, pp. 895–904 (1992).

[40] M. Kronlachner, "Plug-in Suite for mastering the production and playback in surround sound and ambisonics," *Gold-Awarded Contribution to AES Student Design Competition* (2014), URL `http://www.matthiaskronlachner.com/?p=2015`.

[41] S. Zielinski, F. Rumsey, S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451 (2008).

[42] I. Borg, P. J. Groenen, *Modern multidimensional scaling: Theory and applications* (Springer Science & Business Media) (2005), URL `https://doi.org/10.1007/978-1-4757-2711-1`.

[43] J. D. Carroll, J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319 (1970), URL `https://doi.org/10.1007/BF02310791`.

[44] W. L. Martens, N. Zacharov, "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL," presented at the *Audio Engineering Society Convention 109* (2000).

[45] ISO3382-1, "Acoustics-Measurement of Room Acoustic Parameters, Part 1: Performance spaces," Standard, International Organization for Standardization, Geneva, CH (2009).

[46] F. Salmon, É. Hendrickx, N. Épain, Q. George, M. Paquier, "The Influence of the Sound Source on Perceived Differences between Binaurally Rendered Sound Spaces," presented at the *AES international conference on headphone technology* (2019).

[47] E. Parizet, V. Koehl, "Application of free sorting tasks to sound quality experiments," *Applied Acoustics*, vol. 73, no. 1, pp. 61–65 (2012), URL `https://doi.org/10.1016/j.apacoust.2011.07.007`.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Université de Bretagne Occidentale

b com

Titre : Contrôle des impressions spatiales dans un environnement acoustique virtuel.

Mots clés : Perception de l'espace sonore, Acoustique virtuelle, Réverbération, Binaural

Résumé : Les travaux présentés dans cette thèse participent à l'élaboration de nouveaux outils de conception d'un espace sonore pour la production de contenus immersifs. En particulier, ils visent à permettre le contrôle des impressions spatiales : la largeur apparente de source et l'enveloppement. Pour cela, plusieurs études ont été réalisées afin d'évaluer la perception de l'acoustique de salles selon une méthode de reproduction sonore communément employée dans ce contexte : un rendu binaural non-individualisé de scènes sonores encodées en ambisonique. Les contenus immersifs étant généralement composés d'un environnement visuel à 360° et d'une grande diversité de sources sonores, la première partie de cette thèse traite de l'influence de ces deux composantes sur la perception de l'acoustique d'une salle. La seconde partie aborde la paramétrisation de réponses impulsionnelles spatiales de salle (SRIRs) qui caractérisent le trajet acoustique entre la source sonore et l'auditeur. Nous avons cherché à optimiser les résolutions spatiales, fréquentielles et temporelles des SRIRs, permettant ainsi une réduction de la quantité de données et une meilleure appréhension du contrôle perceptif tout en conservant un rendu convenable de l'espace sonore. La dernière partie se concentre sur la modification de SRIRs pour le contrôle des impressions spatiales. Afin d'évaluer le contrôle de la largeur apparente de source, plusieurs transformations spatiales ont été appliquées aux signaux ambisoniques d'ordre 1 utilisés pour décrire les premières réflexions d'une SRIR. Une méthode de contrôle de la sensation d'enveloppement a été proposée en utilisant un réverbérateur artificiel qui permet la modification des amplitudes et pentes de décroissance de l'énergie sonore dans plusieurs bandes de fréquence et secteurs angulaires.

Title: Control of spatial impressions in virtual acoustic environments.

Keywords: Sound space perception, Virtual Acoustics, Reverberation, Binaural

Abstract: The work presented in this thesis contributes to the development of new tools that facilitate the design of immersive sound spaces for the production of Virtual Reality contents. More specifically, we aimed to allow the control of spatial impressions : the apparent source width and the sensation of envelopment. To this end, several studies have been conducted to evaluate the perception of room acoustics using a non-individualized binaural rendering of ambisonic sound scenes, which is common in this context. Since immersive contents usually consist of a 360° visual environment and a wide variety of sound sources, the first part of this thesis focuses on the influence of these two components on the perception of room acoustics. The second part addresses the parameterization of Spatial Room Impulse Responses (SRIRs), which are transfer functions that describe the acoustic path between the sound source and the listener. The spatial, temporal and frequency resolutions of SRIRs were studied to seek a possible reduction in the amount of data and thus a better apprehension of the perceptual control, while maintaining spatial and timbral fidelity. The last part focuses on the modification of SRIRs for the control of spatial impressions. In order to evaluate the control of the apparent source width, several spatial transformations were applied to the first-order ambisonic signals used to describe the early reflections of SRIRs. For the control of the sensation of envelopment, a method was proposed using an artificial reverberator that allows the modification of the sound energy amplitudes and decay slopes in several frequency bands and angular sectors.