Etude de la Perception d'Environnements Acoustiques 3D

Simon Fargeot

▶ **To cite this version:**

Simon Fargeot. Etude de la Perception d'Environnements Acoustiques 3D. Acoustique [physics.class-ph]. AMU - Aix Marseille Université, 2022. Français. NNT : 2022AIXM0480 . tel-04014652

HAL Id: tel-04014652
https://hal.science/tel-04014652v1

Submitted on 4 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Aix*Marseille université
Socialement engagée

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 1er décembre 2022 par

Simon FARGEOT

Etude de la Perception d'Environnements Acoustiques 3D

Discipline
Sciences pour l'ingénieur

Spécialité
Acoustique

École doctorale
ED 353 - SCIENCES POUR L'INGENIEUR : MECANIQUE, PHYSIQUE, MICRO ET NANOELECTRONIQUE

Laboratoire
Laboratoire PRISM
CNRS - AMU - UMR 7061

Composition du jury

Mathieu PAQUIER UBO, Brest	Rapporteur
Olivier WARUSFEL IRCAM - STMS, Paris	Rapporteur
Stefania SERAFIN Université d'Aalborg, Copenhague, Danemark	Examinatrice
Bruno TORRESANI AMU, Marseille	Président du jury
Mitsuko ARAMAKI CNRS - PRISM - AMU, Marseille	Directrice de thèse
Richard KRONLAND-MARTINET CNRS - PRISM - AMU, Marseille	Co-Directeur de thèse

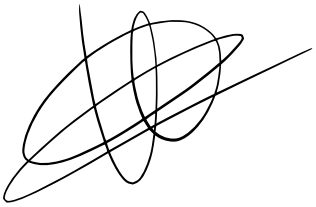
PRISM
Perception
Représentations
Image
Son
Musique

Affidavit

Je soussigné, Simon FARGEOT, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Mitsuko ARAMAKI et Richard KRONLAND-MARTINET, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 28/09/2022



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et participation aux conférences

Liste des publications réalisées dans le cadre du projet de thèse :

1. Fargeot, S., Derrien, O., Parseihian, G., Aramaki, M., & Kronland-Martinet, R. (2019, September). Subjective evaluation of spatial distorsions induced by a sound source separation process. In *EAA Spatial Audio Signal Processing Symposium* (pp. 67-72).
2. Blaise, J.Y., Dudek, I., Pamart, A., Bergerot, L., Vidal, A., Fargeot, S., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2020). Space & sound characterisation of small-scale architectural heritage : an interdisciplinary, lightweight workflow. *IMEKO TC-4 International Conference on Metrology for Archaeology and Cultural Heritage*, pp.263-268.
3. Blaise, J.Y., Dudek, I., Pamart, A., Bergerot, L., Vidal, A., Fargeot, S., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2022). Acquisition & integration of spatial and acoustic features : a workflow tailored to small-scale heritage architecture. *ACTA IMEKO, Vol 11(No 2 (2022))*, pp. 1-14.

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Communication à EAA Spatial Signal Processing Symposium 2019.
2. Communication au 16e Congrès Français d'Acoustique, Marseille 2022.

Résumé

L'étude de la perception des environnements acoustiques est un champ de recherche largement exploré depuis le début du XXème siècle. Ces dernières décennies, l'essor de techniques et outils de captation et d'auralisation de champs acoustiques spatialisés, couplé au développement des technologies de réalité virtuelle permettent d'envisager de nouvelles problématiques relatives à l'immersion et à l'impression spatiale de ces environnements. Les travaux présentés dans cette thèse visent à proposer des nouveaux protocoles expérimentaux, basés sur ces outils et techniques, pour la caractérisation de la perception d'environnements acoustiques en conditions d'immersion 3D.
La première expérience présentée dans cette thèse porte sur l'évaluation perceptive de la qualité spatiale de l'auralisation ambisonique d'acoustiques de salles mesurées, en comparant les performances de localisation de sources sonores en contexte d'auralisation avec celles observées en conditions réelles d'écoute. Cette expérience, basée sur une méthode de report des attributs spatiaux de sources en réalité virtuelle met en évidence des dégradations de la précision spatiale des sources sonores induites par l'auralisation, et révèle également une dépendance de ces attributs saptiaux aux différentes conditions acoustiques à l'étude. Dans un second temps, un travail de caractérisation multi-dimensionnelle (acoustique et perceptive) d'un corpus d'acoustiques mesurées dans une vingtaines d'édifices du patrimoine est présenté. La caractérisation acoustique, basée sur le calcul d'un ensemble de descripteurs acoustiques, permet de proposer une cartographie acoustique des édifices et une catégorisation des paramètres acoustiques en trois grands facteurs. De son côté, la caractérisation perceptive, fondée sur la méthodologie de la première expérience révèle que, dans un contexte d'auralisation HOA, la distance perçue et la précision spatiale des sources semblent affectées par des propriétés acoustiques des salles étudiées. La troisième expérience propose d'étudier l'adéquation audio-visuelle entre des environnements acoustiques et leurs représentations visuelles en contexte d'immersion multi-modale. Cette étude révèle que la cohérence audio-visuelle est principalement évaluée sur une relation simple entre le temps de réverbération de la salle et le volume représenté visuellement. Ces travaux ouvrent des perspectives concrètes autour de l'amélioration des techniques d'auralisation, la caractérisation perceptive des acoustiques du patrimoine et la synthèse d'espaces acoustiques en contexte d'immersion multi-sensorielle.

Mots clés : perception auditive spatiale, perception audio-visuelle, auralisation, acoustique des salles, réalité-virtuelle.

Abstract

Perception of acoustic environments is a research field that has been widely explored since the beginning of the 20th century. In recent decades, the development of techniques and tools for capture and auralization of 3D acoustic fields, coupled with the development of virtual reality technologies, has allowed us to consider new problems related to the immersion and spatial sensation of these environments. The work presented in this thesis aims to develop new experimental protocols based on such tools and techniques to characterize the perception of acoustic environments in immersive conditions.
The first experiment presented in this thesis deals with perceptual evaluations of the spatial quality of ambisonic auralizations of measured room acoustics, based on comparisons between localization performances of sound sources in auralization contexts with those observed in real listening conditions. This experiment, based on a method in which spatial attributes of sources were reported in virtual reality, highlights typical spatial distortions of sound sources induced by the auralization process, and also reveals that these degradations are influenced by the different acoustic conditions under study. In a second chapter, a multi-dimensional characterization (acoustic and perceptual) of an acoustic corpus measured in twenty heritage buildings is presented. The acoustic characterization, based on the calculation of a set of acoustic parameters, allows to propose an acoustic map of the buildings and a categorization of the acoustic parameters in three main factors. The perceptual characterization, based on the methodology of the first experiment, reveals that, in a context of HOA auralization, the perceived distance and the spatial accuracy of sound sources seem to be affected by the acoustic properties of the studied rooms. The third experiment investigates the audio-visual adequacy between acoustic environments and their visual representations in a multi-modal immersive context. This study reveals that the audio-visual coherence of acoustic spaces is mainly evaluated on a simple relation between the reverberation time of the room and its visually represented volume. This work opens concrete prospects regarding the improvement of auralization techniques, the perceptual characterization of heritage acoustics and the synthesis of acoustic spaces in a multi-sensory immersive context.

Keywords: spatial hearing, audio-visual perception, auralization, room acoustics, VR

Remerciements

« Merci, Merci, Merci » - *Cannonball Adderley*

Quelle joie d'ouvrir ce document et dans le même temps, d'en achever la rédaction par une séance de tirages de chapeaux, de serrages de paluches, d'accolades et d'embrassades sur papier. La liste est longue alors allons-y !

Je tiens tout d'abord à remercier mes directeurs de thèse Mitsuko Aramaki et Richard Kronland-Martinet de m'avoir ouvert les portes du laboratoire PRISM et pour la confiance qu'ils m'ont accordée pendant ces quatre années de thèse. Merci Mitsuko pour ta gentillesse, ta patience et pour ton aide précieuse, notamment dans cette fastidieuse étape de rédaction. Merci Richard pour ces échanges passionnés autour du son 3D et du son en général et de m'avoir montré qu'il existe encore des endroits où il est possible de rêver. Merci aussi à Sølvi (Ystad) d'avoir pris part, avec enthousiasme, aux discussions et réunions de chantier tout au long de l'aventure et d'avoir veillé sur moi et mon équilibre psychique depuis Z'.

Je remercie chaleureusement les membres du jury Mathieu Paquier, Olivier Warusfel, Stefania Serafin et Bruno Torresani, d'avoir accepté de s'intéresser à mon travail, pour la qualité de nos échanges lors de la soutenance et pour leurs conseils aguerris qui me permettront de valoriser au mieux mes travaux présents et futurs.

Merci à mon cher acolyte de son 3D, Adrien (Vidal), pour son aide considérable, en première ligne et au quotidien, ainsi que pour tous ces bons moments lors de nos multiples escapades rurales ou à l'occasion de nos innombrables combos Bardoin-Pizza du Cours-Ju. Big-Up à mes ainés Antoine (Bourachot) et Thomas (Bordonné), et à mes frères d'armes Samuel (Poirot), Corentin (Bernard), Théophile (Dupré) et Pierre (Fleurence), pour la bonne ambiance au labo, les parties de coinche endiablées et les apéros de fin de journée, sans lesquels ces quatre années de thèse n'auraient pas eu la même saveur. J'en place une toute spéciale pour le Tonton Etienne (Thoret) pour son précieux soutien, ses conseils de briscard et ses « coups de pieds au cul » toujours dans le tempo. Je ne pouvais pas conclure ce paragraphe sans remercier Matis (Damnon) et Sofiane (Azzouz) pour l'aide importante qu'ils m'ont apportée durant leurs stages respectifs.

Distribution de fières chandelles à tous les prismiens et toutes les prismiennes que j'ai eu la chance de côtoyer tout au long de la thèse. Côté doctorantes : Baptistine (Marcel), Ludmila (Postel) et Caroline (Boé). A la gestion : Claudine (Le Van Phu) et Julie (Perret). Dans les couloirs ou autour d'un café : Jocelyn (Roze), Khoubeib (Kanzari), Salomé (Sudre). Un grand merci à Gaëtan (Parseihian) d'avoir

su, lors de mon stage de Master, éveiller en moi ce goût pour la recherche.

Merci à nos voisins du MAP : Jean-Yves Blaise, Iwona Dudek de m'avoir permis, en temps de pandémie, de m'évader dans la campagne provençale lors des mythiques sessions de mesures du projet Sésames ; Laurent (Bergerot), Anthony (Pamart) et Ariane (Neroulidis) d'y avoir pris part. Enfin, à l'ensemble du staff Joseph Aiguier, que ce soit à l'accueil, à la sécu ou à la cantoche, merci de rendre cette vie de campus si agréable.

Il est l'heure de quitter le campus pour saluer et remercier comme il se doit les copains et les copines. Aux stéphano-phocéens de la team volley/musique/bateau : Lulu, Vinet, Raymond, Romain, Jéjé, merci pour ces moments de plaisir au Prophète, dans les studios Maurel, à Pélu ou sur les flots. Dédicace à mes amours de Grenoble : Adrien, Ponpon, Thomas, Naïma, merci pour votre sens de l'hospitalité. J'embrasse chaleureusement les anciens du Mans : Colas, Louis et Gwen (aka les bobos crados), ainsi que Bapt' le breton et K1000 le bordelais. Et je termine la tournée par un gros bécot à mon Gab, par-dessus l'Atlantique ainsi qu'aux collègues Pabz et Ze.

À ma mère, celle qui, de par son sens hors norme de l'écoute, pourrait inspirer les plus grands spécialistes de l'audition, celle qui comme personne sait m'apporter paix et réconfort quand j'en ai besoin, merci. À mon père, de m'avoir donné dans le biberon ce goût pour la musique et les clés pour le développer, merci. À mon frère, parce qu'il est pour moi une source de motivation et d'inspiration et pour l'immense fierté que je ressens de l'avoir dans mon équipe, merci.

À celle qui navigue à mes côtés, qui m'accompagne et me supporte qu'il vente ou qu'il pleuve. À ma douce Marianne, merci pour ces moments de bonheur, pour ta tendresse et ton infinie patience.

À toutes celles et tous ceux que j'aime, merci.

Table des matières

Site Internet

Ce document est accompagné d'un site internet sur lequel figurent des compléments d'information, notamment les stimuli sonores et visuels des expériences perceptives et les différents jeux de réponses impulsionnelles spatiales au format HOA 4, mesurées à l'Eigenmike (em32) de mh-acoustics, dans plus d'une vingtaines d'environnements acoustiques.

https://www.prism.cnrs.fr/publications-media/TheseFargeot/

Avant-propos

Les travaux présentés dans ce document ont été réalisés au laboratoire PRISM entre fin 2018 et début 2022. Ils abordent différents aspects de la perception en environnements réverbérants en contexte d'auralisation : perception auditive spatiale et perception audio-visuelle. Ces recherches ont été élaborées en lien étroit avec le projet ANR Sésames (*Sémantisation Et Spatialisation d'Artefacts patrimoniaux Multi-Echelles : annotation 3D, Sonification et formalisation du raisonnement*) . Ce projet est le fruit d'une collaboration entre architectes du patrimoine du laboratoire MAP et chercheurs en acoustique du laboratoire PRISM, autour de la question de la caractérisation multi-dimensionnelle d'un corpus de chapelles rurales de la région PACA.

Il est important de noter que les différents travaux décrits ici ont été menés de front plutôt que de façon séquentielle et ce du fait d'une série de contraintes temporelles et techniques imposées par le projet Sésames et le contexte Covid-19. En effet, dans le cadre de Sésames, une vaste campagne de mesure de données métriques (proportions, volumes, ...), visuelles (photos panoramiques, nuages de points) et sonores (réponses impulsionnelles spatiales) a été conduite dans les différents édifices de la collection. Cette campagne, déployée dans une vingtaine de lieux du patrimoine français, s'est trouvée fortement ralentie par les restrictions sanitaires imposées par la pandémie sur la période 2020-2021. Les passations des expériences perceptives ont également été impactées voire rendues carrément impossibles pendant une bonne partie de l'année 2020.

D'autre part, alors que les questions relatives à la perception auditive spatiale en environnements réverbérants ont été envisagées en amont du projet Sésames, c'est grâce à la collecte groupée de données visuelles et sonores que nous avons pu envisager et aborder l'étude la perception audio-visuelle des environnements acoustiques.

Pour toutes ces raisons, l'organisation du manuscrit ne respecte pas toujours la chronologie réelle des recherches et nous invitons les lecteurs et lectrices de ce document à appréhender les différents chapitres de cette thèse comme relativement indépendants les uns des autres.

Introduction

Au quotidien, nous évoluons dans des environnements variés que nous percevons à travers nos différents sens. Un environnement acoustique, au sens où nous l'entendons dans cette thèse, est un espace totalement ou partiellement clos, délimité par des parois capables de restituer une partie de l'énergie produite par une ou plusieurs sources acoustiques, résultant en un phénomène sonore pouvant être entendu par un auditeur. Ce phénomène, appelé réverbération, nous informe par le son sur l'environnement dans lequel nous nous trouvons. En se basant sur différentes propriétés de la réverbération, nous sommes par exemple capables de déterminer si nous nous trouvons plutôt dans une grande cathédrale ou dans un petit cagibi. Alors que la structure spectro-temporelle de la réverbération est porteuse d'informations relatives à la nature de l'environnement (sa taille, les matériaux qui le compose), la dimension spatiale de ce phénomène est essentielle pour nous situer, au sens d'être conscient de sa position et de son orientation, dans cet espace. Certaines personnes sont même capables de déterminer précisément la position d'obstacles présents dans leur environnement, en se basant uniquement sur la modalité auditive. Cette faculté extra-ordinaire, appelée écholocalisation est également observée chez d'autres espèces animales telles que les chauves-souris et les dauphins.

Outre sa capacité à véhiculer de l'information sur notre environnement, la réverbération est empreinte d'une dimension culturelle et esthétique forte. L'acoustique a de tout temps joué un rôle important dans le développement culturel des civilisations. Des études ont par exemple montré une corrélation spatiale entre la présence de peintures rupestres et les propriétés acoustiques des grottes dans lesquelles elles avaient été produites, semblant indiquer qu'au paléolithique déjà, les hommes des cavernes choisissaient l'emplacement de leurs rites en se basant sur des considérations acoustiques [Reznikoff, 2008, Iannace et Trematerra, 2014, Fazenda *et al.*, 2017]. Sur le plan architectural, il ne fait aucun doute qu'une attention forte était portée sur l'acoustique des édifices destinés à véhiculer un message et ce dès leur conception. La forme des théâtres antiques était par exemple conçue pour porter le son jusqu'aux auditeurs les plus éloignés du locuteur situé sur la scène. La longue réverbération des grandes cathédrales donnait un caractère céleste à ces lieux sacrés et aux messages qui y sont diffusés. Aujourd'hui, les chercheurs s'accordent à penser que la compréhension du patrimoine architectural doit passer par la caractérisation de l'acoustique de ces lieux sur le plan objectif, comme cela a déjà été fait par le passé mais également sur le plan de la perception que nous en avons.

D'autre part, les qualités esthétiques de la réverbération ont également suscité un grand intérêt dans le domaine musical et avec l'émergence des technologies électroacoustiques, au début du XX$^{\text{ème}}$ siècle, les ingénieurs ont cherché à capturer

ou à reproduire de manière artificielle les effets produits par la réverbération. Alors que les premiers effets de réverbération étaient obtenus en enregistrant des sources musicales dans des vraies salles réverbérantes, appelées chambre d'échos, Hammond développe une réverb artificielle à ressort et en équipe ses orgues électriques au début des années 40. D'autres systèmes électro-mécaniques du même genre sont ensuite développés pour imiter l'effet de la réverbération. Le principe consiste à exciter une structure mécanique résonante (ressort, plaque, gong, cordes) avec le son que l'on souhaite affecter et à enregistrer le résultat produit par la réponse du système à cette excitation. De nos jours, la plupart des effets de réverbération sont produits numériquement. Dans la lignée des travaux précurseurs de Schroeder [Schroeder, 1962], de nombreux algorithmes de réverbération artificielle ont vu le jour. Parmi les méthodes les plus répandues, on peut citer les méthodes FDN (*Feedback Delay Network*), basées sur l'utilisation de bancs de filtres à boucle de retard [Gerzon, 1976, Jot, 1992], les méthodes par convolution [Gardner, 1994], permettant l'utilisation temps-réel de réponses impulsionnelles mesurées ou bien simulées numériquement par différentes approches par modèles physiques (images sources, lancé de rayon, guides d'ondes), que l'on retrouve dans les principaux outils de simulation acoustiques actuels, tels que CATT-Acoustic ou ODEON. Ces méthodes numériques permettent une émulation de grande qualité du comportement acoustique des salles et offrent aux ingénieurs du son et designers sonores de vastes possibilités de contrôle.

Aujourd'hui de nouvelles problématiques relatives à la perception des environnements acoustiques sont portées par l'essor des technologies immersives et particulièrement de la réalité virtuelle (VR). En effet, l'objectif premier de la VR est d'arriver à procurer chez les utilisateurs un sentiment de présence et d'immersion dans des environnements multi-sensoriels totalement artificiels. Pour cela, un grand soin doit être apporté au réalisme des environnements visuels et sonores présentés mais également à la cohérence entre ces deux modalités. L'auralisation, c'est-à-dire la capacité à capter, simuler et reproduire des environnements acoustiques réalistes, constitue un des piliers de ce nouveau médium et la compréhension de la perception auditive et multimodale des environnements acoustiques en contexte d'auralisation représente un enjeu majeur de ces nouvelles technologies.

D'un côté, nous pensons qu'afin d'accroitre le sentiment de présence et d'immersion dans des environnements virtuels, la prise en compte des propriétés spatiales de la réverbération lors de l'auralisation est essentielle puisque celles-ci permettent de situer le sujet dans son environnement. Il est aujourd'hui possible de simuler des environnements acoustiques 3D d'une grande fidélité, à l'aide d'outils basés sur la modélisation du comportement physiques des salles. Des techniques permettant la mesure et la restitution d'environnements acoustiques réels ont également été développées. Chacune de ces solutions est accompagnée de son lot d'avantages et de limitations et de nouvelles méthodologies doivent être inventées pour permettre

d'évaluer perceptivement la qualité du rendu spatial de ces techniques dans un contexte d'auralisation.

D'un autre côté, en contexte d'immersion multi-sensorielle, la plausibilité d'un espace virtuel est fortement dépendante de la cohérence entre les environnements sonores et visuels présentés. On peut sans nul doute faire l'hypothèse que quelque chose ne tournerait pas rond si vous vous retrouviez immergé(e) dans une scène représentant l'intérieur d'un chalet alpin avec l'acoustique d'un parking sous-terrain. Il semble alors important de comprendre d'avantage les mécanismes multi-modaux nous permettant d'associer ou non un environnement acoustique virtuel à sa représentation visuelle.

Dans ce document, ces différents aspects de la perception des environnements acoustiques seront abordés. Le manuscrit est organisé de la façon suivante :

Dans un premier temps, nous ferons un tour d'horizon de la littérature relative à la perception auditive en environnement réverbérant et aux différentes techniques permettant l'auralisation de ces environnements dans le but de choisir une solution technique appropriée à nos problématiques de recherches.

Dans le deuxième chapitre, la qualité de restitution spatiale d'un système d'auralisation HOA d'environnements acoustiques mesurés sera évaluée. Il s'agira de comparer les performances de localisation de sources sonores dans différents environnements acoustiques réels et auralisés. Une méthode de report des attributs spatiaux de sources sonores (position angulaire, distance et taille apparente) basée sur une interface en réalité virtuelle sera également présentée.

Le troisième chapitre présente une vaste étude visant à caractériser une collection d'environnements acoustiques mesurés dans le cadre du projet ANR Sésames. Après avoir introduit le projet Sésames et le protocole d'acquisition de données acoustiques, métriques et visuelles, déployé dans une vingtaine d'édifices patrimoniaux de la région PACA, une première caractérisation objective de ce corpus sera réalisée sur la base de descripteurs acoustiques calculés pour l'ensemble des réponses impulsionnelles collectées. Ensuite, nous décrirons une expérience perceptive basée sur le protocole de localisation de source en VR présenté dans le chapitre précédent. Cette expérience perceptive vise à étudier les performances de localisation dans plus de vingt environnements acoustiques mesurés et auralisés. Pour finir nous discuterons des relations entre les résultats issus des caractérisations acoustiques et perceptives.

Le dernier chapitre de cette thèse aborde la question de la perception audio-visuelle d'environnements acoustiques virtuels. A partir d'une sélection de données visuelles et sonores mesurées dans le cadre du projet Sésames, une expérience perceptive visant à comprendre les mécanismes perceptifs qui nous permettent de juger de la cohérence entre l'acoustique d'un lieu et sa représentation visuelle sera présentée. Un modèle de la cohérence audio-visuelle d'environnements acoustiques sera proposé.

Nous conclurons finalement en récapitulant les contributions principales de cette

thèse et les multiples perspectives de recherche qui en découlent.

Chapitre 1

Etat de l'art / Travaux connexes

Sommaire

Les recherches présentées dans cette thèse s'inscrivent dans la thématique de l'étude de la perception des acoustiques de salles. Cette vaste thématique d'étude a été abordée dès le début des années 1900, avec un attention particulière portée sur la caractérisation perceptive de la qualité acoustique des salles de spectacles. Depuis une vingtaine d'années, les questions liées à la perception des salles ont connu un essor important facilité par le développement d'outils et techniques de captation, de synthèse et de restitution sonore spatialisée, permettant l'auralisation de ces environnements en laboratoire. Dans ce chapitre nous donnerons en premier lieu une description générale des environnements acoustiques puis présenterons quelques travaux phares associés à l'étude de la perception de ces environnements. Dans un second temps, les grandes techniques de spatialisation et d'auralisation d'environnements acoustiques seront présentées. Enfin nous présenterons les solutions méthodologiques et techniques retenues pour mener à bien les différents travaux de recherches présentés dans ce document.

1.1 Caractéristiques des environnements acoustiques

On entend par environnement acoustique, environnement réverbérant ou plus simplement salle, tout espace tri-dimensionnel clos, soumis à l'excitation acoustique d'une ou plusieurs sources sonores. Le phénomène de réverbération des salles se caractérise par la décroissance temporelle de l'énergie sonore après extinction

de la source, induite par le fait de la propagation et de réflexions multiples de l'onde acoustique initiée par la source, contre les parois et obstacles présents dans l'environnement. Les propriétés d'absorption et de diffraction de ces éléments ainsi que les propriétés géométriques de l'environnement déterminent le comportement de la réverbération. Celle-ci peut être caractérisée sur le plan objectif par la mesure de réponses impulsionnelles, notées RI (RIR en anglais pour *Room Impulse Response*), caractérisant la fonction de transfert entre la source d'une impulsion acoustique et un récepteur, typiquement un microphone.

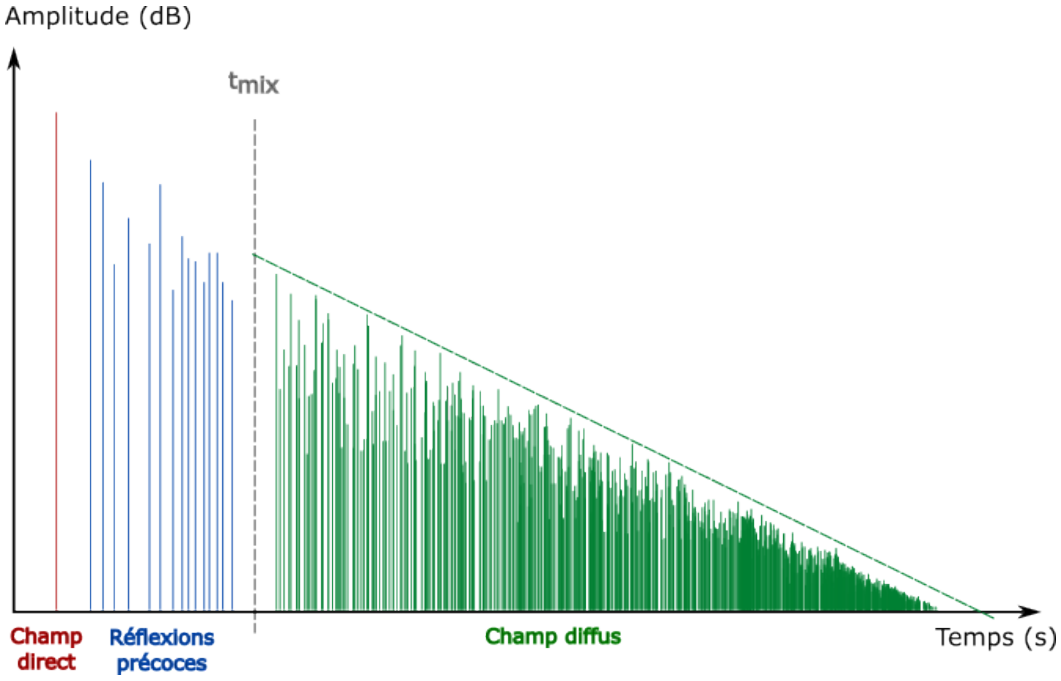


FIGURE 1.1 – Structure temporelle d'une réponse impulsionnelle de salle.

Comme représentée en figure 1.1, la réponse impulsionnelle se caractérise dans le domaine temporel par une série d'impulsions pouvant être décrite selon trois régimes. Le premier régime, correspondant sur la figure à la première impulsion (à gauche), représente le champ direct qui caractérise la propagation directe de l'onde acoustique produite par la source jusqu'au récepteur. Le deuxième régime, représenté en bleu sur la figure, appelé régime précoce, correspond à la réception par le microphone des premières réflexions éparses de l'onde sonore initiée par la source, sur les parois et obstacles rencontrés par l'onde. Au fur est à mesure, la densité des réflexions ou densité d'écho s'intensifie jusqu'à obtention d'un champ acoustique dense appelé champ diffus ou réverbération tardive. La densité d'écho δ_E, exprimée comme la dérivée au cours du temps du nombre de réflexions N_r, peut être estimée grossièrement comme une fonction du temps et du volume V de la salle, pour des environnements acoustiques quelconques [Kuttruff, 2017] :

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

avec c la célérité du son dans l'air. Le temps de mélange (mixing time, en anglais), noté t_{mix}, introduit par Polack [Polack, 1992], représente la frontière temporelle entre les régimes précoce et diffus et peut être exprimé en première approximation comme suit :

$$t_{mix} \approx \sqrt{V}\,\mathrm{ms} \tag{1.2}$$

avec V le volume de la salle. L'absorption de l'énergie du champ acoustique par l'air et par les obstacles que rencontre l'onde acoustique lors de sa propagation résulte, en régime diffus, en une décroissance exponentielle de l'énergie de la réponse au cours du temps. Autrement dit, sur une échelle logarithmique, la décroissance énergétique de la réponse est linéaire en fonction du temps. La courbe de décroissance de la réponse, représentée en figure 1.2, illustre cette relation. D'après cette courbe, le temps de réverbération de la salle, noté RT_{60} peut être calculé. Il correspond au temps mis par l'énergie de la réponse pour décroître de 60 décibels (dB), après extinction de la source acoustique. Il arrive lors de mesures acoustiques que le niveau de bruit ambiant dans la salle, ne permette pas d'observer une dynamique de 60 dB entre le niveau maximum de la réponse et le niveau de bruit. Dans ce cas, le temps de réverbération est estimé par extrapolation du RT_{60} sur une dynamique plus faible, typiquement de 20 ou 30 dB. Les temps de réverbération ainsi calculés sont respectivement notés RT_{20} et RT_{30}.

Le champ diffus est souvent considéré en théorie comme homogène en énergie en tout point de l'espace et dans toutes les directions. En réalité, il existe des cas de figures où ces deux propriétés du champ diffus ne sont pas observées. C'est par exemple le cas de salles couplées, typiquement un espace composé de plusieurs sous espaces, ou de salles avec des propriétés d'absorption non uniformément distribuées. Dans ce cas, certaines directions de propagation acoustique sont privilégiées et on dit alors que la réponse acoustique de la salle est anisotrope [Nélisse et Nicolas, 1997, Alary *et al.*, 2019]. Une autre caractéristique du champ diffus est qu'il est fortement incohérent, car étant composé d'une multitude de versions retardées et décorrélées du signal source, dues aux modifications induites par les multiples réflexions. Le niveau de cohérence de la réponse, aussi appelé diffusivité, varie en fonction des propriétés géométriques et acoustiques de la salle [Cook *et al.*, 1955, Schultz, 1971].

Sur le plan fréquentiel, on distingue deux régimes régis par des phénomènes acoustiques différents. En basses fréquences, le comportement modal de la salle prédomine et se caractérise par la présence éparse de fortes résonances dans le spectre de la réponse, dues à la formation d'onde stationnaires dans l'environnement clos. A l'instar des modes observées sur une corde de guitare par exemple, les modes de salles

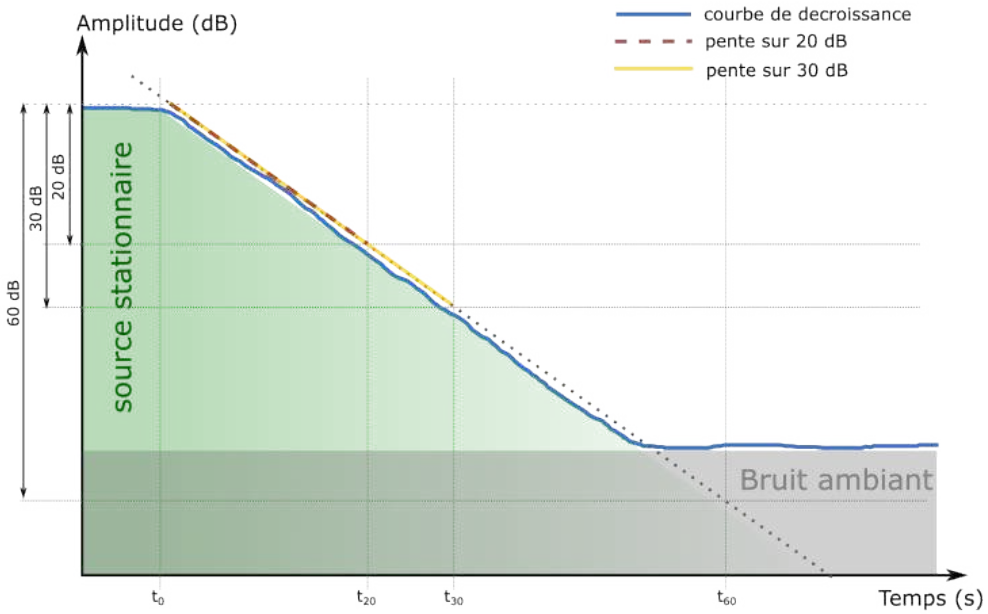


FIGURE 1.2 – Courbe de décroissance de la réponse acoustique d'une salle et calculs des temps de réverbération RT_{20} et RT_{30}.

présentent dans l'espace des nœuds et ventres de vibration, c'est-à-dire des zones de l'espace où la pression acoustique induite par l'onde stationnaire est minimale et d'autres pour lesquelles elle est maximale. Ainsi, dans une salle fortement modale, il est possible d'observer de fortes différences de niveau de pression en fonction des positions dans l'espace de la source et du récepteur. La densité modale δ_f, représente le nombre de modes propres N_f de la salle pour une plage de fréquence donnée. Elle peut être donnée en première approximation comme une fonction quadratique de la fréquence, comme suit [Morse et Ingard, 1986] :

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

avec V le volume de la salle et c la célérité du son dans l'air. Passée une certaine fréquence, les modes propres se superposent et deviennent difficilement discriminables. Schroeder [Schroeder et Kuttruff, 1962] propose une estimation de cette limite fréquentielle, appelée fréquence de Schroeder f_{schr}, en fonction du volume V et du temps de réverbération RT_{60} de l'environnement acoustique, déterminée par la formule suivante :

$$f_{schr} = 2000 \cdot \sqrt{\frac{RT_{60}}{V}} \tag{1.4}$$

Au-delà de cette fréquence limite, la composition modale de la réponse en fréquence de la salle peut être modélisée par un processus stochastique [Schroeder, 1987]. En outre, la décroissance exponentielle de l'énergie au cours du

temps est généralement plus rapide en hautes fréquences qu'en basses fréquences, du fait d'une plus forte absorption des hautes fréquences par l'air et les matériaux constituant les parois et obstacles de la salle.

1.2 Perception auditive spatiale en environnement réverbérant

La perception des environnements acoustiques est un sujet particulièrement complexe à aborder pour plusieurs raisons. D'abord, au quotidien, nous percevons notre environnement avec nos différents sens qui interagissent dans notre appréhension globale du monde [Gibson, 1979]. En ce qui concerne les salles, cette perception multimodale est principalement dictée par la vision et l'audition et les influences réciproques entre ces deux modalités dans la perception des environnements acoustiques est un sujet d'étude en plein essor. Si on laisse la modalité visuelle de côté, la perception auditive en environnements réverbérants est à elle seule multi-dimensionnelle. En effet, elle est d'une part orientée par différentes dimensions physiques du son : dimensions temporelle, fréquentielle et spatiale, mais également par des dimensions culturelles : liées à notre expérience des lieux, à leurs usages, à des considérations esthétiques. Cette perception diffère donc d'un individu à l'autre et varie également au sein d'un même individu en fonction de notre état cognitif au moment de l'écoute (charge mentale, stress, types d'écoute, etc.). Enfin, la perception auditive en environnements réverbérant peut être vue comme la perception du phénomène de l'excitation de celui-ci, par une source acoustique située en son sein. Alors que la présence d'une source sonore est nécessaire pour révéler les caractéristiques acoustique d'un lieu, la perception des sources est fortement dépendante de l'environnement acoustique dans lequel elles se trouvent. Ce dernier point est particulièrement important lorsqu'on s'intéresse à la perception auditive spatiale en situation réverbérante. En effet, la perception auditive spatiale fait référence à l'ensemble des mécanismes perceptifs qui nous permettent, par le biais du son, de nous représenter notre environnement dans l'espace tri-dimensionnel et de nous situer dans celui-ci. Or, ces mécanismes s'appuient sur un certain nombre d'indices acoustiques présent dans le son pouvant être influencés et perturbés par la réverbération. Dans cette section, nous commencerons par présenter brièvement les grands principes et mécanismes de la perception auditive spatiale puis nous verrons comment les environnements acoustiques peuvent influencer cette perception.

1.2.1 Les mécanismes de la perception auditive spatiale

L'homme possède la faculté de localiser des sources sonores dans son environnement avec une précision de quelques degrés. Cette faculté est le résultat d'un traitement de l'information sonore par notre système auditif. En effet, le système auditif est capable d'analyser des indices acoustiques inhérents au comportement du son dans l'espace et d'en déduire la position d'un événement sonore. Ces indices de

localisation sont de deux types :

— **les indices interauraux** (ou binauraux), qui se basent sur les différences entre les signaux arrivant à nos deux oreilles,

— **les indices spectraux** (ou monoraux), correspondants au filtrage du son par le corps de l'auditeur.

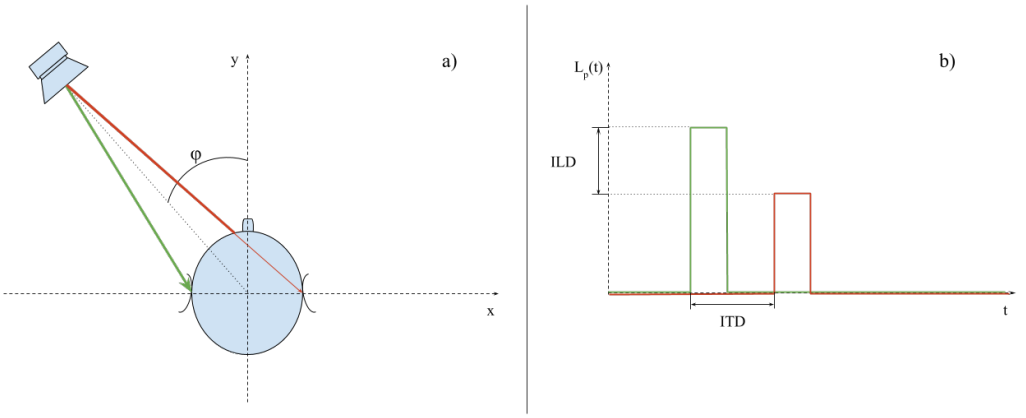


FIGURE 1.3 – Représentation des ILD et ITD. a) Un événement sonore arrivant aux oreilles d'un auditeur avec une incidence φ. b) Niveau de pression acoustique aux deux oreilles en fonction du temps et détermination des indices interauraux (oreille gauche en vert, oreille droite en rouge).

Indices interauraux ou binauraux

D'après Blauert [Blauert, 1997b], qui a synthétisé dans son livre de nombreux travaux sur la perception spatiale des sons, il existe trois principaux indices interauraux : l'indice interaural de niveau : ILD (*Interaural Level Difference*), l'indice interaural de temps : ITD (*Interaural Time Difference*) et l'indice de correlation interaurale : IACC (*InterAural Cross-Correlation*). Les deux premiers indices sont utiles dans la localisation des sources sonores sur un plan horizontal. Ils sont induits par la différence de position sur ce plan entre les deux oreilles, comme le montre la figure 1.3. La différence interaurale de temps est liée à l'angle azimutal φ par une relation trigonométrique simple : $ITD \propto \sin\varphi$. L'ITD intervient de deux manières complémentaires en fonction du type de son émis. Pour les sons continus à basses fréquences ($<$ 1500 Hz), il est déterminé à partir de la différence de phase entre les sons captés par chacune des oreilles. Au-delà de 1500 Hz, cette différence de phase n'est plus caractéristique de la position (phénomène d'aliasing) et l'ITD est plutôt représentatif du retard existant entre les enveloppes des signaux perçus par chacune des oreilles. La différence interaurale de niveau (ILD) est quant à elle due au phénomène d'absorption et de réflexion de l'onde incidente par la tête de l'auditeur, connu sous le nom d'effet d'ombre acoustique. Le son parvenant à l'oreille la plus éloignée de la source est atténué par rapport à celui perçu par l'oreille la plus proche. Ce phénomène est vrai pour des sons de fréquences supérieures à 1500 Hz, pour lesquels la longueur d'onde correspondante est significativement inférieure

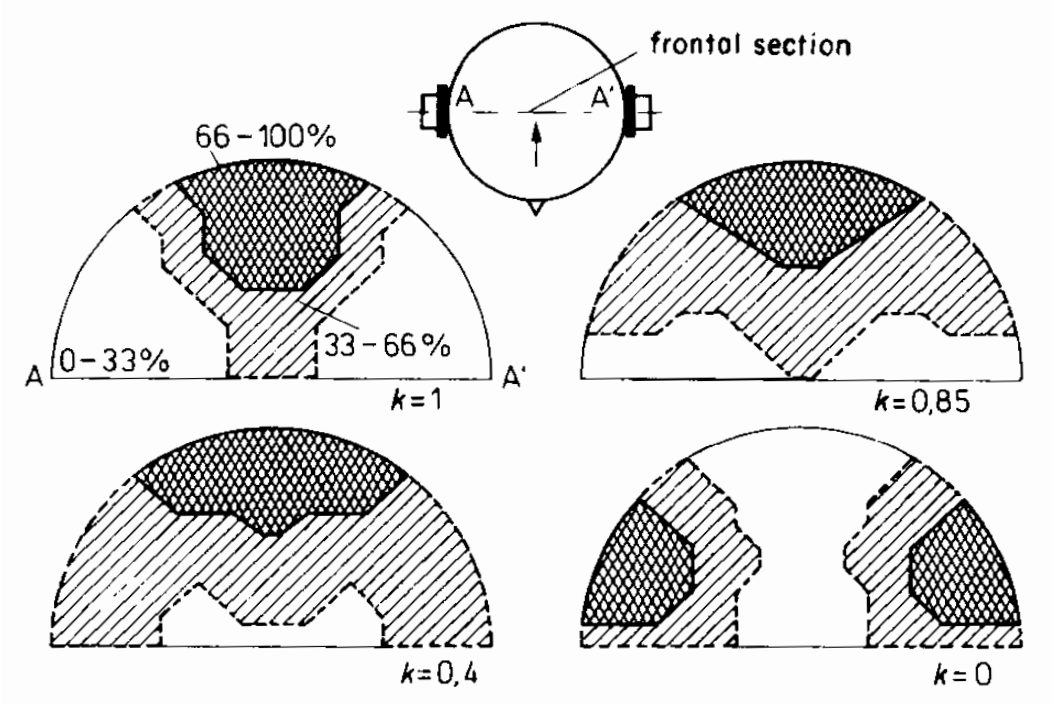


FIGURE 1.4 – Variation de l'image spatiale de l'événement auditif pour deux sources de bruit large-bande présentées au casque, en fonction de leur niveau de cohérence k [Blauert, 1997b, Chernyak, 1968].

au diamètre de la tête. L'erreur de localisation en azimut a été quantifiée par Blauert en 1972 [Blauert, 1997b]. Elle dépend de la nature du son émis : allant de $\pm$ 0.75 $^\circ$ pour des impulsions à $\pm$ 12 $^\circ$ pour des sons purs. Elle est également fonction de la position de la source et varie en moyenne entre $\pm$ 3 $^\circ$ pour les sources sur l'axe frontal de la tête et $\pm$ 10 $^\circ$ pour des sons provenant des côtés.

L'IACC quant à lui est un indice important dans l'évaluation de la largeur apparente des sources sonores (ASW : *Apparent Source Width*). Il représente le niveau de cohérence entre les signaux arrivant à l'entrée des canaux gauche $p_g(t)$ et droit $p_d(t)$ du système auditif et est défini comme le maximum de la fonction d'inter-corrélation (IACF) entre ces deux signaux comme suit :

$$IACC = \max|IACF(\tau)| \tag{1.5}$$

avec, pour une fenêtre temporelle d'analyse donnée, comprise entre t_1 et t_2 :

$$IACF(\tau) = \frac{\int_{t_1}^{t_2} p_g(t) \cdot p_d(t-\tau)\,\mathrm{d}t}{\left(\int_{t_1}^{t_2} p_g(t)^2\,\mathrm{d}t \cdot \int_{t_1}^{t_2} p_d(t)^2\,\mathrm{d}t\right)^{1/2}} \tag{1.6}$$

Si $p_g(t)$ et $p_d(t)$ sont deux signaux de bruit dont on fait varier le niveau de cohérence k ($k = 1$: les deux sources sont parfaitement cohérentes et $k = 0$ les

deux sources sont parfaitement incohérentes), le résultat perceptif peut être de différente nature en fonction de la valeur de k. Comme le représente la figure 1.4, tirée du livre Spatial Hearing [Blauert, 1997b] issue des travaux de [Chernyak, 1968], la diminution de 1 à 0.3 du niveau de cohérence entre les deux sources a pour effet une augmentation de la taille apparente de l'évènement auditif. Pour une cohérence inférieure à 0.3, les deux signaux sont perçus comme deux événements distincts, l'un localisé à gauche et l'autre à droite.

Indices spectraux ou monoraux

Les indices interauraux ne suffisent pas à localiser précisément un événement sonore puisque plusieurs positions de la source peuvent donner le même couple d'ILD et ITD. Toutes les positions possibles pour un couple donné décrivent un cône appelé cône de confusion. D'autres indices sont alors mis en jeu pour lever les ambiguïtés de localisation. L'évaluation de l'élévation ϕ et la discrimination avant-arrière d'une source sonore n'étant pas possible uniquement par la détermination des indices interauraux, ces tâches sont effectuées grâce à des indices spectraux. Un filtrage est opéré sur le son incident par le pavillon des oreilles, les épaules et le buste de l'auditeur. Ce filtrage est largement dépendant de l'orientation de l'auditeur par rapport à la direction incidente de l'événement sonore. Le cerveau parvient donc en analysant ces indices spectraux à remonter à la position en élévation de la source. Le flou de localisation en élévation est toutefois plus grand que celui en azimut et varie entre $\pm\ 9^\circ$ et $\pm\ 22^\circ$.

L'ensemble des indices (interauraux et spectraux) peut être regroupé dans une fonction de transfert, appelée HRTF (*Head-Related Transfer Function*), qui contient ainsi toutes les informations nécessaires à l'identification de la position d'une source. Toutefois ces HRTF varient d'un individu à l'autre puisqu'elles sont intimement liées à la morphologie de chacun, ce qui peut poser des problèmes lorsque l'on souhaite les simuler (*c.f.* paragraphe sur la synthèse binaurale en section 1.3.2).

1.2.2 Influences de l'environnement acoustique sur la perception auditive spatiale

Il a été montré que les environnements acoustiques pouvaient avoir une influence sur notre perception des attributs spatiaux de sources. Ces influences sont multiples, parfois bénéfiques : c'est le cas pour l'évaluation de la distance d'une source, parfois critiques : c'est le cas pour la localisation angulaire des sources. D'autres attributs spatiaux sont également influencés par la réverbération tels que la taille apparente des sources (ASW) ou la sensation d'enveloppement (LEV). Dans cette section nous proposons un rapide tour d'horizon des ces différents effets de l'environnement acoustique sur la perception auditive spatiale.

1.2.2.1 Perception de la distance des sources

La perception auditive de la distance des sources a été largement étudiée. Des états de l'art très complets sur le sujet ont été proposés [Coleman, 1963, Zahorik *et al.*, 2005, Kolarik *et al.*, 2016]. De manière générale, les facultés humaines à évaluer par le son la distance d'une source sont assez mauvaises, avec une tendance à sur-estimer la distance pour des sources sonores proches, situées dans l'espace péri-personnel (< 1 m) [Ashmead *et al.*, 1990, Parseihian *et al.*, 2014] et à la sous-estimer pour des sources lointaines, situées dans l'espace extra-personnel (> 1 m) [Mershon et Bowers, 1979, Bronkhorst et Houtgast, 1999, Zahorik, 2002a]. En se basant sur des travaux antérieurs, Zoharik et al. ont montré que la distance perçue d'une source R_p pouvait être estimée par une loi de puissance de sa distance réelle R :

$$R_p = k \cdot R^a \tag{1.7}$$

avec k et a des paramètres de régression pouvant varier d'une étude à l'autre [Zahorik *et al.*, 2005].

De nombreux facteurs rentrent en compte dans le jugement de la distance d'une source, dépendants des propriétés de la source et de l'environnement dans laquelle elle rayonne. Les différents indices sont présentés ici par ordre d'importance. Le principal indice de la distance d'une source est son niveau sonore. En champ libre, le niveau sonore d'une source est fonction de sa distance au carré. Ainsi, le doublement de la distance séparant la source de l'auditeur fait chuter de 6 dB le niveau sonore parvenant à ses oreilles. Cet indice ne permet toutefois pas d'effectuer un jugement absolu de la distance d'une source puisque l'effet d'une chute de niveau peut également être perçu comme une diminution de l'intensité acoustique de la source, en revanche il permet de comparer les distances entre plusieurs sources [Kolarik *et al.*, 2016]. Lorsque c'est le seul indice disponible, la plus petite différence de distance perçue, notée JND (*Just Noticable Difference*), correspond à une variation de la distance de l'ordre de 6 % pour des trains d'impulsion de bruit large bande [Ashmead *et al.*, 1990] à 25% pour des sinus purs [Jesteadt *et al.*, 1977]. Au contraire lorsque d'autres indices sont disponibles, notamment lorsque la source est située en environnement réverbérant, l'estimation du niveau sonore d'une source peut être effectuée indépendamment de sa distance, montrant que dans certains contextes il est possible de séparer l'information relative au niveau produit par la source de celle relative à sa distance. Ce phénomène appelé "loudness consistancy" en anglais, a été mis en évidence par Zoharik et Wightman [Zahorik et Wightman, 2001].

L'interaction entre la source et l'environnement acoustique dans lequel elle se déploie joue un rôle crucial dans la perception de la distance de cette dernière, En effet, le DRR (Direct to Reverberant Ratio) : rapport entre l'énergie du champ direct, dépendante de la distance source-récepteur et l'énergie

de la réverbération, relativement indépendante de la position de la source, est un indice permettant un jugement absolu de la distance de la source [Nielsen, 1992, Bronkhorst et Houtgast, 1999, Zahorik, 2002b]. Bien que d'après [Zahorik *et al.*, 2005], l'effet seul du niveau sonore soit plus important dans l'évaluation de la distance de source que l'effet seul du DRR, il a été montré que la combinaison de ces deux indices donnait de meilleures performances que dans les deux cas isolés [Bronkhorst et Houtgast, 1999, Kopčo et Shinn-Cunningham, 2011, Ronsse et Wang, 2012]. D'autre part, plusieurs objections ont été faites concernant la saillance du DRR dans le jugement de la distance. La première, formulée par [Zahorik, 2002a] est que pour des stimuli continus, il semble peu probable, d'après eux, que nous soyons en mesure de séparer les flux relatifs à la source et à l'effet de salle, rendant le jugement du DRR difficile voire impossible. D'autre part, comme nous l'avons vu précédemment, l'effet de préséance observé entre le champ direct et les premières réflexions pousse à croire que les contributions énergétiques de ces réflexions contribuent à l'évaluation du niveau de la source. Sur ce principe, [Bronkhorst et Houtgast, 1999] proposent de considérer le rapport d'énergie précoce sur tardif de la réponse comme indice acoustique de la perception de la distance des sources. Plusieurs études ont également révélé un effet marginal de la durée de réverbération sur la distance perçue [Mershon *et al.*, 1989, Altmann *et al.*, 2013], à savoir que l'augmentation du temps de réverbération peut résulter en une augmentation de la distance perçue.

Des indices spectraux sont également employés pour évaluer la distance de sources situées dans l'espace péri-personnel ou bien à une distance élevée (> 15 m). Dans le cas des sources proches, le phénomène de diffraction du signal source par la tête de l'auditeur est responsable de la variation du contenu spectral arrivant aux oreilles de l'auditeur en fonction de la distance de la source. [Brungart, 1999] a mis en évidence l'importance de ce phénomène, notamment aux basses-fréquences. En utilisant trois sources de bruit large-bande (200 Hz -15 kHz), filtré passe-bas (200 Hz - 3 kHz) et filtré passe-haut (3 kHz - 15 kHz), il a révélé qu'un jugement précis de la distance de la source ne pouvait être obtenu que pour des sources présentant des fréquences inférieures à 3 kHz. Une autre étude menée par [Kopčo et Shinn-Cunningham, 2011] proposait aux participants de juger la distance de sources situées en position frontale ou latérale à une distance variant de 0.15 m à 1.7 m. Les stimuli présentés étaient des trains d'impulsions de bruits filtrés de fréquence centrale variant de 300 Hz à 5700 Hz et de largeur de bande variable entre 200 Hz et 5400 Hz. Ils ont révélé que la précision dans le jugement de la distance était négativement impactée par l'augmentation de la fréquence centrale, sans effet de la largeur de bande. L'effet du contenu spectral était plus prononcé pour des sources d'incidence frontale. Pour des sources situées à une distance supérieure à 15m, le contenu spectral de la source est impacté par l'absorption de l'air. En effet, lorsque l'onde acoustique produite par une source parcours de longues distances, l'air, plus absorbant en hautes-fréquences qu'en basses-fréquences, modifie de façon perceptible la balance spectrale du signal source. Ainsi, une source présentant un

contenu hautes-fréquences atténué par rapport à son contenu basses-fréquences aura tendance à être perçue à une distance plus élevée. Ce phénomène, déjà observé en 1938 par von Békésy [von Békésy, 1938], a depuis été confirmé dans d'autres études [Coleman, 1968, Butler *et al.*, 1980, Little *et al.*, 1992].

Dans le cas de sources sonores proches, il a été observé que l'évaluation de la distance était plus précise pour des sources d'incidence latérale que pour des sources d'incidence frontale [Holt et Thurlow, 1969, Kopčo et Shinn-Cunningham, 2011]. Cette observation, corroborée par nombre d'études [Duda et Martens, 1998, Brungart *et al.*, 1999, Shinn-Cunningham *et al.*, 2005], illustre la présence d'indices binauraux dans l'évaluation de la distance de sources situées dans l'espace péri-personnel. En effet, dans cette zone, alors que l'ITD n'est pas ou quasiment pas affecté par la distance de la source, l'ILD est sujet à de grandes variations en fonction de la distance et ce pour des distances inférieures à 1m, distance limite au-delà de laquelle la variation de distance n'affecte quasiment plus l'ILD et plus généralement les HRTFs [Brungart, 1999, Otani *et al.*, 2009]. Il faut noter que ces variations d'ILD sont plus prononcées en basses-fréquences. Un autre phénomène appelé "parallaxe des HRTF" (*HRTF parallax*) représente la différence d'incidences observée entre les fronts d'ondes arrivant à l'entrée des oreilles gauche et droite d'un auditeur. L'angle de parallaxe est d'autant plus grand que la source est proche et frontale. En considérant que cette différence d'angle induit un filtrage différent par les pavillons des oreilles gauche et droite, observable sur une mesure d'HRTF et perceptible par l'être humain, cet indice pourrait permettre d'évaluer la distance de sources proches, dans des cas où l'ILD est nul (sources situées sur le plan médian). C'est ce que suggère Kim et al. dans une étude investiguant la possibilité de contrôler la distance d'un évènement auditif en se basant sur un modèle de parallaxe [Kim *et al.*, 2001]. En effet, à partir de stimuli de synthèse correspondant à une source de bruit rose à des distances allant de 0.1 m à 2 m et sans faire varier les autres indices de distance (niveau et DRR), une augmentation de la distance de la source entre 0.1 et 1 m a bien été perçue par les participants comme une augmentation de la distance de l'évènement auditif. Au-delà de 1 m en revanche, ils n'ont plus été capables de juger correctement de la distance, indiquant un effet négligeable de l'indice de parallaxe au-delà de cette distance.

Notre capacité à évaluer la distance d'une source sonore semble être également influencée par des facteurs cognitifs plus haut-niveau tels que la familiarité de l'auditeur avec le stimulus [Mershon *et al.*, 1989, Shinn-Cunningham, 2000, Brungart et Scott, 2001, Philbeck et Mershon, 2002], ou bien son état émotionnel [Gagnon *et al.*, 2013]. La perception auditive de la distance des sources sonores a également été étudiée en situations dynamiques (sources et/ou auditeurs mobiles) [Schiff et Oldak, 1990, Ashmead *et al.*, 1995], mais n'est pas décrite ici. Enfin, des phénomènes d'interactions multi-modales dans le jugement de la distance des sources en contexte audio-visuel ont également été mis en évidence dans de plusieurs études, ces dernières décennies [Côté *et al.*, 2012, Maempel et Jentsch, 2013,

Anderson et Zahorik, 2014, Paquier *et al.*, 2016]. Ce sujet est traité plus en détail en chapitre 4 du présent ouvrage.

1.2.2.2 Influence sur la localisation angulaire

Plusieurs études ont été menées pour tenter de caractériser les performances de localisation de sources en environnement réverbérant. Dans de telles conditions, bien que la composante directe du champ sonore soit suivie de nombreuses réflexions, dans la plupart des cas, le système auditif est capable de supprimer l'information directionnelle véhiculée par ces réflexions. Ce phénomène est connu sous le nom d'effet de préséance (*precedence effect*), largement décrit par [Litovsky *et al.*, 1999]. Lorsque le système auditif est confronté à deux stimuli cohérents décalés temporellement, différents mécanismes cognitifs entrent en jeu afin de permettre à l'auditeur de se concentrer sur les informations significatives en termes de direction. Lorsque le retard entre les deux stimuli est inférieur à 1 ms, les deux stimuli sont cognitivement fusionnés en un seul événement sonore et l'incidence perçue de la source correspond à l'incidence moyenne des deux stimuli. Pour des retards compris entre 1 et 50 ms, les stimuli sont également fusionnés mais l'incidence perçue de la source est déterminée par la position du premier stimulus, avec toutefois un élargissement de la source pour des retards de 5 à 50 ms. Au dela de 50 ms, les deux stimuli sont perçus comme deux évènements différents. En s'intéressant à ce principe, [Hartmann, 1983] a montré que les premières réflexions pouvaient avoir un rôle bénéfique ou bien critique en fonction de leur incidence. Ainsi, les premières réflexions d'incidence proche de celle de la source avaient tendance à améliorer les performances de localisation, tandis que les réflexions latérales avaient tendance à dégrader ces performances.

A travers une série de mesures binaurales, à l'aide d'une tête artificielle KEMAR, [Shinn-Cunningham, 2001] a mis en évidence des perturbations de l'ILD et ITD induites par la réverbération, pouvant troubler la localisation des sources sur le plan azimutal. En effet, les indices binauraux de localisation se trouvent significativement dégradés par l'augmentation du temps de réverbération et la diminution du DRR (augmentation de la distance de la source), résultant en une légère dégradation des performances de localisations observées. L'impact du temps de réverbération sur la localisation angulaire a également été observée par [Hartmann, 1983], à savoir que pour des réverbérations longues la localisation est plus difficile. D'autre part, il a été montré que la nature de la source était déterminante dans notre capacité à localiser les sources en champ réverbérant. Plusieurs facteurs tels que son contenu spectral [Ihlefeld et Shinn-Cunningham, 2011], son temps d'attaque et sa durée [Rakerd et Hartmann, 1986] ont été révélés. [Ihlefeld et Shinn-Cunningham, 2011] ont par exemple montré qu'en milieu réverbérant, les sources de bruits contenant des hautes-fréquences dans leur contenu spectral étaient mieux localisées. De leur côté, [Rakerd et Hartmann, 1986] ont révélé que pour des tons purs, la diminution du temps d'attaque et de la durée du signal avaient un effet bénéfique sur la

localisation de sources.

En résumé, les travaux d'Hartmann et Rakerd ont permis, entre autres, de mettre en évidence une influence des premières réflexions sur les performances de localisation. D'autres facteurs tels que le temps de réverbération, la distance de la source ou encore sa nature semble également jouer un rôle dans notre capacité à localiser précisément les sources. Toutefois, de plus amples recherches semblent nécessaires pour quantifier davantage l'influence de ces différents facteurs.

1.2.2.3 Influence sur la taille apparente des sources

Il a été montré que la taille apparente de sources dépend de plusieurs facteurs relatifs à la source comme son contenu fréquentiel (*i.e.* l'ASW est plus élevée pour des signaux basses-fréquences), son niveau sonore (*i.e.* plus le niveau de la source est élevé plus la source est perçue comme large) ou bien de sa durée (*i.e.* plus la source produit un son long, plus elle est perçue comme étendue) [Perrott et Buell, 1982]. Elle est également grandement influencée par l'environnement acoustique dans lequel elles se déploient et principalement par son comportement précoce. En effet, la réponse d'une salle à une stimulation acoustique peut être vue comme la superposition du signal d'excitation avec de multiples versions atténuées, retardées et plus ou moins cohérentes du signal source, ce qui a plusieurs conséquences sur notre perception de la taille des sources.

D'abord, comme évoqué précédemment, l'effet de préséance produit par la fusion des premières réflexions avec le champ direct peut produire une sensation de source étendue lorsque ces réflexions interviennent entre 5 et 50 ms après le signal source [Litovsky *et al.*, 1999]. L'effet de la latéralité de ces réflexions sur l'ASW a été étudié dans plusieurs expériences [Barron et Marshall, 1981, Morimoto *et al.*, 1993, Morimoto *et al.*, 2008]. [Barron et Marshall, 1981] ont par exemple mené une expérience en faisant varier l'amplitude, le retard et l'incidence de deux réflexions. Cette expérience a révélé que l'élargissement de la source était le plus important pour des réflexions ayant une incidence perpendiculaire au plan médian et les auteurs concluent que l'ASW pourrait être proportionnelle au cosinus de l'angle entre l'incidence de la réflexion et l'axe passant par les deux oreilles de l'auditeur. Sur cette base, la norme ISO 3382 propose une mesure objective de l'ASW, appelée fraction latérale précoce , notée LF_E ou J_{LF} et définie par la formule suivante :

$$LF_E = \frac{\int_{5\,\mathrm{ms}}^{80\,\mathrm{ms}} p_8(t)^2\,\mathrm{d}t}{\int_{0\,\mathrm{ms}}^{80\,\mathrm{ms}} p(t)^2\,\mathrm{d}t} \tag{1.8}$$

avec $p_8(t)$ la pression mesurée à l'aide d'un microphone bidirectionnel orienté dans l'axe porté par les oreilles de l'auditeur.

Lors d'une large étude s'intéressant à l'impression spatiale (ASW et LEV) dans 16 environnements acoustiques différents [Bradley *et al.*, 2000] ont montré que le

gain latéral précoce de la salle LG_E défini par l'équation 1.9 était un bon indicateur de la taille apparente de la source.

$$LG_E = 10 \cdot \log \frac{\int_{5\,\mathrm{ms}}^{80\,\mathrm{ms}} p_8(t)^2 \,\mathrm{d}t}{\int_0^{\infty} p_{A\,10\,\mathrm{m}}(t)^2 \,\mathrm{d}t} \,, \mathrm{dB} \tag{1.9}$$

avec $p_{A\,10\,\mathrm{m}}(t)$, pression de référence, mesurée en condition anéchoïque avec un microphone omnidirectionnel placé à 10 m de la source. Ils ont aussi révélé que la quantité de réverbération tardive, décrite par le gain latéral tardive LG_L avait une influence sur l'ASW et ont quantifié l'influence relative de ces deux composantes par un rapport d'environ 6 dB, à savoir pour compenser l'effet d'une augmentation de 1 dB du LG_E sur l'ASW, il faut augmenter le LG_L de 6 dB.

D'autre part, la présence de multiples réflexions partiellement cohérentes avec le signal source a une influence négative sur l'IACC, reconnu comme un indice fort de la taille de sources [Perrott et Buell, 1982, Morimoto *et al.*, 1993, Hidaka *et al.*, 1995, Okano *et al.*, 1998, Mason *et al.*, 2005]. Beranek propose un indicateur largement corrélé à la taille apparente de source, le BQI (*Binaural Quality Index*, calculé à partir de la valeur moyenne des IACC précoces pour les bandes d'octave 500 Hz, 1kHz et 2kHz [Beranek, 2004], tel que :

$$BQI = 1 - IACC_{E3} \tag{1.10}$$

Du fait du fort impact des basses fréquences sur l'ASW, [Okano *et al.*, 1998] montrent qu'une combinaison linéaire du BQI et du gain basse-fréquence ($<$ 250 Hz) de la salle G_{Low} donne une meilleure prédiction de la taille apparente perçue que le simple BQI.

D'autres propositions plus complexes d'indicateurs objectifs de la taille apparente de sources ont été formulées, en se basant notamment sur une décomposition spatiale plus fine que celle obtenue avec un microphone bi-directionnel, comme le B_{LF} [de Vries *et al.*, 2001], équivalent du LF_E, ou le prédicteur RAP_{ASW}, issu du modèle RAP de la perception spatiale en environnements réverbérants, proposé par Klockgether et van de Par [Klockgether et van de Par, 2014], basée sur des calculs d'IACC.

En résumé, de nombreux facteurs interviennent dans la perception de la taille des sources sonores. Celle-ci est grandement influencée par l'environnement acoustique et notamment par son comportement précoce. Deux familles d'indicateurs objectifs sont généralement utilisés pour représenter la taille apparente de sources : ceux prenant en compte la quantité d'énergie latérale restituée par la salle, à l'instar du gain latéral précoce LF_E et ceux basés sur la mesure de la cohérence interaurale caractérisée par l'IACC. D'autres paramètres de salles comme la réverbération tardive ou sa réponse fréquentielle (notamment en basses fréquences), peuvent également influencer notre jugement de l'ASW.

1.2.2.4 Influence sur la sensation d'enveloppement

L'enveloppement, noté LEV (*Listener Envelopment*), est défini comme la sensation perceptive d'être entouré par la réverbération [Lindau *et al.*, 2014a]. Comme le fait remarquer [Zacharov *et al.*, 2016], cette définition est spécifique à l'effet de salle alors que la sensation d'enveloppement peut également être créée par une source étendue ou par un ensemble de sources entourant l'auditeur. Toutefois, nous nous intéressons ici à l'enveloppement par la réverbération. Le terme d'enveloppement est pour la première fois mentionné par Michael Barron dans ces travaux sur l'effet des premières réflexions sur l'impression spatiale des réponses acoustiques de salles [Barron, 1971, Barron et Marshall, 1981]. Dans [Barron et Marshall, 1981], l'impression spatiale est décrite comme la sensation d'espace associée aux premières réflexions latérales, toutefois les auteurs, Barron et Marshall, distinguent deux effets principaux produits par la modification des réflexions latérales. A travers plusieurs expériences jouant sur le niveau, l'incidence et le retard des réflexions latérales et sur le contenu spectral de la source, ils révèlent qu'en basses-fréquences l'effet produit par ces modifications est décrit comme une sensation d'enveloppement par la réverbération, tandis qu'en hautes-fréquences (au-delà de 1000 Hz) il est perçu comme un élargissement de la source sonore. Alors que ces premiers travaux révèlent un impact de la partie précoce de la réponse, plusieurs études ont depuis montré que le rapport entre énergies latérales précoce et tardive de la réponse affecte la sensation d'enveloppement [Bradley *et al.*, 2000, Hanyu et Kimura, 2001, Berg et Nyberg, 2008] et que le comportement spatial de la réponse tardive a un effet plus déterminant que celui du champ précoce [Bradley *et al.*, 2000]. Du fait de la forte dépendance du LEV au comportement tardif de la réponse et de la latéralité des réflexions, plusieurs paramètres acoustiques prenant en compte ces deux aspects ont été établis afin d'être capables à partir de mesures objectives de prédire l'enveloppement associé à une mesure de réponse acoustique. Bradley et Soulodre ont observé une forte corrélation entre l'enveloppement et l'énergie latérale de la réponse tardive de la salle qu'ils caractérisent par une grandeur appelée gain latéral tardif de la salle [Bradley et Soulodre, 1995b], défini par l'équation 1.11 et qui correspond au rapport de l'énergie captée par un microphone bidirectionnel $p_8(t)$ à partir de 80 ms après le champ direct et une valeur de référence $p_{A\,10\,\mathrm{m}}(t)$ correspondant à l'énergie mesurée par un microphone omni placé à 10 m de la source en champ libre. Ce paramètre est noté LG_L, GLL, LG_{80}^{∞}, ou L_J :

$$LG_L = 10 \cdot \log \frac{\int_{80}^{\infty} p_8(t)^2 \,\mathrm{d}t}{\int_0^{\infty} p_{A\,10\,\mathrm{m}}(t)^2 \,\mathrm{d}t} \,, \mathrm{dB} \tag{1.11}$$

Une autre mesure de l'enveloppement, proposée dans plusieurs études [Bradley et Soulodre, 1995a, Barron, 2001] est la fraction latérale d'énergie tardive, notée LF_L ou LLF et définie par l'équation suivante :

$$LF_L = \frac{\int_{80}^{\infty} p_8(t)^2 \,\mathrm{d}t}{\int_0^{\infty} p(t)^2 \,\mathrm{d}t} \tag{1.12}$$

Alors que le LG_L est implicitement constitué d'une composante d'énergie latérale et d'une composante relative à l'énergie globale tardive de la mesure G_L, induite par la division par une mesure de référence à 10 m, le LF_L, en divisant la composante d'énergie latérale par l'énergie de la mesure omnidirectionnelle de la même réponse, ne prend en compte que la composante d'énergie latérale. Ainsi, LG_L peut être exprimée en fonction du gain tardif global G_L et de la fraction latérale d'énergie tardive LF_L :

$$LG_L = G_L + 10 \cdot \log(LF_L) \tag{1.13}$$

avec :

$$G_L = 10 \cdot \log \frac{\int_{80}^{\infty} p(t)^2 \, \mathrm{d}t}{\int_{0}^{\infty} p_{A\,10\,\mathrm{m}}(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{1.14}$$

A partir de l'analyse de 17 salles de concert, [Barron, 2001] montre une faible variabilité dans la mesure du LF_L au sein de et entre les différentes salles, et conclut qu'en condition réelle, la sensation d'enveloppement serait principalement portée par la quantité d'énergie tardive globale G_L de la salle et que l'enveloppement est, de façon générale, élevé dans les petites salles et faible dans des grandes salles.

De nombreuses études ont également révélé que les contributions tardives venant d'autres directions, notamment du plafond et de derrière l'auditeur, avaient une influence sur la sensation d'enveloppement [Evjen *et al.*, 2001, Furuya *et al.*, 2001, Hanyu et Kimura, 2001, Wakuda *et al.*, 2003]. En introduction de son mémoire de thèse, David A. Dick, en se basant sur cette littérature, relève un certain nombre de points d'ombre concernant l'élaboration d'une métrique objective de la sensation d'enveloppement. Il constate notamment que la valeur de 80 ms choisie pour le calcul des paramètres relatifs à l'enveloppement est somme toute arbitraire et montrera dans [Dick et Vigeant, 2015] que des contributions antérieur à 80 ms peuvent affecter la sensation d'enveloppement. D'autre part, il constate que le recours à un microphone bi-directionnel pour le calcul des paramètres représentant le LEV est également arbitraire. Pour pallier cela, Dick propose un nouveau prédicteur du LEV basé sur une décomposition spatiale complexe des mesures réalisées à l'aide d'un microphone sphérique ambisonique d'ordre 4 [Dick, 2017].

1.3 Auralisation d'environnements acoustiques 3D

Etant donné un intérêt grandissant pour les technologies immersives ces dernières décennies d'une part et la difficulté technique d'étudier la perception des environnements acoustiques dans des conditions réelles d'écoute d'autre part, le recours à l'auralisation s'est aujourd'hui imposé dans le paysage de la recherche en audio [Rindel et Christensen, 2003, Cabrera *et al.*, 2005, Yadav *et al.*, 2011, Johnson, 2018]. Dans son livre sur l'auralisation [Vorländer, 2020], Vorländer en

donne une définition très large. Il la définit comme "la technique consistant à créer du contenu audible à partir de données numériques (simulées, mesurées, synthétisées)". Ce terme est néanmoins le plus souvent employé selon sa définition donnée par Wikipédia : "L'auralisation est un procédé visant à recréer un environnement acoustique à partir de données mesurées ou simulées". L'auralisation s'appuie donc d'une part sur des mesures ou des techniques de modélisation de réponses impulsionnelles spatiales de salles, notées SRIRs (*Spatial Room Impulse Responses*), visant à capter ou à simuler le comportement acoustique des salles en trois dimensions, et d'autre part sur des systèmes et techniques de spatialisation sonore permettant une restitution en trois dimensions des contenus mesurés ou synthétisés. Dans cette section, nous passerons en revue les principales techniques pour la mesure et la restitution 3D d'environnements acoustiques.

1.3.1 Mesure de réponses impulsionnelles de salles

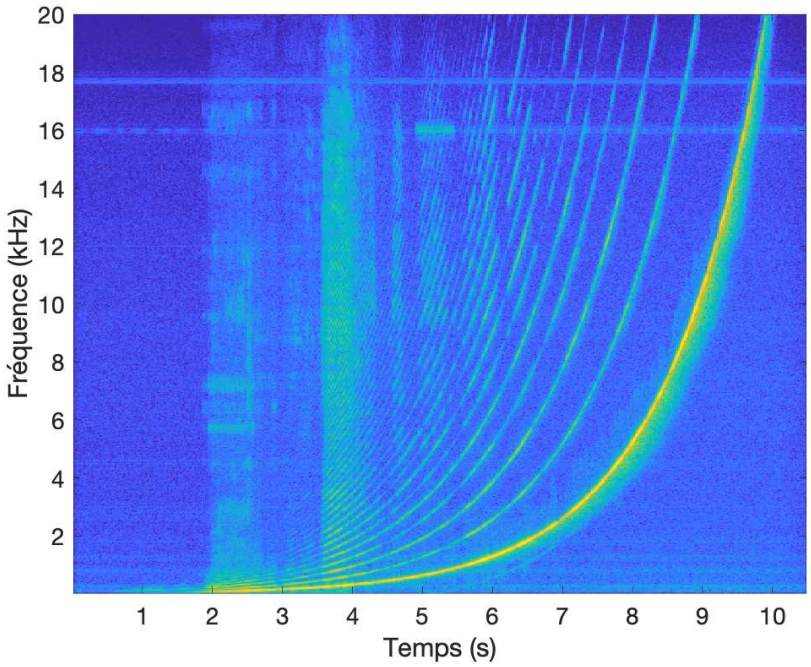
Comme évoqué en section 1.1, le comportement acoustique des salles peut être caractérisé par la mesure de réponses impulsionnelles. La solution la plus directe pour mesurer une réponse impulsionnelle de salle est d'enregistrer à l'aide d'un microphone l'excitation de la salle par une source impulsive telle que l'explosion d'un ballon de baudruche ou la détonation d'un pistolet. Cependant, cette solution se confronte en pratique à une mauvaise reproductibilité et à un manque de contrôle lors de la mesure. Pour pallier ces problèmes, les acousticiens ont développé d'autres méthodes plus ou moins complexes et plus ou moins adaptées à différents cas pratiques. D'après [Stan *et al.*, 2002], il existe trois grandes méthodes qu'il s'agit de présenter ici.

- La méthode MLS (*Maximum Length Sequence*) proposée par Schroeder en 1979 [Schroeder, 1979], consiste à utiliser comme signal d'excitation une source de bruit blanc pseudo-aléatoire. Il est ensuite possible de remonter à la réponse impulsionnelle par un processus de déconvolution basée sur le calcul de la correlation circulaire entre le signal mesuré et le signal d'excitation. Cette méthode présente l'avantage de pouvoir être déployée dans des environnements bruyants puisqu'elle est peu sensible aux sources de bruits (stationnaires ou impulsifs) non corrélées avec la séquence source. En revanche, elle est relativement sensible aux non-linéarités du système de mesure principalement portées par le haut-parleur source et pouvant résulter en la présence de distorsions impulsives sur la réponse impulsionnelle. La méthode IRS (*Inverse Repeated Sequence*) est dérivée de la méthode MLS et permet de limiter les distorsions produites par les non-linéarités du système de mesure. Cette méthode est également résiliante aux bruits ambiants [Dunn et Hawksford, 1993].
- [Aoshima, 1981] propose une méthode basée sur une technique d'expansion et de compression temporelle d'un signal impulsif, appelée *Time-Stretched Pulses*. Elle consiste à utiliser en signal d'excitation une impulsion temporellement étirée, permettant d'augmenter la puissance du signal en conservant une amplitude d'excitation identique à celle d'une impulsion classique et ainsi

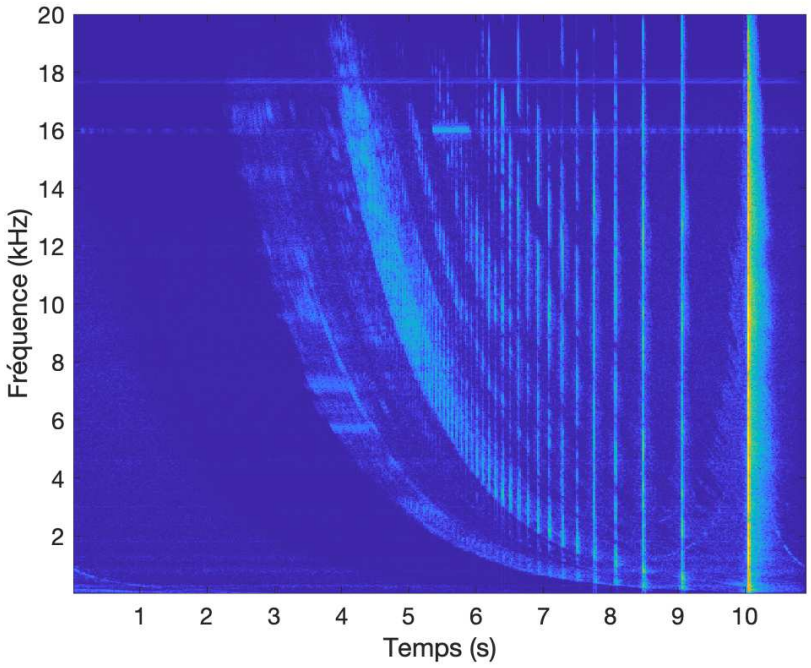
augmenter le rapport signal sur bruit SNR de la mesure tout en limitant les distorsions induites par les non-linéarités du haut-parleur. La réponse impulsionnelle de la salle est obtenue en appliquant au signal mesuré un filtre de compression, correspondant au filtre inverse du filtre utilisé pour l'étirement temporel de l'impulsion. Bien que nettement atténuées par rapport aux méthodes MLS et IRS, des distorsions induites par les non-linéarités de la chaine de mesure peuvent être observées sur la réponse impulsionnelle finale. Pour limiter ce phénomène une calibration en niveau du système doit être réalisée. D'autre part, malgré des bonnes performances en terme de rapport signal à bruit, cette méthode est très sensible aux sources de bruits parasites impulsifs durant la mesure, rendant son utilisation compromise dans des environnements bruités.

- La méthode par sinus glissant exponentiel proposée par Farina [Farina, 2000] est aujourd'hui la méthode de mesure de réponses impulsionnelles la plus répandue pour la caractérisation acoustique des salles. Le signal d'excitation employé par cette méthode est un sinus balayant la gamme de fréquence audible de façon exponentielle au cours du temps. La déconvolution du signal mesuré est réalisée par convolution linéaire de la réponse avec un filtre inverse. Le filtre inverse est produit par : 1. renversement temporel du signal source, 2. égalisation de la réponse fréquentielle du filtre inverse afin de compenser la coloration rose du sinus glissant exponentiel. Alors que les méthodes précédentes s'appuient sur l'hypothèse que le système de mesure est un système linéaire invariant dans le temps, les rendant sensibles aux non-linéarités du système, la présente méthode permet de séparer temporellement la partie linéaire du signal (*i.e.* la réponse de la salle) des non-linéarités présentes lors de la mesure . A titre d'exemple, la figure 1.5) illustre une mesure par sinus glissant exponentiel présentant de fortes non-linéarités induites par les distorsions harmoniques de la source et la réponse impulsionnelle obtenue après déconvolution. Cette méthode présente donc l'immense avantage de pouvoir d'une part extraire la réponse impulsionnelle linéaire de la salle et d'autre part de caractériser les non-linéarités du système. A l'instar de la méthode *Time-Stretched Pulses*, la présence de bruits impulsifs durant la mesure peut sévèrement impacter la qualité de la réponse impulsionnelle finale. Toutefois, plusieurs solutions sont envisagées pour réduire cet impact. [Farina, 2007] propose d'appliquer a posteriori un filtre passe-bande étroit, autour de la fréquence excitée par le sinus glissant à l'instant de l'occurrence du bruit impulsif. La méthode la plus courante est d'effectuer plusieurs réalisations de la même mesure et de sélectionner une réalisation ne présentant pas de bruits impulsifs. Divers algorithmes de détection de bruits impulsifs ont été proposés [Guski et Vorländer, 2015, Prawda *et al.*, 2022] afin de faciliter la sélection.

D'autre part, la mesure de réponses impulsionnelles spatiales de salles, notées SRIR (*Spatial Room Impulse Response*) s'est largement démocratisée au cours des 20 dernières années, en témoigne une littérature foisonnante sur le sujet



(a)



(b)

FIGURE 1.5 – Mesure d'une réponse impulsionnelle par sinus glissant [Farina, 2000]. (a) Spectrogramme de la mesure, en présence de fortes non-linéarités induites par les distorsions harmoniques du haut-parleur. (b) Spectrogramme de la réponse impulsionnelle après déconvolution. Les non-linéarités sont temporellement séparées de la réponse utile (située à 10 s).

[Farina *et al.*, 2011, Khaykin et Rafaely, 2012, Barre *et al.*, 2014, Embrechts, 2015, Alary *et al.*, 2019, Massé *et al.*, 2020, McCormack *et al.*, 2020]. Ces mesures permettent d'une part d'analyser les propriétés spatiales des environnements acoustiques et d'autre part d'en proposer une écoute spatialisée sur haut-parleurs ou au casque. Cet essor a notamment été rendu possible par la formalisation de l'ambisonie d'ordres supérieurs (HOA) par Daniel en 2000 [Daniel, 2000], suivi de la commercialisation, au début des années 2010, de réseaux sphériques de microphones, permettant la captation de contenus audio 3D au format HOA. Les réseaux sphériques de microphones sont des antennes constitués de plusieurs microphones réparties autour d'une sphère de faible diamètre (*e.g.* 8,4 cm pour l'Eigenmike 32 de mh acoustics), permettant une captation du champ acoustique à 360°. Il est également possible de réaliser des mesures de réponses impulsionnelles porteuses d'informations spatiales à l'aide d'une tête artificielle. On parle de réponses impulsionnelles binaurales, notées BRIR (*Binaural Room Impulse Responses*). Elles sont principalement utilisées pour une auralisation binaurale statique. En effet, à la différence des mesures par réseaux sphériques de microphones, le positionnement et l'orientation de la tête lors de la mesure de BRIR détermine le positionnement et l'orientation de l'auditeur dans la scène sonore lors de la restitution au casque. Si l'auditeur tourne la tête l'ensemble de la scène sonore tournera avec lui. Davantage de détails sur les techniques ambisoniques et binaurales sont donnés par la suite.

1.3.2 Restitution spatialisée d'environnements acoustiques

La restitution spatiale d'environnements acoustiques mesurés ou simulés peut être réalisée par le biais d'environnements acoustiques virtuels, notés VAE (*Virtual Acoustic Environments*). Les VAE ont pour objectifs de présenter de façon virtuelle, par le biais d'un casque audio ou de réseaux de haut-parleurs, des scènes sonores préalablement captées ou simulées de manière à recréer chez l'auditeur une sensation de plausibilité, de réalisme de la scène présentée, en reproduisant les indices acoustiques relatifs à l'audition spatiale en conditions réelles d'écoute. Ils s'appuient sur un ensemble de techniques de spatialisation sonore, variées dans leur manière de décrire les environnements sonores. En effet, on peut distinguer ces techniques selon trois grandes approches :

- Les techniques dites "physiques", basées sur une description physique du champ de pression acoustique produit par des sources sonores dans leur environnement. Les principales techniques de reconstitution de champ sont la WFS (*Wave Field Synthesis*) et l'ambisonie de premier ordre [Gerzon, 1985] et d'ordres supérieurs (HOA) [Daniel, 2003].
- Les techniques dites "signal", visant à recréer à l'entrée des conduits auditifs de l'auditeur l'ensemble des indices acoustiques de la perception auditive spatiale, *i.e.* les HRTFs. C'est le cas des techniques binaurales et transaurales.
- Les techniques dites "perceptives", dont l'objectif est de recréer les effets perceptifs de l'audition spatiale, en reproduisant les indices perceptifs les

plus saillants dans la perception des attributs spatiaux des scènes sonores. Cette approche regroupe les techniques de panning telles que la stéréophonie, le VBAP (*Vector Base Amplitude Panning*) [Pulkki, 1997], les techniques *Surround* utilisées au cinéma, mais également des techniques hybrides du type DirAC (*Directional Audio Coding*) [Pulkki, 2007] ou encore SDM (*Spatial Decomposition Method*) [Tervo *et al.*, 2013] et (HO)-SIRR (*(Higher Order) Spatial Impulse Response Rendering*) [Merimaa et Pulkki, 2004, McCormack *et al.*, 2020], dédiées à la spatialisation de réponses impulsionnelles.

Dans cette section, les techniques les plus couramment utilisées pour l'auralisation d'environnements acoustiques sont brièvement présentées.

1.3.2.1 Synthèse Binaurale

Dans la section 1.2, nous avons vu que l'ensemble des indices acoustiques utilisés pour localiser une source sonore à une position donnée peut être modélisé par deux fonctions de transfert entre la source et chacun des conduits auditifs de l'auditeur : on parle de couple d'HRTF (Head-Related Transfer Function). Ce couple d'HRTF permet de modéliser les filtrages et délais opérés par la morphologie de l'auditeur (cheveux, nez, pavillon d'oreille, buste, etc.) sur le signal qui arrive à l'entrée de ses oreilles. Le principe de la synthèse binaurale consiste à reproduire l'écoute naturelle, en appliquant le couple d'HRTF à un signal mono. Le signal ainsi filtré est ensuite délivré à chaque oreille à l'aide d'un casque stéréophonique [Begault et Trejo, 2000]. Le cerveau décode naturellement les indices acoustiques contenus par ces filtres et replace la source dans l'espace à la position où elle aurait été pour fournir ces indices. Afin de reproduire des sources virtuelles provenant de toutes les directions de l'espace, il est nécessaire de disposer d'une base de données d'HRTF (ou jeu d'HRTF) couvrant tout l'espace environnant l'auditeur. Un jeu d'HRTF peut être mesuré soit à l'aide d'une tête artificielle, soit en plaçant des microphones intra-auriculaires à l'entrée des conduits auditifs d'une personne. On mesure ensuite des réponses impulsionnelles pour une source placée à différentes positions autour de l'auditeur [Gardner et Martin, 1995]. Les mesures d'HRTF se réalisent en chambre anéchoïque. En général, pour couvrir l'espace, les mesures sont espacées de 5 à 15° en azimut et en élévation. Pour pouvoir placer des sources virtuelles entre les points de mesure, il est nécessaire d'interpoler les mesures [Carlile *et al.*, 2000].

En pratique, l'obtention d'une synthèse binaurale de qualité nécessite des HRTF individuelles. Or, la mesure d'HRTF est un processus laborieux du fait du grand nombre de positions spatiales à mesurer pour mailler l'espace avec une précision suffisante. C'est pourquoi les recours à des bases d'HRTF génériques est souvent préféré, ce qui peut résulter en une dégradation de l'image spatiale perçue par l'auditeur (mauvaise localisation en élévation, confusions avant-arrières, internalisation). Toutefois il a été montré que la précision de localisation avec des HRTF génériques pouvait devenir acceptable à condition d'entraîner les sujets à la localisation avec ces HRTF [Parseihian et Katz, 2012].

En ce qui concerne l'auralisation d'environnements acoustiques, nous avons vu dans la section précédente qu'il était possible d'effectuer des mesures de réponses impulsionnelles binaurales de salles (BRIR) à l'aide d'une tête artificielle, permettant une spatialisation statique de l'environnement mesuré. En revanche, les logiciels de modélisation d'environnements acoustiques tels que CATT-Acoustic et ODEON sont munis de moteurs de spatialisation binaurale, basée sur la synthèses de BRIR, permettant une écoute dynamique temps réel et une navigation de l'auditeur dans l'environnement acoustique modélisé [Vorländer, 2020].

1.3.2.2 Ambisonie et HOA

Dans son manuscrit de thèse, Jérome Daniel définit l'ambisonie comme "un ensemble de techniques de synthèse et de reproduction du son destinées à la spatialisation. Elle est basée sur une représentation physique du champ sonore dans le domaine des harmoniques sphériques ou circulaires qui permet de nombreuses opérations telles que la rotation ou la distorsion de la perspective" [Daniel, 2000]. Cette technique s'appuie sur la description du champ de pression acoustique p en tout point $\overrightarrow{r}$ de l'espace par l'équation d'onde dans le système de coordonnées sphériques ($\overrightarrow{r}$ est décrit selon ses coordonnées sphériques θ, ϕ et r, respectivement les angles en azimuth et élévation et la distance), conduisant au développement en séries de Fourier-Bessel suivant :

$$p(\overrightarrow{r}) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0;\,\sigma=\pm 1}^{m} \sum B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\phi) \tag{1.15}$$

où :

- k représente le nombre d'onde ;
- $j_m(kr)$ sont des fonctions de Bessel sphériques, dépendantes du nombre d'onde et de la distance $\overrightarrow{r}$;
- $Y_{mn}^{\sigma}(\theta,\phi)$ sont les harmoniques sphériques d'ordre n. Chacune d'elle est associée à un coefficient B_{mn}^{σ} qui représente la projection de la pression acoustique sur la base des harmoniques sphériques.

Les harmoniques sphériques $Y_{mn}^{\sigma}(\theta,\phi)$ sont une base de figures de directivités (représentées en figure 1.6 pour les 5 premiers ordres de la série), pouvant être décrites en fonction des deux coordonnées sphériques angulaires θ et ϕ :

$$Y_{mn}^{\sigma}(\theta,\phi) = \tilde{P}_{mn}(sin\phi) \times \begin{cases} cos(n\theta) & \text{si } \sigma = 1 \\ sin(n\theta) & \text{si } \sigma = -1 \end{cases} \tag{1.16}$$

où $\tilde{P}_{mn}(sin\phi)$ sont les versions semi-normalisées des fonctions de Legendre associées, m et n sont des entiers positifs tels que $n \leq m$, m étant l'ordre de l'harmonique et n le degré. σ prend les valeurs 1 et -1.

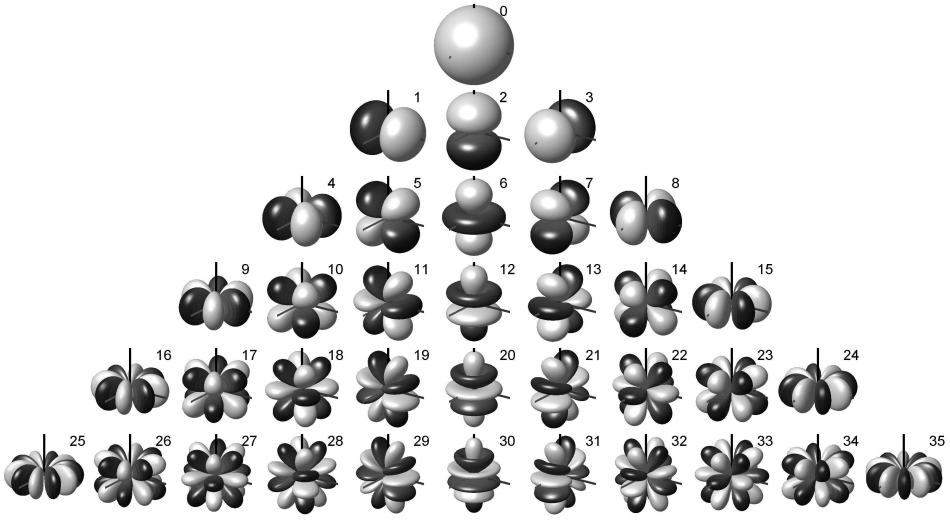


FIGURE 1.6 – Figures de directivité des harmoniques sphériques pour les 5 premiers ordres de la base (figure extraite de [Zotter *et al.*, 2012]).

Sur ce principe, la spatialisation ambisonique est réalisée en deux étapes : une étape d'encodage spatial du champ acoustique dans le domaine des harmoniques sphériques puis une étape de décodage permettant une restitution du champ à partir d'un réseau de sources acoustiques réelles, typiquement un réseau de haut-parleurs. Côté encodage, il est possible d'encoder des sources monophoniques de synthèse et de les positionner virtuellement dans l'espace acoustique, en s'appuyant sur l'équation 1.15. D'autre part, ce formalisme permet également d'encoder un flux audio capté à l'aide d'un réseau sphérique de microphones [Meyer et Elko, 2002, Moreau *et al.*, 2006, Rafaely, 2015]. Côté restitution, l'équation 1.15 implique une bonne reconstruction du champ en un point de l'espace, qui se matérialise en pratique, lors d'une restitution sur réseau de haut-parleurs, par l'existence d'une zone d'écoute restreinte appelée "*sweet spot*", généralement située au centre du dispositif de spatialisation.

Comme indiqué par l'équation 1.15, la description parfaite du champ acoustique au point $\overrightarrow{r}$ est basée sur une somme infinie des composantes sphériques. En pratique, la série est tronquée à un ordre M, ne permettant qu'une représentation partielle du champ spatial. En effet, l'ordre M définit le nombre N de composantes sphériques qu'il faut être capable de décrire et donc le nombre de canaux du signal ambisonique, par la relation $N = (M+1)^2$ pour une représentation sphérique (3D) et $N = 2M+1$ pour une représentation cylindrique (2D). Cela se traduit en pratique par une limitation en ordre induite par le nombre de canaux disponibles pour la captation et la restitution. Ainsi, une restitution ambisonique à l'ordre M = 4 implique d'avoir un réseau de microphones composé d'au moins 25 microphones pour la captation et un réseau de haut-parleurs composé d'au moins 25 haut-parleurs pour la restitution. D'autre part, la dépendance en kr du champ

de pression sphérique implique une limite fréquentielle induite par la troncature à l'ordre M, au-delà de laquelle le champ acoustique n'est plus correctement représenté. Ainsi, la fréquence de coupure d'un système ambisonique d'ordre 4 est $f_c = 2500$ Hz [Bertet, 2009]. L'augmentation de l'ordre de troncature M se traduit donc par : 1. une description spatiale plus précise du champ acoustique 2. une augmentation de la plage de fréquences pouvant être correctement reproduite.

Les premiers développements de la théorie ambisonique appliquée à l'audio ont été proposés par Gerzon [Gerzon, 1985] et étaient limités à l'ordre 1. Le signal ambisonique était alors constitué de 4 composantes sphériques : une composante omnidirectionnelle W, correspondant à la représentation sphérique du signal à l'ordre 0 et trois composantes bi-directionnelles X, Y et Z, correspondant aux trois composantes de l'ordre 1 (*c.f.* figure 1.6). Ce format est connu sous le nom de B-Format. La captation d'un flux acoustique 3D au B-Format peut être effectué grâce à un microphone tétraédrique, tandis que la restitution nécessite un réseau de quatre haut-parleurs. En 2000, Daniel propose une généralisation de l'ambisonie aux ordres supérieurs, maintenant connue sous l'acronyme HOA (*Higher-Order Ambisonics*).

La technologie HOA présente un certain nombre d'avantages et d'inconvénients. Tout d'abord, la représentation du champ acoustique dans le domaine des harmoniques sphériques est indépendante de sa complexité (sources multiples, réverbération). De plus, le passage par une étape d'encodage et de décodage permet de découpler le processus de captation de celui de restitution, ce qui en pratique permet une grande flexibilité (captation et restitution sur des systèmes variés). Cette technique offre également une grande scalabilité : il est possible de représenter grossièrement la scène en se plaçant à un ordre faible tandis que la montée en ordre permet une augmentation de la précision spatiale de la scène. D'autre part, des transformations spatiales de la scène sonores telles que des rotations ou des distorsions de perspectives peuvent être facilement appliquées dans le domaine des harmoniques sphériques [Daniel, 2000]. Enfin, plusieurs techniques ont été mises au point afin d'effectuer une réduction binaurale d'un flux HOA [Meyer et Elko, 2002, Noisternig *et al.*, 2003, Bernschütz *et al.*, 2014, Routray *et al.*, 2021], ce qui rend cette technique également adaptée à des situations d'écoute au casque. En revanche, l'ambisonie nécessite un nombre important de haut-parleurs pour pouvoir monter en ordre et ainsi décrire le champ avec précision. Dans un contexte d'auralisation d'environnements acoustiques, les composantes fortement directives de la réponse impulsionnelle spatiale, *i.e.* le champ direct et les premiers échos, sont reproduites avec une résolution spatiale limitée par la troncature en ordre de la série, pouvant résulter en un flou de localisation de ces composantes. En revanche le caractère diffus de la réverbération tardive, ne nécessitant pas une grande résolution spatiale, est d'un point de vue perceptif bien restitué. D'autre part, à la différence de la WFS, autre technique de synthèse de champ, l'ambisonie HOA est soumise à la notion de *sweet spot* : la reconstruction du champ s'opère en un point précis de l'espace

(centre du système pour un dispositif sphérique). La conséquence de cela est que cette technologie n'est pas bien adaptée pour des situations d'écoute à plusieurs ou pour des situations d'écoute dynamique (possibilité de s'orienter dans la scène mais pas de se déplacer).

1.3.2.3 Techniques perceptives

La plus ancienne technique de spatialisation est également la plus répandue aujourd'hui. Il s'agit de la stéréophonie développée par Blumlein en 1931. Son fonctionnement est basé sur le principe de panoramique d'amplitude ou sur la différence de phase appliquée à deux haut-parleurs (gauche L et droite R) formant idéalement un angle de 60° sur un cercle dont le centre est donné par la position de la tête de l'auditeur. Le panoramique d'amplitude consiste à créer des sources virtuelles en jouant sur la différence de gain entre les sources réelles émettant le même signal. Dans le cas de la stéréophonie, cela revient à synthétiser l'ILD d'une source dont la position serait entre L et R. D'autres systèmes multicanaux destinés au grand public utilisent le même principe, comme par exemple le 5.1 (5 HP répartis sur un plan horizontal) et le 22.2 (22 HP répartis dans les 3 dimensions de l'espace) [Hamasaki *et al.*, 2004]. Pour un effet de spatialisation optimal, l'auditeur doit être placé au sweet spot, zone dans laquelle toutes les sources sont à peu près à équidistance de l'auditeur.

La technique VBAP est une généralisation de la stéréophonie étendue à la 3D, où l'on utilise non plus une paire mais un triplet de haut-parleurs formés en triangle. Elle est développée par Ville Pulkki [Pulkki, 1997] et tire son nom du formalisme vectoriel utilisé pour calculer les gains à appliquer à chaque HP. En effet, si l'on considère $l_i = [l_{i1}\, l_{i2}\, l_{i3}]^T$ la position d'un haut-parleur i par rapport à la position de l'auditeur, et $p = [p_1\, p_2\, p_3]^T$ la position de la source virtuelle, on a pour un triplet de HP l_1, l_2, l_3 :

$$p = g_1 l_1 + g_2 l_2 + g_3 l_3 \tag{1.17}$$

avec g_1, g_2 et g_3 les gains respectivement appliqués aux haut-parleurs l_1, l_2, et l_3 pour obtenir une source virtuelle en p et $g = [g_1 g_2 g_3]$ tel que

$$g = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{33} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1} \tag{1.18}$$

Les gains sont ensuite normalisés en amplitude (VBAP) ou en énergie (VBIP, Vector Based Intensity Panning) [Jot *et al.*, 1999].

Un des grands atouts du VBAP réside dans sa simplicité théorique. Sur le plan perceptif, les performances de localisation de sources ponctuelles statiques en VBAP sont très proches de celles obtenues pour des sources réelles [Pulkki et Hirvonen, 2004]. De manière générale, le VBAP est une technique

adaptée à la mise en espace de multiples sources sonores monophoniques. En revanche, à la différence de la technologie HOA, le VBAP ne dispose pas de moyens satisfaisants de captation de scènes sonores. Son utilisation pour la mesure et la restitution d'environnements acoustiques n'est donc pas directement envisageable.

Toutefois, des méthodes de restitution hybride de SRIR s'appuient sur les bonnes performances du VBAP en terme de précision spatiale pour pallier les performances mitigées de l'HOA lors de la restitution des composantes directives de la SRIR. C'est le cas notamment de la méthode SDM (*Spatial Decomposition Method*) proposée par [Tervo *et al.*, 2013] et des méthodes (HO)-SIRR (*(Higher Order) Spatial Impulse Response Rendering*) initialement proposée par [Pulkki, 2007] et améliorée par [McCormack *et al.*, 2020]. Ces méthodes sont dites "paramétriques" puisqu'elles s'appuient sur une première étape d'estimation de paramètres spatiaux à partir de l'analyse de SRIR captées par réseaux de microphones, puis par une re-spatialisation en VBAP des composantes directives, basée sur les estimations de la première étape. Plus précisément, la méthode SDM considère que chaque échantillon de la RI peut être vu comme une composante spatiale provenant d'une direction moyenne. Ainsi, pour chaque grain temporel, une estimation de la direction d'arrivée (DoA : *Direction of Arrival*) est effectuée à partir du signal multi-canal. Chaque grain temporel est ensuite re-spatialisé en VBAP, suivant la direction qui lui a été attribuée. De son côté la méthode SIRR propose de paver la réponse impulsionnelle dans le plan temps-fréquence et considère que chaque élément du pavage peut être vu comme le fruit d'une composante directionnelle et d'une composante isotrope diffuse. Sur ce principe, l'analyse consiste à estimer deux paramètres pour chaque pavé du plan TF : 1. la direction de la composante directionnelle et 2. l'énergie relative de cette composante par rapport à l'énergie totale de la zone. Ce deuxième paramètre est représentatif de la "diffusivité/diffusion ?" de la zone. Pour chaque pavé, la composante directionnelle est re-spatialisée en VBAP, tandis que la composante diffuse est générée à l'aide d'un banc de filtres décorrélateurs. La balance en niveau entre ces deux composantes est ensuite réglée grâce à l'estimation du coefficient de "diffusivité" compris entre 0 et 1 (0 signifiant parfaitement directif et 1 parfaitement diffus). Alors que ces deux méthodes s'appuient sur l'analyse de SRIR ambisoniques d'ordre 1, [McCormack *et al.*, 2020] proposent une amélioration de la méthode SIRR, appelée HO-SIRR, exploitant les informations spatiales supplémentaires du format HOA. La méthode HO-SIRR est sensiblement identique à la méthode SIRR, à ceci près qu'une décomposition spatiale plus fine du champ est réalisée, permettant notamment de gérer des champ diffus anisotropes ainsi que des situations de coprésence de composantes directives (*i.e.* échos).

1.4 Solutions retenues

La présente thèse est motivée par différents enjeux relatifs à la perception des environnements acoustiques 3D. Comme nous avons pu le voir dans les sections pré-

cédentes, la perception auditive des environnements acoustiques est un processus cognitif complexe qui a suscité l'intérêt des chercheurs en acoustique depuis plus d'un siècle. Lorsqu'on se penche plus spécifiquement sur la question de la perception auditive spatiale en environnement réverbérant, on s'aperçoit que les environnements acoustiques jouent un rôle déterminant dans notre appréhension des attributs spatiaux des sources sonores, tels que sa distance, sa taille apparente ou encore sa localisation angulaire. Ces multiples influences ont souvent été étudiées de manière séparée et de nouvelles recherches semblent nécessaires afin de quantifier de façon plus globale l'impact de l'acoustique sur l'image spatiale des scènes sonores. D'autre part, les recherches menées tout au long de la thèse doivent permettre de caractériser objectivement et perceptivement des environnements acoustiques réels. Or ces dernières décennies ont été le théâtre de grandes avancées technologiques dans le domaine de la captation et de la restitution tri-dimensionnelle d'environnements acoustiques, permettant une auralisation de qualité d'environnements réels. Leur utilisation à des fins de caractérisation objective et subjective des salles est aujourd'hui courante puisqu'elles permettent de collecter, analyser et reproduire une grande diversité d'environnements dans des conditions contrôlées en laboratoire. Toutefois, ces technologies sont imparfaites et présentent un certain nombre d'avantages et d'inconvénients. Dans cette section nous présentons les choix méthodologiques et techniques retenus pour mener à bien les différents travaux de la thèse.

1.4.1 Choix de l'étude de la perception d'environnements acoustiques mesurés

Nous avons vu que l'étude in-situ de la perception des environnements acoustiques est souvent difficilement réalisable. Le passage par l'auralisation est donc retenu dans la plupart des études perceptives actuelles [Cabrera *et al.*, 2005, Ahrens *et al.*, 2016, Johnson, 2018, Luizard *et al.*, 2018]. Les deux approches les plus répandues pour l'auralisation sont la simulation/modélisation/synthèse d'une part et la mesure acoustique d'autre part. Le recours à la simulation acoustique est une option intéressante pour répondre à la première problématique, puisqu'elle présente entre autres l'avantage de pouvoir générer un corpus d'environnements acoustiques réalistes d'une grande diversité. Toutefois lorsqu'il s'agit de caractériser des environnements acoustiques réels, la modélisation se confronte à un certain nombre de limitations. Premièrement, une lourd travail de description de la géométrie des édifices à l'étude est requis pour pouvoir mener à bien la simulation acoustique. Ce travail devient conséquent et difficilement envisageable lorsqu'on s'intéresse à un grand nombre d'environnements acoustiques dont les données géométriques ne sont pas forcement connues. D'autre part, les différentes approches de modélisation proposées par les logiciels de simulation (approches géométriques, ondulatoires et hybrides) sont imparfaites et les modèles qui en découlent doivent la plupart du temps être ajustés par une campagne de mesures in-situ. Dans ses travaux de thèse, Postma a dû développer une méthode de calibration de ses modèles géométriques à partir de mesures, ainsi qu'une évaluation perceptive des modèles calibrés, par

une comparaison entre le modèle et la mesure [Postma, 2017]. Cette dernière sert donc souvent de référence lorsqu'il s'agit de caractériser des lieux réels. De plus, la réalisation de mesures ne demande pas de connaissances précise sur la géométrie et les matériaux des lieux. Enfin, comme nous le verrons en chapitre 3, une collaboration avec des chercheurs spécialistes du patrimoine architectural a permis de nous ouvrir les portes d'une vingtaine de lieux du patrimoine de la région PACA et nous offrait ainsi la possibilité de constituer, par la mesure, un corpus important d'environnements acoustiques, destiné à l'étude de la perception auditive spatiale.

1.4.2 Choix de l'ambisonie (HOA)

Précédemment, différentes techniques de spatialisation ont été décrites. Deux d'entre elles sont particulièrement adaptées à la mesure puis à l'auralisation d'environnements acoustiques. Il s'agit de la technique binaurale et de la technique ambisonique. D'un côté, le binaural est une technique largement répandue permettant, le plus souvent avec une tête artificielle, de réaliser des mesures de réponses impulsionnelles binaurales de salles, porteuses d'informations spatiales. Ces réponses peuvent ensuite être restituées directement aux oreilles d'un auditeur grâce à un casque. Cette technique a l'avantage d'être très directe et de ne nécessiter aucun traitement entre la mesure et la restitution. Les réponses binaurales sont bi-voies et donc légères, ce qui les rend particulièrement adaptées pour des applications web par exemple. En revanche, l'orientation de l'auditeur par rapport à la scène sonore est figée et déterminée par la position de la tête artificielle lors de la mesure. De plus, la qualité du rendu binaural est très dépendante de la morphologie de l'auditeur. D'un autre côté, la technique ambisonique permet la mesure de réponses impulsionnelles 3D par un réseau de microphones, et une restitution de ces réponses sur un réseau de haut-parleurs. Les avantages du formalisme ambisonique, notamment la décomposition du champ sur une base d'harmoniques sphériques, permettent une grande indépendance entre le dispositif de mesure et le système de restitution du champ acoustique. A l'inverse du binaural, l'ambisonie caractérise le champ acoustique dans les trois dimensions de l'espace et la restitution sur haut-parleurs permet à l'auditeur de s'orienter à sa guise dans la scène sonore. En revanche une bonne précision spatiale ne peut être atteinte que pour des ordres élevés de l'ambisonie (HOA). Or, l'ordre dépend du nombre de capsules du réseau de microphones utilisé pour la captation et du nombre de haut-parleurs du dispositif de restitution. Une mesure à l'ordre N = 4 par exemple nécessite un microphone composé d'au moins $(N+1)^2 = 25$ capsules. Ce qui signifie que les données mesurées avec un microphone ambisonique d'ordre 4 sont plus de 10 fois plus lourdes que celles obtenues par un enregistrement binaural. Toutefois, de nombreuses méthodes et optimisations ont vu le jour permettant de réaliser une réduction binaurale d'un flux ambisonique [Noisternig *et al.*, 2003, Vennerød, 2014, Zaunschirm *et al.*, 2018, McKenzie *et al.*, 2019], ce qui rend également possible l'utilisation de contenus ambisoniques pour des applications à faible bande passante. Enfin, le laboratoire PRISM est équipé d'un système de spatialisation multicanale de 42 haut-parleurs répartis sur une structure géodésique de 4m de

diamètre et placé en chambre semi-anéchoïque. Ce système permet une auralisation HOA à l'ordre 5, dans des conditions acoustiques contrôlées [1]. Côté captation, le laboratoire PRISM est également équipé d'un réseau sphérique de microphones : l'Eigenmike 32 de mh acoustics, constitué de 32 microphones répartis le long d'une sphère rigide de 8,4 cm de diamètre. Des enregistrements au format HOA d'ordre 4 peuvent être obtenus par le biais de ce microphone. Pour toutes ces raisons, nous avons fait le choix de l'ambisonie HOA pour la mesure et la restitution d'environnements acoustiques 3D.

1. Une description générale du système de spatialisation du laboratoire PRISM est donnée en annexe A.

Chapitre 2

Expérience 1 : Evaluation perceptive d'un système d'auralisation ambisonique d'acoustiques 3D mesurées

Sommaire

Ce chapitre présente une expérience perceptive visant à quantifier le degré de fidélité de restitution du système ambisonique d'auralisation d'acoustiques mesurées, sur des critères spatiaux. Dans cette étude, les performances de localisation de sources sonores dans différents environnements acoustiques en conditions réelles et virtuelles sont comparées. Le chapitre est organisé de la manière suivante. D'abord, un récapitulatif des problématiques et hypothèses ayant conduit à la réalisation de ce test est présenté. Dans un second temps, nous présentons une méthode de report de la localisation de sources sonores en réalité virtuelle, développée pour permettre aux participants de reporter leur perception des sources en termes de position angulaire, de distance et de taille apparente. Puis, l'expérience perceptive et ses résultats seront décrits et discutés. Enfin, nous présenterons les conclusions et perspectives de ces travaux.

2.1 Contexte et enjeux de l'étude

Il a été montré à plusieurs reprises que l'environnement acoustique influence notre perception spatiale des sources sonores. Des études ont notamment révélé que l'acoustique jouait un rôle décisif dans la perception de la distance [Zahorik *et al.*, 2005, Kolarik *et al.*, 2016] et de la taille apparente des sources sonores [Ihlefeld et Shinn-Cunningham, 2011, Kasbach *et al.*, 2015, Wang *et al.*, 2020] et dans une moindre mesure sur la perception de leur position angulaire [Hartmann, 1983, Shinn-Cunningham, 2001, Rychtáriková *et al.*, 2009]. Toutefois, l'étude groupée de l'influence de l'acoustique sur ces trois grands attributs spatiaux de la source n'a à notre connaissance pas été traitée, faute de protocole expérimental adapté.

D'autre part, un système d'auralisation HOA d'ordre 4 a été retenu pour la mesure et la restitution d'environnements acoustiques réels. Ce système a notamment pour vocation de pallier la complexité d'étudier la perception des salles en conditions réelles, en permettant de conduire des expériences sur un ensemble d'environnements mesurés, dans des conditions d'écoute contrôlées. La qualité spatiale d'un rendu HOA à l'ordre 4 a été jugée bonne dans plusieurs études [Bertet *et al.*, 2009, Braun et Frank, 2011, Bertet *et al.*, 2013]. Un tel système semble donc adapté à l'étude de l'influence de l'acoustique sur la perception spatiale des sources sonores (position angulaire, distance et taille apparente). Cependant, pour pouvoir conclure sur la pertinence du recours à ce dispositif d'auralisation, il est tout de même nécessaire d'évaluer la fidélité du rendu selon ces trois attributs de sources.

En résumé, l'objectif principal de l'expérience présentée dans ce chapitre est donc d'évaluer et de caractériser perceptivement les potentielles dégradations spatiales induites par le système. La méthodologie mise en oeuvre pour cette étude est illustrée par la figure 2.1. Pour ce faire, nous comparerons les performances de localisation de sources sonores dans différents environnements acoustiques, en conditions

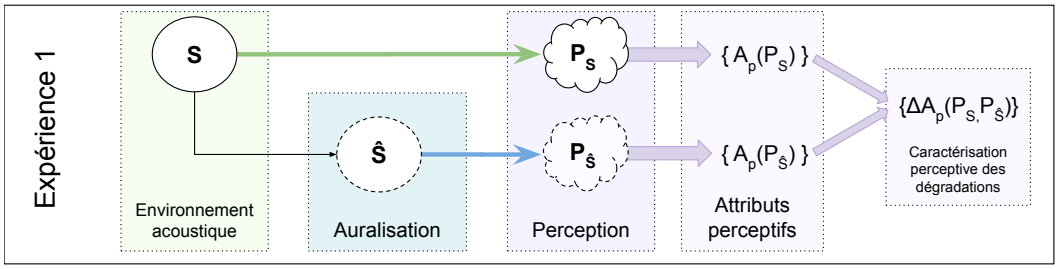


FIGURE 2.1 – Diagramme bloc de la méthodologie pour la caractérisation perceptive de la qualité spatiale de l'auralisation d'environnements acoustiques mesurés. S et $\hat{S}$ représentent respectivement les environnements acoustiques et leur auralisation. P représente la perception de ces environnements et $\{A_p\}$ un jeu d'attributs perceptifs permettant de caractériser la perception spatiale des sources sonores (position angulaire, taille apparente et distance de source). $\{\Delta A_p\}$ représente la mise en relation des attributs perceptifs obtenus en condition réelle d'écoute et en condition auralisée, permettant la caractérisation perceptive des dégradations induites par l'auralisation.

réelles et auralisées. En parallèle, il s'agit de proposer une nouvelle méthode de report de la perception spatiale de sources sonores et d'évaluer sa pertinence pour la caractérisation perceptive d'environnements réverbérants.

2.2 Interface VR pour le report d'attributs perceptifs spatiaux de sources sonores

2.2.1 Méthodes de report

Demander à un auditeur d'indiquer la position d'une source sonore peut paraître trivial. Pourtant il existe bien des façons de réaliser cette tâche. Dans son mémoire de thèse [Bertet, 2009], Bertet fait un état de l'art assez exhaustif des différentes méthodes de report de la position perçue d'une source sonore, répertoriées à l'époque (2009). Une première méthode consiste à décrire oralement ou par le biais d'une interface la position de la source, en indiquant les deux angles des coordonnées sphériques, azimut et élévation. Cette méthode est appelée méthode du jugement absolu [Wightman et Kistler, 1989]. Elle présente l'avantage de pouvoir indiquer la position de sources situées à l'arrière de l'auditeur mais requiert une lourde phase d'entraînement. Manson et al. suggèrent également que l'élicitation verbale pour indiquer la position d'un objet est sujette à une plus grande variabilité dans les réponses par rapport à des méthodes d'élicitation non-verbale [Mason *et al.*, 2000].

Parmi les méthodes d'élicitation non-verbale, les méthodes dites de pointage consistent à pointer avec le doigt ou la tête ou bien à manipuler des objets (e.g. une canne, un pistolet) [Haber *et al.*, 1993, Gröhn *et al.*, 2007, Bahu *et al.*, 2016], dans

la direction perçue de la source sonore. C'est une méthode répandue, qui pourtant peut induire des biais en fonction du pointeur utilisé. Le pointage du doigt, par exemple, induit un biais dépendant de la main utilisée [Majdak *et al.*, 2008], tandis que le pointage avec la tête ou le nez induisent un biais pour les directions de haute élévation notamment, dû aux limites physiologiques de l'orientation de la tête [Djelani *et al.*, 2000]. Ces méthodes sont relativement bien adaptées pour des tâches de localisation en boucle fermée (présentation d'un stimulus sonore en continu ou autant de fois que le souhaite l'auditeur) et lorsque l'auditeur est autorisé à se tourner en direction de la source. En revanche, lorsque le stimulus est bref et pour des positions de sources éloignées de la position frontale, ces méthodes sont moins adaptées puisqu'elles demandent un travail de mémorisation de la position du stimulus, puis d'orientation de l'auditeur dans la direction mémorisée. D'autres méthodes égocentrées de pointage ont été développées pour pallier certains de ces biais. Elles consistent également à pointer dans la direction de la source mais cette fois-ci de façon indirecte, en manipulant un pointeur visuel ou acoustique. Les méthodes ayant recours à un pointeur visuel (e.g. pointeur laser, positionné sur la tête ou dans la main de l'auditeur), ont été employées dans plusieurs études [Seeber, 2002, Majdak *et al.*, 2010, Winter *et al.*, 2017]. Elles consistent à projeter sur une surface, la plupart du temps entourant l'auditeur, la position perçue de la source à l'aide du pointeur visuel. Cette technique propose donc un report visuel de la perception auditive, ce qui, pour des sources situées dans le champ visuel de l'auditeur, s'avère être une méthode présentant un faible biais [Seeber, 2002]. Elle ne permet toutefois pas de reporter la position de sources situées en dehors du champ visuel et est donc inappropriée pour des expériences où le sujet n'est pas autorisé à bouger la tête. Les méthodes basées sur la manipulation d'un pointeur acoustique ont pour but d'éviter de mêler deux modalités dans le report de la localisation [Langendijk et Bronkhorst, 1997, Pulkki et Hirvonen, 2004, Bertet *et al.*, 2009]. Elles consistent à utiliser un haut parleur monté sur un bras, dont l'utilisateur peut contrôler la position. Ce dernier doit alors faire coïncider la position du pointeur acoustique avec celle de la source à localiser. Cette méthode requiert une instrumentation assez lourde et est la plupart du temps utilisée pour de la localisation de sources situées dans le plan azimutal. En revanche elle présente l'avantage de permettre au sujet de reporter leur jugement pour des positions de sources à 360° sans qu'il ait besoin de se tourner.

Enfin, il existe des méthodes exocentrées, pour lesquelles l'auditeur doit reporter son jugement sur un objet ou une interface graphique représentant l'espace auditif [Gilkey *et al.*, 1995, Braasch et Hartung, 2002, Schoeffler *et al.*, 2014]. L'un des exemples les plus cités est la méthode GELP (God's Eye Localization Pointing) consistant à reporter la position de la source sur une sphère de 20cm de diamètre placée devant l'auditeur [Gilkey *et al.*, 1995]. Ces méthodes ont pour principal atout de permettre un report rapide de la position perçue. En revanche, en comparant la méthode GELP à des méthodes de pointage avec le doigt et la tête, Bahu et al. ont révélé une erreur de localisation systématiquement plus élevée induite par la

méthode exocentrée [Bahu *et al.*, 2016].

En résumé, de nombreuses méthodes permettent le report de la perception de la position angulaire de sources sonores : verbales ou non-verbales, égocentrées ou exocentrées, avec pointeurs visuels ou acoustiques. Les méthodes induisant le moins de biais sont celles ayant recours à un pointeur visuel (pointeur laser par exemple), allant dans le sens d'une intégration multi-sensorielle de l'espace. Ces méthodes sont bien adaptées à la localisation de sources situées dans le champ visuel de l'auditeur ou lorsque celui-ci est autorisé à s'orienter dans la scène sonore. En revanche, elles ne le sont pas pour des tests de localisation à l'aveugle et pour des sources positionnées sur l'ensemble de la sphère auditive, lorsque l'auditeur doit rester statique. Toutes ces méthodes s'intéressent uniquement au report de la position angulaire des sources sans prendre en compte d'autres attributs spatiaux des sources tels que la distance et la taille apparente de la source.

2.2.2 Recours à la réalité virtuelle

Les technologies VR offrent aujourd'hui de nouvelles possibilités pour questionner la perception [Rumsey, 2018, Stecker, 2019]. En ce qui concerne la perception auditive spatiale, de nouvelles interfaces pour la localisation de source en VR ont vu le jour et se sont révélées adéquates pour la réalisation de cette tâche. Majdak et al. ont par exemple étudié la localisation de source via une interface en réalité virtuelle [Majdak *et al.*, 2010]. Les participants devaient pointer en direction de la source, avec la tête ou avec un contrôleur tenu dans la main, et ce avec ou sans retour visuel de l'environnement virtuel. L'étude révèle des performances de localisation accrues avec le retour visuel, par rapport à celles obtenues par une localisation à l'aveugle, indépendamment du membre utilisé pour pointer (tête ou main).

D'un autre côté, le port d'un casque de VR a un impact sur les HRTF de l'auditeur qui le porte [Gupta *et al.*, 2018]. Les performances de localisation avec un casque de VR ont été récemment étudiées par Huisman et al. pour une restitution sonore ambisonique 2D, à différents ordres, avec ou sans retour visuel [Huisman *et al.*, 2021]. Dans cette étude, la tâche de localisation a été réalisée en boucle-ouverte, i.e. les participants devaient garder leur tête orientée dans une direction de référence à 0° lors de la présentation du stimulus, puis dès la fin de la présentation, ils pouvaient s'orienter dans la direction de la source pour effectuer le report. Le port du casque induit une latéralisation de la position perçue des sources d'environ 2° en moyenne, pour des sources situées de part et d'autre du plan frontal. En revanche aucun effet n'a été observé pour les sources autour de la position frontale. Ces résultats indiquent que l'utilisation d'un casque de VR peut être envisagée sans crainte d'induire un biais, pour des tâches de localisation en boucle-fermée (lorsque le sujet est autorisé à s'orienter pendant la présentation du stimulus).

Enfin, le recours à la VR permet d'envisager de nouveaux protocoles donnant une description plus globale de l'image spatiale perçue de scènes sonores [Fargeot *et al.*, 2019]. Pour cette étude, une interface en réalité virtuelle a permis de mettre en évidence les dégradations spatiales induites par un remix spatial de sources sonores issues d'un processus de séparation de sources. La tâche de localisation consistait à entourer les sources composant la scène sonore et ainsi définir leur(s) position(s) dans l'espace ainsi que leur forme et leur taille perçue. Pour en revenir aux problématiques abordées dans ce chapitre, à savoir : l'évaluation de l'impact de l'acoustique sur les caractéristiques spatiales des sources ainsi que l'évaluation de la qualité du rendu spatial d'un système d'auralisation ambisonique 3D à l'ordre 4, celles-ci peuvent être traitées avec le même type de méthodes que celle utilisée dans cette étude. Toutefois, en plus du report de la position angulaire et de la taille, nous souhaitons également être en mesure de quantifier la perception de la distance source-auditeur. Nous avons donc repensé l'interface afin de pouvoir intégrer ces informations complémentaires.

2.2.3 Méthode proposée et interface

L'interface proposée se présente de la manière suivante. L'auditeur est immergé dans un espace virtuel illustré en figure 2.2, constitué d'un sol quadrillé (carreaux d'un mètre de côté) à perte de vue et d'une demi-sphère semi-transparente entourant l'auditeur. A l'aide d'une manette tenue dans la main droite, le sujet manipule un pointeur de la même façon qu'il manipulerait un pointeur laser. Un rayon lumineux part de la manette, dans la direction pointée et l'intersection entre ce faisceau et la surface de la demi-sphère est marquée par une petite boule lumineuse, faisant office de pointeur. Avec tous ces éléments l'auditeur peut reporter sa perception des sources en termes de position (azimut, élévation), de taille et de distance de la source.

Pour régler la distance, le sujet peut, à l'aide du joystick présent sur la manette, régler le rayon de la sphère qui l'entoure. La distance du pointeur (au point d'intersection entre la demi-sphère et le faisceau du pointeur) est donc également affectée. Afin de mieux représenter la distance entre l'auditeur et la projection du pointeur, une petit cercle lumineux est tracé au sol, à l'aplomb de ce dernier. Une fois la distance réglée et validée, l'auditeur peut dessiner sur les parois de la sphère en manipulant le pointeur tout en maintenant enfoncée la gâchette de la manette. De façon analogue à [Fargeot *et al.*, 2019], il est demandé au sujet de définir la position et la taille de la source en entourant une zone dans laquelle il estime que la source se trouve. Plus la source est perçue comme large ou difficile à localiser, plus la zone entourée est grande. Une fonction de "retour en arrière" est également implémentée, permettant au sujet d'effacer son tracé ou de revenir à l'étape du réglage de la distance s'il n'est pas satisfait de sa réponse. Lorsqu'il est satisfait de son tracé, l'utilisateur doit valider sa réponse pour passer à la suite.

Cette interface VR a été développée à l'aide du moteur de jeu Unity, pour un

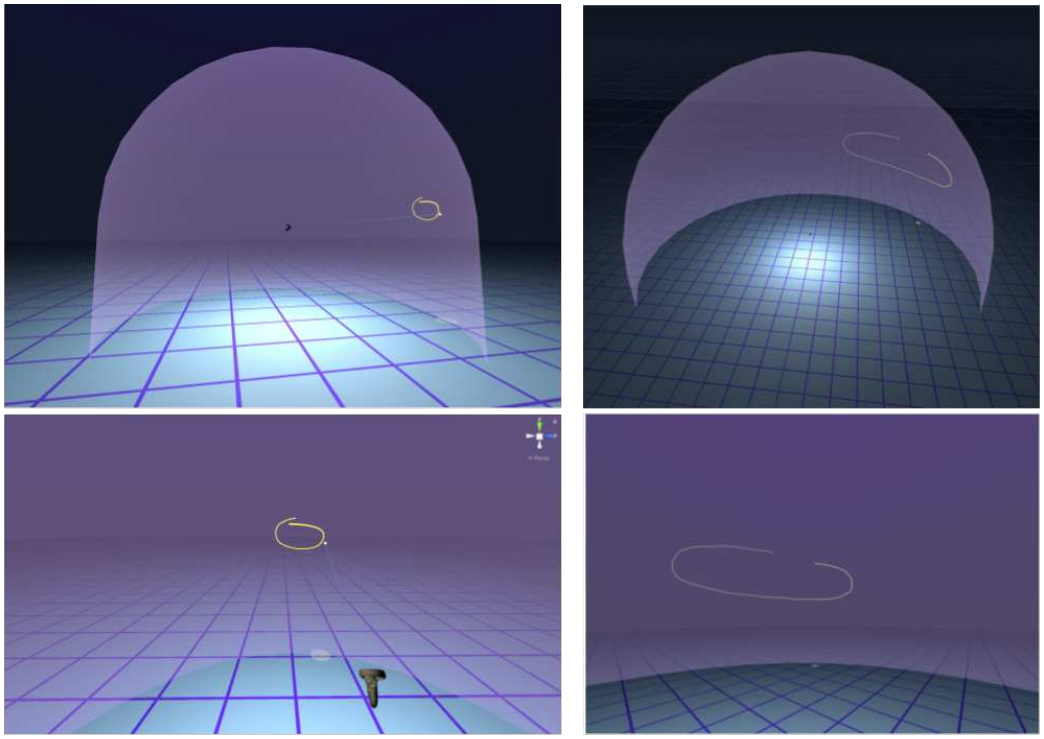


FIGURE 2.2 – Interface VR pour le report d'attributs spatiaux de sources sonores. A gauche : report d'une source proche, de petite taille. A droite : report d'une source éloignée, perçue comme large. En haut : point de vue spectateur (de l'extérieur de la sphère). En bas : point de vue utilisateur.

usage sur un casque de VR. Elle est compilée sous la forme d'une application Androïd, permettant son utilisation sur des casques nomades, en l'occurence ici sur Oculus Quest. L'intérêt d'utiliser un casque nomade est de créer un workflow simple et léger, en concentrant la gestion de l'interface graphique graphique et des contrôles dans le casque et ainsi de pouvoir réaliser l'expérience de localisation dans différents environnements acoustiques réels, sans nécessité de recourir à un ordinateur tiers.

D'autre part, afin de faire correspondre l'orientation de l'environnement virtuel avec celle de l'environnement réel, une étape de calibration est proposée au lancement de l'application. L'Oculus Quest étant équipé de deux manettes, celle des deux qui n'est pas utilisée par l'utilisateur pour le report est placée à une position de référence dans l'environnement réel, considérée comme le 0 degré sur le plan azimutal. Sa position est marquée dans l'espace virtuel par un petit cube vert. Lors de la phase de calibration un viseur apparaît dans le casque, au centre de l'écran. L'utilisateur doit alors orienter sa tête pour pointer avec le viseur en direction du cube vert. Au bout de 2 secondes maintenue dans cette position, l'orientation de l'environnement virtuel est réinitialisée et le zéro degré en azimut de l'environnement virtuel est à présent coïncident avec le zéro degré de l'environnement réel.

2.3 Méthodologie de l'expérience

L'expérience consiste donc à comparer les performances de localisation dans différents environnements acoustiques, en conditions d'écoute réelle par rapport à celles obtenues lors d'une auralisation de mesures de ces mêmes environnements acoustiques.

2.3.1 Choix des conditions acoustiques

Trois acoustiques différentes sont à l'étude ici et ont été sélectionnées pour leur diversité et leur proximité géographique. En effet, dans l'idée de mesurer l'impact de l'acoustique sur les performances de localisation, en condition réelle d'écoute, nous voulions être capables de tester plusieurs environnements acoustiques dans une seule et même session de test, dans un souci de simplicité et de confort pour les participants à l'expérience. Ces environnements devaient aussi dans l'idéal se situer à proximité de notre plateforme d'auralisation ambisonique, pour que les participants puissent dans le même temps réaliser la session de test en conditions auralisées. Trois salles vides notées $S1$, $S2$ et $S3$, de caractéristiques géométriques et acoustiques différentes ont ainsi été sélectionnées au sein d'un même bâtiment du campus Joseph Aiguier à Marseille, à proximité immédiate du laboratoire PRISM. La figure 2.3 présente les temps de réverbération par bande d'octave des 3 environnements sélectionnés.

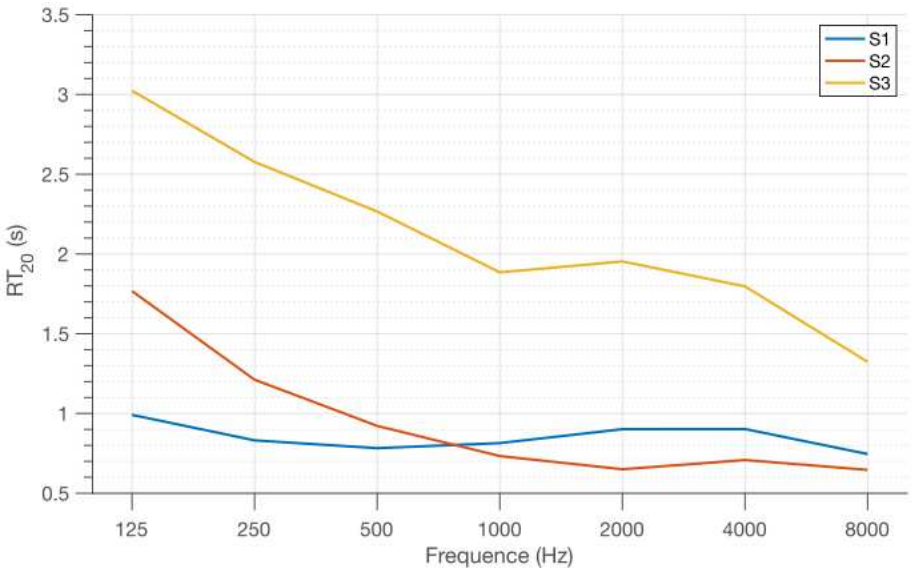


FIGURE 2.3 – Temps de réverbération par bandes d'octave des trois salles sélectionnées pour l'expérience.

Afin de quantifier l'influence de la distance source-auditeur sur la localisation,

deux sources (Genelec 8020A) ont été positionnées dans chacune des salles à l'étude, aux distances de 2 et 4m de la position d'écoute, notées respectivement D2 et D4. La figure 2.4 donne à voir la géométrie et la configuration spatiale des sources et position d'écoute dans chacune des salles. La position angulaire des sources est choisie arbitrairement, de façon à ne pas créer d'attente sur la configuration spatiale dans les différentes salles.

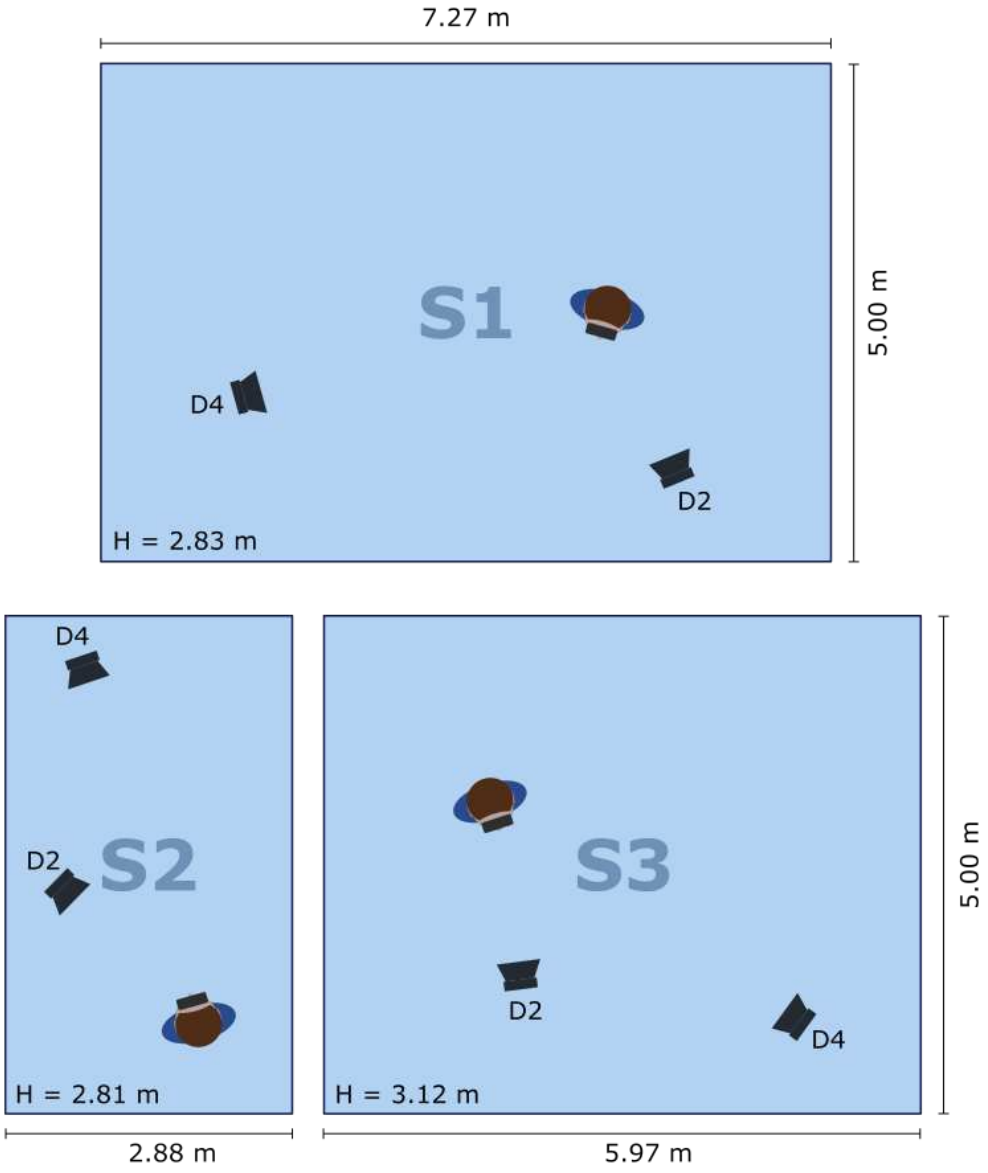


FIGURE 2.4 – Dimensions et configurations spatiales des sources et positions d'écoutes des trois salles à l'étude.

2.3.2 Mesures acoustiques et auralisation

Dans chacune des salles sélectionnées pour l'expérience et pour les deux configurations source-auditeur, des mesures acoustiques ont été réalisées avec le microphone sphérique eigenmike (em32), placé à la position d'écoute. La mesure consiste à collecter les réponses impulsionnelles spatiales des 6 conditions acoustiques de test (3 salles $\times$ 2 distances), avec la méthode par sinus glissant proposée par [Farina, 2000]. Le signal utilisé pour la mesure est un sinus glissant exponentiel d'une durée de 10 s, balayant la gamme de fréquence [20Hz; 22kHz]. Les réponses impulsionnelles ainsi mesurées seront ensuite convoluées avec les stimuli anéchoïques choisis, pour une auralisation dans la sphère.

2.3.3 Choix des stimuli

Les performances de localisation sont affectées par la nature du stimulus auditif présenté, en particulier son contenu spectral [Ihlefeld et Shinn-Cunningham, 2011]. Lorsqu'il s'agit d'évaluer la taille de la source, d'autres facteurs cognitifs plus hauts niveaux peuvent également rentrer en ligne de compte. On peut par exemple penser que pour des sources écologiques, i.e. des sons du quotidien (voix humaine, instruments acoustiques, bruit de la mer, etc...) nous avons a priori une représentation de la taille de la source, qui peut influencer notre jugement de la taille effective. D'un autre côté, lorsqu'il s'agit d'étudier la perception d'acoustiques de salles, il peut être intéressant de choisir des stimuli en lien avec les usages des lieux. Ici nous avons donc choisi trois stimuli sonores pour représenter ces différents cas de figures. Deux stimuli ont été choisis en lien avec l'usage des lieux : un stimulus de parole (durée : 3s), identifié "Parole" et un extrait de guitare classique (durée : 1 min) identifié "Guitare". Le troisième stimulus choisi est un train d'impulsions de bruit blanc (durée : 1s), identifié "Burst". C'est un stimulus abstrait, large bande, couramment utilisé dans des expériences de localisation de sources [Macé *et al.*, 2012].

2.3.4 Plan d'expérience

Afin de comparer les performances de localisation de sources en conditions réelles et virtuelles et pour plusieurs acoustiques de salles, un plan factoriel à quatre facteurs a donc été élaboré pour cette expérience. Les quatre facteurs en question sont : la condition d'écoute (COND : réelle RE, virtuelle V), l'environnement acoustique (SALLE : S1, S2, S3, décrites en section 2.3.1), la distance de la source (DIST : D2, D4, également définies en section 2.3.1) et le stimulus sonore (STIM : Parole, Guitare et Burst, présentés en section 2.3.3). Tous les sujets doivent évaluer toutes les combinaisons du plan factoriel. L'expérience est organisée en deux sessions, correspondant aux deux conditions d'écoute RE et V. Pour éviter les biais de présentation, une moitié des participants a commencé par la session d'écoute RE, tandis que l'autre moitié a commencé par la session V. Au sein de la session V, toutes les conditions ont été présentées aléatoirement. En revanche la session RE est organisée en 3 sous-sessions relatives aux trois environnements acoustiques S1, S2, S3. L'ordre

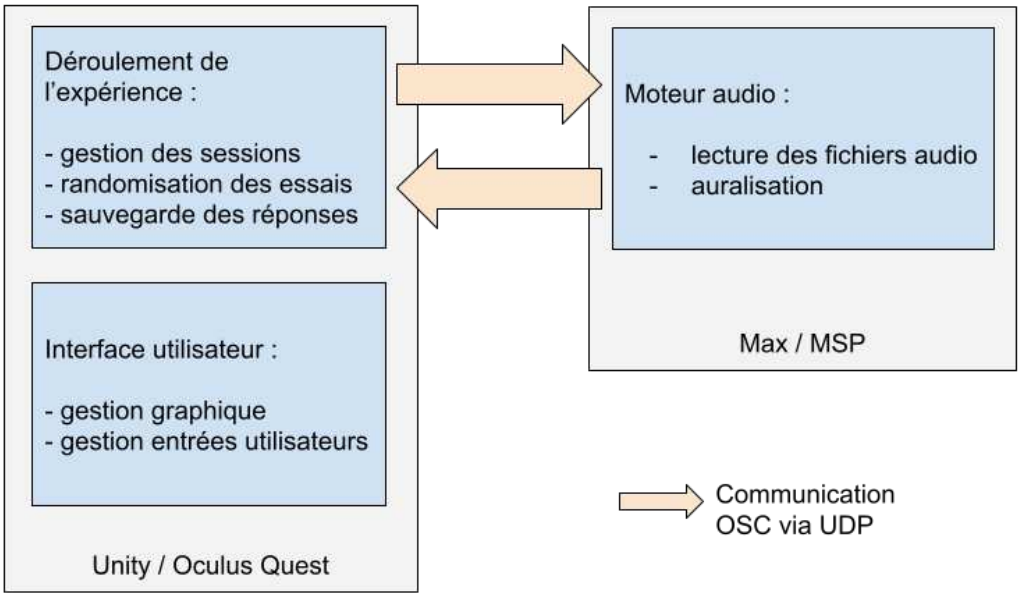


FIGURE 2.5 – Schéma bloc du dispositif expérimental.

des conditions au sein de chaque sous-session ainsi que l'ordre des sous-sessions sont également déterminés par tirage aléatoire pour chaque participant.

2.3.5 Participants

21 participants (15 hommes, 6 femmes) ont accepté de prendre part à cette étude. Ils sont en moyenne âgés de 28.6 ans (SE = 6.5 ans) et ne présentent pas de problèmes d'audition. Tous les sujets étaient naïfs quand à la configuration spatiale des sources et des environnements acoustiques, que ce soit en condition virtuelle ou réelle.

2.3.6 Dispositif expérimental

Le dispositif expérimental est conçu en 3 blocs réalisant les différentes fonctions nécessaires au bon déroulement de l'expérience, à savoir : l'interface utilisateur (gestion graphique, gestion des commandes utilisateurs), le moteur audio (lecture des fichiers audio et auralisation), le déroulement de l'expérience (organisation des sessions, randomisation des essais, formatage et sauvegarde des réponses). L'interface utilisateur, présentée en section 2.2.3 ainsi que le déroulement de l'expérience sont gérés par l'application Android développée avec Unity pour tourner sur Oculus Quest.

La gestion de l'audio est assurée grâce au logiciel Max/MSP. Elle est diffère légèrement en fonction des deux conditions d'écoute RE et V. En condition réelle, chaque salle à l'étude est équipée d'une carte son (Motu UltraLite MK3) connectée

aux deux haut-parleurs présents dans la salle (*c.f.* section 2.3.1). Un ordinateur portable est utilisé pour pouvoir naviguer facilement de salle en salle. Le patch Max assure simplement la lecture du fichier audio correspondant à l'un des 3 stimuli sélectionnés pour l'expérience, par l'une ou l'autre des sources présente dans la salle. Dans la condition virtuelle, Max/MSP assure l'auralisation via le système de diffusion multicanale du laboratoire PRISM (c.f. annexe A). Pour ce faire, les stimuli anéchoïques sont convolués aux SRIR mesurées dans les configurations présentées en section 2.3.2, encodées au format ambisonique. Le décodage sur le réseau de HP est aussi réalisé par Max/MSP, grâce aux outils de la librairie spat5. Il est effectué avec la méthode "energy preserving", sans optimisation.

Une communication OSC est établie entre le casque de VR et le moteur audio via le protocole UDP. Elle permet de transmettre du casque vers Max/MSP les informations concernant le déroulement de l'expérience, en particulier la clé du stimulus à jouer en début de chaque nouvel essai. Cette clé se présente sous la forme suivante {id_salle ; id_dist ; id_stim} respectivement les indices de la salle (S1, S2, S3), de la distance de la source (D2, D4) et du stimulus (Parole, Guitare, Burst).

2.3.7 Procédure

Avant de commencer l'expérience, les participants sont invités à lire une feuille de consigne décrivant le déroulement de l'expérience et la tâche à réaliser. Après lecture, ils peuvent poser leurs questions à l'expérimentateur s'ils en ressentent le besoin. L'expérience est composée de trois sessions : une courte session de familiarisation et deux sessions de test définies par les deux conditions d'écoute, réelle ou virtuelle. La session de familiarisation est une session informelle, réalisée en début d'expérience, où le participant est amené découvrir l'interface afin de se familiariser avec la tâche à réaliser et les différents contrôles proposés par l'interface. Le sujet est positionné sur une chaise, dans une salle ne faisant pas partie des environnements acoustiques à l'étude. La tâche à réaliser est une tâche de localisation de sources dont la méthode est détaillée en section 2.2.3. Pour rappel, elle se décompose en deux étapes : (1) reporter la distance perçue de la source, en réglant le rayon d'une demi-sphère entourant l'auditeur, (2) entourer le plus précisément possible la zone dans laquelle la source est entendue, à l'aide du pointeur. Si la source est étendue ou difficilement localisable, les participants sont encouragés à entourer une zone large qui doit circonscrire la source. Dans la phase de familiarisation le sujet est placé sur une chaise et l'expérimentateur se positionne dans la salle de façon aléatoire. Le sujet doit alors localiser la voix de l'expérimentateur, puis valider sa réponse. L'essai est répété jusqu'à ce que le sujet se sente à l'aise avec la tâche à réaliser.

Pour la session en condition réelle d'écoute, les participants sont conduits, yeux bandés, dans chacune des salles du test et sont accompagnés jusqu'à une chaise pivotante placée à la position d'écoute. Une fois confortablement installé sur la

chaise, le sujet est équipé avec le casque de réalité virtuelle, sur lequel l'application présentant l'interface de l'expérience a été préalablement lancée. Dès cet instant, les participants sont autorisés à ouvrir les yeux et sont invités à ajuster la position du casque sur la tête pour un port confortable et pour que l'interface présentée soit visible et nette. Avant de commencer l'expérience, une calibration est réalisée telle qu'elle est décrite en section 2.2.3. La manette gauche, servant de repère pour la calibration, est positionnée dans l'axe du haut-parleur le plus proche (position D2). Une fois la calibration effectuée, les sujets peuvent lancer le début de la session in-situ. Un premier essai est tiré aléatoirement parmi les 6 combinaisons possibles "position source * stimulus sonore". Il est joué en boucle, durant le temps nécessaire au sujet pour réaliser la tâche de localisation. Lorsque celui-ci valide sa réponse, l'essai suivant est lancé et la tâche est répétée jusqu'à ce que les 6 combinaisons de test aient été présentées. A la fin de la session dans une première salle, le sujet est raccompagné à l'extérieur, les yeux bandés. L'opération est répétée dans les deux autres salles sélectionnées pour le test.

En ce qui concerne la session en condition d'écoute virtuelle, la logistique est plus simple. Les participants sont invités à s'asseoir sur une chaise pivotante placée au milieu du système d'auralisation, puis, une fois installés, à enfiler et régler le casque de VR à leur convenance. La calibration est ensuite réalisée. Enfin, la phase de test consiste à effectuer la tâche de localisation pour les 18 conditions auralisées. Pour chaque essai le stimulus est présenté en boucle jusqu'à validation de la réponse. De plus, afin d'éviter toute attente perceptive quant à la position de la source, à chaque nouvel essai, une rotation dans le plan azimutal est appliquée à la scène sonore de façon à ce que la source à localiser soit présentée avec une incidence en azimut aléatoire, comprise entre -90 et 90 degrés. Les participants peuvent mettre en pause la session à tout moment s'ils le désire.

2.3.8 Formatage, sauvegarde et traitement des données brutes

Au lancement de l'expérience, un fichier texte vierge est crée pour chaque nouveau sujet. Ce fichier est complété au fur est à mesure de l'expérience. Pour chaque essai, les informations de l'essai en cours sont inscrites dans le fichier, à savoir, le numéro de l'essai, et les indices de niveau des différents facteurs : la condition d'écoute (COND), la salle (SALLE), la distance de la source (DIST) et le stimulus sonore (STIM) ainsi que l'incidence théorique en azimut de la source (sachant que la position en élévation est invariante selon les différentes conditions). La réponse du sujet est enregistrée sous la forme d'un vecteur contenant les coordonnées de l'ensemble des points constituant le tracé, en coordonnées cartésiennes. Quelques exemples typiques de réponses sont présentés en figure 2.6. Enfin, le temps effectué pour réaliser la tâche est également sauvegardé. Chaque réponse est ensuite traitée pour déduire des données sauvegardées la position angulaire en azimut et élévation, la distance et la taille apparente perçues de la source.

Les tracés sont approximés par des ellipses, dans le plan azimut - élévation.

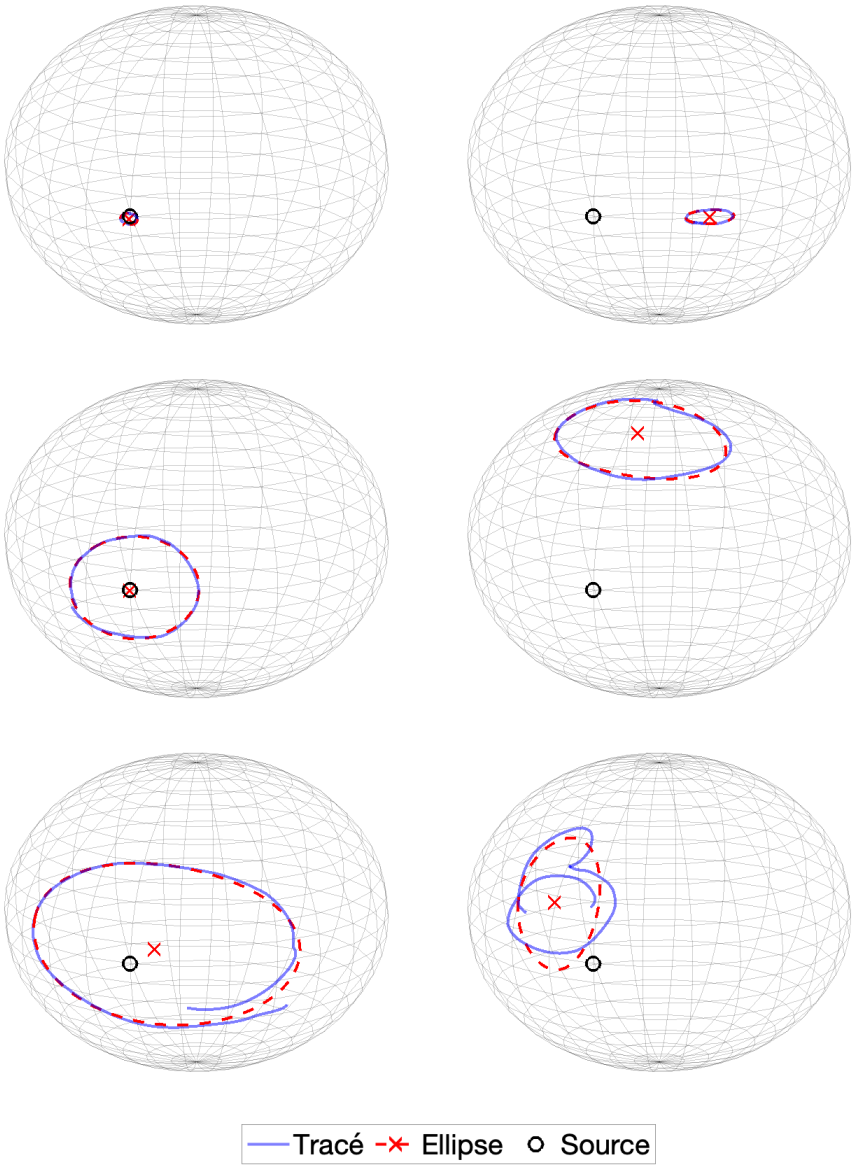


FIGURE 2.6 – Exemple de tracés réalisés pour différentes conditions de test et approximation des tracés en ellipse. La position angulaire de la source est repérée par un rond noir. Le tracé est marqué d'un trait plein bleu, l'ellipse est représentée en trait pointillé rouge et le centre de l'ellipse est marqué d'une croix rouge. Ces exemples illustrent la diversité des tracés en termes d'erreur angulaire et de taille reportée ainsi que la pertinence du recours à une approximation des tracés par une ellipse. L'exemple en bas à droite illustre les limites de cette approximation.

La position angulaire perçue en azimut θ_p et élévation ϕ_p et la taille apparente reportée sont déduites de cette approximation. La position angulaire est donnée par les coordonnées polaires du centre de l'ellipse. La taille apparente de l'ellipse, quant à elle, peut être caractérisée par un angle solide équivalent S_{eq} donné par la formule suivante :

$$S_{eq} = 4\pi \cdot sin\left(\frac{A}{2}\right) \cdot sin\left(\frac{B}{2}\right)$$

avec A et B, respectivement les petit et grand rayons de l'ellipse, exprimés en radian. L'unité relative aux angles solides est le stéradian, noté (sr).

Pour l'estimation de la distance perçue, le tracé étant effectué sur une demi-sphère centrée à la position d'écoute O, tous les points qui le compose sont équidistants de cette position centrale O. La distance perçue R_p est donc donnée par norme du vecteur entre OP_1 avec P_1 : la position du premier point du tracé.

A partir des grandeurs relatives à la position angulaire reportée par les participants (azimut θ_p, élévation ϕ_p), et connaissant les positions angulaires théoriques des sources dans les salles (θ_{theo}, ϕ_{theo}), les erreurs associés à ces grandeurs peuvent être calculées. Ainsi, pour chaque essai l'erreur en azimut $\varepsilon_\theta = \theta_p - \theta_{theo}$, en élévation $\varepsilon_\phi = \phi_p - \phi_{theo}$ sont calculées. Les valeurs absolues de ces erreurs $|\varepsilon_\theta|$ et $|\varepsilon_\phi|$ sont également rapportées. Les erreurs signées sont révélatrices d'une dissymétrie dans la perception angulaire des sources. Dans le cas où les participants localisent en moyenne correctement la source, cette erreur est nulle. L'erreur absolue quant à elle représente la quantité d'erreur moyenne commise lors de la tâche de localisation, indépendamment du signe de l'erreur.

2.3.9 Analyses statistiques

Le tableau 2.1 récapitule les variables étudiées lors de l'analyse statistique. Les données de certaines de ces variables ont du subir une transformation logarithmique pour respecter la condition de normalité, nécessaire à l'utilisation de modèles linéaires pour l'analyse statistique.

Les effets des différents facteurs sur ces variables et leurs interactions sont analysés par modèle linéaire mixte. en considérant trois effets fixes et deux effets aléatoires. La condition d'écoute (RE, V), l'environnement acoustique (S1, S2, S3), La distance de la source (D2, D4), et leurs interactions sont traités comme des effets fixes. Le stimulus sonore (Parole, Guitare, Burst) ainsi que les sujets sont traités comme des effets aléatoires. Une ANOVA à mesures répétées est réalisée sur le modèle pour mesurer la significativité des effets fixes. La taille des effets fixes est donnée par un estimateur de la quantité de variance expliquée, notée η^2. Des analyses post-hoc ont également été réalisées avec un ajustement de Tukey, afin de comparer les différents niveaux des facteurs à l'étude.

Grandeur	Variable (unité)	Expression	Trans. Log
Erreur azimut (±)	ε_θ (°)	$\theta_p - \theta_{theo}$	
Erreur azimut (abs.)	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\theta_p - \theta_{theo}\rvert$	✓
Erreur élévat. (±)	ε_ϕ (°)	$\phi_p - \phi_{theo}$	
Erreur élévat. (abs.)	$\lvert\varepsilon_\phi\rvert$ (°)	$\lvert\phi_p - \phi_{theo}\rvert$	✓
Distance perçue	R_p (m)	$\lvert\lvert OP_1 \rvert\rvert$	✓
Taille apparente	S_{eq} (sr)	$4\pi \cdot sin\left(\frac{A}{2}\right) \cdot sin\left(\frac{B}{2}\right)$	✓

TABLE 2.1 – Récapitulatif des variables analysées pour l'expérience. La colonne Trans.Log indique les variables pour lesquelles les données ont subi une transformation logarithmique, afin de respecter la condition de distribution normale des données, nécessaire à l'analyse statistique au travers d'un modèle linéaire.

Les réponses ayant une erreur absolue en azimut supérieure à 45° sont considérées comme aberrantes et retirées de l'analyse des grandeurs ε_θ et $|\varepsilon_\theta|$. De la même manière, toute réponse ayant une erreur absolue en élévation supérieure à 60° n'est pas prise en compte dans l'analyse de ε_ϕ et $|\varepsilon_\phi|$. Ainsi, 1.2 % des réponses ont été retirées de l'analyse des grandeurs azimutales et 3.6 % des réponses pour les grandeurs relatives à l'élévation.

Ces analyses ont été conduites avec R, à l'aide de la librairie 'LmerTest' [Kuznetsova *et al.*, 2015]. Le recours à cette méthodologie d'analyse est inspirée d'une analyse statistique conduite par Ahrens et. al, sur des données de localisation de sources sonores dans différentes conditions audio-visuelles [Ahrens *et al.*, 2019].

2.4 Résultats de l'expérience perceptive

Les résultats bruts pour chaque variable dépendante étudiée sont présentés en tableau 2.2. Les résultats des analyses statistiques sont organisés par variable dépendante. Les paragraphes qui suivent présentent ainsi les effets des différents facteurs et de leurs interactions sur les erreurs en azimut et élévation ainsi que sur la distance perçue et la taille apparente reportée des sources. Sur chacune des figures illustratives des résultats, la distribution des réponses est représentée par des boites à moustaches (valeurs médianes, premier et troisième quartiles) superposées à des diagrammes en violon. Les valeurs moyennes sont représentées par des points et les résultats des tests post-hoc sont indiqués par les labels suivants : $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

2.4.1 Effets sur la localisation en azimut

La localisation en azimut est évaluée selon deux grandeurs : une erreur signée ε_θ représentative d'une quantité de dissymétrie dans la position perçue des sources sur le plan azimutal et une erreur absolue $|\varepsilon_\theta|$ représentant la quantité d'erreur commise dans la localisation azimutale de la source. Concernant ε_θ, on observe un léger effet de la condition d'écoute $COND$: $F(1, 706.23) = 9.8091$, $p = 0.001808$, $\eta^2 = 0.01$, indiquant que dans le cas auralisé V, les sources ont été reportées légèrement plus à gauche ($+1.77°$ en moyenne) que dans le cas réel RE (*c.f.* Figure 2.7(a)). Une large interaction entre l'environnement acoustique et la distance de la source est également observée $SALLE \times DIST$: $F(2, 706.24) = 87.3471$, $p < 0.0001$, $\eta^2 = 0.20$ (*c.f.* Figure 2.7(b)). Enfin, la triple interaction des trois effets fixes a un également un effet significatif sur ε_θ $COND \times SALLE \times DIST$: $F(2, 706.17) = 33.1693$, $p < 0.0001$, $\eta^2 = 0.09$ (*c.f.* Figure 2.7(c)).

Comme illustré en Figure 2.8, l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$ est significativement affectée par (a) la condition d'écoute $COND$: $F(1, 706.59) = 173.0953$, $p < 0.0001$, $\eta^2 = 0.20$), (b) l'environnement acoustique $SALLE$: $F(2, 706.69) = 8.5683$, $p = 0.0002$, $\eta^2 = 0.02$). La condition d'écoute et l'environnement acoustique interagissent de manière significative $COND \times SALLE$: $F(2, 706.50) = 7.4978$, $p = 0.0006$, $\eta^2 = 0.02$) (c). Cette interaction décrit le fait qu'en condition réelle d'écoute, l'erreur absolue en azimut n'est pas dépendante de l'acoustique tandis qu'en condition virtuelle, l'erreur absolue est significativement moins élevée dans le cas de la salle S1 que dans les deux autres. Inversement, la distance de la source semble également avoir un effet significatif sur la précision de localisation $DIST$: $F(1, 706.63) = 7.5523$, $p = 0.0061$, $\eta^2 = 0.01$) (d). Une interaction entre la condition d'écoute et la distance de la source $COND \times DIST$: $F(1, 706.41) = 9.7136$, $p = 0.0019$, $\eta^2 < 0.01$) (e), indique que l'effet de la distance de la source sur $|\varepsilon_\theta|$ n'est observé qu'en écoute réelle. En effet, en conditions réelles, les participants ont en moyenne commis une erreur de localisation plus élevée pour les sources à 4 m, que pour celles placées à 2 m, alors qu'aucun effet de la distance n'est observée en conditions auralisées.

2.4.2 Effets sur la localisation en élévation

Une différence significative entre les différentes conditions d'écoute est observée au niveau de l'erreur en élévation signée ε_ϕ $COND$: $F(1, 695.58) = 356.8744$, $p < 0.0001$, $\eta^2 = 0.34$, ainsi que pour l'erreur absolue $|\varepsilon_\phi|$ $COND$: $F(1, 695.81) = 359.5922$, $p < 0.0001$, $\eta^2 = 0.34$. La figure 2.9 illustre ces résultats et laisse voir en (a) une large dissymétrie vers le haut de l'erreur signée, dans le cas virtuel et en (c) une erreur absolue plus élevée dans le cas virtuel également. D'autre part, l'erreur signée ε_ϕ est très légèrement affectée par l'environnement acoustique $SALLE$: $F(2, 695.07) = 4.1102$, $p = 0.0168$, $\eta^2 = 0.01$. Les tests post-hoc révèlent une différence dans le jugement de l'élévation entre la salle S2 et S1 (*c.f.* figure 2.9(b)).

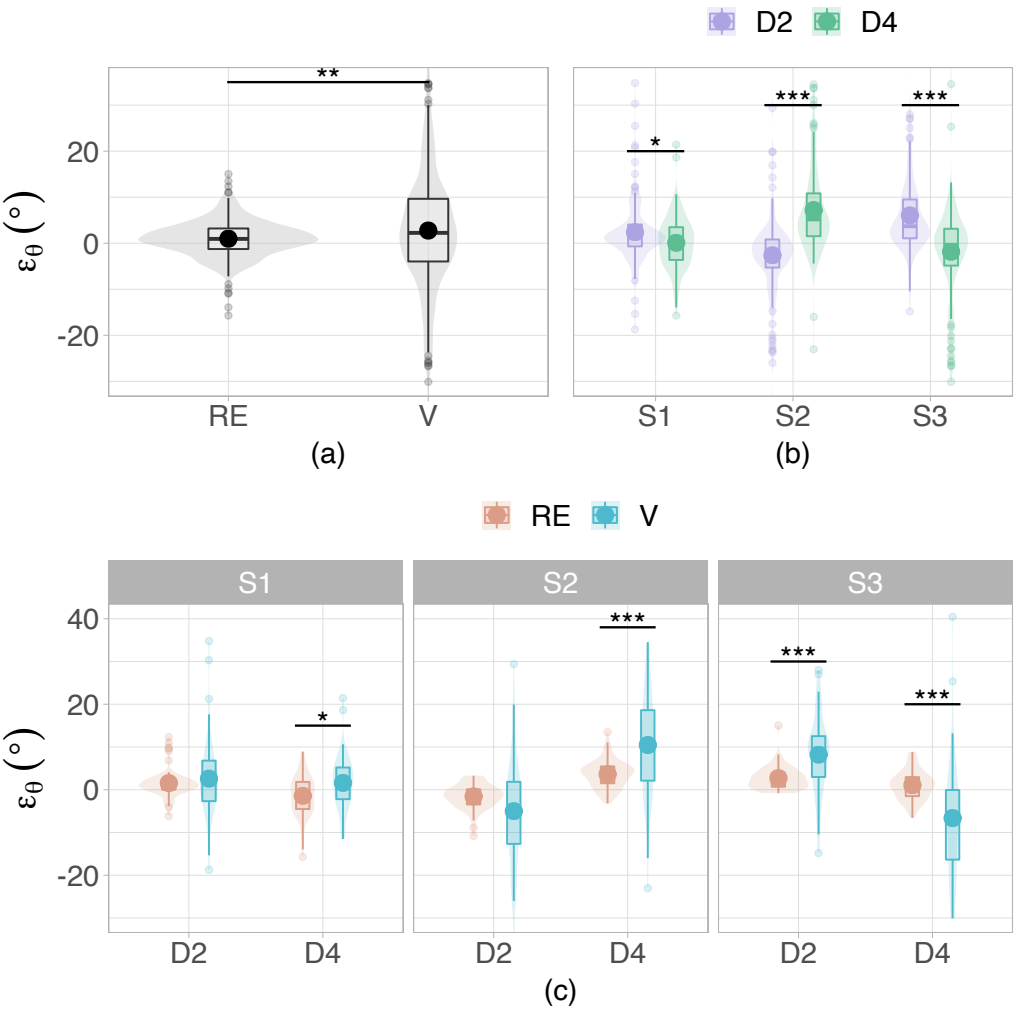


FIGURE 2.7 – Erreur de localisation en azimut ε_θ, en fonction de (a) la condition d'écoute $COND$ (RE : écoute réelle, V : auralisation HOA4), (b) l'interaction $SALLE \times DIST$ entre la salle $SALLE$ (S1, S2, S3) en abscisse et la distance $DIST$ (D2 : source à 2 m, D4 : source à 4 m) en couleur, (c) la triple interaction $COND \times SALLE \times DIST$ entre la condition d'écoute en couleur, la distance en abscisse et la salle.

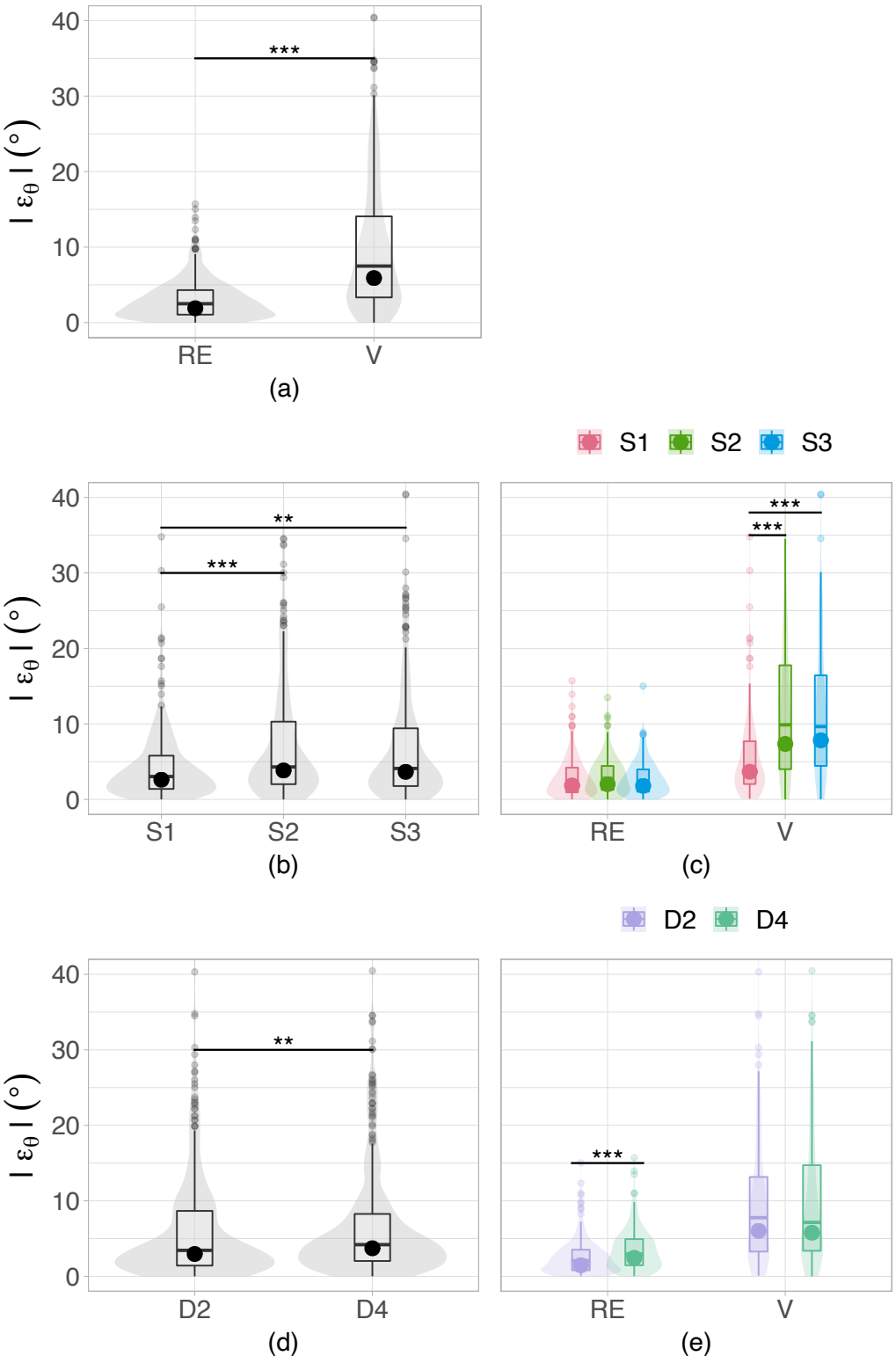


FIGURE 2.8 – Erreur absolue de localisation en azimut $|\varepsilon_\theta|$, en fonction de (a) la condition d'écoute $COND$ (RE : écoute réelle, V : auralisation HOA4), (b) la salle $SALLE$ (S1, S2, S3), (c) l'interaction $COND \times SALLE$ entre la condition d'écoute en abscisse et la salle en couleur, (d) la distance de la source $DIST$ (D2 : source à 2 m, D4 : source à 4 m), (e) l'interaction $COND \times DIST$ entre la condition d'écoute en abscisse et la distance de la source en couleur.

En effet, les participants semblent avoir en moyenne perçu les sources de la salle S2 légèrement plus hautes que celles de la salle S1.

2.4.3 Effets sur la distance perçue

La distance perçue des sources R_p est principalement dépendante de leur distance réelle $DIST$: $F(1,722) = 126.6416$, $p < 0.0001$, $\eta^2 = 0.15$. En moyenne, les sources placées à 2 m (D2) ont été perçues plus proches que celles placées à 4 m (D4) (*c.f.* Figure 2.10(d)). La distance perçue est également affectée, dans une moindre mesure, par l'acoustique de la salle dans laquelle se trouvent les sources $SALLE$: $F(2,722) = 15.8712$, $p < 0.0001$, $\eta^2 = 0.04$ (*c.f.* Figure 2.10(a)).

Bien que la condition d'écoute ne semble pas avoir d'influence globale sur le jugement de la distance, elle interagit significativement avec la salle $COND \times SALLE$: $F(2,722) = 11.5848$, $p < 0.0001$, $\eta^2 = 0.03$, comme illustré en figure 2.10(b)-(c). Les résultats des tests post-hoc révèlent qu'en condition réelle d'écoute, la salle n'a pas eu d'influence significative sur le jugement de la distance alors qu'en condition d'auralisation, les sources ont en moyenne été perçues plus loin dans la salle la plus réverbérante (S3) que dans les deux autres ($p < 0.0001$) et légèrement moins loin dans la salle S2 que dans la salle S1 ($p = 0.02$) (figure 2.10(b)). Cette interaction se manifeste aussi par le fait que les effets de l'auralisation sont différents pour chacune des salles à l'étude (figure 2.10(c)). D'après les tests post-hoc, la condition d'écoute n'a pas eu d'effet sur l'évaluation de la distance des sources dans la salle S1. En revanche, les sources de la salle S2 ont été perçues plus proches en condition auralisée qu'en condition réelle ($p = 0.0073$) tandis que celles de la salle S3 ont été perçues plus loin en condition auralisée qu'en condition réelle ($p = 0.0001$).

L'analyse révèle également la présence d'une faible interaction entre la condition d'écoute et la distance réelle de la source $COND \times DIST$: $F(1,722) = 7.5482$, $p = 0.0062$, $\eta^2 = 0.01$ (figure 2.10(e)). En effet, la distance est en moyenne correctement reproduite par l'auralisation pour la source à 4 m (D4), tandis que les sources placées à 2 m (D2) ont été perçue significativement plus loin en condition auralisée (V) qu'en condition réelle d'écoute (RE).

2.4.4 Effets sur la taille apparente de la source

On observe un large effet de la condition d'écoute sur la taille perçue des sources $COND$: $F(1,722) = 362.6654$, $p < 0.0001$, $\eta^2 = 0.33$, illustré par la figure 2.11(a). La taille apparente reportée est significativement plus élevée dans les conditions virtuelles qu'en écoute réelle. Un effet plus léger de la salle est également observé $SALLE$: $F(2,722) = 5.9941$, $p = 0.0.0026$, $\eta^2 = 0.02$, en figure 2.11(b). Une interaction significative entre la condition d'écoute et la salle $COND \times SALLE$: $F(2,722) = 6.2536$, $p < 0.0022$, $\eta^2 = 0.02$ est également observée (*c.f.* figure 2.11(c)). Les tests post-hoc révèlent en fait que l'effet de l'acoustique sur la taille

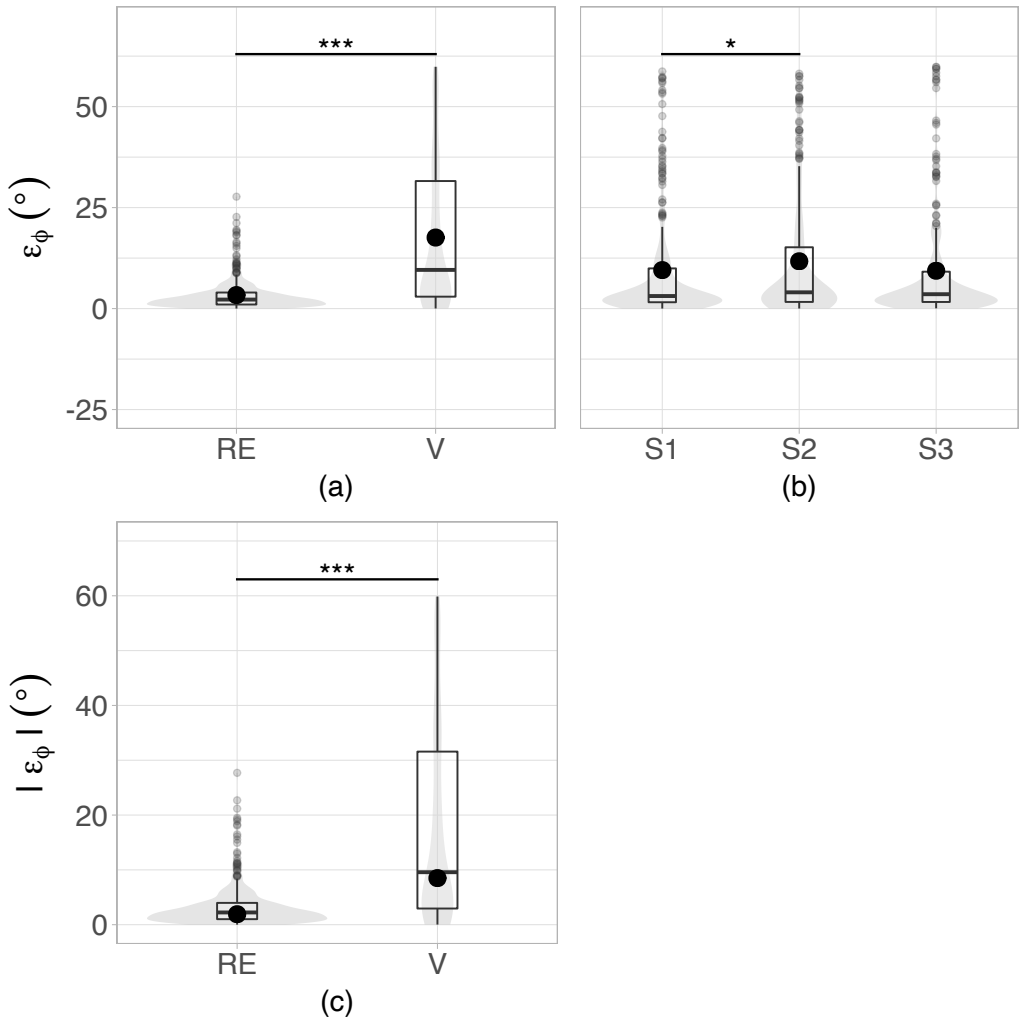


FIGURE 2.9 – Erreurs de localisation en élévation ε_ϕ, en fonction de (a) la condition d'écoute $COND$ (RE : écoute réelle, V : auralisation HOA4), (b) l'environnement acoustique $SALLE$ (S1, S2, S3), (c) erreurs absolues de localisation en élévation $|\varepsilon_\phi|$ en fonction de la condition d'écoute $COND$.

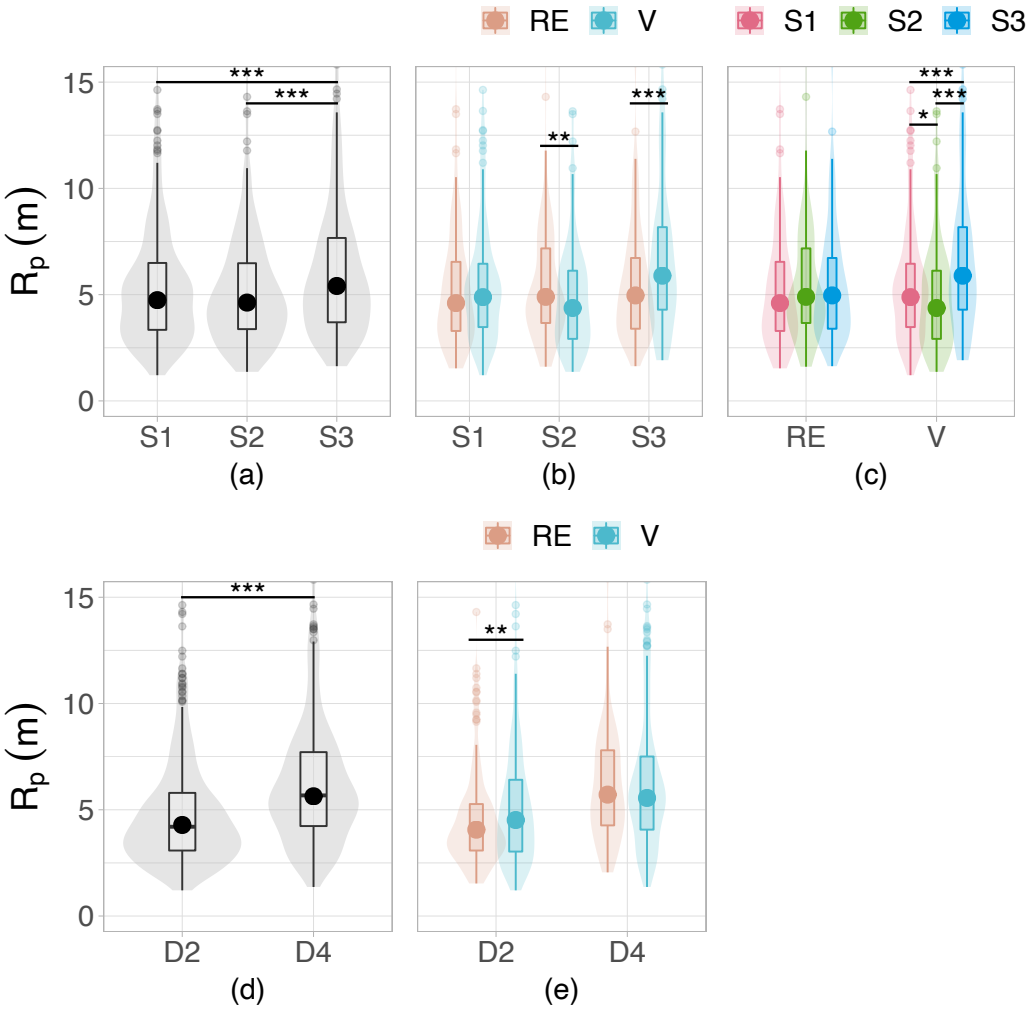


FIGURE 2.10 – Distance perçue en fonction de (a) l'environnement acoustique $SALLE$ (S1, S2, S3), (b) l'interaction $COND \times SALLE$ entre la condition d'écoute $COND$ (RE : écoute réelle, V : auralisation HOA4) en abscisse et l'environnement acoustique en couleur (c) l'interaction $COND \times SALLE$ entre l'environnement acoustique en abscisse et la condition d'écoute en couleur, (d) la distance réelle de la source $DIST$ (D2 : source à 2 m, D4 : source à 4 m), (e) l'interaction $COND \times DIST$ entre la distance réelle de la source en abscisse et la condition d'écoute en couleur.

apparente reportée n'est présent que lors de l'auralisation. Dans les cas auralisés, la taille apparente des sources pour la salle S1 est significativement plus faible que celle reportée pour les deux autres acoustiques S2 et S3.

Enfin, l'analyse révèle également un effet de la distance de la source sur la taille apparente de la source $DIST : F(1, 722) = 8.2625$, $p < 0.0042$, $\eta^2 = 0.01$, comme illustré en figure 2.11(d). Une fois n'est pas coutume, ce résultat est à pondérer par la présence d'une interaction entre la condition d'écoute et la distance de la source $COND \times DIST : F(1, 722) = 6.1858$, $p < 0.0131$, $\eta^2 < 0.01$ (*c.f.* figure 2.11(e)). Les résultats des tests post-hoc indiquent que dans le cas d'une écoute réelle, la taille apparente reportée des sources est plus faible pour les sources à 4 m que pour celles à 2 m, alors qu'en condition virtuelle, aucune différence de taille apparente n'est observée en fonction de la distance de la source.

2.5 Discussions des résultats

2.5.1 Performances de localisation dans les cas réels

Les résultats de l'étude révèlent que dans des conditions réelles d'écoutes, pour des sources placées à 2 et 4m dans trois environnements acoustiques plus ou moins réverbérants, les performances de localisation sont globalement bonnes et comparables à celles obtenues dans d'autres études sur le sujet [Blauert, 1997b]. Les participants ont en moyenne perçu les sources légèrement plus à gauche que leur position réelle (+0.9°) et ont commis une erreur absolue moyenne de localisation en azimut de 1.9°. Les sources ont été localisées en moyenne à une élévation de +0.6°, soit très légèrement au-dessus de la position réelle des sources, avec une erreur absolue moyenne de 1.90°.

L'acoustique de la salle ne semble pas perturber les performances de localisation en condition réelle. En revanche la distance de la source a une influence multiple sur ces performances. Sans surprise, les sources positionnées à 2 m ont été perçues plus proches que celles positionnées à 4 m, avec une distance moyenne reportée de 4.06 m pour les sources à 2 m contre 5.71 m pour celles à 4 m. Ces résultats indiquent néanmoins que la distance des sources a été globalement surestimée, ce qui diverge de la plupart des résultats sur la perception auditive de la distance, qui indiquent plutôt une tendance à sous-estimer la distance pour des sources situées au-delà de 1 m [Kolarik *et al.*, 2016]. Cette différence avec la littérature peut être la conséquence d'un biais induit par le report de la distance via l'interface VR et est discutée plus en détail en section 2.5.3.

D'autre part, la distance réelle de la source semble influencer la précision de localisation en azimut, puisque les participants ont en moyenne commis une erreur absolue plus grande pour les sources à 4 m (2.4°) que pour celles à 2 m (1.5°). Bien que l'écart d'erreur soit faible (< 1°), il ressort de l'analyse comme très significatif

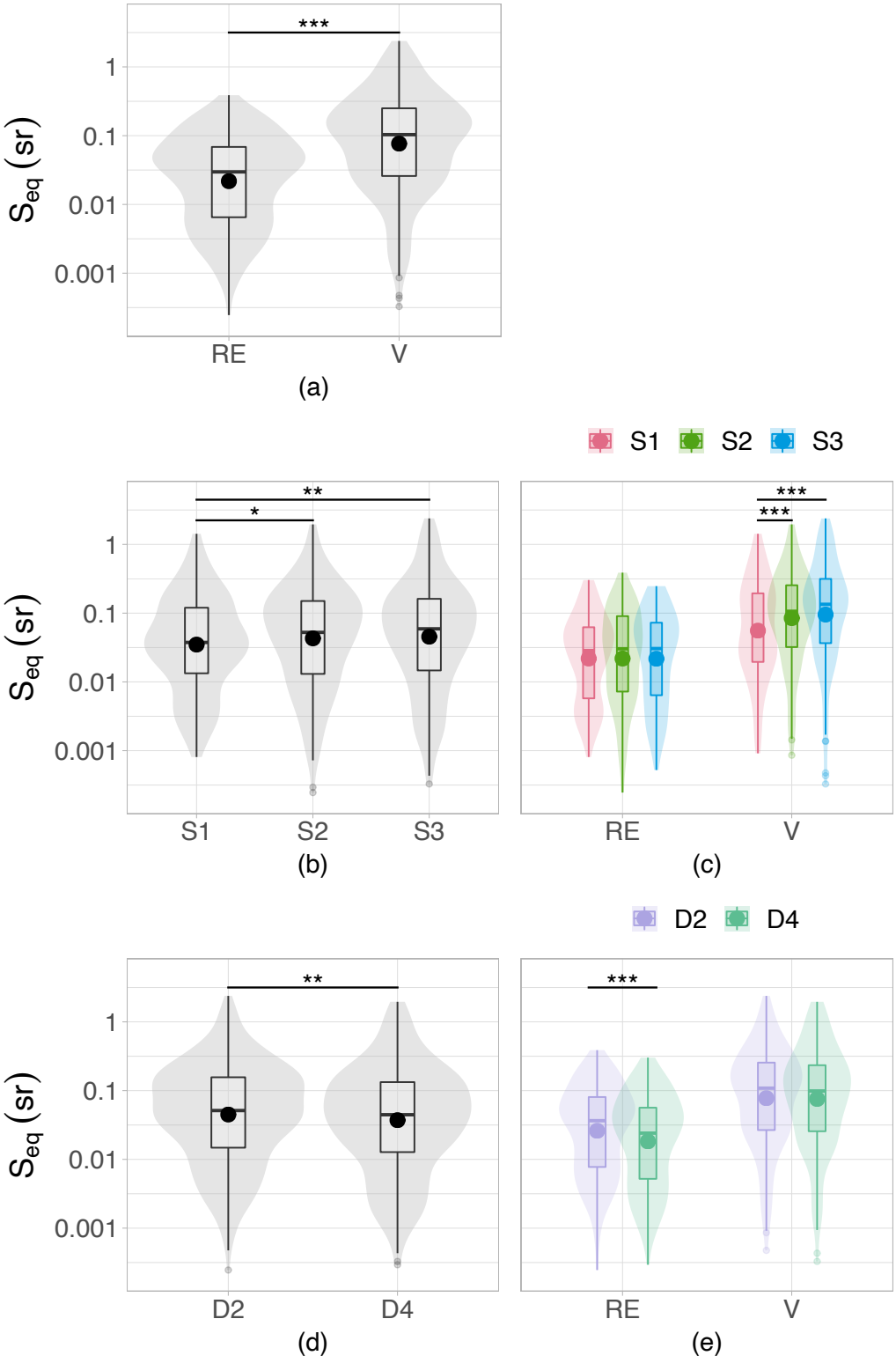


FIGURE 2.11 – Taille apparente reportée en fonction de (a) la condition d'écoute $COND$ (RE : écoute réelle, V : auralisation HOA4), (b) la salle $SALLE$ (S1, S2, S3), (c) l'interaction $COND \times SALLE$ entre la condition d'écoute en abscisse et la salle en couleur, (d) la distance de la source $DIST$ (D2 : source à 2 m, D4 : source à 4 m), (e) l'interaction $COND \times DIST$ entre la condition d'écoute en abscisse et la distance de la source en couleur.

($p < 0.0001$). Plusieurs éléments pourraient permettre d'expliquer cette différence. D'abord, le niveau de la source aux oreilles de l'auditeur diminue lorsque la distance source auditeur augmente. Toutefois, [Vliegen et Van Opstal, 2004] semble indiquer que le niveau de la source n'a pas d'influence sur les performances de localisation en azimut. En revanche, l'étude étant menée dans des salles non traitées et en considérant que le niveau de la réponse de la salle à l'excitation de la source est peu sensible à la position de la source, le rapport d'énergie du champ direct sur l'énergie réverbérée par la salle, appelé Direct-to-Reverberant Ratio (DRR), diminue également lorsque la source s'éloigne. Il est plausible que les performances de localisation en azimut soient impactées par le DRR, à savoir, plus cette grandeur est faible, plus la source est difficile à localiser et donc plus l'erreur de localisation est grande.

Pour finir, on observe également une influence de la distance de la source sur la taille apparente reportée par les participants. En effet, la taille apparente reportée pour les sources à 4 m (0.018 sr) est significativement plus faible que celle reportée pour les sources à 2 m (0.026 sr). Cette différence semble indiquer que les participants ont perçu et reporté la taille effective de la source. Celle-ci peut être calculée en multipliant la taille apparente moyenne reportée $\overline{S_{eq}}$ par la distance moyenne reportée $\overline{R_p}$, pour chacune des deux distances à l'étude : $\overline{S_{eq}}(D2) \times \overline{R_p}(D2) = 0.106\ m^2$ et $\overline{S_{eq}}(D4) \times \overline{R_p}(D4) = 0.103\ m^2$. Elle est sensiblement identique pour les deux distances et correspond à la surface d'un disque d'environ 9 cm de rayon.

2.5.2 Dégradations de l'image spatiale des sources induites par l'auralisation

D'après les résultats présentés précédemment, il est clair que la perception spatiale des sources sonores est lourdement impactée par l'auralisation ambisonique des environnements acoustiques mesurés. On constate en moyenne une large dégradation des performances de localisation dans les cas auralisés par rapport aux cas réels.

Nature des dégradations

De manière générale, ces dégradations se manifestent par une erreur de localisation accrue en azimut et élévation et par une augmentation significative de la taille apparente des sources reportée. Plus précisément, l'erreur absolue de localisation en azimut est en moyenne de 1.9° en écoute réelle tandis qu'elle est de 5.9° en condition virtuelle, soit une augmentation moyenne de l'erreur de 4°. Les résultats obtenus en conditions auralisées sont toutefois comparables à ceux présentées dans la littérature sur la localisation de sources pour différents systèmes de captation et restitution ambisonique [Braun et Frank, 2011, Bertet *et al.*, 2013, Huisman *et al.*, 2021]. Une cause probable de la dégradation de la localisation en azimut en ambisonique est une mauvaise reproduction des indices binauraux (ILD et ITD), nécessaires à la localisation dans le plan azimutal. D'autre part, les résultats concernant l'évalua-

tion de l'élévation des sources indiquent non seulement une nette augmentation de l'erreur absolue induite par l'auralisation (1.9° d'erreur absolue moyenne dans le cas réel, contre 8.5° dans le cas virtuel), mais également que cette erreur est largement signée vers le haut. En moyenne les sources ont été perçues à une élévation de 15.7°, contre 0.7° pour les conditions réelles. Ce phénomène d'attraction vers le haut des sources lors d'une restitution ambisonique 3D semble assez commun mais n'a que peu été étudié à notre connaissance. L'élévation étant jugée à partir d'indices fréquentiels monoraux situés dans la gamme de fréquence 4-16kHz, l'aliasing spatial à hautes-fréquences, imposé par les caractéristiques du réseau de microphones (au delà de 8kHz pour l'Eigenmike) est une cause possible de la difficulté à localiser les sources dans le plan vertical, lors d'une restitution ambisonique. De plus, bien que le port d'un casque de VR semble avoir un impact limité sur les performances de localisation en azimut lors d'une restitution ambisonique 2D [Huisman *et al.*, 2021], la modification des HRTFs par le casque de VR pourrait avoir un effet bien plus dramatique sur le jugement de l'élévation des sources. Enfin, les dégradations induites par le système d'auralisation se manifestent également par une augmentation significative de la taille apparente des sources reportée par les participants. Cette observation est à mettre en regard des observations précédentes qui tendent à indiquer que la localisation des sources est plus difficile et imprécise lors d'une écoute des auralisations. En effet, les participants ayant plus de difficultés à localiser les sources, ils ont effectué des tracés plus large afin d'y inscrire les positions réelles des sources. L'interprétation de ces résultats, en regard de la méthode de report est plus amplement discutée par la suite (*c.f.* section 2.5.3).

Des dégradations différentes en fonction de la distance de la source

En analysant les interactions entre la condition d'écoute et la distance des sources, on remarque que la plupart des différences perçues pour les deux positions de sources en écoute réelle ne sont pas observées dans les cas auralisés. C'est notamment le cas de l'augmentation avec la distance de l'erreur absolue de localisation en azimut et de la diminution de la taille apparente reportée des sources. La difficulté en condition virtuelle à localiser précisément les sources quelle que soit leur distance semble prévaloir et conduit à 1) une large augmentation de la valeur moyenne de ces deux grandeurs 2) une augmentation de la variabilité dans le report de ces grandeurs. Ainsi, les différences subtiles reportées dans la perception des phénomènes réels semblent masquées par les dégradations induites par l'auralisation. En revanche, l'impression globale de distance de la source semble assez bien reproduite par le système. En effet, une augmentation de la distance perçue, lorsque la distance réelle de la source augmente, est observée dans l'une et l'autre des conditions d'écoute. Toutefois, alors que le report de la distance perçue pour les sources à 4 m est sensiblement identique dans les deux conditions d'écoute, l'analyse révèle que les sources à 2 m ont été perçues significativement plus loin dans les cas auralisés (4.52m en moyenne) que dans les cas réels (4.06m en moyenne). D'après la littérature sur la perception auditive de la distance, les indices acoustiques les plus

saillants pour juger de la distance d'une source en environnement réverbérant sont d'une part le niveau sonore du champ direct et d'autre part le DRR, qui diminuent tous les deux lorsque la distance de la source augmente. On peut donc supposer que ces deux grandeurs sont relativement bien restituées par l'auralisation des sources à 4m, mais légèrement dégradées dans les cas des sources à 2 m, en faisant l'hypothèse que pour cette distance, soit le champ direct est atténué, soit le champ réverbéré est boosté lors de l'auralisation.

Des dégradations différentes en fonction de l'environnement acoustique mesuré

Alors que l'étude n'a pas révélé d'effet significatif de l'environnement acoustique sur les réponses des participants en condition réelle d'écoute, des différences significatives entre les salles ont été observées en condition virtuelle. D'abord, la précision de localisation angulaire est plus ou moins bonne en fonction des salles et les performances de localisation en termes d'erreur absolue en azimut et de taille apparente de sources sont significativement meilleures dans la salle S1 que dans les deux autres salles. Or la salle S1 se distingue des deux autres, notamment par le fait qu'elle est moins réverbérante et plus particulièrement en basses fréquences (*c.f.* Figure 2.3). L'auralisation pourrait donc entre autres être perturbée par la présence d'un champ diffus à basse fréquence. Toutefois, étant donné que l'information de position angulaire de la source est principalement portée par le champ direct de la réponse impulsionnelle, il est possible que la présence d'un fort champ diffus en BF ait un effet négatif sur la restitution de ce champ direct. La distance perçue des sources est également impactée par l'acoustique en condition virtuelle alors qu'aucune différence n'est observée en conditions réelles. En effet, lors de l'auralisation, la distance moyenne reportée par les participants pour les sources de la salle S3 est significativement supérieure à celles reportées dans les deux autres. Une légère différence est également observée entre la salle S1 et S2, à savoir que les sources de la salle S2 ont été en moyenne perçues légèrement plus proches que celles de la salle S1. Et bien que la distance moyenne sur l'ensemble des conditions réelles d'écoute ne soit pas différente de celle reportée sur l'ensemble des conditions auralisées, l'auralisation a un effet différent sur la distance perçue en fonction de l'acoustique de l'environnement auralisé. En effet, la distance perçue dans la salle S1 est inchangée tandis qu'elle est légèrement plus faible dans le cas auralisé pour la salle S2 et largement plus élevée pour la salle S3. L'absence de différence en moyenne semble indiquer que les indices permettant le jugement de la distance (notamment le DRR) sont en moyenne bien reproduits par l'auralisation, toutefois, les différences observées entre les salles suggèrent que ces indices ne sont pas également reproduits dans toutes les situations. L'hypothèse que le DRR est d'avantage perturbé pour des environnements ayant un champ diffus important permettrait d'expliquer la sur-estimation de la distance dans le cas auralisé pour la salle S3 mais ne permet pas d'expliquer les observations faites pour la salle S2.

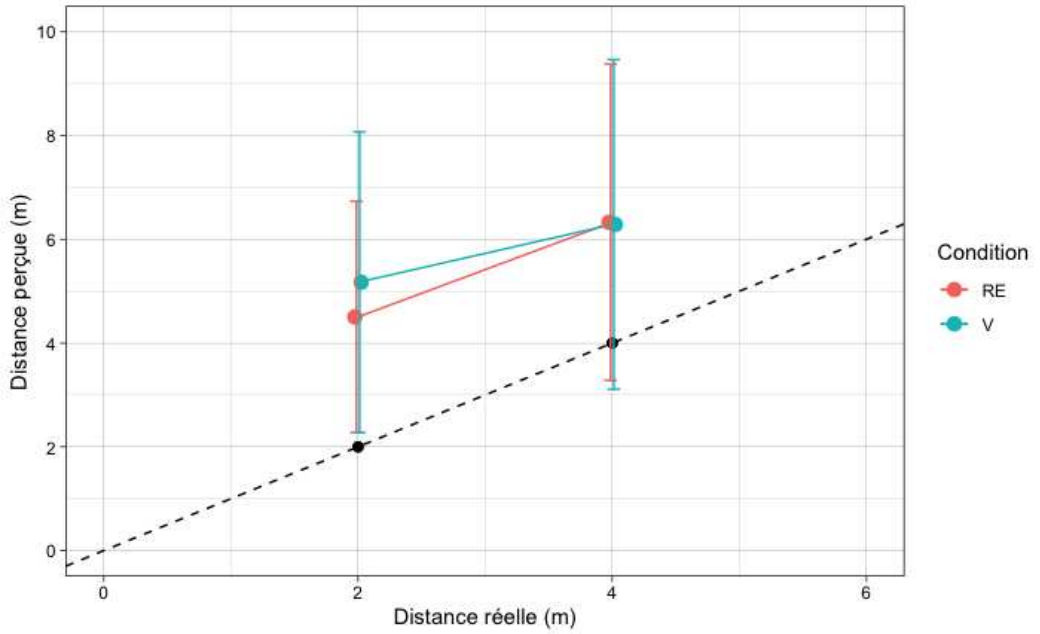


FIGURE 2.12 – Distances moyennes reportées pour les distances réelles de 2 et 4m, en condition d'écoute réelle (RE) et virtuelle (V). Les moustaches représentent l'intervalle de confiance à 95%. La droite pointillée noire représente la distance réelle des sources. Tous les points au-dessus de cette droite correspondent à une sur-estimation de la distance. Tous les points en-dessous correspondent à une sous-estimation de la distance.

2.5.3 Retours sur la méthode de report

La méthode de report employée lors de cette expérience a l'avantage de caractériser la perception spatiale des sources sonores selon différents attributs. Toutefois, cette méthode peut être source d'un certain nombre de biais et de confusions dans le jugement de ces attributs.

Biais de la méthode sur l'évaluation de la distance

En comparant les distances reportées avec les distances réelles dans les différentes conditions d'écoute (*c.f.* figure 2.12), on remarque une surestimation globale de la distance par les participants, quelle que soit la condition d'écoute. Ces résultats sont en désaccord avec ceux de la littérature sur la perception auditive de la distance, qui ont montré que la distance de sources sonores situées au-delà d'un mètre sont généralement sous-estimées et que plus la distance augmente plus la sous-estimation est élevée [Mershon et Bowers, 1979, Zahorik *et al.*, 2005, Kolarik *et al.*, 2016].

La méthode de report semble donc être à l'origine de cette sur-estimation. La cause la plus probable de ce biais de report est une mauvaise estimation visuelle de la distance dans l'interface VR. En effet, plusieurs études comparant

la perception visuelle de la distance en conditions réelles et virtuelles ont révélé une sous-estimation dans le jugement visuel de la distance, en réalité virtuelle [Loomis et Knapp, 2003, Armbrüster *et al.*, 2008, Kelly *et al.*, 2018]. Les participants auraient donc sous-estimé visuellement la distance du pointeur permettant le report de la distance perçue, conduisant à une sur-estimation de la distance reportée. La compréhension de ce phénomène de sous-estimation de la distance en VR est encore aujourd'hui un sujet de recherche important dans le domaine de la réalité virtuelle. En croisant les résultats de pas moins de 40 études sur le sujet, [Feldstein *et al.*, 2020] ont montré une nette amélioration des technologies du virtuel et une minimisation de ce biais au cours des 10 dernières années. Ceci étant dit, parmi les facteurs pouvant influencer ce biais, il a notamment été montré que le jugement de la distance est meilleur pour des scènes virtuelles représentant des environnements clos, par rapport à des environnements ouverts [Armbrüster *et al.*, 2008]. Le réalisme de l'environnement virtuel semble également avoir une influence sur l'évaluation de la distance [Feldstein *et al.*, 2020], en particulier le réalisme du sol ou encore la présence d'un avatar représentant le corps de l'utilisateur. La présence d'éléments représentant la perspective (e.g. des lignes de fuites) améliore également le jugement de la distance. Ces études donnent donc des pistes d'amélioration de l'environnement virtuel afin de permettre un meilleur report de la distance perçue.

Confusions sur le report de la taille apparente

Le report de la taille apparente est effectué en entourant les sources sonores. Pour rappel, la consigne concernant la tâche à réaliser était : "Entourer le plus précisément possible, à l'aide du pointeur, la zone dans laquelle la source est entendue". Toutefois, d'après les résultats précédents, on peut s'interroger sur le sens que porte ce report et cette consigne. En effet, dans les cas réels, il semblerait que le report de la taille de la source soit réellement lié à la taille perçue puisque les différences de tailles apparentes observées entre les deux distances de source conduisent à une même taille effective (*c.f.* section 2.5.1). En revanche, en condition d'auralisation, les participants ont reporté une taille de source plus large que dans les cas réels, mais cette fois-ci, la forte corrélation entre taille reportée et erreur de localisation en azimut semble indiquer que ce report soit plutôt le reflet d'un flou dans la localisation des sources. Ainsi, les différences de taille observées entre les salles dans la condition virtuelle (*c.f.* figure 2.11) sont davantage représentatives de la quantité de flou induite par l'auralisation de chacune des salles. Ce report permet donc dans un cas de mettre en évidence des différences de taille perçue des sources et dans l'autre de mesurer un niveau d'imprécision dans la localisation des sources. Bien que dans ces deux cas présents il soit possible de conclure sur le sens de ce report, il est néanmoins important de garder à l'esprit qu'une confusion dans l'interprétation de cette grandeur est possible dans de nouvelles expériences ayant recours à cette méthode de report.

2.5.4 Conclusions et perspectives

Dans cette étude, les performances de localisation dans différents environnements acoustiques ont été évaluées en condition réelle d'écoute et en condition d'auralisation via un système d'auralisation HOA 3D d'ordre 4 d'acoustiques mesurées. Une méthode de report de l'image spatiale des sources en réalité virtuelle a été utilisée et a permis de décrire la perception des sources en termes de position angulaire, de distance et de taille apparente. L'étude révèle des performances de localisation dégradées dans les conditions auralisées, notamment en termes de précision angulaire des sources. Ces dégradations sont variables et semblent plus prononcées dans le cas d'acoustiques réverbérantes. Il a également été observé que le report de la taille apparente est tantôt représentatif d'une taille perçue de source (en condition réelle) et tantôt relatif à une quantité de flou dans la localisation (en condition d'auralisation). Dans ce dernier cas, l'acoustique a un effet sur cette sensation de flou et on observe une augmentation du flou de localisation dans les cas les plus réverbérants. Enfin, bien que d'un point de vue global la distance semble être correctement restituée par le dispositif, la présence d'interactions révèle qu'en condition d'auralisation, l'acoustique et la distance réelle de la source peuvent avoir une influence sur la distance perçue. De plus, une surestimation systématique de la distance perçue a été observée. Ce résultat diffère de ceux présentés dans la littérature sur la perception de la distance de sources sonores, ce qui révèle la présence d'un biais dans la méthode de report de la distance, probablement induit par une mauvaise évaluation visuelle de la distance en réalité virtuelle. A partir de ces résultats, des perspectives peuvent être envisagées :

1) La nature des dégradations de l'image spatiale semble dépendre des propriétés acoustiques de l'environnement mesuré et auralisé, toutefois il est difficile de conclure précisément sur l'origine physique de ces dégradations. Pour préciser les hypothèses formulées précédemment, une analyse objective de la mesure paraît indispensable. La figure 2.13 représente la méthodologie de la présente étude ainsi que les éléments méthodologiques envisagés pour la caractérisation objective des dégradations induites par l'auralisation. En effet, en comparant les mesures réalisées in-situ et une re-mesure en condition d'auralisation, sur une base de descripteurs objectifs, choisis en connaissance des dégradations perçues, des pistes concrètes d'amélioration du dispositif pourraient être envisagées. A titre d'exemple, la dégradation des performances de localisation angulaire, pourrait être traitée à travers l'étude de la qualité de restitution des indices de localisation ILD, ITD et IACC ainsi que par la visualisation de la répartition spatiale de l'énergie au cours du temps, grâce à des méthodes de formation de voies (beamforming).

2) La méthodologie de cette expérience a permis l'évaluation perceptive de la qualité de restitution spatiale d'un système d'auralisation ambisonique 3D d'ordre 4 et peut tout à fait être déployée pour caractériser d'autres systèmes et méthodes d'auralisation telles que les méthodes hybrides HO-SIRR [McCormack *et al.*, 2020] ou SDM [Tervo *et al.*, 2013], censées améliorer la précision spatiale des composantes directives du signal, ou encore des méthodes d'auralisation binaurale avec head-

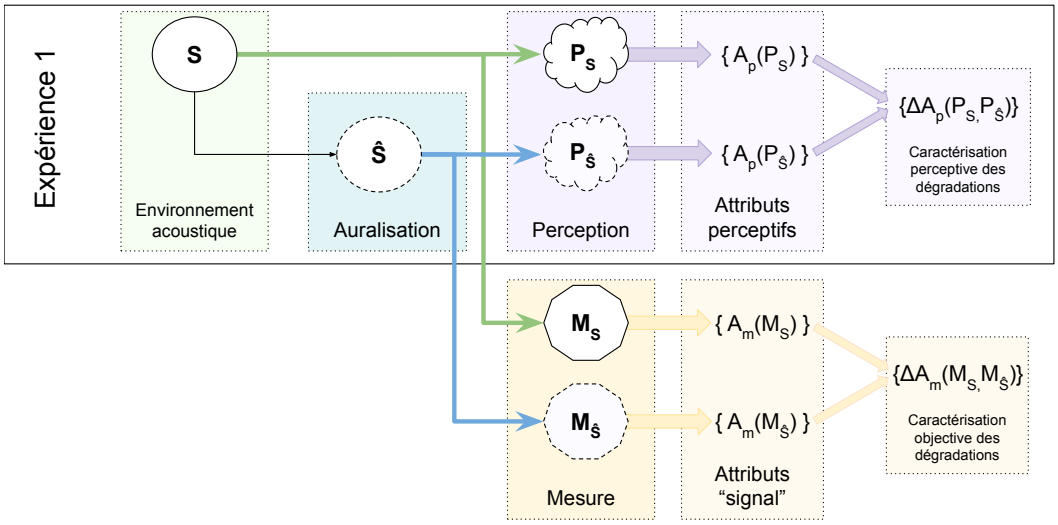


FIGURE 2.13 – Rappel de la méthodologie de l'expérience (*c.f.* section 2.1) et proposition d'une méthodologie pour la caractérisation objective des dégradations de l'image spatiale, induite par l'auralisation. M_S et $M_{\hat{S}}$ représentent respectivement la mesure in-situ de l'environnement acoustique et la mesure de l'auralisation. A_m est un jeu de descripteurs acoustiques calculés à partir des mesures. ΔA_m représente la mise en relation des descripteurs calculés dans le cas réel et dans le cas auralisé, permettant la caractérisation objective des dégradations spatiales induites par l'auralisation.

tracking [Stitt *et al.*, 2016, Romanov *et al.*, 2017].

3) Le caractère portable et léger du dispositif VR de report a permis de réaliser des expériences en conditions réelles d'écoute et d'obtenir des informations précises sur la perception réelle dans trois environnements acoustiques et pour deux distances source-auditeur. De nouvelles études in-situ peuvent être conduites avec ce dispositif, permettant ainsi de diversifier les conditions acoustiques et de tendre vers une caractérisation plus générale de la perception spatiale réelle des sources en environnements réverbérants. Bien que dans les configurations étudiées ici, l'acoustique et la position de la source ne semblent pas avoir d'influence sur la perception de la position angulaire de la source, on peut par exemple faire l'hypothèse que pour des acoustiques plus réverbérantes et pour des distances de sources plus importantes, un masquage du champ direct par le champ réverbérant, vienne perturber notre capacité à localiser précisément la source.

4) En ayant conscience des limites du système d'auralisation d'acoustiques mesurées et des biais induits par la méthode de report, il n'en reste pas moins que le dispositif et la méthodologie employés ici permettent bel et bien de mettre en évidence des différences de perception spatiale de sources dans divers environnements acoustiques auralisés. Par conséquent, les outils proposés dans cette étude peuvent être employés pour caractériser la perception spatiale dans un ensemble plus vaste

d'acoustiques mesurées. Dans le prochain chapitre nous verrons comment ce protocole a permis d'effectuer une caractérisation perceptive d'un corpus d'une vingtaine d'environnements acoustiques mesurés et auralisés.

			ε_θ (°)	$\lvert\varepsilon_\theta\rvert$ (°)	ε_ϕ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
RE	S1	D2	1.56 (0.72 :2.39)	1.38 (1.00 :1.90)	-0.15 (-1.21 :0.91)	2.04 (1.56 :2.68)	3.86 (3.46 :4.30)	2.73 (1.94 :3.84)
		D4	-1.42 (-2.61 :-0.24)	2.50 (1.92 :3.27)	0.10 (-0.89 :1.09)	1.60 (1.18 :2.17)	5.49 (4.87 :6.20)	1.75 (1.21 :2.52)
	S2	D2	-1.56 (-2.26 :-0.86)	1.41 (0.99 :2.00)	2.47 (0.81 :4.12)	2.01 (1.37 :2.96)	4.03 (3.60 :4.50)	2.60 (1.73 :3.90)
		D4	3.64 (2.82 :4.45)	2.94 (2.34 :3.69)	1.34 (0.13 :2.55)	1.93 (1.42 :2.63)	5.95 (5.33 :6.65)	1.84 (1.24 :2.72)
	S3	D2	2.71 (2.03 :3.40)	1.63 (1.10 :2.40)	0.16 (-1.09 :1.42)	2.17 (1.71 :2.75)	4.32 (3.88 :4.82)	2.48 (1.71 :3.60)
		D4	1.06 (0.21 :1.91)	1.99 (1.55 :2.56)	0.02 (-1.10 :1.13)	1.78 (1.38 :2.29)	5.70 (5.13 :6.32)	1.90 (1.26 :2.85)
V	S1	D2	3.21 (0.83 :5.60)	3.93 (2.92 :5.28)	15.82 (10.79 :20.85)	8.72 (6.08 :12.52)	4.37 (3.86 :4.95)	6.04 (3.94 :9.27)
		D4	1.62 (0.09 :3.14)	3.43 (2.70 :4.36)	13.53 (9.15 :17.92)	7.07 (5.07 :9.85)	5.46 (4.87 :6.12)	5.12 (3.43 :7.64)
	S2	D4	-3.67 (-7.02 :-0.32)	6.61 (4.94 :8.86)	17.02 (11.37 :22.66)	9.81 (6.78 :14.18)	3.89 (3.45 :4.39)	7.98 (5.37 :11.84)
		D4	10.94 (7.89 :14.00)	8.22 (5.71 :11.84)	18.68 (13.61 :23.76)	9.64 (6.51 :14.27)	4.91 (4.35 :5.55)	8.99 (6.03 :13.41)
	S3	D2	9.48 (6.87 :12.10)	8.59 (6.95 :10.62)	15.88 (10.77 :20.99)	9.02 (6.51 :12.51)	5.42 (4.78 :6.15)	9.68 (6.34 :14.77)
		D4	-5.12 (-8.96 :-1.29)	7.11 (4.99 :10.12)	13.10 (8.04 :18.17)	7.19 (4.99 :10.35)	6.40 (5.61 :7.30)	9.38 (5.79 :15.19)

TABLE 2.2 – Résultats bruts de l'expérience. Moyennes et intervalles de confiance à 95% entre parenthèses, sur les 21 sujets et les 3 stimuli sonores, pour les 6 variables dépendantes analysées : ε_θ et $|\varepsilon_\theta|$: erreurs relatives et absolues de localisation en azimut, ε_ϕ et $|\varepsilon_\phi|$: erreurs relatives et absolues de localisation en élévation, R_p : distance perçue, S_{eq} : taille apparente de la source. Conditions d'écoutes : RE = écoute réelle, V = écoute virtuelle (auralisation HOA4). Environnements acoustiques : S1 = Salle 1, S2 = Salle 2, S3 = Salle 3 (*c.f.* sec. 2.3.1). Distance source-auditeur : D2 = source à 2 m, D4 = source à 4 m.

Chapitre 3

Caractérisations objective et subjective d'un corpus d'environnements acoustiques

Sommaire

Les travaux présentés dans ce chapitre sont motivés par deux problématiques. La première, relative au sujet principal de cette thèse, est celle de l'étude de la perception en environnements acoustiques 3D. Dans le précédent chapitre, nous avons pu évaluer la qualité spatiale d'un système d'auralisation HOA d'acoustiques mesurées et avons conclu que malgré un certain nombre de dégradations induites par le système, il était possible de mettre en évidence des différences perceptives entre les trois environnements acoustiques de l'étude. Cette méthodologie peut alors être employée pour investiguer plus largement la perception spatiale des sources en contexte d'auralisation, sur un plus large ensemble d'acoustiques mesurées. La deuxième problématique, formulée dans le cadre du projet ANR Sésames, est celle de la caractérisation multi-dimensionnelle d'un corpus d'édifices du patrimoine, et

FIGURE 3.1 – Présentation du corpus Sésames : situation géographique et enveloppe intérieure des 15 édifices du corpus. Image issue du site internet du projet : http://anr-sesames.map.cnrs.fr.

plus spécifiquement ici, de caractérisations acoustique et perceptive. Ces deux problématiques peuvent être vues comme miroirs l'une de l'autre puisque la première vise à une meilleure compréhension de la perception en environnements réverbérants, à travers la diversification des conditions acoustiques tandis que la deuxième aspire à intégrer des dimensions acoustiques et perceptives dans un processus de caractérisation de corpus d'environnements acoustiques. Dans un premier temps, nous verrons comment le projet ANR Sésames adresse ces deux problématiques, à travers l'élaboration et le déploiement d'un protocole original de mesures dans 15 petites et moyennes chapelles rurales de la région PACA, par la suite étendu à d'autres collections. Dans un second temps, une caractérisation objective des corpus, basée sur le calcul et l'analyse de descripteurs acoustiques, sera présentée. Enfin, une caractérisation perceptive de ce même corpus sera proposée, en s'appuyant sur la méthodologie expérimentale proposée dans le chapitre précédent.

3.1 Le protocole Sésames

Alors que les spécialistes du patrimoine sont habitués à caractériser de larges collections d'artefacts patrimoniaux, à travers la collecte et l'étude de données architecturales, les dimensions sonores et perceptives sont souvent laissées de côté. Du point de vue des acousticiens, l'acquisition groupée de données architecturales, visuelles et sonores dans un ensemble varié d'environnements acoustiques, représente une réelle opportunité de constituer une base de données exploitable pour l'étude de

la perception des salles non seulement sonore mais également audio-visuelle. Le projet ANR Sésames, lancé en 2018, est motivé par ces deux observations. Il prend pour objet d'étude une quinzaine de petites et moyennes chapelles rurales de la région PACA (représentées en Figure 3.1), sélectionnées pour leur cohérence en termes d'usage initial et pour leur diversité en termes de caractéristiques architecturales et de situation géographique. Avec une volonté de croiser les regards et les disciplines autour d'un corpus d'édifices souvent ignorés, des architectes et spécialistes du patrimoine, du laboratoire MAP (Modèles et simulations pour l'Architecture et le Patrimoine) se sont associés à des chercheurs en acoustique et perception du laboratoire PRISM, dans le but de proposer des méthodes collaboratives d'acquisition et d'analyse de données architecturales, acoustiques et perceptives, pour la caractérisation multi-dimensionnelle de ce corpus. Le projet est articulé en deux grandes étapes de travail i.e. :

1. L'élaboration d'un protocole d'acquisition de données multi-dimensionnelles (architecturales et acoustiques) et son déploiement dans les 15 chapelles du corpus. Dans les prochains paragraphes, une description générale de ces travaux est présentée, en mettant l'accent sur la mesure acoustique [1].
2. L'analyse des données récoltées et la caractérisation multi-échelle du corpus. Dans ce chapitre, seules les caractérisations acoustiques et perceptives seront présentées.

3.1.1 Protocole d'acquisition groupée de données architecturales et acoustiques

La caractérisation du corpus s'appuie sur l'acquisition dans les différents lieux de deux jeux principaux de données :

- des nuages de points de l'intérieur des lieux, obtenus par relevé photogrammétrique, pour une caractérisation spatiale des lieux,
- des réponses impulsionnelles spatiales, en plusieurs points, pour une caractérisation acoustique et perceptive des lieux.

Plusieurs séances d'échanges entre architectes et acousticiens, en amont de la campagne de mesure, ont permis d'identifier un certains nombre de points devant garantir le bon déroulement de l'acquisition dans les différents édifices du corpus :

- Malgré la diversité du corpus en termes d'architecture, les données récoltées dans les différents édifices doivent être comparables sur les plans acoustique et perceptif. Pour cette raison, les mesures acoustiques sont réalisées selon une grille spatiale fixe, reproduite à l'identique dans tous les édifices du corpus.
- Le protocole doit être respectueux de l'usage initial ou présent de ce type d'édifice. Le choix des positions de mesures acoustiques (positions des sources et des microphones) doit en tenir compte.

1. Ces travaux sont également détaillés dans un article de 2021 [Blaise *et al.*, 2021], avec notamment des informations complémentaires à propos des relevés réalisés par les architectes.

- Les mesures acoustiques doivent permettre d'une part une écoute spatialisée sur réseau de haut-parleurs ou au casque, afin de conduire des expériences perceptives en laboratoire et d'autre part, une analyse acoustique et donc le calcul d'un certain nombre de grandeurs acoustiques (détaillées en section 3.2 du présent chapitre).
- Afin de pouvoir conduire des expériences perceptives en condition d'immersion multi-sensorielle (audio-visuelle), l'acquisition de photos panoramiques aux positions de mesures acoustiques doit être conduite.
- Une précision relativement importante doit être garantie pour la mesure des grandes dimensions des édifices et des positions des instruments utilisés pour les mesures acoustiques. Une plus grande tolérance est permise concernant la précision des nuages de points.
- Une durée complète de l'intervention de 3h maximum par lieu, installation, acquisitions, vérifications et rangement compris, permettant la caractérisation de deux édifices dans une même journée.
- Une instrumentation relativement légère et compacte, certains lieux n'étant accessibles qu'à pieds.
- Une autonomie en termes d'alimentation en énergie, certains lieux ne disposant pas d'électricité.

A partir de cette série de spécifications, une phase de développement, calibration, vérification et validation du protocole d'acquisition a pu être réalisée, en amont de la campagne de mesure dans les 15 chapelles du corpus. Pour ce faire, une salle renommée pour l'occasion "fake-chapel", située sur le campus Joseph Aiguier, a été choisie pour sa facilité d'accès et pour ses dimensions, comparables à celles des édifices du corpus. Plusieurs sessions de travail collaboratif autour du processus d'acquisition ont pu être menées dans ce lieu et ont permises d'aboutir au protocole, défini par une grille spatiale (représentée en figure 3.2) et par une succession chronologique de tâches :

1. Positionnement de deux niveaux laser Huepar 3D Cross Line Self-Leveling, de façon à segmenter l'espace selon quatre plans de référence, comme illustré en figure 3.2.
2. Positionnement des supports pour les instruments de mesure acoustique (microphones et haut-parleurs) selon la grille spatiale représentée en figure 3.2, en s'appuyant sur le maillage de référence définie par les niveaux laser. Davantage de détails sur la grille de mesure acoustique sont donnés par la suite, en section 3.1.2.
3. Positionnement d'un télémètre laser Leica DISTO S910 pour un relevé métrique précis de différents éléments. Le télémètre doit être placé de manière à ce qu'il puisse pointer et mesurer chaque point d'intersection de la grille de référence ainsi que chaque instrument de mesure acoustique.
4. Relevé au télémètre laser de la position exacte des instruments de mesure acoustique.

5. Relevé au télémètre laser des positions marquées par les intersections de la grille de référence (nc, nh, nb, ng, nd, cc, ch, cb, cg, cd). Cette mesure est utilisée pour la mise à l'échelle du nuage de points.
6. Relevé au télémètre laser de la position exacte de plusieurs points jugés comme signifiants par les architectes (une clé de voûte, l'entrée de la chapelle, une corniche, etc.).
7. Relevé photogrammétrique, réalisé avec une caméra panoramique "low-cost" YI 360 VR. Davantage d'informations sur le protocole de relevé photogrammétrique et les étapes de traitements permettant la génération de nuages de points peuvent être trouvées dans l'article de référence [Blaise *et al.*, 2021].
8. Relevé de photos panoramiques aux positions des microphones MA, MC, MD et MG, avec la YI 360 VR.
9. Mesures acoustiques, selon le protocole présenté en section 3.1.2.

Les étapes 1, 2, 4, 5, sont reproduites invariablement d'un édifice à l'autre. Les étapes 3, 6, 7 nécessite une adaptation aux conditions trouvées sur place.

3.1.2 Quelques précisions sur la mesure acoustique

A partir des contraintes spécifiques à l'acquisition acoustique formulées précédemment, un certain nombre de choix concernant les équipements, la méthode et la grille de mesure ont été faits et sont discutés ici. La mesure acoustique réalisée dans les 15 chapelles du corpus est basée sur la méthode de mesure de sinus glissant proposée par [Farina, 2000]. Pour rappel, la méthode s'appuie sur la mesure de sinus glissants d'une durée de 10 s, balayant exponentiellement toutes les fréquences de la gamme audible (20 Hz - 20 kHz). Cette méthode présente un certain nombre d'avantages par rapport à d'autres méthodes, incluant un bon rapport signal à bruit (SNR) et une robustesse aux non linéarités de la source. En revanche, cette méthode est sensible à la présence de bruits impulsifs durant la mesure. Pour cette raison, la mesure est répétée trois à cinq fois, en fonction de l'exposition du lieu au bruit. La réponse la moins bruitée est conservée.

Les réponses impulsionnelles spatiales sont mesurées à l'aide d'un réseau sphérique de microphones Eigenmike 32 de mh acoustics, permettant la conversion des réponses au format HOA à l'ordre 4 pour une restitution sur réseau de haut-parleurs. Les sources employées ici sont des Genelec 8020A. Ce haut-parleur présente l'avantage d'être compact et facile à transporter, tout en ayant une réponse en fréquence acceptable ($\pm 2.5dB$, 66 Hz - 20 kHz) pour une finalité d'auralisation. Il n'est en revanche pas omni-directionnel (comme la norme ISO 3382-1 sur la mesures de paramètres acoustique le préconise). Toutefois, ce point n'est pas considéré comme critique pour l'auralisation, dans la mesure où les sources que l'on souhaite auraliser (voix, instruments de musique, etc.) ont de toute façon leurs propres propriétés de directivité, souvent bien différentes de l'omnidirectionnalité. De plus, en admettant que les paramètres acoustiques mesurés avec les 8020A en guise de sources puissent être biaisés, ils permettent tout de même d'effectuer une

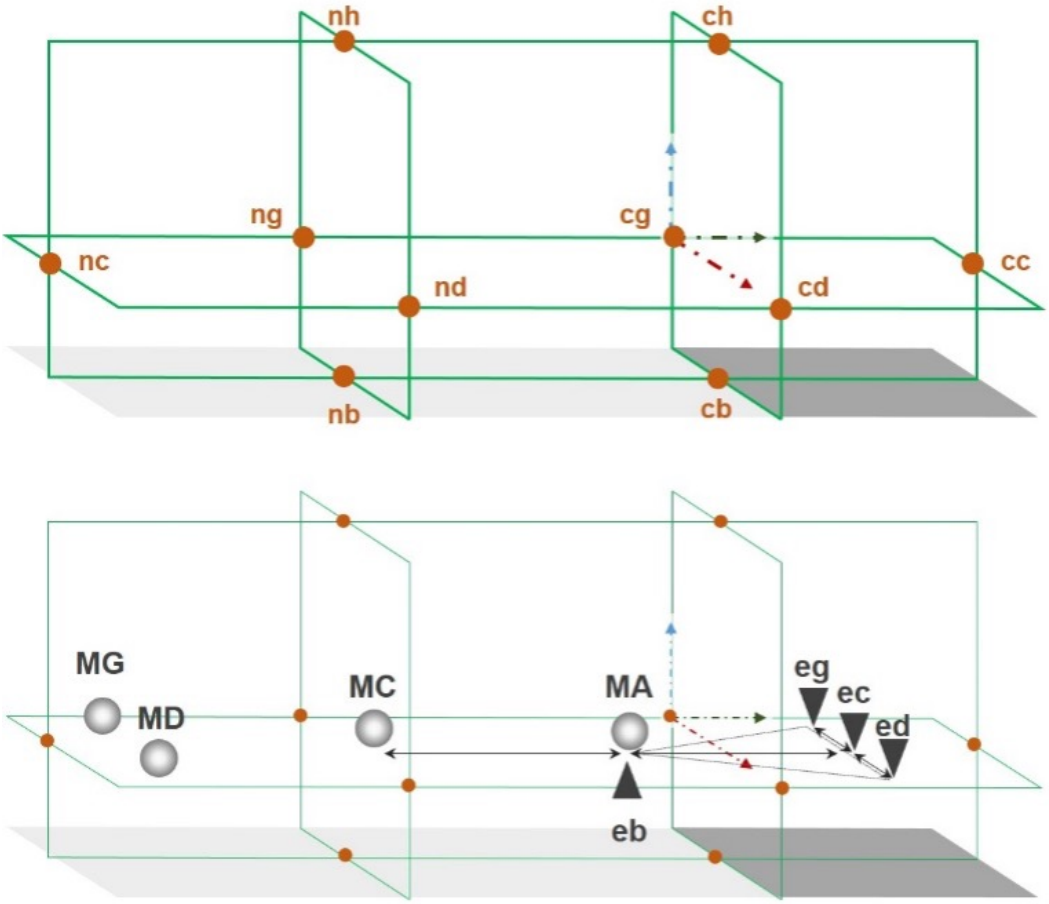


FIGURE 3.2 – Grille spatiale pour l'acquisition de données métriques et acoustiques. Le rectangle gris clair représente la nef de la chapelle, le rectangle gris foncé représente le chœur.
En haut, système spatial de référence : les faisceaux laser (lignes vertes) et leurs intersections (cercles bruns) visibles sur les surfaces du bâtiment et relevés à l'aide du télémètre.
En bas, grille de mesure acoustique : les triangles sombres (notés *eg*, *ec*, *ed*, *eb*), représentent les haut-parleurs. Les points gris clairs (notés MA, MC, MG, MD) représentent les quatre positions de microphones. Trois haut-parleurs *eg*, *ec*, *ed* sont placés dans le chœur, en configuration "concert". Un quatrième haut-parleur, *eb*, est placé juste en dessous du microphone MA. En MA et MC, microphones sphériques em32, alignés dans le plan longitudinal de la chapelle. En MG et MD, microphones omnidirectionnels. Le système formé par tous les points de mesure, à l'exception de MG et MD est fixe et invariant dans les chapelles. Les microphones MG et MD sont placés au tout début de la nef, à une distance fixe (un mètre) des murs, leur position (relativement au reste de la grille) dépend des dimensions de la chapelle.

comparaison objective des différents édifices à partir du moment où les mesures ont été effectuées avec un dispositif commun. En revanche, ces mesures ne seront pas directement comparables à d'autres mesures extérieures au corpus d'étude.

Le choix d'une grille fixe (représentée en figure 3.2), reproduite à l'identique dans toutes les chapelles du corpus a été privilégié par rapport à celui d'une grille proportionnelle aux dimensions des différents édifices. Il a été motivé par la volonté de pouvoir comparer la perception des acoustiques, indépendamment de la position d'écoute de l'auditeur. Deux niveaux laser Huepar 3D Cross Line Self-Leveling servent de base pour établir la grille de mesures acoustiques. L'alignement longitudinal et vertical des instruments de mesures acoustiques est assuré grâce au maillage défini par les faisceaux laser. Trois haut-parleurs Genelec 8020A, sont placés en ligne dans le chœur, face à la nef, en configuration "concert", pour correspondre à un usage vraisemblable des lieux. Le haut-parleur central (noté *ec* sur la figure 3.2) est aligné sur le plan vertical longitudinal. Les haut-parleurs *eg* et *ed* sont placés à une distance de 1.25m, respectivement à gauche et à droite de l'enceinte centrale. Une mesure de réponses impulsionnelles spatiales est réalisée aux points de mesures MA et MC avec le réseau de microphone em32. Ces deux positions de microphones sont également alignées dans le plan vertical longitudinal tracé par les niveaux laser. MA et MC sont donc situés dans la nef, dans l'axe du haut-parleur central *ec*, à des distances respectives de 2.17 m et 5.50 m. La distance du microphone MC a été déterminée par la distance maximum qui puisse être obtenue dans la plus petite des chapelles du corpus. La distance de MA a été choisie pour observer un triangle équilatéral (de 2.5 m de coté) entre le microphone MA, et les enceintes *eg* et *ed*.

Pour compléter la mesure, un quatrième haut parleur, noté *eb*, est placé 40 cm en dessous de l'em32 et orienté vers le plafond de la chapelle, pour les deux positions de mesures MA et MC. Cette mesure a pour but de proposer une situation d'auralisation où la source et le récepteur sont quasi-coïncidents, proche de celle d'un instrumentiste s'écoutant jouer de son instrument. L'auralisation de cette condition de mesure peut être réalisée afin de proposer aux auditeurs une situation d'interaction en temps-réel avec l'environnement mesuré. Une expérience visant à quantifier l'influence de l'acoustique sur l'implication corporelle des musiciens a notamment été envisagée.

Enfin, bien que l'essentiel des mesures soit réalisé selon une grille fixe, une mesure représentative de la diversité de dimensions des édifices à l'étude est réalisée en plaçant deux microphones omnidirectionnels, marqués MG et MD sur la figure 3.2, de part et d'autre de l'entrée de la nef, à une distance de 1 m par rapport aux murs latéraux et au mur du fond de la nef.

3.1.3 Données acoustiques et traitements

Les mesures acoustiques selon le protocole présenté en section 3.1.2 ont été réalisées dans les 15 petites et moyennes chapelles rurales du corpus Sésames. Ce protocole a également été déployé dans 3 salles du Palais des Papes d'Avignon : la Grande Chapelle, le Tribunal et la Chapelle St-Martial. La Chapelle St-Martial est une petite chapelle comparable à celles du corpus Sésames en termes de dimensions, tandis que la Grande Chapelle et le Tribunal sont de très grands espaces. Le protocole, pensé pour caractériser des petits et moyens volumes est donc moins adapté pour ces deux derniers, du fait de leur grande taille par rapport à celle de la grille de mesure. Toutefois, le choix de conserver cette grille fixe permet de pouvoir comparer les données mesurées dans ces lieux à celles récoltées dans le reste du corpus. Enfin, une mesure de référence basée sur la même grille spatiale a été effectuée en chambre anéchoïque afin de caractériser le dispositif de mesure en champ libre. Au total, 19 environnements acoustiques ont été caractérisés selon ce protocole. Des informations relatives à chaque lieu sont données en tableau 3.1. Pour chaque salle, 8 réponses impulsionnelles spatiales ont été mesurées : 2 positions de microphones (MA et MC) $\times$ 4 positions de source (eg, ec, ed et eb).

Afin d'homogénéiser les mesures sur l'ensemble du corpus, une étape de normalisation des réponses impulsionnelles a été réalisée. Pour chacune des 8 fonctions de transfert $RI_i(n,m)$ mesurée dans une salle i, liant une position de haut-parleur n à une position de microphone m, l'énergie RMS du champ direct de la réponse impulsionnelle omnidirectionnelle $E_{RMS}(RI_i(n,m))$ est calculée. En comparant cette énergie à celle mesurée en champ libre dans la même configuration $E_{RMS}(RI_{anech}(n,m))$, un gain de normalisation par rapport à la référence anéchoïque g_{RMS} peut être calculé et appliqué à chacune des réponses :

$$g_{RMS}(RI_i(n,m)) = \frac{E_{RMS}(RI_{anech}(n,m))}{E_{RMS}(RI_i(n,m))} \tag{3.1}$$

soit un gain en décibels g_{dB} :

$$g_{dB}(RI_i(n,m)) = 20 * log(g_{RMS}) \tag{3.2}$$

Pour rendre compte de la variabilité relative à cette étape de normalisation, un gain moyen et son écart-type ont été calculés : $g_{dB} = -0.27 \pm 0.48$ dB. On observe un gain moyen proche de 0 dB et une faible variabilité du gain < 1 dB, confirmant qu'en moyenne la reproductibilité des mesures tout au long de la campagne est bonne.

Pour chaque mesure le niveau de bruit L_{bruit} et le rapport d'énergie entre le premier pic de la RI et le niveau de bruit de fond, noté $PSNR$ (Peak Signal to Noise Ratio), sont calculés dans le but d'avoir un aperçu de la qualité de la mesure. L'estimation de ces deux paramètres est réalisée avec la librairie *ITA-toolbox* [Berzborn *et al.*, 2017], grâce à la méthode proposée par Lundeby et. al

Corpus	**Code**	**Commune / Site**	**Code Postal**	**Nom**	**Type**	**Volume (m3)**	L_{bruit} **(dBFS)**
Sesames	Bru	Brue-Auriac	83119	Prieuré Notre-Dame	Chapelle	530	-125.6
	StMaPa	St-Martiin-de-Pallières	83160	St-Etienne	Chapelle	170	-120.8
	PuyStMa	Puyloubier	13114	Ste-Marie	Eglise	650	-123.2
	PuyStPce	Puyloubier	13114	St-Pancrace	Chapelle	380	-121.0
	Esp	Esparron	83560	N.D. du Reverst	Chapelle	990	-124.1
	LaVdre	La Verdière	83560	St-Roch	Chapelle	290	-118.0
	Peyn	Peynier	13790	St-Pierre	Chapelle	350	-118.4
	Peyrl	Peyrolles-en-Provence	13860	N.D. d'Astors	Chapelle	550	-116.8
	TvNDSalt	Tourves	83170	N.D. de la Salette	Chapelle	920	-116.8
	Pvrt	Pierrevert	04860	St-Patrice	Chapelle	300	-121.6
	Prue	Pierrerue	04300	St-Pierre	Chapelle	280	-122.1
	LMSHo	Les Mées	04190	St-Honorat	Chapelle	480	-122.9
	LMSRch	Les Mées	04190	St-Roch	Chapelle	290	-119.0
	TvStPbce	Tourves	83170	St-Probace	Chapelle	590	-119.8
	Bras	Bras	83149	N.D. de Bethléem	Chapelle	280	-112.5
Avignon	AviGC	Avignon (Palais des Papes)	84000	Grande Chapelle	Chapelle	12000	-122.0
	AviStMa	Avignon (Palais des Papes)	84000	St-Martial	Chapelle	200	-125.8
	AviTri	Avignon (Palais des Papes)	84000	Tribunal de la Rota	Salle	10000	-131.1
Référence	Anécho	Marseille (PRISM)	13009	Chambre anéchoïque	Chambre anéchoïque	360	-130.0

TABLE 3.1 – Informations sur les salles des corpus à l'étude. Le niveau de bruit L_{bruit} correspond au niveau de bruit large-bande mesuré en moyenne sur les 8 mesures spatiales réalisées dans chaque salle. Il est calculé avec la méthode proposé par Lundeby et. al [Lundeby *et al.*, 1995].

[Lundeby *et al.*, 1995]. Le tableau 3.1 donne le niveau de bruit moyen pour chacune des salles des différents corpus. Le tableau 3.2 répertorie le PSNR large bande pour chaque position de mesure, moyenné sur les 19 environnements acoustiques mesurés. Comme l'illustre ce tableau, le PSNR dépend de la distance source-microphone. Pour la distance maximum ($MC-eg$ ou $MC-ed$) de la grille, le PSNR est en moyenne de 85 dB, indiquant une dynamique raisonnable même dans ces cas extrêmes.

Position em32	**Position source**	**PSNR (dB)**
MA	ec	98.3
	eg	90.7
	ed	94.1
	eb	105.1
MC	ec	89.7
	eg	86.5
	ed	85.3
	eb	106.1

TABLE 3.2 – Rapports signal à bruit PSNR large-bande pour chaque position de mesure, moyennés sur les 19 salles de l'étude.

3.2 Caractérisation acoustique du corpus

La caractérisation acoustique de corpus de salles est un sujet largement traité dans la littérature. De nombreuses études ont été réalisées dans le but de décrire les salles et la perception que l'on en a, à partir de mesures de paramètres objectifs. La dimension la plus étudiée à ce sujet est probablement celle de la qualité acoustique des salles de concert, avec pour enjeu principal celui de trouver un jeu restreint de paramètres acoustiques indépendants permettant de décrire au mieux la préférence subjective des individus. Or, comme le mentionne [Cerdá *et al.*, 2009], il existe un "nombre considérable de paramètres proposés par différents chercheurs pour déterminer la qualité acoustique des salles de concerts" [Schroeder *et al.*, 1974, Ando, 2014, Kahle, 1995, Hidaka et Beranek, 2000, Beranek, 2003, Cerdá *et al.*, 2012, Giménez *et al.*, 2014, Cerdá *et al.*, 2015]. Afin de limiter ce nombre, la norme ISO 3382-1, dans ses annexes, fait état d'une liste réduite des paramètres jugés comme les plus pertinents pour la caractérisation acoustique des salles [ISO, 2009].

3.2.1 Choix et calcul de descripteurs acoustiques

En se basant sur ces travaux et recommandations, une sélection d'une vingtaine de descripteurs acoustiques a pu être réalisée, dans le but de proposer une carac-

térisation acoustique objective du corpus d'environnements acoustiques à l'étude. L'objectif de la présente section n'est pas de dresser un état de l'art de tous les descripteurs acoustiques existants mais plutôt de définir succinctement ceux qui ont été retenus pour l'analyse acoustique présentée par la suite et d'en justifier les choix.

Une pratique courante pour le calcul des paramètres acoustiques d'une salle est de décomposer la mesure en plusieurs sections temporelles et fréquentielles et de calculer quand cela est possible une valeur du paramètre pour chaque portion spectro-temporelle de la réponse. Concernant la segmentation temporelle, on distingue deux régimes : précoce (*early* : généralement les 80 premières millisecondes de la réponse, relatif au champ direct et aux premières réflexions de la réponse) et tardif (*late* : au-delà de 80 ms, représentatif du comportement diffus de la réponse). En ce qui concerne la segmentation fréquentielle, une analyse par bande d'octave est réalisée. Les descripteurs sont calculés pour 7 bandes d'octaves centrées sur les fréquences suivantes : 125, 250, 500, 1000, 2000, 4000 et 8000 Hz. Une valeur large-bande, calculée sur l'ensemble du spectre audible, est également exprimée.

D'après [Cerdá *et al.*, 2009], les descripteurs acoustiques peuvent être classés selon quatre catégories :

- les descripteurs de *réverbération* : décrivent le comportement temporel et fréquentiel de la réverbération de la salle.
- les descripteurs d'*énergie* : représentatifs de la quantité d'énergie restituée par la salle.
- les descripteurs d'*intelligibilité* : généralement des rapports d'énergie entre différentes sections temporelles de la réponse, ils quantifient l'intelligibilité à la position de mesure du message véhiculé par une source.
- les descripteurs *spatiaux* : relatifs au comportement spatial de la réponse.

3.2.1.1 Les descripteurs de réverbération

Ces descripteurs permettent de caractériser le comportement temporel et spectral d'une salle. Le temps de réverbération TR est exprimé ici par le calcul du RT_{20} qui est le temps de décroissance de 60 dB de la réponse, estimé sur la courbe de décroissance comprise entre -5 dB et -25 dB. Le temps de décroissance précoce EDT (*Early Decay Time*), correspond au temps de réverbération estimé cette fois-ci sur la courbe de décroissance précoce entre 0 et -10 dB. Ces deux paramètres sont calculés en large bande et par bande d'octave. Le temps central T_c correspond à l'instant où la moitié de l'énergie de la réponse est diffusée. Il est exprimé par la formule suivante :

$$T_c = \frac{\int_0^\infty t \cdot p(t)^2 \, \mathrm{d}t}{\int_0^\infty p(t)^2 \, \mathrm{d}t} \tag{3.3}$$

Cette grandeur est parfois associée aux indices de clarté et d'intelligibilité [Bradley, 2011] puisque sa valeur dépend grandement du rapport entre énergies précoce et tardive de la réponse. Le *Bass Ratio* BR et la *brillance* Br (aussi appelée *Treble Ratio*) représentent respectivement la quantité de basses et d'aiguës de la réponse acoustique de la salle. Ils sont exprimés à partir du calcul du RT_{20} par bande d'octave de la manière suivante :

$$BR = \frac{RT_{20,\,125} + RT_{20,\,250}}{RT_{20,\,500} + RT_{20,\,1k}} \text{ , s} \tag{3.4}$$

$$Br = \frac{RT_{20,\,2k} + RT_{20,\,4k}}{RT_{20,\,500} + RT_{20,\,1k}} \tag{3.5}$$

où le chiffre en indice correspond à fréquence centrale de la bande d'octave concernée. L'équivalent fréquentiel du temps central est donné par le centroïde spectral f_c calculé comme suit :

$$f_c = \frac{\int_{20\,\text{Hz}}^{20\,\text{kHz}} f \cdot |p(f)|\,\mathrm{d}f}{\int_{20\,\text{Hz}}^{20\,\text{kHz}} |p(f)|\,\mathrm{d}f} \text{ , Hz} \tag{3.6}$$

Enfin, la fréquence de Schroeder indique la séparation fréquentielle de la réponse entre le comportement modal (à basse fréquence) de la salle et son comportement diffus. Elle dépend du temps de réverbération RT_{20} de la salle et de son volume V et est exprimée par la formule suivante :

$$f_{schr} = 2000\sqrt{\frac{RT_{20}}{V}} \text{ , Hz} \tag{3.7}$$

3.2.1.2 Les descripteurs d'énergie

La force sonore G (*Strength Factor* en anglais) correspond à l'énergie de la réponse de la salle $p(t)$ à l'excitation d'une source (théoriquement omnidirectionnelle) mesurée à une position donnée, normalisée par l'énergie de la réponse $p_{A\,10\text{m}}(t)$ mesurée à 10 m de cette même source en situation anéchoïque. Il est représentatif de l'intensité sonore perçue par un auditeur situé à la position de mesure.

$$G = 10 \cdot \log \frac{\int_0^\infty p(t)^2\,\mathrm{d}t}{\int_0^\infty p_{A\,10\text{m}}(t)^2\,\mathrm{d}t} \text{ , dB} \tag{3.8}$$

Dans notre cas, la mesure anéchoïque choisie pour le calcul de G correspond à la fonction de transfert anéchoïque liant la position de microphone MC et la position de source ec de la grille de mesure définie en section 3.1.2, soit à une distance source-microphone de 5.6 m. De plus, la source utilisée pour la mesure (Genelec 8020A) n'est pas omnidirectionnelle. La valeur de G telle que nous l'avons calculée n'est donc pas comparable à celles obtenues en respectant la définition initiale de G. Toutefois, elle est représentative du même phénomène et est calculée de la même manière pour chacune des réponses du corpus étudié ici. Pour éviter toute confusion, nous appellerons ce paramètre de force sonore G' :

$$G' = 10 \cdot \log \frac{\int_0^\infty p(t)^2 \, \mathrm{d}t}{\int_0^\infty p_{A\,5.6\,\mathrm{m}}(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{3.9}$$

Ce paramètre est également calculé par bandes d'octave et pour les deux régimes temporels : précoce ($_E$) et tardif ($_L$). Leurs expressions sont données par les équations suivantes :

$$G'_E = 10 \cdot \log \frac{\int_0^{80\,\mathrm{ms}} p(t)^2 \, \mathrm{d}t}{\int_0^\infty p_{A\,5.6\,\mathrm{m}}(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{3.10}$$

$$G'_L = 10 \cdot \log \frac{\int_{80\,\mathrm{ms}}^\infty p(t)^2 \, \mathrm{d}t}{\int_0^\infty p_{A\,5.6\,\mathrm{m}}(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{3.11}$$

3.2.1.3 Les descripteurs d'intelligibilité

Ils sont représentatifs de la compréhension par un auditeur du message véhiculé par une source au sein d'un environnement acoustique. Ils dépendent principalement de deux facteurs : le niveau de réverbération de la salle et la distance source-auditeur. Le DRR (Direct-to-Reverberant Ratio) correspond au rapport d'énergie entre le champ direct et le reste de la réponse (eq. 3.12). Il joue un rôle important dans la perception de la distance d'une source sonore.

$$DRR = \frac{\int_{t_0-2.5\mathrm{ms}}^{t_0+2.5\mathrm{ms}} p(t)^2 \, \mathrm{d}t}{\int_{t_0+2.5\mathrm{ms}}^\infty p(t)^2 \, \mathrm{d}t} \tag{3.12}$$

avec t_0 correspondant à l'instant du pic du champ direct de la RI. D'après les travaux initiaux de Wallach et al. et de Haas, un phénomène d'intégration cognitive entre un signal et une version retardée de ce même signal est observé pour un retard maximum de 5 ms (pour un signal impulsif [Wallach *et al.*, 1949]), de 50 ms (pour un signal de parole [Haas, 1951]) et jusqu'à 100 ms (pour un signal de musique [Blauert, 1997a]). Ce phénomène d'intégration est connu sous le nom d'*effet de préséance*. Sur ce principe, il a par la suite été montré que les premiers échos d'une réponse pouvaient être intégrés au champ direct et ainsi améliorer l'intelligibilité d'une source en conditions réverbérantes [Bradley *et al.*, 2003]. C'est pourquoi les indices de clarté, prenant en compte ces premières réflexions, sont préférés au DRR pour représenter l'intelligibilité de la parole ou de la musique. Les indices de clarté C_x sont des rapports d'énergie entre deux sections temporelles précoce et tardive de la réponse. x représente la délimitation temporelle entre les deux régimes tel que :

$$C_x = 10 \cdot \log \frac{\int_0^x p(t)^2 \, \mathrm{d}t}{\int_x^\infty p(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{3.13}$$

Le C_{50} ($x = 50$ ms) correspond au rapport d'énergie entre les 50 premières millisecondes et le reste de la réponse. Il est communément employé pour représenter l'intelligibilité d'un message de parole. De son côté, le C_{80} ($x = 80$ ms) est plutôt

utilisé pour quantifier la clarté d'un contenu musical. Enfin, le STI (Speech Transmission Index) est un autre indicateur de l'intelligibilité de la parole. Sa valeur est comprise entre 0 et 1 (0 représentant un discours inintelligible et 1 : un discours parfaitement intelligible). Il peut être calculé à partir de la mesure d'une réponse impulsionnelle comme suggéré par la norme IEC 60268-16 [IEC, 2003].

3.2.1.4 Les descripteurs spatiaux

Afin de décrire l'impression spatiale en conditions réverbérantes, deux attributs perceptifs sont couramment employés : le niveau d'enveloppement LEV (Listener envelopment), correspondant à la sensation subjective de l'auditeur d'être enveloppé par le champ sonore et la taille apparente de source ASW (Apparent Source Width). D'après [Hidaka *et al.*, 1995] ces attributs spatiaux sont fortement liés à deux paramètres objectifs. D'abord, le niveau de corrélation interaurale $IACC$ (Interaural Cross-Correlation), correspond au maximum de la fonction d'inter-correlation $IACF$ entre les signaux arrivant aux oreilles gauche $p_g(t)$ et droite $p_d(t)$ d'un auditeur[2], calculée sur une fenêtre temporelle $[t_1, t_2]$:

$$IACC_{t_1,t_2} = \max|IACF(\tau)| \tag{3.14}$$

avec :

$$IACF_{t_1,t_2}(\tau) = \frac{\int_{t_1}^{t_2} p_g(t) \cdot p_d(t-\tau)\,\mathrm{d}t}{\left(\int_{t_1}^{t_2} p_g(t)^2\,\mathrm{d}t \cdot \int_{t_1}^{t_2} p_d(t)^2\,\mathrm{d}t\right)^{1/2}} \tag{3.15}$$

D'un autre côté la fraction d'énergie latérale LF (Lateral Fraction) représente la proportion d'énergie venant des côtés par rapport à l'énergie totale de la réponse, exprimée par la formule suivante :

$$LF = \frac{\int_{t_1}^{t_2} p_8(t)^2\,\mathrm{d}t}{\int_{t_1}^{t_2} p(t)^2\,\mathrm{d}t} \tag{3.16}$$

avec $p_8(t)$, l'énergie mesurée par un microphone bidirectionnel dont le nœud est orienté vers la source. Une valeur globale de l'$IACC$ et du LF est obtenue en prenant pour fenêtre temporelle la durée totale de la RI. Des valeurs précoce et tardive, notées par les indices $_E$ et $_L$ sont également calculées pour ces deux descripteurs, en prenant pour fenêtres temporelles respectives $[t_{1E} = 0\,\mathrm{ms}, t_{2E} = 80\,\mathrm{ms}]$ et $[t_{1L} = 80\,\mathrm{ms}, t_{2L} = \infty]$.

2. La mesure de l'IACC est une mesure binaurale, réalisée avec une tête artificielle ou avec la tête d'un auditeur, équipée de microphones dans chaque oreille. Or dans notre cas, les mesures ont été conduites avec un réseau sphérique de microphones (em32) pour un formalisme HOA. Une conversion des réponses HOA au format binaural a dû être effectuée afin de pouvoir calculer cet indicateur. Pour ce faire, les réponses HOA sont décodées sur un réseau de HP virtuels, dont les positions sont données par une base de HRTF (le jeu de HRTF utilisé est issu de la base SCUT [Xie, 2018]). Une fois le décodage sur haut-parleurs réalisé, la conversion binaurale est effectuée par convolution matricielle avec le jeu de HRTF choisi.

Enfin, la force latérale tardive LG_L (Late Lateral Strength) de la réponse, introduit par [Bradley et Soulodre, 1995b] semble être un bon indicateur de la sensation d'enveloppement (LEV). La force latérale LG est une mesure latérale de la force sonore G de la salle.

$$LG = 10 \cdot \log \frac{\int_{t_1}^{t_2} p_8(t)^2 \, \mathrm{d}t}{\int_0^{\infty} p_{A\,10\,\mathrm{m}}(t)^2 \, \mathrm{d}t} \, , \mathrm{dB} \tag{3.17}$$

Dans sa formulation mathématique il s'agit d'un indicateur hybride entre LF et G. De la même manière que pour le calcul de G, LG' représente une version de LG, normalisée par la mesure anéchoïque d'une source placée à une distance de 5.6 m du microphone. A l'instar des deux autres descripteurs spatiaux, 3 versions LG', LG'_E et LG'_L sont calculées et représentent respectivement la force latérale globale, précoce et tardive de la réponse.

3.2.1.5 Récapitulatif

Le tableau 3.3 récapitule les informations sur les différents indicateurs utilisés pour l'analyse acoustique de la collection d'acoustiques mesurées.

Bien qu'un calcul par bande d'octave de certains descripteurs ait été réalisé, l'analyse acoustique présentée par la suite a été conduite à partir de leurs valeurs large-bande, afin de limiter la quantité de données à analyser.

3.2.2 Analyse factorielle

L'analyse factorielle permet, à partir d'une large jeu de variables, de définir un espace de dimensions réduites repésentatif de la variabilité des données étudiés. Elle est basée sur l'étude de la corrélation entre les variables mesurées et s'appuie sur les outils de l'analyse en composantes principales (ACP). A la différence de l'ACP, qui opère une réduction de dimension en exploitant uniquement la variance observée entre les individus, l'analyse factorielle définit ces dimensions comme des facteurs latents portés par des groupes de variables observées. Cette méthode a pour principal avantage de facilité l'interprétabilité des dimensions. Dans notre cas, il s'agit donc de trouver un nombre réduits de facteurs, représentatifs de la variabilité des conditions acoustiques mesurées et caractérisées par 23 descripteurs. Afin de définir un espace caractéristique des environnements réverbérants qui ont été mesurés, nous avons décider de ne pas inclure dans l'analyse les mesures réalisées en champ libre (trop différentes des autres conditions). En effet, les espaces issues d'analyses de ce type sont extrêmement dépendants du jeu de données en en exploitant les informations de variances, ce qui a pour conséquence qu'un cas particulier comme le cas anéchoïque aurait un poids fort, par rapport aux autres données, dans la détermination de l'espace réduit. D'autre part, les conditions de coïncidence source-microphone (pour la source en position *eb*) sont également écartées du fait du cas particulier qu'elles constituent. Au total, 108 conditions acoustiques (6 positions de mesures $\times$ 18 environnements acoustiques)

Type	Descripteurs	Bandes d'octave	Notes
Réverberation	RT	✓	Temps de réverbération (diffus)
	EDT	✓	Temps de réverbération (précoce)
	T_c	✓	Barycentre temporel
	BR		Quantité de basses
	Br		Quantité d'aigues
	f_c		Barycentre fréquentiel
	f_{schr}		Séparation fréq. modal / diffus
Energie	G'	✓	Représentatif du niveau sonore perçu
	G'_E	✓	
	G'_L	✓	
Intelligibilité	DRR	✓	Représentatif de la distance perçue
	C_{50}	✓	Clarté pour de la parole
	C_{80}	✓	Clarté pour de la musique
	STI	✓	Intelligibilité de la parole
Spatial	$IACC$	✓	
	$IACC_E$	✓	Représentatif de l'ASW
	$IACC_L$	✓	
	LF	✓	
	LF_E	✓	Représentatif de l'ASW
	LF_L	✓	
Spatial/Energie	LG'	✓	
	LG'_E	✓	
	LG'_L	✓	Représentatif du LEV

TABLE 3.3 – Récapitulatif des 23 descripteurs acoustiques calculés pour chacune des réponses impulsionnelles spatiales mesurées d'après le protocole présenté en section 3.1. La colonne "Bandes d'octave" indique les descripteurs qui ont été calculés par bande d'octave (marqués par le signe ✓).

sont exploitées pour l'analyse.

3.2.2.1 Etude de corrélation entre les variables

En se basant sur ces 108 observations, l'étude de la corrélation croisée entre les descripteurs donne un premier aperçu des relations de dépendance ou d'orthogonalité entre les variables. D'après la figure 3.3, représentant la valeur absolue des coefficients de corrélation de Pearson, il est possible de distinguer des groupes de variables, relativement indépendants entre eux. On constate notamment une forte indépendance entre les variables relatives à la réverbération (RT_{20}, EDT, Br) et le reste des variables. Deux autres groupes paraissent également relativement indépendants, l'un relatif aux descripteurs d'énergie (G', LG' et leurs déclinaisons temporelles) et l'autre relatif à l'intelligibilité (STI, C_{80}, C_{50}, DRR). A noter que, les grandeurs spatiales $IACC$ et LF, globales et précoces, semblent fortement liés aux paramètres d'intelligibilité, tout comme le centroïde spectral (F_c). Enfin, trois descripteurs ($IACC_L$, LF_L et BR) semblent relativement indépendants. Concernant, l'$IACC_L$ et le LF_L ces grandeurs varient faiblement d'une condition à l'autre, en comparaison de leurs équivalents globaux et précoces, ce qui s'explique par le fait qu'en théorie, le champ acoustique tardif dans une salle est fortement incohérent et homogène dans toutes les directions et en tout point de l'espace, quel que soit l'environnement acoustique. Leur pertinence dans l'analyse est discutable. Le Bass-Ratio BR quant à lui varie bel est bien d'une condition à l'autre et la faible corrélation avec les autres descripteurs semble plutôt indiquer une forte orthogonalité entre cette grandeur et le reste des variables à l'étude. D'après cette étude de corrélation, on peut également noter une forte redondance dans l'information portée par certains descripteurs. D'abord, les énergies tardive omnidirectionnelle et latérale, respectivement représentées par G'_L et LG'_L évoluent de façon quasiment identique entre les conditions, et la fréquence de Schroeder f_{schr}, grandeur empirique, semble être un excellent indicateur de ces descripteurs plus complexes. D'autre part, une grande redondance est observée entre les indicateurs d'intellibilité et de clarté STI, C_{50} et C_{80}. Enfin, les expressions globales et précoces des grandeurs spatiales LF et $IACC$ semblent fortement redondantes, indiquant que ces deux paramètres, lorsqu'ils sont calculés globalement, sont principalement régis par le comportement précoce de la réponse.

3.2.2.2 Analyse en composantes principales et choix du nombre de composantes

A la suite de cette première analyse qualitative de la corrélation entre les variables, une analyse en composantes principales (PCA) est réalisée, afin de déterminer le nombre de dimensions indépendantes nécessaires pour représenter la variabilité des données. D'après l'observation des valeurs propres issues de l'analyse

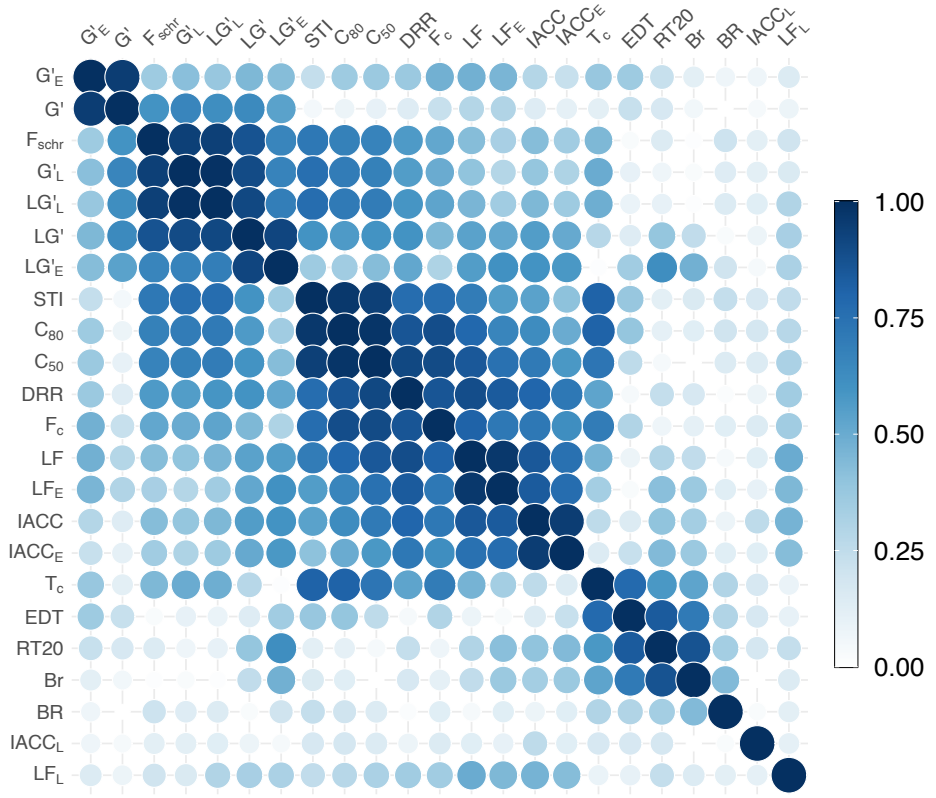


FIGURE 3.3 – Matrice de corrélation entre les indicateurs acoustiques, calculée à partir de l'ensemble des observations . La valeur absolue du coefficient de corrélation de Pearson est indiquée par la couleur et la taille des cercles.

il est possible de représenter la quantité de variance expliquée par chacune des dimensions principales, comme l'illustre la figure 3.4. On constate sur ce diagramme que les trois premières dimensions portent la grande majorité de l'information contenue dans le jeu de données puisqu'elles permettent d'en exprimer plus de 80 % de la variance. Ce résultat indique que l'espace composé des trois premières composantes de l'ACP est satisfaisant pour représenter la diversité des conditions acoustiques à l'étude.

3.2.2.3 Analyse factorielle : étude des variables

A partir de la sélection de ces trois dimensions issues de l'analyse en composantes principales, une analyse factorielle à trois facteurs est réalisée. Pour ce faire, une rotation de type *varimax* est appliquée aux trois premières dimensions de l'ACP. Cette rotation a pour but d'extraire trois nouvelles dimensions orthogonales, les facteurs, expliquant la même quantité de variance dans les données tout en maximisant la corrélation entre les dimensions et les variables étudiées. La figure 3.5 représente la

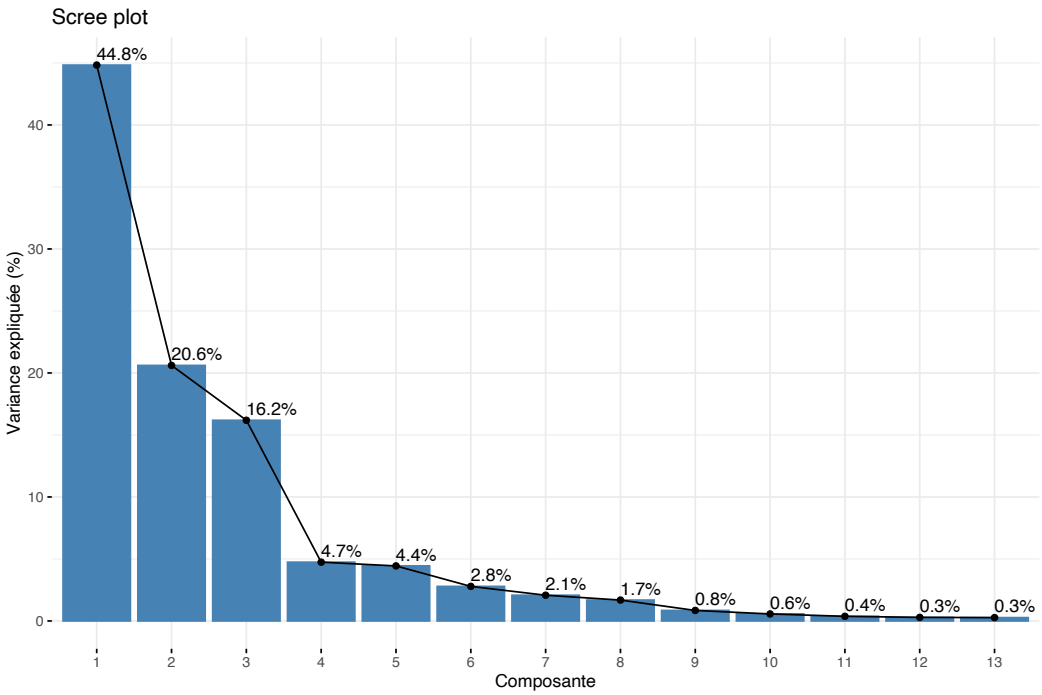


FIGURE 3.4 – Variance expliquée par les 13 premières composantes de l'ACP. Les trois premières dimensions représentent à elles trois, 81.6 % de la variance des données.

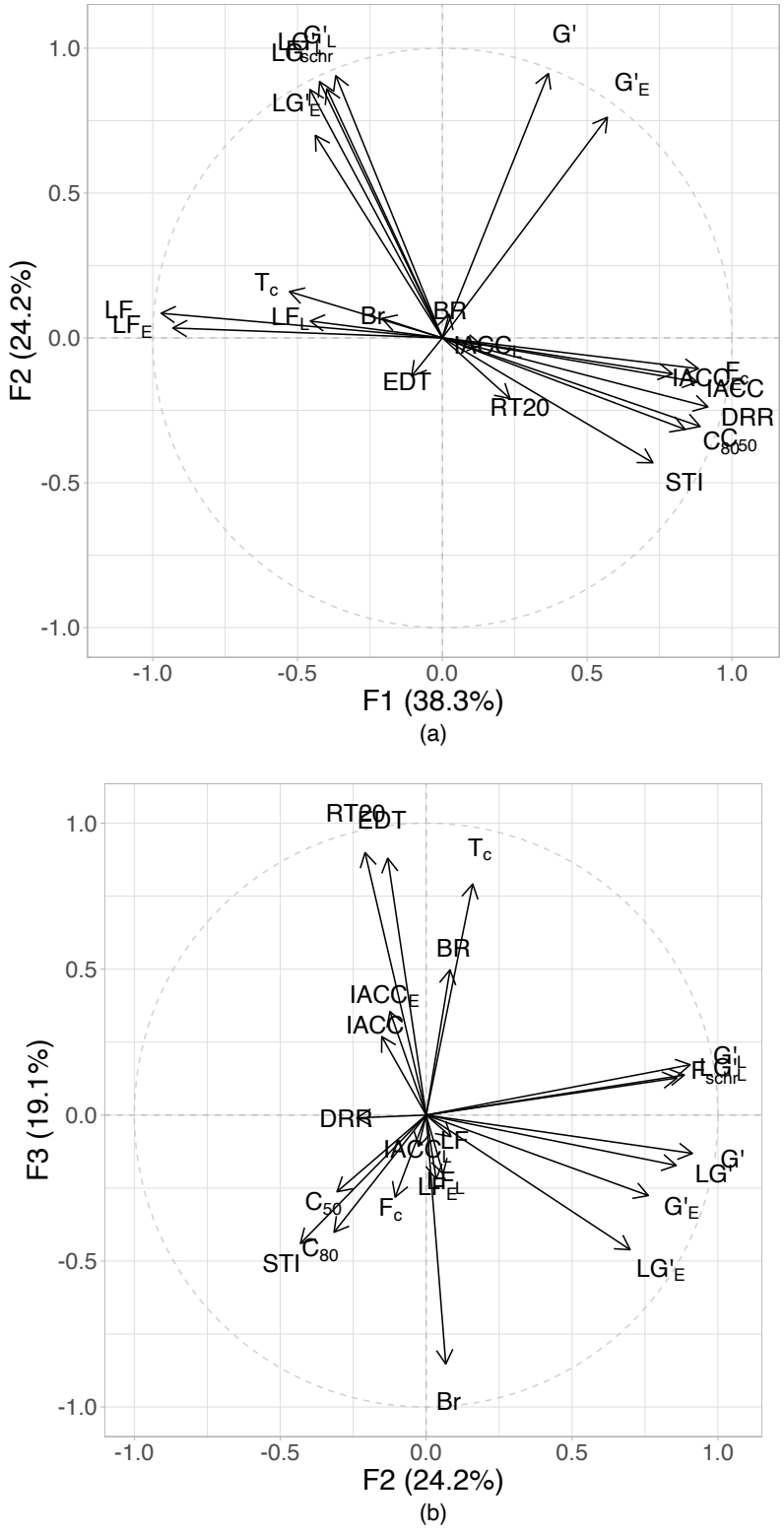


FIGURE 3.5 – Projection des vecteurs propres dans l'espace des facteurs (a) F1 vs. F2, (b) F2 vs. F3. Entre parenthèses est donnée la quantité de variance expliquée par chacun des facteurs.

projection des variables dans l'espace porté par les trois facteurs F1, F2 et F3. Le tableau 3.4 indique les coefficients de corrélation de Pearson, calculés entre chaque descripteur et les facteurs. D'après ces deux données, trois groupes de variables portés par chacun des facteurs sont clairement définis. Le facteur F1 est largement décrit par, d'une part les indices d'intelligibilité et de clarté (DRR,C_{50}, C_{80}, STI) et d'autre part par des grandeurs connues pour être représentatives de la taille apparente de la source ($IACC_E$ et LF_E). Il semble donc que ce facteur soit indicatif de la "précision" de la source, à ce titre nous l'appellerons "Précision". Le facteur F2 est quant à lui principalement porté par les paramètres de force sonore (G et LG, globaux, précoces et tardifs). Comme discuté précédemment la fréquence de Schroeder est également porté par ce facteur que nous renommerons "Niveau". Enfin, le troisième facteur F3 est porté par les indices de réverbération de la salle (RT_{20}, EDT, Br et T_c). Ce facteur est donc renommé "Réverbérance", terme décrivant la sensation de réverbération dans une salle [Weinzierl *et al.*, 2018].

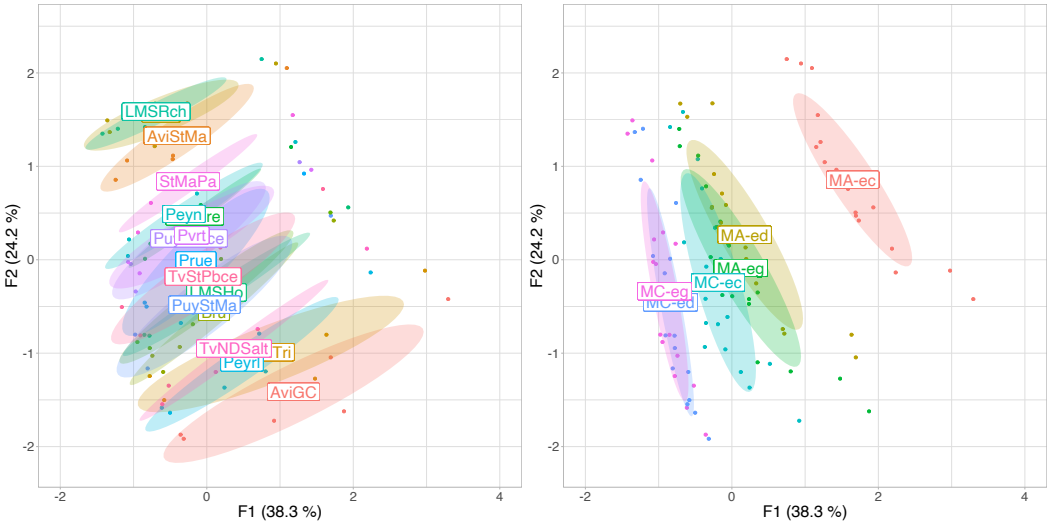
3.2.2.4 Analyse factorielle : étude des conditions acoustiques

Les 108 conditions acoustiques à l'étude peuvent être représentés dans l'espace factoriel fraîchement obtenu. Les figures 3.6 et 3.7 illustrent respectivement la projection des données sur les dimensions F1-F2 (Précision-Niveau) et F2-F3 (Niveau-Réverbérance). Les conditions à l'étude y sont regroupées (a) par environnement acoustique afin d'illustrer la variabilité inter-édifices et (b) par position de mesure afin d'illustrer la variabilité intra-édifice (inter-positions de mesure). A partir de la figure 3.6, on constate une certaine orthogonalité entre les environnements acoustiques et les positions de mesure. En outre, la variabilité intra-édifice est principalement portée par le facteur F1 "Précision", tandis que le facteur F2 "Niveau" permet davantage de discriminer les édifices. Une autre manière d'expliquer cette première figure serait de dire que la précision d'une source est principalement affectée par la distance qui la sépare du récepteur (meilleure précision en $MA-ec$, correspondant à la plus faible distance source-microphone) et que le niveau sonore dans une salle est principalement déterminé par la capacité de la salle à restituer une grande quantité de l'énergie produite par une source. Toutefois l'orientation légèrement oblique des édifices et perpendiculairement des positions de mesures dans cet espace (Précision / Niveau) suggère une interdépendance entre ces deux paramètres. En effet, les salles présentant le plus fort niveau sont également celles pour lesquelles la précision est la moins bonne, laissant penser à un phénomène de masquage de la source par la salle. A l'inverse les positions de mesures pour lesquelles la précision est la meilleure sont également celles pour lesquelles le niveau est le plus élevé, du fait de la distance source-auditeur réduite.

D'après la figure 3.7, le facteur F3 "Réverbérance" est également discriminant pour les environnements acoustiques, en revanche il ne l'est pas pour les positions de mesures. Cela s'explique par le fait que les descripteurs de réverbération associés à ce facteur (notamment RT, EDT et Br) sont censés être relativement iden-

	F1	F2	F3
LF	-97.1		
LF_E	-93.1		
DRR	91.7		
C_{50}	89.0		
f_c	88.4		
$IACC$	88.0		
C_{80}	83.8		
$IACC_E$	79.4		
STI	72.7		
G'		91.3	
G'_L		90.5	
LG'_L		88.5	
f_{schr}		86.0	
LG'		85.7	
G'_E		76.2	
LG'_E		69.9	
RT_{20}			89.9
EDT			88.0
Br			-85.3
T_c			79.2
BR			
$IACC_L$			
LF_L			

TABLE 3.4 – Corrélation de Pearson en %, entre les descripteurs et les trois facteurs. Dans le tableau, sont uniquement représentés les coefficients dont la valeur absolue est supérieure à 60 %.

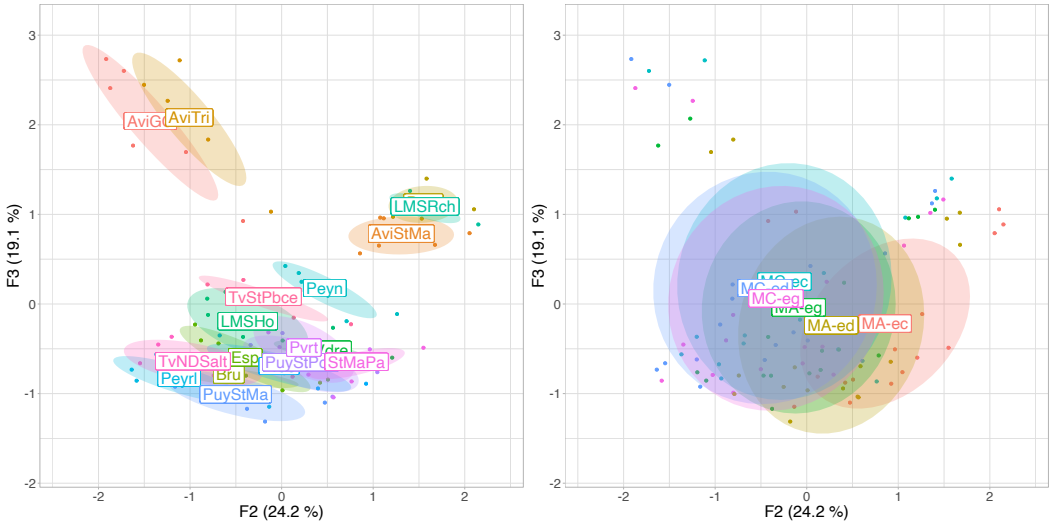


(a) (b)

FIGURE 3.6 – Représentation des individus dans l'espace des facteurs F1 vs. F2. Les points représentent les différentes conditions acoustiques mesurées. Les ellipses représentent la dispersion des points au sein d'un même groupe. (a) Regroupement par environnement acoustique, représenté par les ellipses et le code couleur. (b) Regroupement par position de mesure, représentée par les ellipses et le code couleur.

tiques en tout point d'une salle mais peuvent en revanche varier d'une salle à l'autre.

L'espace 3D représentatif de la variabilité des données mesurées peut donc être décomposé entre deux sous-espaces, un espace 1D caractéristique de la variablité intra-édifice, donné par le facteur F1 "Précision" et un espace 2D caractéristique de la variabilité inter-édifices, porté par les dimensions F2 et F3 "Niveau" et "Réverbérance", adapté à l'objectif premier de ces travaux qui est celui de caractériser du point de vue acoustique les édifices d'un corpus. En s'attardant alors sur cet espace, il est possible en figure 3.7(a) de distinguer trois groupes principaux d'individus. Le premier est composé des deux grandes salles du Palais des Papes d'Avignon. Situé en haut à gauche de l'espace F2-F3, il correspond aux édifices de grande réverbérance et de faible niveau. Le deuxième groupe est constitué de trois chapelles de petite taille : la chapelle St-Martial du Palais des Papes d'Avignon, la chapelle St-Roch des Mées et la chapelle Templier de Bras, toutes deux issues du corpus Sésames. Ce groupe est situé en haut à droite sur la figure 3.7(a), ce qui indique des chapelles de réverbérance et de niveau élevés. Le troisième et dernier groupe est constitué de la grande majorité des édifices du corpus Sésames et se situe au centre de l'espace 2D, correspondant à des salles de niveau et de réverbérance moyens. Malgré la diversité architecturale des chapelles de ce dernier groupe, il semble donc d'après cette étude qu'elles soient assez semblables sur le plan acoustique.



(a) (b)

FIGURE 3.7 – Représentation des individus dans l'espace des facteurs F2 vs. F3. Les points représentent les différentes conditions acoustiques mesurées. Les ellipses représentent la dispersion des points au sein d'un même groupe. (a) Regroupement par environnement acoustique, représenté par les ellipses et le code couleur. (b) Regroupement par position de mesure, représentée par les ellipses et le code couleur.

3.2.3 Conclusions sur la caractérisation acoustique

En résumé, un travail de caractérisation acoustique d'un corpus de édifices patrimoniaux a été réalisé. Il est basé sur la réalisation d'une campagne de mesures acoustiques dans plus de 18 lieux, puis sur la sélection et le calcul d'un ensemble de 23 descripteurs acoustiques de révérbération, d'intelligibilité, de niveau et d'espace pour l'ensemble des mesures et enfin sur une analyse factorielle des mesures sur la base de ces descripteurs. L'analyse factorielle a permis de mettre en évidence un espace acoustique basé sur trois facteurs : un facteur de "précision" caractéristique de la précision du stimulus sonore véhiculé par une source en environnement réverbérant, un facteur de "niveau" représentatif du niveau sonore restitué par la source et l'environnement acoustique et un facteur de "réverbérance" relatif à la durée de réverbération des différents lieux à l'étude. A partir de cette espace, il a été montré que la première dimension de précision met principalement en évidence des différences intra-édifice et plus précisément permet de discriminer les différentes positions de mesures au sein d'un même édifice, d'après la distance séparant la source du récepteur. L'espace formé par les deux autres dimensions de niveau et de réverbérance est quant à lui caractéristique de la diversité inter-édifices. Dans cette espace trois groupes de chapelles ont pu être révélés : un groupe correspondant à des acoustiques très réverbérantes et de faible niveau, représenté par deux grandes salles du Palais des Papes d'Avignon, un deuxième

groupe correspondant à des environnements réverbérants et de niveau élevé, porté par trois petites chapelles fortement réberbérantes et un troisième groupe caractéristique des acoustiques de révervération et de niveau moyen, intégrant une grande majorité des édifices du corpus Sésames. Il est possible de conclure à partir de cette analyse que sur le plan acoustique, la majeure partie des environnements étudiés présente de grandes similarités tandis que quelques cas particuliers émergent.

Pour aller plus loin dans ce travail de caractérisation, une première piste serait d'élargir le corpus d'étude, tant sur le plan des environnements acoustiques étudiés que sur celui de la grille de mesure déployée, ce qui permettrait de définir un nouvel espace plus largement représentatif des acoustiques du patrimoine. A l'inverse afin de mieux discriminer les environnements acoustiques du corpus Sésames, une nouvelle analyse factorielle pourrait être conduite en isolant les individus du troisième groupe. Enfin, l'étude du lien entre cet espace acoustique et les dimensions architecturale et perceptive est un des objectifs majeurs du projet Sésames. Sur le plan perceptif plusieurs expériences sont envisagées pour créer des cartographies perceptives des édifices à l'étude. Une première consiste à réaliser un test de dissemblance sur l'ensemble de la collection afin de définir de façon globale des distances perceptives entre les différents lieux. La deuxième étude envisagée est celle qui est présentée par la suite. Elle a pour objectif de définir un espace perceptif représentatif de la perception auditive spatiale en environnements réverbérants et s'appuie sur l'instrumentation et la méthodologie de l'expérience 1 présentée en chapitre 2.

3.3 Expérience 2 : Caractérisation de la perception auditive spatiale dans un corpus d'environnements acoustiques mesurés

Cette expérience, basée sur le protocole d'évaluation de la perception spatiale de sources sonores en environnements réverbérants, présenté en chapitre 2, vise à quantifier les distorsions de l'image spatiale des sources sonores perçues par un auditeur dans chacun des 19 environnements acoustiques de la collection lors de leur auralisation. En effet, d'après les résultats de la première expérience, il a été montré que le protocole mis en place à cette occasion permettait de mettre en évidence un certain nombre de différences perceptives, concernant les caractéristiques spatiales de sources sonores (position angulaire, distance et taille apparente), induites par divers environnements acoustiques mesurés et auralisés. Cette méthodologie semble donc bien adaptée pour répondre au double objectif de la présente étude qui est celui de proposer une caractérisation perceptive du corpus d'étude, selon des critères perceptifs spatiaux et dans un même temps de mieux comprendre comment l'acoustique influence notre perception spatiale des sources sonores. Dans cette section, nous présenterons en premier lieu les éléments méthodologiques propres à cette étude. Dans un second temps, nous proposerons en guise de premiers résultats de l'étude une projection des différents édifices du corpus dans un espace représentatif

de la perception spatiale des sources sonores en contexte d'auralisation. Une étude de corrélation entre les données perceptives obtenues et les grandeurs acoustiques mesurées sera également menée, afin de tenter de trouver un ensemble de paramètres acoustiques caractéristiques de la perception auditive spatiale. Enfin, une discussion et des conclusions sur ces résultats seront proposées.

3.3.1 Méthodologie

Une bonne partie de la méthodologie de la présente étude est commune à celle de l'expérience 1 présentée en chapitre 2, notamment en termes d'instrumentation, de méthode de report, de tâche à réaliser et de traitement des données. Nous rappellerons donc succinctement ces points méthodologiques et nous attarderons davantage sur les éléments spécifiques à cette expérience, à savoir le choix des conditions acoustiques et des stimuli sonores, les participants, le déroulement de l'expérience et l'analyse des résultats.

3.3.1.1 Rappels sur le protocole expérimental de l'expérience 1

L'interface en réalité virtuelle et la tâche de report des caractéristiques spatiales de sources sonores (position, distance et taille apparente) restes inchangées par rapport à celle présentée pour l'expérience précédente (*c.f.* section 2.2.3 du chapitre 2). Pour rappel, il s'agit d'une tâche de localisation de source composée de deux étapes : une première étape permettant à un utilisateur de reporter sa perception de la distance de la source, en réglant dans l'interface VR le rayon d'une demi-sphère qui l'entoure et une deuxième étape permettant le report de la position angulaire et de la taille perçue de la source, en entourant sur la surface de la demi-sphère la zone dans laquelle il perçoit la source. Ces trois attributs de source sont évalués dans les différentes conditions acoustiques du corpus, auralisées dans la sphère de 42 haut-parleurs du laboratoire (présentée en annexe A). L'auralisation des mesures ambisoniques (HOA d'ordre 4) réalisées à l'em32 est assurée sur un poste fixe, grâce aux outils de la librairie spat5 pour Max/MSP. Le décodage sur réseau de haut-parleurs est effectué avec la méthode "energy preserving", sans optimisation. L'interface VR, le déroulement de l'expérience et l'enregistrement des données utilisateurs sont gérés dans une application android portée sur casque de VR (Oculus Quest). L'analyse des données brutes (coordonnées des points des tracés pour chaque condition de test) est la même que pour l'expérience 1 et conduit à l'extraction, pour chaque essai, des grandeurs suivantes : erreurs de localisation en azimut ε_θ et en élévation ε_θ, distance perçue R_p et taille apparente reportée S_{eq}.

3.3.1.2 Conditions acoustiques et choix des stimuli

Pour la présente expérience, les 19 environnements acoustiques du corpus présenté précédemment ont été retenus à savoir les 15 acoustiques des petites et moyennes chapelles du corpus Sésames, les 3 acoustiques du Palais des Papes d'Avignon et enfin la chambre anéchoïque. Nous avons choisi d'intégrer cette dernière

dans l'étude afin de pouvoir évaluer les performances de localisation en champ libre, souvent données comme cas de référence de la perception auditive spatiale. D'autre part, 4 configurations source-récepteur ont été sélectionnées, à savoir les trois configurations pour la position "proche" MA du microphone : $MA-ec$, $MA-ed$ et $MA-eg$ et la configuration centrale pour la position éloignée MC du microphone : $MC-ec$. Au total donc, 76 (19×4) réponses impulsionnelles sont à l'étude. Lors de l'auralisation chaque réponse est convoluée avec un train d'impulsion de bruit-blanc (durée 1s), identique à celui employé dans l'expérience 1.

3.3.1.3 Participants

15 participants normaux-entendants (4 femmes, 11 hommes) âgés en moyenne de 32 ans ($\pm$7 ans) ont accepté de prendre part à cette expérience. Les sujets n'avaient aucune connaissance à priori de la diversité du corpus ni de la configuration spatiale de la grille de mesure sur laquelle s'est basée la campagne (*c.f.* figure 3.2 en section 3.1.2 de ce chapitre).

3.3.1.4 Procédure

Avant de commencer l'expérience, les participants doivent prendre acte des consignes de l'expérience. La tâche à réaliser est décrite sur la fiche de consignes comme suit : "Dans chacune des conditions acoustiques qui vous sera présentée, vous devrez tenter de localiser le plus précisément possible la source sonore. La tâche de localisation devra être effectuée à travers l'interface de réalité virtuelle. Pour chaque essai, il vous sera demandé de :

1. reporter votre jugement de la distance vous séparant de la source, en contrôlant la distance d'un objet dans l'environnement virtuel.
2. reporter votre jugement de la position et de la taille de la source en entourant *le plus précisément possible* une zone dans laquelle vous pensez que la source se situe."

Il est également précisé : "La zone que vous délimitez par votre tracé devra toujours inscrire la source. Plus la source vous semblera nette/précise/ponctuelle, facile à localiser, plus la zone tracée autour de la source devra être petite. A l'inverse, plus la source vous semblera large, diffuse, difficile à localiser, plus la zone tracée autour de celle-ci devra être grande." Après la lecture des consignes, l'opérateur est présent pour répondre aux questions du sujet et doit s'assurer que les instructions ont été comprises.

L'expérience est décomposée en trois sessions distinctes, une session d'apprentissage afin de permettre aux participants de se familiariser avec le dispositif et la tâche à réaliser, et deux sessions de test. Avant de commencer la première session, une étape de calibration, décrite en section 2.2.3 du chapitre précédent, est réalisée et permet de faire correspondre l'orientation des environnements virtuel et réelle. La session de familiarisation est analogue à celle proposée pour l'expérience 1. Le

sujet est positionné, comme pour le reste de l'expérience, au centre du dispositif de spatialisation, sur une chaise pivotante, le casque de réalité virtuelle sur les yeux, dans lequel l'interface de test est projetée. Il est équipé d'un contrôleur qui lui permet de réaliser la tâche de localisation. L'opérateur se positionne arbitrairement dans la salle d'expérimentation. Le sujet doit alors localiser la voix de l'opérateur, en suivant les consignes qui lui ont été données par écrit. Lorsque le participant est satisfait de sa réponse, il doit la valider. Le contenu visuel projeté dans le casque est retranscrit sur un écran visible par l'expérimentateur, permettant à ce dernier de s'assurer que le sujet réalise la tâche correctement. L'opération est répétée jusqu'à ce que le sujet se sente à l'aise avec la tâche et les contrôles. Les données de cette session ne sont pas enregistrées.

La phase de test consiste donc à réaliser la tâche de localisation d'un signal de bruit blanc, dans chacune des 76 conditions acoustiques auralisées. L'ordre de présentation des stimuli est déterminé par tirage aléatoire et les 76 essais sont répartis dans deux sessions successives de 38 essais, afin de limiter la durée des sessions et permettre aux participants de faire une pause entre les deux phases de test. Pour chaque essai, une rotation d'un angle aléatoire compris entre -60° et 60° est appliqué à la scène sonore. Le participant doit alors réaliser la tâche de localisation de la source en deux étapes, puis valider sa réponse, qui est automatiquement enregistrée sur un fichier texte. Une fois la réponse validée et lorsqu'il se sent prêt, il peut alors lancer l'extrait suivant, et ainsi de suite jusqu'à la fin de la session en cours. Toutes phases comprises, l'expérience dure en moyenne 30 minutes par sujet.

3.3.1.5 Analyses statistiques

Les grandeurs ε_θ, ε_ϕ, R_p et S_{eq}, respectivement les erreurs de localisation en azimut et élévation, la distance perçue et la taille apparente de la source, sont analysées. Afin de respecter les conditions de normalité nécessaires à l'analyse statistique, toutes les grandeurs étudiées ont subi une transformation "*log*". Les résultats bruts de l'expérience sont donnés en tableau B.1 de l'annexe B. Une étude de corrélation entre les attributs perceptifs spatiaux reportés pour chaque condition du test et les descripteurs acoustiques calculés sur l'ensemble des réponses impulsionnelles du corpus (*c.f.* section 3.2) est conduite, dans le but de comprendre sur quelles grandeurs physiques les différents attributs perceptifs sont évalués. D'autre part, une analyse statistique par modèle linéaire mixte est conduite pour chacun des attributs mesurés, en prenant pour facteur principal les édifices du corpus *SALLE*. Les participants et les positions d'écoute sont considérés dans ce modèle comme facteurs aléatoires. Une ANOVA est ensuite réalisée sur le modèle mixte, afin de mesurer l'influence des environnements acoustiques sur les attributs spatiaux de sources. Des tests post-hoc avec ajustement de Tukey ont également été conduits sur chacune des grandeurs à l'étude. Les résultats de ces tests post-hoc sont consultables en tableaux B.2, B.3 et B.4 de l'annexe B. L'objectif de cette analyse est de révéler parmi les attributs perceptifs mesurés, ceux qui sont soumis à une influence de l'acoustique, afin de

proposer, à partir de ces grandeurs, un espace perceptif représentatif de la façon dont les acoustiques modifient notre perception spatiale des sources sonores. A partir de cet espace, une classification hiérarchique des individus est réalisée afin de mettre en évidence des groupes d'individus semblables sur le plan perceptif étudié. Cette classification hiérachique est basée sur la méthode de Ward [Ward Jr, 1963] et consiste à calculer les distances entre chaque individu représenté dans l'espace étudié et de les regrouper de telle sorte à maximiser l'inertie inter-groupes.

3.3.2 Résultats

Les résultats de l'expérience sont présentés en deux sections. La première présente les résultats de l'étude de corrélation entre les attributs spatiaux et les paramètres acoustiques objectifs analysés dans la première partie de ce chapitre. La seconde section, présente les résultats de l'analyse de variance et les effets de l'environnement acoustique $SALLE$, sur les différents attributs spatiaux de source reportés par les participants.

3.3.2.1 Résultats de l'étude de corrélation

L'étude de corrélation porte d'une part sur la corrélation entre les grandeurs perceptives collectées lors de cette expérience et les grandeurs acoustiques étudiées dans la section précédente dans le but premier de révéler l'existence d'un jeu d'indices objectifs représentatif des attributs perceptifs spatiaux des sources sonores en environnements réverbérants. De plus, les trois facteurs ($F1$: "Précision", $F2$: "Niveau" et $F3$: "Réverbérance") issus de l'analyse factorielle sur les paramètres acoustiques sont également inclus dans cette analyse de corrélation, afin de mesurer la pertinence de ces dimensions vis à vis de la perception spatiale des sources. Enfin, la corrélation entre les quatre attributs perceptifs de source est également étudiée permettant de révéler des liens potentiels dans l'évaluation de ces grandeurs. Les résultats de cette étude sont donnés en tableau 3.5.

3.3.2.2 Résultats de l'ANOVA

Un effet significatif de l'environnement acoustique est observé sur les quatre grandeurs mesurées. Un effet modéré de la salle sur l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$ est révélé par l'analyse : $F(18, 1042) = 2.2065$, $p = 0.00262$, $\eta^2 = 0.04$. L'erreur absolue de localisation en élévation $|\varepsilon_\phi|$ est également soumis à un effet marginal de la salle : $F(18, 1035) = 1.8102$, $p = 0.02009$, $\eta^2 = 0.03$. La distance perçue R_p est plus significativement impactée par l'environnement acoustique : $F(18, 1050) = 38.8600$, $p < 0.0001$, $\eta^2 = 0.40$. Enfin, la taille apparente de source reportée S_{eq} est également largement influencée par l'acoustique de la salle : $F(18, 1050) = 6.4088$, $p < 0.0001$, $\eta^2 = 0.10$.

Afin de créer une cartographie des édifices du corpus selon la manière dont ils influencent notre perception spatiale des sources, seuls les trois attributs perceptifs

	$\lvert\varepsilon_\theta\rvert$	$\lvert\varepsilon_\phi\rvert$	R_p	S_{eq}
G'_E	0.04	0.34	-0.34	**0.53**
G'	0.16	0.31	-0.10	**0.63**
f_{schr}	0.41	0.07	**0.51**	**0.60**
G'_L	0.25	0.28	**0.64**	0.37
LG'_L	0.25	0.28	**0.64**	0.38
LG'	0.39	0.23	0.41	**0.69**
LG'_E	0.40	0.20	0.20	**0.69**
STI	-0.33	-0.13	**-0.83**	-0.35
C_{80}	-0.25	-0.22	**-0.75**	-0.27
C_{50}	-0.29	-0.20	**-0.76**	-0.29
DRR	-0.45	-0.06	**-0.78**	-0.43
f_c	-0.36	0.09	**-0.69**	-0.29
LF	0.42	0.06	**0.67**	0.42
LF_E	0.45	-0.02	0.50	0.44
$IACC$	-0.35	0.07	**-0.51**	-0.38
$IACC_E$	-0.31	0.06	-0.40	-0.33
T_c	0.21	-0.10	**0.78**	0.16
EDT	-0.11	-0.04	**0.56**	-0.12
RT_{20}	-0.30	0.02	0.31	-0.26
Br	0.28	0.06	-0.04	0.22
BR	-0.17	-0.26	**-0.53**	-0.22
$IACC_L$	-0.08	-0.03	-0.49	-0.12
LF_L	-0.06	-0.12	-0.26	0.03
$F1$	-0.38	-0.03	**-0.78**	-0.31
$F2$	0.28	0.32	**0.47**	**0.56**
$F3$	-0.11	-0.01	**0.54**	-0.15
$\lvert\varepsilon_\theta\rvert$	-			
$\lvert\varepsilon_\phi\rvert$	0.02	-		
R_p	0.21	-0.05	-	
S_{eq}	**0.57**	0.29	0.02	-

TABLE 3.5 – Résultats de l'étude de corrélation entre les attributs perceptifs spatiaux de sources (erreurs absolues de localisation en azimut $|\varepsilon_\theta|$ et élévation $|\varepsilon_\phi|$, distance perçue R_p et taille apparente de source S_{eq}) et les différents paramètres et facteurs acoustiques étudiés en section 3.2. Les valeurs représentent les coefficients de corrélation de Pearson. Les valeurs ayant une p-value $p < 0.001$ sont indiquées en gras.

soumis à un effet important de la salle sont retenus comme dimensions pertinentes de l'espace de projection. Par ordre de significativité et de taille d'effet, les attributs retenus comme dimensions de cet espace sont : (1) la distance perçue R_p, (2) la taille apparente S_{eq}, (3) l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$. Ainsi, les individus du corpus peuvent être représentés dans l'espace porté par ces trois grandeurs, comme l'illustre la figure 3.8.

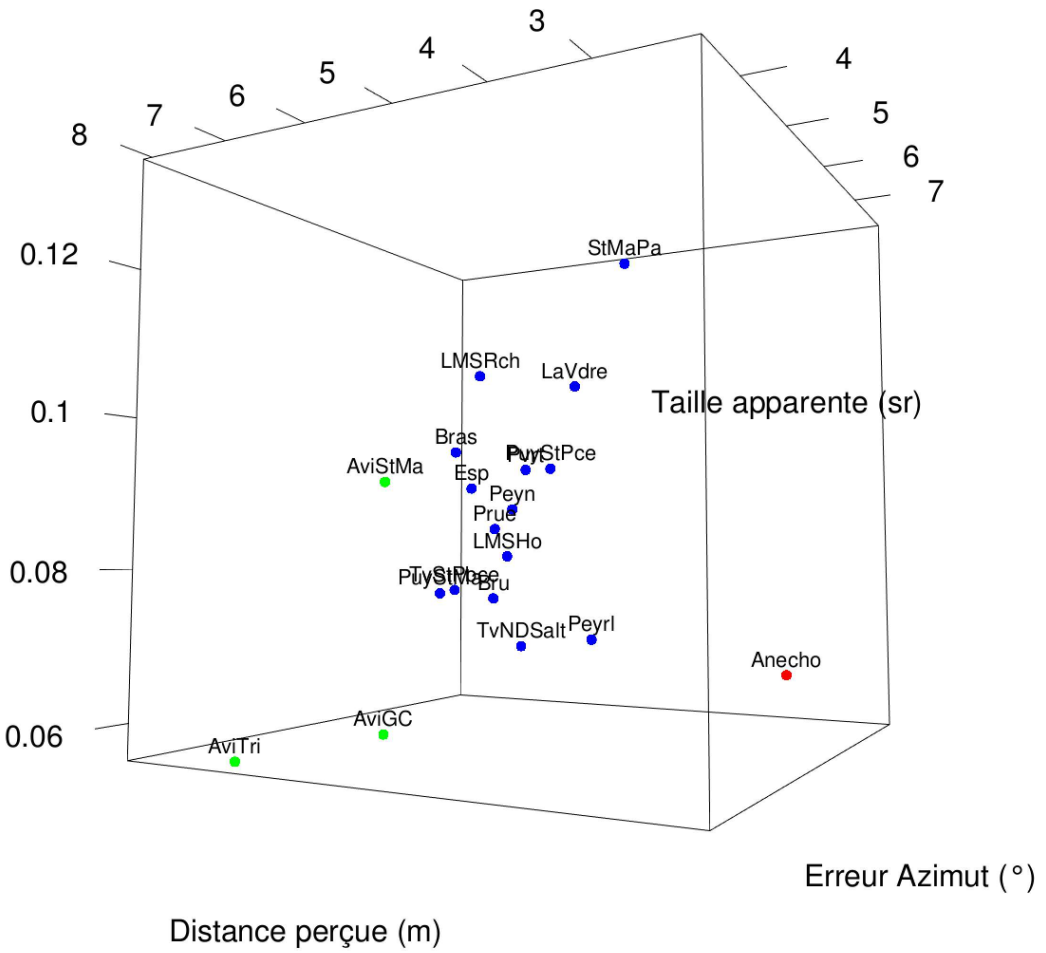


FIGURE 3.8 – Projection des individus du corpus dans un espace 3D représentatif de la perception spatiale des sources en environnement réverbérant. Les trois dimensions représentées sont la distance perçue R_p, la taille apparente de la source S_{eq} et l'erreur absolue de localisation en azimut $|\varepsilon\theta|$. Les points représentent les valeurs moyennes de chaque individu sur l'ensemble des participants et des positions d'écoute. La couleur des points indique la collection à laquelle appartient chaque individu (bleu : corpus "Sésames", vert : corpus "Avignon", rouge : chambre anéchoïque).

A partir de cet espace 3D, une classification hiérarchique des édifices, basée sur le calcul des distances inter-individus, est réalisée et présentée en figure 3.9. Cette

classification nous renseigne sur la façon dont peuvent être regroupés les individus du corpus, d'après leurs ressemblances ou dissemblances en termes d'impact sur la perception spatiale des sources.

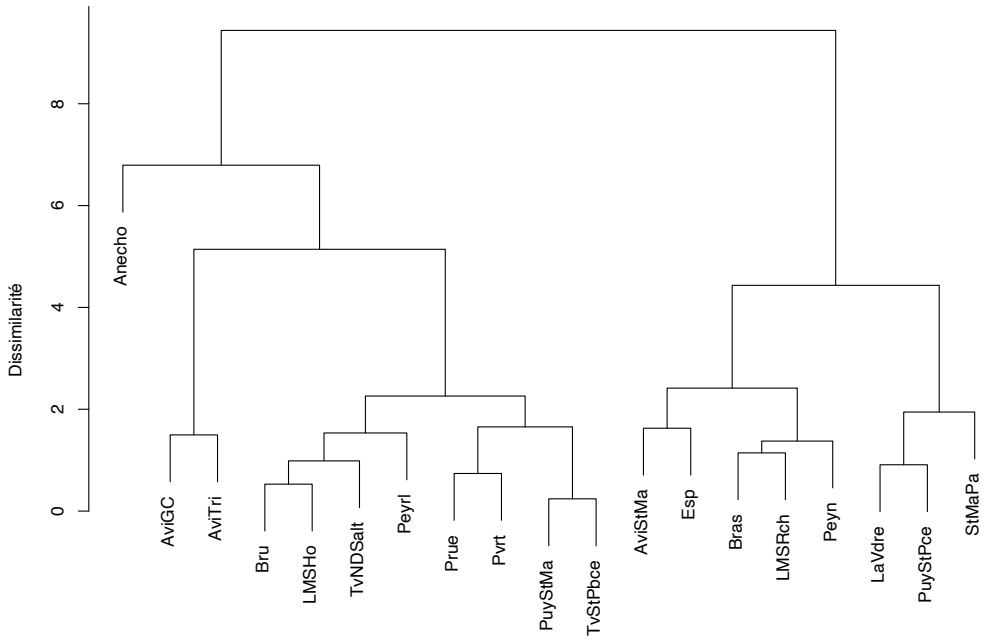


FIGURE 3.9 – Dendrogramme représentant la classification hiérarchique des individus du corpus à partir de l'espace perceptif porté par les dimensions R_p, S_{eq}, $|\varepsilon\theta|$ (*c.f.* figure 3.8).

3.3.3 Discussions

3.3.3.1 Lien entre attributs perceptifs spatiaux de source et paramètres acoustiques

D'après l'étude de corrélation entre les attributs perceptifs de sources reportés et les paramètres acoustiques calculés, un lien fort entre la distance reportée R_p et les indices de clarté et d'intelligibilité est révélé. Ce lien s'exprime également par la forte corrélation de la distance reportée avec le facteur F1 "Précision", porté en grande partie par ces grandeurs STI, DRR, C_{50}, C_{80}, représentant pour la plupart un rapport d'énergie entre le champ direct ou précoce et le champ diffus de la salle. Ce résultat est cohérent avec ceux de la littérature sur la perception de la distance [Kolarik *et al.*, 2016], indiquant que le rapport d'énergie direct/réverbérant

(DRR) est un indice fort pour l'évaluation de la distance d'une source. Une forte corrélation entre la distance perçue et le temps central T_c est également observée. Or cette grandeur est aussi reconnue comme étant un bon estimateur de la clarté d'une source sonore [Bradley, 2011], elle-même fortement dépendante de la distance source-récepteur. On note aussi une corrélation de -0.69 entre la distance reportée et le barycentre spectral f_c. Ce résultat va également dans le sens de plusieurs études montrant que le jugement de la distance peut être influencé par le contenu spectral d'un stimulus [Little *et al.*, 1992, Blauert, 1997b]. Une corrélation positive de 0.64 est également observée entre la distance reportée et les forces tardives omnidirectionnelles G'_L et latérales LG'_L de la salle, ce qui pourrait aussi expliquer la corrélation significative de R_p avec le facteur F2 "niveau". Ces grandeurs sont relativement insensibles à la configuration spatiale de la source et de l'auditeur dans une salle et donc ne donne a priori pas d'information sur la distance source-auditeur. Toutefois, à distance égale, une salle présentant une force sonore tardive élevée devrait avoir un rapport d'énergie direct sur réverbérant faible par rapport à une salle de force sonore tardive faible. Or un faible DRR est perceptivement associé à une source éloignée, ce qui pourrait expliquer le lien qui existe entre la distance reportée et ces deux paramètres acoustiques. Pour finir, la distance perçue est significativement corrélée au facteur F3 "Réverbérance", suggérant que le jugement de la distance peut être influencé par la quantité de réverbération d'une salle. Un effet de la durée de réverbération a d'ailleurs été mis en évidence dans plusieurs études [Mershon *et al.*, 1989, Altmann *et al.*, 2013], à savoir qu'à distance source-auditeur égale, plus la salle est réverbérante plus la source sera perçue comme distante.

Une corrélation de 0.57 est observée entre l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$ et la taille apparente S_{eq} reportée par les participants. Ce résultat va dans le sens des observations reportées pour l'expérience 1, à savoir que la taille de la forme tracée par les participants est plutôt représentative de la précision de localisation, dans des conditions d'auralisation HOA sur notre réseau sphérique de haut-parleurs. Un axe représentatif de la précision de localisation est tracé sur la figure 3.10, illustrant cette corrélation, . Les individus en bas à gauche de la figure sont ceux pour lesquels la précision de localisation est bonne (faible erreur de localisation en azimut et petite taille de source reportée), tandis que les individus en haut à droite sont ceux pour lesquels la précision de localisation est mauvaise (erreur de localisation en azimut élevée et grande taille de source reportée).

D'autre part, un grand nombre de travaux sur la perception de la taille apparente de source (ASW) ont révélé que celle-ci serait principalement influencée par les deux indices acoustiques spatiaux $IACC_E$ et LF_E [Okano *et al.*, 1998, Mason *et al.*, 2005, Käsbach *et al.*, 2014, Wang *et al.*, 2020]. Or ce lien de corrélation entre la taille du tracé S_{eq} et ces deux indices objectifs n'est pas observée dans notre étude. Ces premières remarques conduisent à penser que la difficulté que rencontrent les sujets pour localiser précisément les sources,

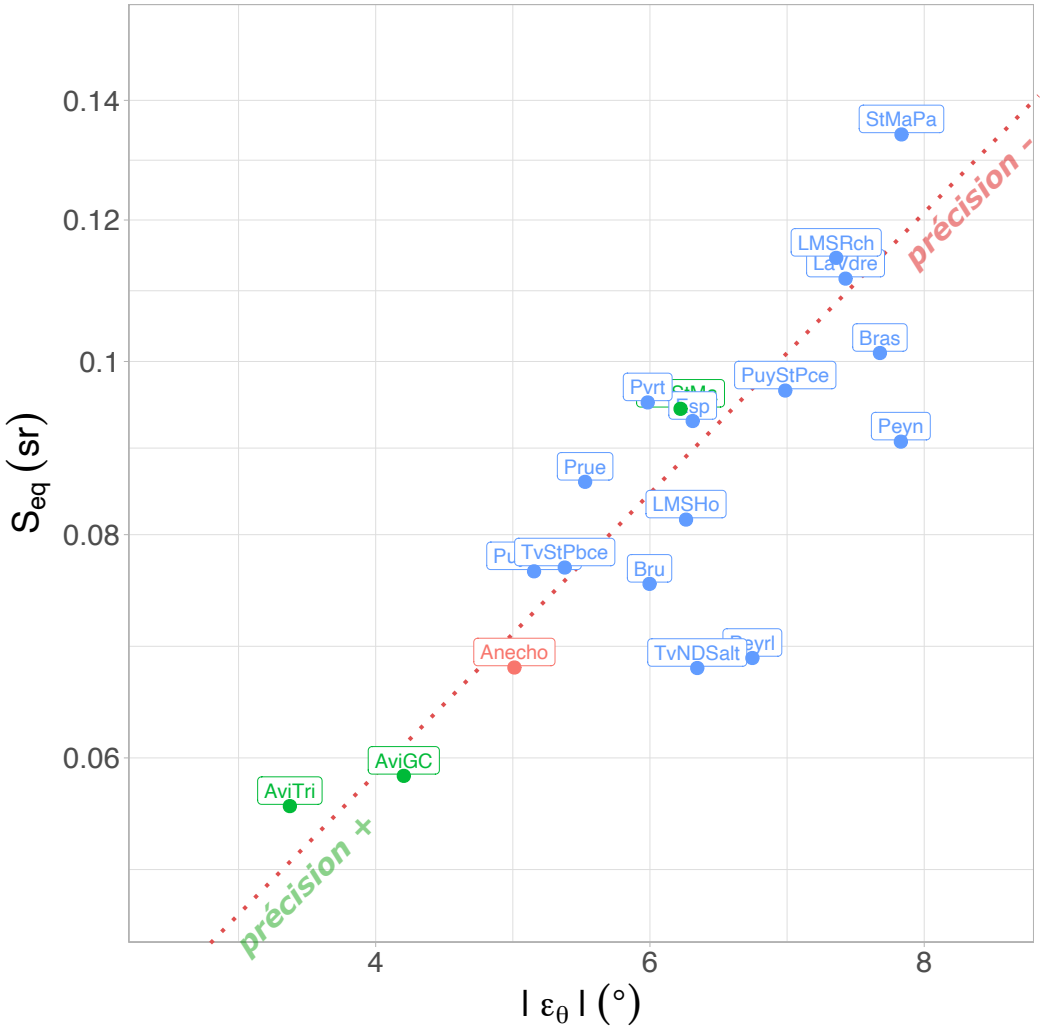


FIGURE 3.10 – Réponses moyennes des sujets pour chaque individu du corpus en fonction de l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$ et de la taille de source reportée S_{eq}. L'axe pointillé représente la notion de précision de localisation. Le code couleur des points indique l'appartenance de chaque individu à une collection (Bleu : Corpus "Sesames", Vert : Corpus "Avignon", Rouge : corpus "Anecho").

du fait de l'auralisation ambisonique, prend le pas sur leur capacité à évaluer la taille apparente effective des sources et opère à un masquage ou a une dégradation des indices $IACC_E$ et LF_E. On observe en revanche une corrélation significative entre la taille reportée et les grandeurs acoustiques de niveau : LG', LG'_E, f_schr, G' et G'_E ainsi qu'avec le facteur F2 "Niveau". Deux interprétations peuvent être envisagées pour tenter d'expliquer ce lien. La première explication est que le flou de localisation ou la difficulté à localiser une source en environnement réverbérant, représenté par S_{eq}, serait réellement largement influencée par la force sonore et principalement par ses contributions latérales (traduites par LG' et LG_E). Un

tel lien pourrait être expliqué par le fait que de fortes contributions énergétiques de la salle notamment latérales, pourrait avoir tendance à masquer les indices de localisation de source porté par le champ direct de la réponse. Cependant, ce lien n'a a notre connaissance jamais été rapporté dans la littérature sur le sujet. Ce qui nous conduit à formuler une seconde seconde hypothèse qui est que la qualité et la précision spatiale de l'auralisation ambisonique de la mesure est impactée par la force sonore de la salle à auraliser. Ainsi, la précision spatiale de l'auralisation serait d'autant plus dégradée que le niveau notamment latéral LG' de la salle est élevé. Une solution simple pour investiguer davantage la question de la dépendance de la qualité d'une restitution ambisonique aux caractéristiques acoustiques de la salle auralisée serait de comparer les mesures réalisées in-situ dans l'ensemble du corpus, avec une mesure de leur restitution par notre système d'auralisation, sur la base des descripteurs acoustiques présentés en section 3.2.

Pour finir, si l'on s'intéresse à l'adéquation entre les attributs perceptifs mesurés et l'espace acoustique constitué des trois facteurs $F1$, $F2$ et $F3$, il semble que, bien que des corrélations fortes entre les dimensions perceptives et acoustiques aient été observées (e.g. R_p vs. $F1 : 0.78$), la mise en correspondance de ces deux espaces ne soit pas pertinente. En effet, le manque de corrélation forte entre les erreurs de localisation et les dimensions acoustiques montre clairement une correspondance faible entre ces deux espaces. A l'inverse, la corrélation significative de la distance perçue avec chacune des trois dimensions de l'espace acoustique indique que l'information relative à l'évaluation de la distance est répartie entre ces trois facteurs. Enfin, il est difficile de trancher sur l'origine du lien observé entre la taille de source reportée S_{eq} et le facteur F2 représentatif du niveau acoustique. Nous pouvons alors conclure que les deux espaces décrivent des phénomènes différents et sont donc complémentaires pour la caractérisation du corpus à l'étude.

3.3.3.2 Discussions sur la caractérisation perceptive du corpus

L'objectif premier de cette étude était de proposer une caractérisation perceptive des édifices du corpus, en étudiant la manière dont ils modifient notre perception des attributs spatiaux de sources sonores. D'après les résultats de l'analyse de variance, trois dimensions perceptives sont sensibles à l'acoustique, à savoir la distance R_p et la taille apparente reportées S_{eq} et dans une moindre mesure l'erreur de localisation en azimut $|\varepsilon_\theta|$. L'étude de corrélation montre que l'évaluation de ces deux premiers attributs de source peut être reliée à des grandeurs physiques caractéristiques de l'acoustique des lieux. De plus, une corrélation significative entre S_{eq} et $|\varepsilon_\theta|$ indique une certaine redondance entre ces deux attributs, tous deux représentatifs dans le cas présent de la difficulté à localiser précisément la source sonore. Pour ces deux raisons, l'espace perceptif 3D représenté en figure 3.8 peut être réduit à ses deux premières dimensions R_p et S_{eq}, tel qu'illustré en figure 3.11.

D'après cette projection et en s'appuyant sur la classification hiérarchique et

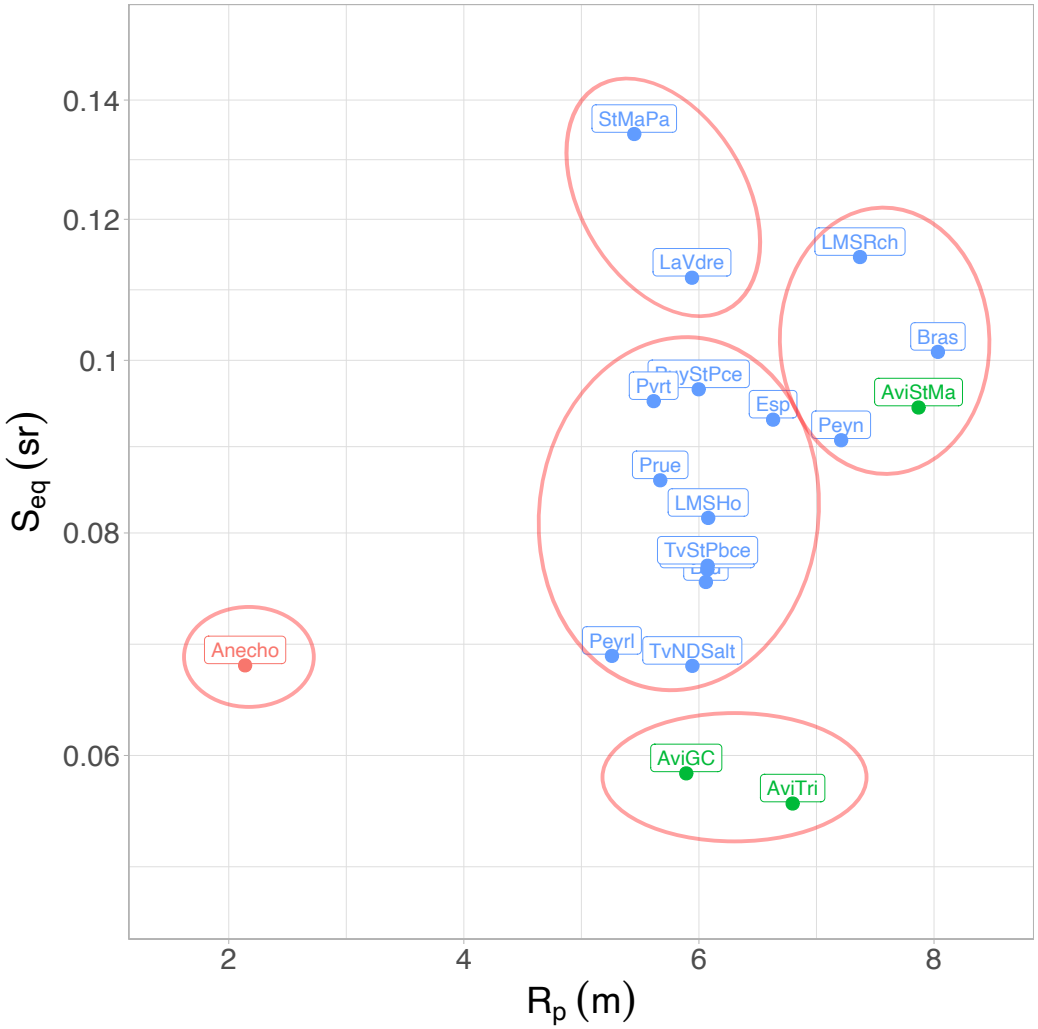


FIGURE 3.11 – Représentation des édifices des différents corpus en fonction de la distance perçu R_p et de la taille apparente reportée S_{eq} des sources en leur sein. Les points représentent les valeurs moyennes sur l'ensemble des participants et des configurations d'écoute. La couleur des points indique l'appartenance aux collections à l'étude (Bleu : Corpus "Sesames", Vert : Corpus "Avignon", Rouge : corpus "Anecho"). Les cercles rouges représentent 5 groupes d'individus en grande partie basés sur la classification hiérarchique, ajustés empiriquement à partir de l'analyse des tests post-hocs conduits sur les différentes dimensions perceptives (*c.f.* tableaux B.2, B.3 et B.4 en annexe B).

les test post-hocs conduits sur chacune des grandeurs étudiées, il est possible de distinguer un certain nombre de groupes d'individus. Le premier groupe constitué du cas anéchoïque se distingue nettement du reste du corpus, avec une distance perçue nettement inférieure à la moyenne de l'ensemble. Cette différence peut être expliquée par le fait que dans le cas anéchoïque, l'évaluation de la distance est principalement

basée sur l'intensité sonore de la source, contrairement aux cas réverbérants où le DRR est un indice accessible et largement exploité pour juger de la distance d'une source.

Un deuxième groupe, formé par les deux grandes salles du Palais des Papes d'Avignon "AviGC" et "AviTri", est identifiable. Pour ces deux édifices, la précision de localisation est globalement meilleure que pour le reste de la collection. Ces deux individus sont très semblables autant sur le plan acoustique (*c.f.* l'analyse acoustique présentée en section 3.2 de ce chapitre), que sur le plan architectural. Ils sont largement plus volumineux que les autres édifices du corpus et ont un temps de réverbération également plus élevé. On peut noter d'après la figure 3.10, que la précision moyenne dans ces deux cas apparaît comme meilleure que dans le cas anéchoïque. Les tests post-hocs conduits sur les attributs représentatifs de la précision de localisation révèlent que cette différence observée entre les grandes salles Palais des Papes et la chambre anéchoïque n'est pas significative. Cet écart de valeurs moyennes peut toutefois être expliqué par le lien qui existe entre distance et taille apparente. En effet, si on fait l'hypothèse que la précision de l'auralisation de ces trois environnements acoustiques est suffisamment bonne pour permettre aux participants d'évaluer la taille effective de la source (identique dans tous les environnements), alors la taille apparente des sources en condition anéchoïque serait globalement plus élevée que dans les deux conditions de grandes salles, du fait que la distance reportée dans le premier cas est nettement inférieure à celle reportée dans les deux autres.

Un troisième groupe formé par les 10 individus inscrits dans le rond central sur la figure 3.11, peut être identifié comme représentatif des individus dont l'influence sur l'image spatiale des sources sonores est moyenne. En effet, la précision de localisation et la distance perçue dans ces acoustiques sont moyennes vis à vis de l'ensemble étudié ici. D'après la caractérisation acoustique, les édifices constituant ce groupe moyen sont également très proches sur le plan acoustique.

Le quatrième groupe, représenté sur la figure 3.11 par le cercle situé en haut à droite, est composé de quatre individus ayant tendance à amplifier la distance perçue des sources en leur sein. Ces édifices sont plutôt des salles de petite taille et fortement réverbérantes. D'après la forte corrélation de la distance perçue avec les indices de clarté et d'intelligibilité ainsi qu'avec les grandeurs de niveau tardif, on peut également dire de ces édifices que de façon générale, ils sont également sujets à une mauvaise intelligibilité ainsi qu'à un niveau d'enveloppement assez élevé. Trois des quatre individus de ce groupe forment également un groupe clair dans l'espace acoustique proposé précédemment (AviStMa, LMSRch, Bras).

Enfin un dernier groupe composé de deux des plus petits édifices du corpus "StMaPa" et "LaVdre" se distingue des autres individus du corpus du fait des particulièrement mauvaises performances de localisations qui y ont été observées. En effet, il semble que dans ces deux édifices, ainsi que dans la salle "LMSRch", la précision de localisation, représentée par l'erreur absolue de localisation en azimut et la taille apparente reportée, soit particulièrement mauvaise. D'après l'étude de corrélation, ces édifices présentent également de fortes valeurs de niveau latéral (LG' et

LG'_E). Or, comme nous l'avons évoqué précédemment, il est probable que la qualité spatiale de l'auralisation ambisonique soit négativement affectée par ces grandeurs.

3.3.4 Conclusions sur l'expérience 2

Pour résumer, l'expérience perceptive présentée dans cette section adressait un objectif double basé sur l'hypothèse que l'environnement acoustique influence notre perception de l'image spatiale des sources sonores. Le premier objectif était de proposer une caractérisation perceptive d'un corpus d'environnements acoustiques mesurés et auralisés. Le second était celui d'investiguer le lien entre la perception et les propriétés acoustiques des salles, afin d'être en mesure de mieux appréhender la perception spatiale des sources dans des conditions acoustiques variées. Pour répondre à ces deux objectifs, le protocole de report en VR des principaux attributs spatiaux de sources (position angulaire, distance et taille apparente), proposé pour l'expérience 1 en chapitre 2, a été appliqué à l'étude d'un corpus de 19 environnements acoustiques mesurés et restitués en laboratoire par auralisation ambisonique 3D à l'ordre 4 sur un réseau sphérique de 42 haut-parleurs.

L'étude révèle que trois des quatre attributs perceptifs observés sont clairement impactés par l'acoustique des lieux étudiés. Les trois attributs en question sont la distance perçue R_p et la taille apparente S_{eq} des sources et dans une moindre mesure l'erreur absolue de localisation en azimut ε_θ. Une corrélation importante entre l'erreur ε_θ et la taille apparente S_{eq} a été révélée poussant à interpréter ces deux grandeurs comme représentatives d'un même effet perceptif qui est celui de la précision de localisation. En représentant la perception moyenne de chaque environnement acoustique de la collection dans l'espace formé par ces trois dimensions perceptives d'intérêt, une cartographie perceptive de ces environnements a été proposée, à partir de laquelle il a été possible d'identifier un certain nombre de groupes d'individus semblables dans leur façon d'influencer notre perception de l'image spatiale des sources sonores. Un des points forts de cette méthodologie est qu'elle propose une méthode de jugement et un espace perceptif absolue (sans comparaison direct) permettant a posteriori de comparer entre eux les édifices d'une collection, ce qui permet d'envisager la caractérisation perceptive d'une collection plus vaste ou du moins différente de celle étudiée ici et de pouvoir la comparer sur une base commune les résultats d'études ayant eu recours à cette méthodologie. Elle ne se cantonne pas non plus à l'étude d'environnements mesurés mais peut également être appliquée pour investiguer la qualité spatiale d'acoustiques 3D issues de simulations ou de processus de synthèse.

D'autre part, l'analyse des corrélations entre les attributs perceptifs étudiés et les paramètres acoustiques calculés pour les différentes conditions acoustiques du test a permis de retrouver un certain nombre de liens connus de la littérature entre la perception spatiale des sources et les propriétés acoustiques des salle et d'en mettre en évidence d'autres qui n'ont à notre connaissance jamais été formulés. Une

forte corrélation a par exemple été observée entre la distance perçue et le rapport entre les énergies du champ direct et du champ diffus (DRR) reconnu pour être un indice fort dans le jugement de la distance d'une source sonore. La corrélation importante observée entre la distance perçue et les indices de niveau tardif G'_L etLG'_L est moins intuitive du fait que ces deux grandeurs ne sont pas sensibles à la position de la source. Elle traduit le fait que pour une distance donnée, les salles ayant un niveau tardif plus élevé ont également un DRR plus faible perceptivement associé à une source plus éloignée. Concernant l'attribut de taille apparente de source, le lien entre cette grandeur et les indices spatiaux précoce LF_E et $IACC_E$ rapporté dans plusieurs études n'a pas été observé ici, appuyant l'idée qu'en conditions auralisés il a été difficile pour les participants d'évaluer la taille effective des sources et que la taille reportée est plutôt représentative de la confiance des participants en leur jugement de la position de la source. La corrélation importante de cet attribut avec les paramètres de niveau latéral LG' et LG'_E a permis de soulever l'hypothèse d'une potentielle influence de ces paramètres sur la précision de l'auralisation ambisonique. Pour investiguer davantage cette question, il est proposé en perspectives de réaliser des mesures de l'auralisation des différents environnements à l'étude et de voir : (1) parmi les paramètres acoustiques calculés, lesquels sont impactés par le processus d'auralisation, (2) parmi les environnements acoustiques, lesquels sont impactés par ce processus.

Enfin, la mise en relation des espaces acoustiques et perceptifs présentés dans les sections précédentes est une perspective qui sera abordée en conclusion de ce chapitre.

3.4 Conclusions et perspectives

Ce chapitre fait état d'un travail conséquent allant de l'élaboration d'un protocole de mesures groupées de données architecturales et acoustiques puis au déploiement de ce protocole dans près d'une vingtaines d'édifices, dans le cadre du projet ANR Sésames, jusqu'à la caractérisation acoustique et perceptive des données collectées. Les deux caractérisations proposées dans ce chapitre se présentent sous la forme des cartographies des édifices du corpus dans des espaces multi-dimensionnels, l'un représentatif de la diversité acoustique des mesures réalisées tout au long du projet, l'autre caractéristique de l'influence de l'environnement acoustique sur la perception spatiale des sources sonores (en contexte d'auralisation). L'espace acoustique est bâti sur l'analyse des individus à travers le calcul de 23 paramètres acoustiques de réverbération, d'intelligibilité, d'énergie et d'espace. A partir de l'analyse factorielle de cet ensemble de données, trois facteurs principaux permettant d'expliquer plus de 80 % de la variabilité acoustique observée ont pu être identifiés. Un premier facteur relatif à la précision, la clarté et l'intelligibilité de source en environnement réverbérant met avant tout en exergue les variabilités intra-individu, c'est-à-dire les différences entre les positions de mesures au sein des édifices étudiés. Le deuxième

facteur est représentatif du niveau acoustique observé dans la salle et nous renseigne principalement sur le rapport entre le temps de réverbération et le volume des lieux. Un troisième facteur de réverbérance discrimine les individus selon la durée de réverbération. Les deux derniers facteurs permettent de bien représenter la variabilité acoustique inter-individus, indépendamment des positions de mesures. De son côté, l'espace perceptif issu de l'expérience 2 est caractéristique des performances de localisation au sein des différents édifices à l'étude et plus particulièrement en termes de distance perçue et de précision de localisation des sources dans ces environnements. L'étude de corrélation entre cet espace et les grandeurs acoustiques mesurées, a permis d'une part de retrouver des indices acoustiques importants dans le jugement de la distance de la source et d'autre part de mettre en évidence une difficulté potentielle du système d'auralisation ambisonique à reproduit précisément certains environnements acoustiques, ayant pour points communs d'être des petits volumes et de renvoyer une grande quantité d'énergie latérale.

En outre, l'étude de corrélation s'est également intéressée aux liens entre les espaces acoustiques et perceptifs et nous permet de conclure que bien que ces espaces ne soient pas orthogonaux, ils permettent bel et bien d'observer la collection d'environnements acoustique sous des angles différents. Des groupes d'individus communs aux deux types de caractérisation ont été mis en évidence. On note par exemple qu'un groupe composé des grandes salles du Palais des Papes d'Avignon se distingue autant du point de vue acoustique que perceptif. Sur le plan acoustique il s'agit d'édifices de longue réverbération et de faible niveau acoustique. Du point de vue perceptif, cela ce traduit par des salles dans lesquelles, pour les positions de mesures observées (distance source-auditeur comprise entre 2 et 6m), la précision de localisation est significativement meilleure que dans le reste du corpus. Un autre groupe commun aux deux caractérisations est formé par trois à quatre individus qui, du point de vue acoustique sont fortement réverbérants et de niveau acoustique élevé, ce qui se traduit par une distance perçue élevée et une mauvaise précision de localisation. Enfin les deux espaces présentent un groupe majoritaire d'individus proches sur le plan acoustiques et semblables sur le plan perceptif. L'étude perceptive a en revanche permis d'identifier un groupe d'individus qui, bien qu'indissociables du groupe majoritaire sur le plan acoustique, se distinguent sur le plan perceptif de part une précision de localisation particulièrement mauvaise.

Ces travaux ouvrent de nombreuses perspectives de recherches, orientées par les deux thématiques de ce chapitre, à savoir (1) l'étude de la perception des environnements acoustiques 3D, (2) la caractérisation multi-échelle de corpus d'édifices patrimoniaux, . Concernant l'étude de la perception en environnements acoustiques 3D, la méthodologie d'investigation de la perception des attributs spatiaux de sources initialement proposée lors de l'expérience 1, s'est également montrée efficace pour mettre en évidence des différences perceptives sur un large ensemble d'environnements acoustiques auralisés. Par rapport à l'expérience 1, de nouvelles hypothèses concernant l'origine des dégradations induites par la restitution ambisonique ont pu être formulées, notamment le fait que la précision spatiale de la restitution pourrait

dépendre des propriétés acoustiques de l'environnement que l'on souhaite auraliser. Comme mentionné précédemment, une comparaison acoustique entre les mesures réalisées in-situ et les mesures des auralisations sur notre système semble maintenant indispensable pour investiguer d'avantage cette question. La diversification des conditions acoustiques à l'étude est également une piste intéressante pour tenter de caractériser plus largement la perception spatiale de l'auralisation de sources sonores en environnements réverbérants. Enfin, alors que nous nous sommes intéressés jusqu'à présent à l'étude de la perception auditive spatiale, les données spatiales et visuelles collectées par nos collègues architectes et prenant la forme de nuages de points et de photos panoramiques de l'intérieur des lieux sont des ressources précieuses permettant d'envisager d'étudier la perception multi-sensorielle de ces mêmes environnements acoustiques en contexte d'immersion 3D. Dans ce sens, nous proposerons dans le prochain chapitre d'exploiter ces différents jeux de données (spatiales, visuelles et acoustiques), afin de questionner la perception de l'adéquation entre environnements acoustiques et représentations visuelles. Concernant la caractérisation de corpus, les travaux présentés ici s'inscrivent dans une démarche de caractérisation plus large, prenant également en compte des données architecturales collectées par nos collègues architectes. Un travail de croisement de ces données avec les données acoustiques et perceptives produites ici est envisagé pour prétendre à une meilleure compréhension du patrimoine architectural. Pour aller plus loin dans la caractérisation perceptive du corpus, le recours à d'autres méthodologies expérimentales est une piste à creuser. Par exemple, la réalisation d'un test de dissemblance sur l'ensemble du corpus semble être intéressante, permettant de mettre en évidence des différences perceptives entre les individus à partir d'un jugement plus global. Enfin, les outils et méthodologies présentées ici forment un base originale et exploitable pour la caractérisation d'autres collections d'environnements acoustiques. Néanmoins, une réflexion doit être entretenue autour des protocoles d'acquisition in-situ, des outils d'auralisation et des protocoles expérimentaux afin d'adapter au mieux ces éléments à des cas d'études plus variés.

Chapitre 4

Etude de la perception visuo-auditive d'environnements acoustiques virtuels

Sommaire

Ce chapitre décrit une expérience perceptive multimodale, s'intéressant à la perception visuo-auditive d'environnements acoustiques. Les différents travaux effectués en amont de cette étude sont également présentés. Dans un premier temps, le contexte d'étude est formulé à travers les motivations et les problématiques. Ensuite, le travaux préalables à la réalisation de l'étude seront explicités, en particulier, le choix du corpus et la modélisation des rendus visuels à l'étude. Enfin, la méthodologie, les résultats et les conclusions de l'expérience seront présentés.

4.1 Contexte d'étude

L'étude de la perception visuo-auditive d'environnements acoustiques est un sujet d'intérêt tant sur le plan fondamental que sur le plan applicatif. Du côté fondamental, bien que la perception des acoustiques de salles ne soit pas un sujet d'étude récent, le processus perceptif multisensoriel est encore assez méconnu. L'essor des technologies du virtuel, offre des possibilités d'étude jusqu'alors difficilement envisageables, notamment des conditions d'expérimentations contrôlées en laboratoire et la capacité d'étudier des conditions audio-visuelles non-congruentes. Du point de vue applicatif, une meilleure connaissance des interactions visuo-auditives dans la perception d'environnements acoustiques aurait des retombées directes dans les différents secteurs de la création immersive tels que le cinéma, le jeu vidéo ou encore le design et la simulation architecturale. En effet, un enjeu commun à ces champs d'application est d'être capable de créer des environnements audio-visuels virtuels qui soient cohérents pour les utilisateurs. Ceci étant dit, l'étude menée dans ce chapitre est animée par plusieurs problématiques liées à la perception visuo-auditive d'environnements acoustiques, présentées dans les deux prochaines sections.

4.1.1 Perception multimodale de salles en environnements virtuels

L'étude de la perception des espaces est, depuis la deuxième moitié du 20ème siècle, un thème de recherche important de la psychophysique. Elle est le plus souvent abordée soit à travers le prisme de la perception auditive, soit celui de la vision. Pourtant, comme l'avait déjà remarqué Gibson dans son approche écologique de la perception visuelle [Gibson, 1979], notre perception de l'environnement est multimodale par essence. Depuis, de nombreuses études ont révélé des interactions entre les modalités auditive et visuelle dans le processus perceptif, [Regan et Spekreijse, 1977, Macdonald et McGurk, 1978, Beerends et De Caluwe, 1999]. La perception de l'espace ne déroge pas à cette règle. Un des effets les plus célèbres de cette interaction dans l'espace est l'effet "ventriloque" [Thurlow et Jack, 1973], largement utilisé au cinéma. Il décrit la fusion dans l'espace d'une source sonore avec une source visuelle lorsque leur contenu est cohérent, même lorsque ces deux sources sont physiquement géographiquement séparées.

Perception visuo-auditive des salles

L'étude de la perception multimodale des environnements acoustiques, en revanche, est un sujet relativement nouveau, facilité par le développement des technologies du virtuel depuis la fin des années 90s. De nombreux travaux de recherche ont depuis été menés. Maempel et Jentsch ont par exemple réalisé une étude visant à quantifier l'impact des modalités visuelles et auditives et les interactions visuo-auditives sur le jugement de la distance source-auditeur et de la taille de la salle [Maempel et Jentsch, 2013] . Pour ce faire, un plan d'expérience comportant

des conditions unimodales, audio et visuelles, et bimodales (visuo-auditives) congruentes et non-congruentes a été réalisé. Les auteurs concluent que la distance source-auditeur est principalement évaluée sur des critères acoustiques et que l'évaluation de la taille de la salle est plutôt guidée par les informations visuelles. Il n'ont toutefois pas pu conclure sur la présence d'une interaction entre les deux modalités. Toujours d'après cette étude, en demandant au sujet de juger la longueur (L), largeur (l) et hauteur (h) ainsi que la taille globale des environnements, il a été montré, en comparant la taille perçue directement reportée et le volume perçu, calculé indirectement à partir des 3 dimensions reportées (L*l*h), qu'il existait une relation cubique entre ces deux grandeurs. Ainsi, il semblerait que la taille d'un environnement 3D soit plutôt représentée par une grandeur 1D correspondant à une valeur moyenne des trois dimensions principales, plutôt qu'un concept en trois dimensions.

Une influence du retour visuel sur la perception de la distance source-auditeur a été depuis révélée dans plusieurs études [Larsson *et al.*, 2001, Gorzel *et al.*, 2012, Postma et Katz, 2017a]. [Valente et Braasch, 2010] reportent quant à eux un effet de l'environnement visuel sur la perception de la taille apparente de source (ASW) et du sentiment d'enveloppement (LEV) en environnements réverbérants. En demandant aux participants de régler la quantité de champ direct et de champ diffus de la simulation acoustique, en accord avec les différentes conditions visuelles (différentes distances source-auditeur), ils ont également montré que le rapport champ-direct/champ-diffus (D/R) était toujours sur-estimé par rapport à la réalité. En 2020, dans un travail de thèse de Master, Greif tente de répondre à la question : "Peut-on entendre la forme d'une salle de concert ?", à travers une expérience multimodale [Greif et Ackermann, 2020]. Une modélisation acoustique et visuelle de 4 salles de concerts de formes différentes ("Shoebox", "Fer à cheval", "Eventail" et "Vineyard"), avec des caractéristiques géométriques (volume) et acoustiques (RT) équivalentes, a été réalisée. L'environnement visuel et l'interface de test étaient présentés via un casque de RV et la restitution de l'environnement acoustique était effectuée en binaural dynamique. Par un choix forcé, les participants devaient sélectionner parmi les 4 acoustiques proposées, celle qui correspondait le mieux à la salle représentée dans le casque. L'hypothèse d'un apprentissage de la tâche à réaliser a également été testée en proposant une phase d'entraînement à la moitié des sujets. L'étude montre qu'en moyenne les sujets n'ont pas été capables d'associer les différentes acoustiques à leur représentation visuelle. Toutefois elle révèle qu'avec un apprentissage, les scores d'association sont légèrement meilleurs, bien que statistiquement toujours proches du hasard. D'après ces résultats, nous ne serions donc pas auditivement sensibles à la forme des salles.

Dans le cadre du consortium SEACEN [1], Maempel s'est également employé à

1. Le consortium SEACEN (Simulation et Evaluation of Acoustical ENvironments), regroupe depuis 2011 des chercheurs de différentes universités allemandes autour des thématiques de la cap-

définir une méthodologie de l'étude de la perception visuo-auditive [Maempel, 2017]. Dans ce document, il fait le constat que dans le domaine de l'étude de la perception multimodale, il existe une grande diversité de protocoles expérimentaux, rendant parfois les études difficilement comparables entre elles. Il propose un cadre théorique permettant de mener au mieux ce genre d'études et présente une plate-forme dédiée à ces expérimentations : "the Virtual Concert Hall".

Influence des environnements virtuels sur la perception des salles

Des efforts ont été fournis pour construire des environnements virtuels audio-visuels propices à la recherche sur la perception visuo-auditive [Seeber *et al.*, 2010, Maempel et Horn, 2017]. Toutefois, le recours aux technologies de virtualisation n'est pas sans biais. Les suédois Larsson, Västfjäll et Kleiner montraient en 2001 [Larsson *et al.*, 2001] que la perception en environnement virtuel audio-visuel diffère grandement de celle en condition réelle, en termes de taille perçue d'une salle et de distance perçue d'une source sonore. Dans un test comparant la perception de ces deux grandeurs dans quatre conditions différentes : Audio seul, Audio + photographie, Audio + Environnement virtuel (EV), Audio + Environnement Réel, les auteurs ont montré que la condition Audio + EV est plutôt comparable à celle obtenue pour les conditions audition seule et audition + Photo. Ces résultats indiquent qu'à l'époque le rendu visuel en VR de la distance et de la taille d'un environnement n'était pas convainquant et que les participants dans la condition virtuelle se sont principalement basés sur des caractéristiques acoustiques pour évaluer ces grandeurs. Plus récemment, Maempel et Horn ont publié une vaste étude intitulée "Audiovisual perception of real and virtual rooms" [Maempel et Horn, 2018]. Les participants devaient évaluer une vingtaine de critères géométriques, esthétiques et de sensation de présence, dans les domaines auditif, visuel et visuo-auditif, soit en condition réelle, soit en environnement virtuel. Le dispositif virtuel optique consiste en une projection stéréoscopique sur un écran courbé occupant un champ de vision de 160° et la restitution sonore est réalisée par synthèse binaural dynamique grâce a un système de Head-Tracking. Les résultats de l'étude montrent une différence significative entre conditions réelle et virtuelle dans l'évaluation des critères géométriques bi-modaux (distance perçue de la source, taille apparente de source, taille perçue de la salle), notamment pour les conditions "visuel seul" et "visuo-auditive". Les auteurs concluent à ce sujet que leur système virtuel optique n'est pas adapté à l'étude de ces grandeurs géométriques. En revanche, le recours au dispositif virtuel visuo-auditive est valide pour l'étude des critères mono-modaux (acoustiques et visuels), tels que la réverbérance (acoustique) ou la luminosité de la salle (visuel).

La plupart des études menées sur le sujet de la perception visuo-auditive d'environnements acoustiques a recours à des technologies similaires de restitution visuelle (TV, écran courbe ou en U) [Valente et Braasch, 2010, Postma et Katz, 2017a,

tation, simulation, restitution et perception d'environnements acoustiques. Hans-Joachim Maempel y est responsable de la recherche sur la perception visuo-auditive des environnements acoustiques [Lindau *et al.*, 2014b].

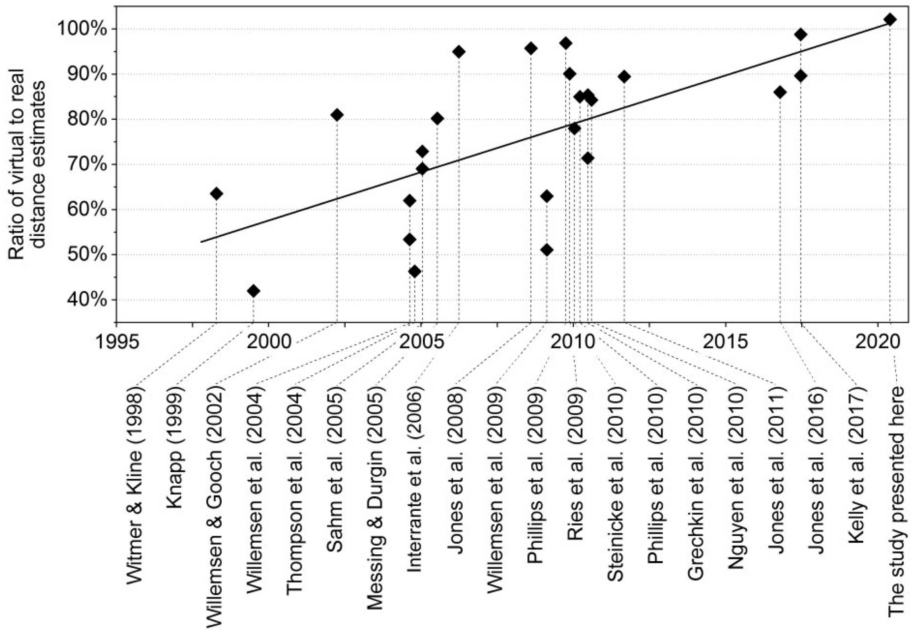


FIGURE 4.1 – Distance perçue en environnement virtuel par rapport à la distance perçue en condition réelle, en fonction de différentes études, ordonnées par ordre chronologique. Les études présentes sur cette figure ont toutes été réalisées avec un casque de RV [Feldstein *et al.*, 2020].

Postma et Katz, 2017b, Maempel et Horn, 2018]. Pourtant, la technologie dominante en termes d'environnement virtuel optique est le casque de réalité virtuelle. Or, un article de 2020, comparant les résultats obtenus dans différentes études sur la perception visuelle de la distance en environnements virtuels et réels, révèle que les études récentes à ce sujet montrent peu voire aucune distorsion de la distance perçue en RV par rapport à la réalité [Feldstein *et al.*, 2020]. La figure 4.1 représente le rapport de la distance perçue via un casque de RV sur la distance perçue en condition réelle, selon plusieurs études classées chronologiquement. Ce résultat suggère que les performances des casques de VR en termes de restitution de la géométrie (notamment la distance et la profondeur), ont nettement progressé au cours de ces dix dernières années, rendant aujourd'hui favorable leur utilisation dans l'étude de la perception visuo-auditive d'environnements acoustiques. Ce constat est également partagé par [Vorländer *et al.*, 2015] dans un article présentant une vue d'ensemble de la VR pour l'acoustique architecturale et par [Ahrens *et al.*, 2019], dans une étude de 2019 sur les performances de localisation de source dans différentes conditions visuo-auditiveles, réelles et virtuelles.

4.1.2 Motivations et problématiques

De nombreuses études ont révélé l'importance d'étudier la perception des salles dans leur aspect multimodal, au cours des 10 dernières années. La plupart du temps, ces études s'intéressent à l'influence d'une stimulation visuelle sur la perception de grandeurs relatives à l'audition (e.g. paramètres d'acoustique de salle, paramètres de la source sonore). En revanche, la cohérence globale entre une acoustique et sa représentation visuelle a été peu étudiée et présente pourtant un intérêt majeur notamment dans les secteurs du jeu vidéo, du cinéma ou de la simulation architecturale. D'autre part, les stimuli visuels utilisés dans ces études sont soit des images de la réalités re-projetées (photos, vidéos 3D), soit des modèles 3D virtuels des lieux à l'étude. Or, la quantité d'informations importantes tels que la profondeur, la texture et le niveau de détails varie grandement d'un rendu à l'autre. Il nous parait donc important de connaître l'influence du choix de ces représentations sur la perception visuo-auditive de l'environnement. Ainsi, l'expérience que nous présentons ici est motivée par deux problématiques :

- Sur quels attributs acoustiques et visuels nous-basons nous pour juger de la cohérence visuo-auditive d'environnements acoustiques ?
- Dans quelle mesure le type de rendu visuel influence-t-il ce jugement ?

4.2 Choix et construction du corpus

Pour tenter de répondre à ces deux questions, il est nécessaire de constituer un corpus d'environnements acoustiques et visuels pertinent. Un travail de modélisation des environnements visuels est également décrit dans la présente section.

4.2.1 Choix du corpus

Les environnements acoustiques et visuels à l'étude ici sont issus du corpus "Sésames" présenté en section 3.1.3 du chapitre 3. L'une des forces de ce corpus est que, pour chaque individu, des données visuelles (photos 360 aux positions de mesures acoustiques et nuages de points de l'intérieur des édifices) ont été collectées en parallèle des mesures acoustiques, le rendant parfaitement adapté à l'étude de la perception visuo-auditive. Une sélection de 5 chapelles du corpus a été effectuée sur la base de leurs temps de réverbération et de leurs volumes. Comme illustré en figure 4.2, les 4 individus extrêmes du corpus en termes de RT et de volume ont été choisis (Peyrl, StMaPa, Bras, Esp). Un 5ème environnement (TvNDSalt) tiré du même corpus a également été sélectionné car il se distinguait des autres de par son architecture complexe. Il est toutefois de temps de réverbération et de volume comparable à celui de Esp. Une réponse impulsionnelle par chapelle est sélectionnée, correspondant à la position micro-source "MC-ec" (cf : figure 3.2 en section 3.1.2 du chapitre 3 : source positionnée au centre de la nef et auditeur à 5,6m de la source).

Les temps de réverbération et volumes des 5 lieux sélectionnés sont donnés en tableau 4.1.

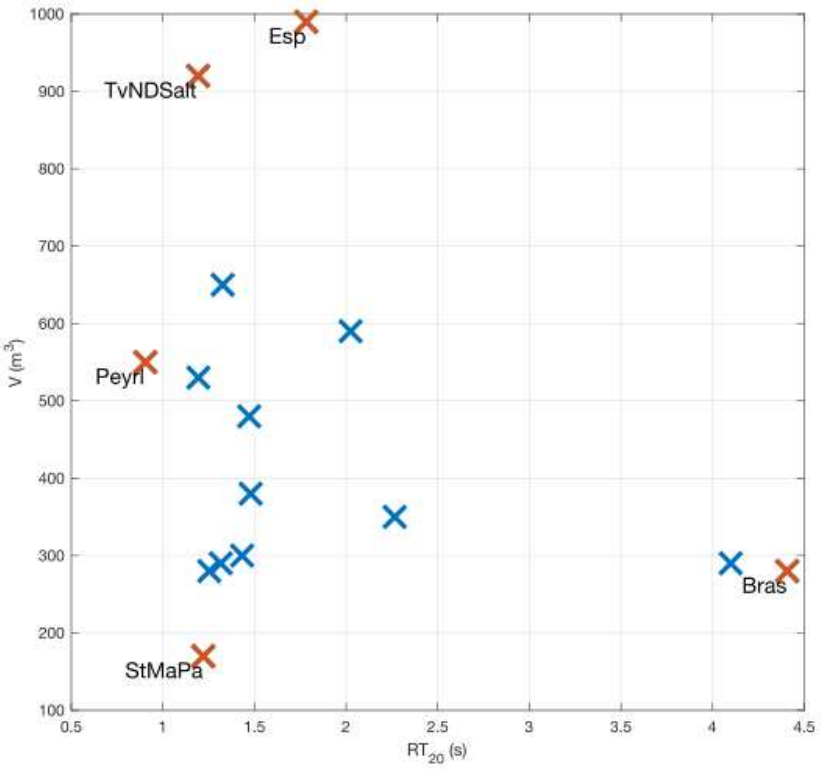


FIGURE 4.2 – Visualisation des 5 environnements sélectionnées pour cette étude, en fonction de leur RT_{20} et de leur Volume V.

	Grandeurs	
Environnements	RT_{20} (s)	V (m^3)
Bras	4.41	280
Esp	1.78	990
StMaPa	1.22	170
TvNDSalt	1.19	920
Peyr	0.90	550

TABLE 4.1 – Temps de réverbération et Volumes des 5 environnements à l'étude.

4.2.2 Choix des représentations visuelles

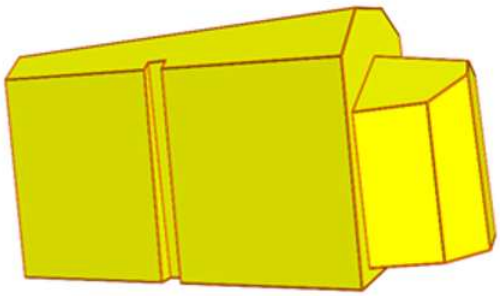
Pour quantifier l'influence du rendu visuel sur la cohérence visuo-auditive perçue d'environnements acoustiques, 3 types de rendus ont été choisis (*cf* figure 4.3). Ils possèdent chacun des caractéristiques différentes relativement aux critères de

profondeurs, de texture et de niveau de détail :

- **Photos 360** : captées à la position de mesure des réponses impulsionnelles sélectionnées. Ces photos panoramiques ne sont pas stéréoscopiques, elles sont restituées par projection sphérique. elles représentent correctement les informations de textures et de couleurs, mais peinent à restituer la sensation de profondeur.
- **Nuages de points** (texturés) : issus d'un relevé photogrammétrique des environnements. Ils contiennent des informations de textures et de couleurs (plus éparses que pour les photos 360), mais également des informations de profondeur (représentation 3D).
- **Modèles simples** : représentant les environnements étudiés sous une forme polygonale simplifiée et de texture uniforme. Ces modèles donnent une représentation grossière de la géométrie des lieux mais ne contiennent pas d'information sur leurs textures.

(a) Photo 360

(b) Nuage de points



(c) Modèle simple

FIGURE 4.3 – Illustration des 3 types de rendus proposés pour l'expérience. Exemple pour l'individu "Bras".

Les photos 360 sont directement issus de la campagne de mesure réalisée dans

le cadre du projet ANR Sésames, en coopération avec des chercheurs du laboratoire MAP (Modèles et simulations pour l'Architecture et le Patrimoine), porteur du projet. Elles ont été systématiquement relevées aux positions des microphones, correspondant également à la position de l'auditeur lors de l'auralisation. De cette façon, une première représentation multimodale des environnements, coïncidente dans l'espace, est obtenue. Il faut noter que la source sonore n'est pas visible sur les photos 360. En effet ces dernières ont été prises une fois les sources sonores retirées de leur pied. D'autre part, la captation ayant été réalisée avec un appareil monoscopique, les photos panoramiques obtenues sont également monoscopiques et représentent donc l'espace en deux dimensions (azimuth, élévation). Pour en proposer une visualisation, ces images 2D sont projetées sur sphère 3D, ne reproduisant pas deux éléments importants pour la perception de la profondeur, à savoir la disparité binoculaire et la parallaxe du mouvement [Howard et Rogers, 2002].

Les nuages de points nous sont également fournis par les chercheurs du MAP. Ils ont été obtenus à partir d'une méthode de relevé photogrammétrique basé sur la captation d'un ensemble de photos panoramiques à l'intérieur du bâti, décrite par [Barazzetti *et al.*, 2018]. Le protocole de captation groupée de données acoustiques, métriques et visuels, développé et déployé dans le cadre du projet Sésames, est présenté dans cet article [Blaise *et al.*, 2021]. Il détaille notamment la méthode d'acquisition photogrammétrique et métrique ainsi que les étapes de traitements permettant la construction à l'échelle de modèles 3D sous la formes de nuages de points. La position du microphone et des sources dans les chapelles ayant également été relevées, il est possible lors de la visualisation de placer précisément le sujet de manière coïncidente à la position d'écoute. Cette fois-ci, les indices de profondeur sont correctement restitués. De plus, le procédé photogramétrique permet également de faire correspondre à chaque point du maillage obtenu une information de couleur, résultant en une représentation raisonnable de la texture à l'échelle globale. Toutefois, étant donnée la nature discrète du nuage de points, le niveau de détail, notamment en termes de texture est plus bas que celui représenté sur les photos 360.

La troisième représentation visuelle choisie pour cette étude est un modèle géométrique 3D simplifié de l'intérieur des édifices et de texture unie. En proposant une condition visuelle représentant la géométrie des lieux de façon grossière et sans informations de texture et de matériaux, il sera possible de vérifier si la cohérence visuo-auditive est plutôt jugée sur une impression visuelle globale et géométrique, ou bien si la présence de détails et de textures influence ce jugement. Toutefois, les données dont nous disposons pour chaque environnement sont les photos 360 et les nuages de points. Une solution consistant à construire ces modèles simples à partir des nuages de point a donc été choisie.

4.2.3 Génération de modèles simples à partir de nuages de points

La reconstruction 3D d'environnements intérieurs à partir de nuages de points est encore aujourd'hui un problème ouvert [Li *et al.*, 2016b]. Parmi les techniques existantes, la solution retenue est celle proposée par [Nan et Wonka, 2017]. Elle permet la reconstruction d'un modèle sous la forme d'un polyèdre composé de surfaces planes et fonctionne en deux étapes : la génération de faces candidates et la sélection des faces candidates. La première étape prend en entrée un nuage de points et consiste à générer un grand nombre de faces candidates à partir de ces données. L'extraction de ces primitives planaires est réalisée par l'algorithme RANSAC [Schnabel *et al.*, 2007], initialement proposé par [Fischler et Bolles, 1981], puis une étape de rafinement, permettant de fusionner des faces partageant un grand nombre de points, est réalisée par un algorithme proposé par [Li *et al.*, 2016a]. Enfin la sélection des faces candidates consiste à choisir un sous ensemble de sections planes décrivant au mieux la géométrie de la scène, par une fonction de minimisation d'énergie. Cette fonction objective de minimisation d'énergie combine trois paramètres d'optimisation des énergies relatives à l'ajustement des données E_f, la couverture ponctuelle E_c et la complexité du modèle E_m. L'énergie associée à l'ajustement des données correspond à la distance euclidienne entre une face et l'ensemble des points qui lui sont associés. Elle représente la confiance que l'on peut avoir en un ensemble de points et est minimal pour des nuages de points non-bruitées. La couverture ponctuelle correspond à la différence d'aire entre la surface de la face candidate et celle représentée par les points inscrits dans ce plan. L'énergie correspondante est minimale lorsque les nuages de points sont homogènes et ne possèdent pas de zones non couvertes (possiblement générées par des occlusions). Enfin, le terme de complexité du modèle a été introduit afin de favoriser les modèles simples avec un nombre limité de faces. La sélection des faces candidates est donc effectuée par une fonction de minimisation de ces trois grandeurs pondérées, donnée en équation 4.1.

$$\min_x \lambda_f \cdot E_f + \lambda_c \cdot E_c + \lambda_m \cdot E_m. \tag{4.1}$$

Les coefficients de pondération λ_f, λ_c, λ_m sont réglables dans l'outil Polyfit. Leurs valeurs typiques pour l'obtention d'un modèle correct sont : $\lambda_f = 0.3$, $\lambda_c = 0.4$ $\lambda_m = 0.3$. Une vérification visuelle est effectuée en superposant le nuage de points sur le modèle obtenu.

Comme l'illustre la figure 4.4, cette méthode ne permet pas de générer des modèles avec des sections courbes, toutefois les voûtes et autres sections courbes de la géométrie exacte des lieux sont approximées par un ensemble de sections planes. Les modèles ainsi obtenues sont simples et donnent une représentation grossière mais correcte du volume réel.

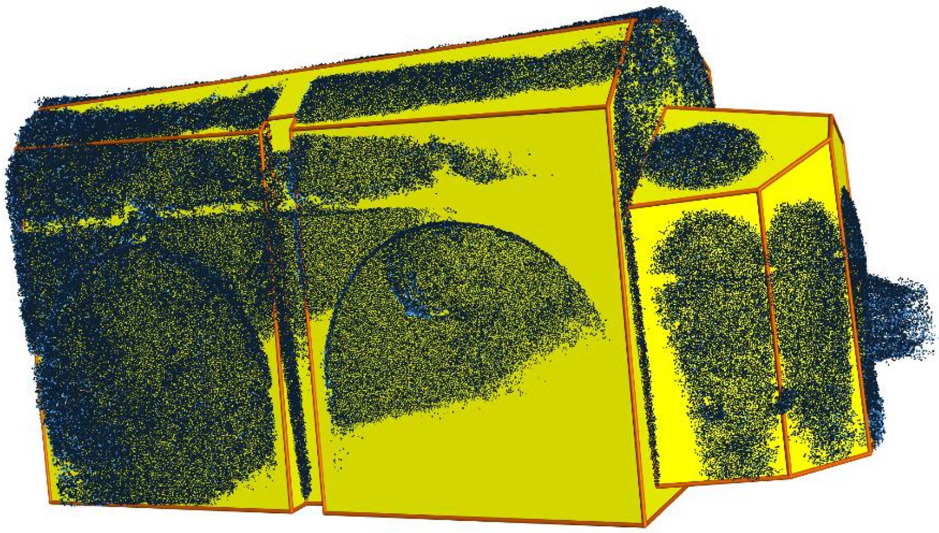


FIGURE 4.4 – Superposition du nuage de points et du modèle simplifié pour la chapelle "Bras".

4.3 Méthodologie

4.3.1 Participants

21 participants normo-entendants (14 hommes et 7 femmes) âgés en moyenne de 30 ans (std : 10,1 ans) ont pris part à cette expérience. L'expertise des sujets en matière d'acoustique des salles est variable et ne constitue pas un facteur d'analyse pour cette étude.

4.3.2 Stimuli

Deux stimuli sonores ont été choisis, en cohérence avec l'usage courant des lieux étudiés (des petites et moyennes chapelles rurales de la région PACA) : un court extrait de parole (5 sec) suivi de quelques secondes de silence, et un extrait de guitare classique (3 min), tous deux enregistrés en condition anéchoïque. Ces deux extraits sont convolués par les réponses impulsionnelles des 5 environnements étudiés, résultant en un total de 10 conditions sonores différentes.

Pour la modalité visuelle, les 3 types de visualisation présentés en section 4.2.2 sont testés (photos 360, nuages de points et modèles simples), pour chacun des 5 environnements visuels sélectionnés, soit 15 conditions visuelles différentes. Dans l'expérience, une condition multimodale correspond à la présentation d'un des 10 stimuli sonores avec un des 15 stimuli visuels. Le plan d'expérience ainsi obtenu est un plan factoriel à 4 facteurs avec au total 150 conditions de test :

- Environnement acoustique (5),
- Environnement visuel (5),
- Type de visualisation (3),
- Stimulus sonore (2),

avec le nombre de niveaux de chaque facteur est indiqué entre parenthèses.

4.3.3 Dispositif expérimental

Pour cette expérience, les participants sont placés au centre du dispositif de spatialisation multicanal du laboratoire (sphère de 42 haut-parleurs). Les acoustiques de salles sont restituées en HOA d'ordre 4 grâce aux outils de la librairie spat5 pour Max8. L'interface de test est une interface en réalité virtuelle, développée dans le moteur de jeu Unity et est restituée via un casque de réalité virtuelle (Oculus Quest). Le déroulement de l'expérience, la présentation des environnements visuels et les retours utilisateurs sont gérés dans Unity tandis que la gestion de l'audio est faite dans Max8. Une communication OSC entre les deux logiciels permet la transmission des informations de stimulus sonore, depuis Unity vers Max8.

4.3.4 Procédure

Les 150 conditions de test sont réparties en 3 sessions (une session par type de représentation visuelle). Dans chaque session, les 50 conditions visuo-auditiveles sont présentées aléatoirement et l'ordre des sessions est également aléatoire, afin d'éviter des biais de présentation. En début d'expérience, une phase d'apprentissage d'environ 5 minutes est prévue, afin de permettre aux sujets de prendre en main le dispositif expérimental et de se familiariser avec la tâche à réaliser. Pour chaque essai, les participants doivent juger de la cohérence entre l'acoustique et le visuel présentés. Pour ce faire, il doivent répondre par "oui" ou par "non" à la question suivante : "Selon vous, le son que vous entendez a-t-il été produit dans ce lieu ?". La réponse est donnée via l'interface utilisateur dans l'environnement virtuel, comme illustré en figure 4.5. Les stimuli sonores sont joués en boucle jusqu'à ce que le participant ait répondu à la question. Il est libre de tourner à 360°, sans contrainte de temps, avant d'émettre son jugement. Chaque session dure en moyenne 15 minutes. Entre les sessions, une pause de 10 minutes minimum est imposée, pour limiter la fatigue liée au dispositif de réalité virtuelle. Les participants sont également invités à prendre une pause ou arrêter l'expérience à tout moment, s'ils le désirent.

4.3.5 Analyses des données

Pour chaque essai du test, la réponse-utilisateur est enregistrée dans un fichier texte sous la forme d'une réponse binaire : 1 pour la réponse "oui", 0 pour la réponse "non". Les réponses des 21 participants sont sommées pour chacune des 150 conditions de test et stockées dans une matrice de dimension 4, correspondant aux quatre facteurs expérimentaux (cf tableau C.1 en annexe C). Pour chaque cellule

FIGURE 4.5 – Capture de l'interface utilisateur dans le casque de VR. A l'aide d'un contrôleur, l'utilisateur peut sélectionner la réponse souhaitée à la question posée à l'écran.

du tableau le taux de positivité est calculé. Il correspond au pourcentage de réponses positives par rapport au nombre de réalisation de la condition considérée.

Une première analyse visant à évaluer le niveau de cohérence perçue entre environnements acoustiques et visuels est réalisée. Cette analyse permet notamment de répondre à la question "est-on capable d'associer une acoustique avec sa représentation visuelle ?". Pour mesurer un effet statistique, ces résultats sont mis en regard de la distribution théorique du hasard, d'un test à réponse binaire. En effet, pour savoir si les participants ont répondu significativement différemment du hasard, un intervalle de confiance associé à l'hypothèse H_0 "Les participants ont répondu au hasard" peut être calculé par la formule suivante :

$$\left[\hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1-\hat{p})}{n}}\right]$$

avec :

— $\hat{p} = 0.5$: probabilité du hasard pour un test binaire,
— $z_{\alpha/2} = 1.96$: score standard pour une confiance de test à 0.95,
— n : taille de l'échantillon, correspond ici au nombre de répétitions d'une condition.

Les valeurs de taux de positivité en dehors de cette intervalle de confiance ne respectent pas l'hypothèse H_0 et sont donc considérées comme statistiquement différentes du hasard ($p_{val} < 0.05$).

D'autre part, l'influence des facteurs "stimulus sonore" et "type de visualisation" est étudiée. Pour mesurer l'effet d'un facteur, les réponses selon la dimension de l'autre facteur sont sommées, résultant en un tableau 3D de dimensions

"Environnement acoustique" × "Environnement visuel" × "Facteur étudié". Un test d'indépendance du χ^2 est ensuite réalisé entre les niveaux du facteur d'intérêt.

Pour aller plus loin et tenter de comprendre sur quels critères les participants se sont basés pour effectuer la tâche, ces résultats sont mis en relation avec les caractéristiques acoustiques et géométriques des lieux à l'étude. Une régression des résultats est réalisée en fonction des temps de réverbération RT et volumes V des environnements du test.

4.4 Résultats et discussions

Les résultats et discussions sont traités en trois sections distinctes. La première section décrit la cohérence visuo-auditive perçue de manière qualitative. La deuxième section traite des effets du stimulus sonore et du type de visualisation. Enfin, la troisième section présente un modèle de régression permettant d'estimer le niveau de cohérence attendu entre environnements acoustiques et visuels à partir de caractéristiques acoustiques et géométriques des lieux.

4.4.1 Etude qualitative de la cohérence perçue entre environnements acoustiques et visuels

La cohérence visuo-auditive perçue est donnée par le taux de positivité τ_+ (rapport du nombre de réponses positives sur le nombre d'essais de la condition). Le tableau 4.2 représente donc la cohérence visuo-auditive de chaque paire acoustique/visuel, sans considération du stimulus sonore et du type de visualisation.

		V				
		Bras	Esp	StMaPa	Peyr	TvNDSalt
A	Bras	14,29%	70,63%	7,94%	23,81%	50,79%
	Esp	46,03%	53,97%	22,22%	52,38%	59,52%
	StMaPa	61,11%	37,30%	45,24%	61,90%	44,44%
	Peyr	62,70%	19,05%	64,29%	42,06%	16,67%
	TvNDSalt	59,52%	12,70%	64,29%	54,76%	19,84%

TABLE 4.2 – Cohérence visuo-auditive perçue (taux de positivité τ_+), pour les paires d'environnements acoustiques (A) et visuels (V). Les cellules en couleurs représentent les résultats significativement différents du hasard. En rouge : les conditions jugées non congruentes, en vert : les conditions jugées congruentes. Intervalle de confiance à 0.95 : $[0.41; 0.59]$.

Dans 8 cas sur 25, les réponses obtenues sont statistiquement proches du ha-

sard. A l'inverse, dans 68% des cas (17 conditions sur 25), les taux de positivité sont statistiquement différents de ceux qui auraient été obtenus par une réponse au hasard. Plus précisément, parmi ces 17 conditions, 8 ont été jugées comme fortement cohérentes (en vert) et 9 ont été jugées comme incohérentes (en rouge). Le valeurs minimales et maximales de cohérence sont respectivement de 7.94% pour la condition visuo-auditivele "Bras-StMaPa" et de 70.63% pour la condition "Bras-Esp".

Discussions

En observant les taux de positivité obtenus dans la diagonale du tableau, correspondants aux conditions visuo-auditiveles congruentes, il est possible de répondre à la question "Sommes-nous capables d'associer un environnement acoustique avec sa propre représentation visuelle ?". On remarque qu'aucune des conditions congruentes n'a été jugée par les participants comme cohérente. Parfois même (dans le cas de Bras et TvNDSalt), elles ont été jugées comme fortement incohérentes, ce qui permet de conclure que même dans des conditions de tests favorables (immersion audio-visuelle 3D), les participants n'ont pas été capables d'associer les différentes acoustiques à leur représentation visuelle. Ces observations sont comparables à celles issues d'une étude multimodale conduite récemment à TU Berlin, intitulée "Can you hear the shape of concert halls ? An audiovisual test in simulated 3D environments" [Greif et Ackermann, 2020]. Toutefois, ce résultat ne signifie pas que les participants ont été incapables de juger de la pertinence entre environnements acoustiques et visuels. Au contraire, dans 68% des cas, le jugement de la cohérence visuo-auditive est significatif, ce qui indique que dans l'ensemble, les sujets se sont basés sur des attributs acoustiques et visuels spécifiques pour évaluer cette cohérence. Cette question est abordée plus en détails en section 4.4.3.

On peut également noter que le taux de positivité ne dépasse pas 70.63%. A l'inverse, les réponses ont été plus unanimes dans le jugement de la non cohérence entre environnements acoustiques et visuels. De fait, 6 réponses sur les 9 jugées non cohérentes ont un taux de positivité inférieur à 20% et le taux de positivité minimum obtenu est de 7.94%. Les explications à ces observations sont multiples. En effet, même si les lieux ont été choisis pour représenter la diversité du corpus dont ils sont extraits - un corpus de petites et moyennes chapelles -, le groupe formé est relativement hétérogène et il n'est pas représentatif de l'ensemble des environnements auxquels nous pouvons être confrontés dans la vie courante. On peut notamment remarquer sur la figure 4.2 que les environnements les plus volumineux sont relativement mats et que l'environnement le plus réverbérant est de faible volume. En revanche, le corpus ne présente pas d'environnement de grand volume et de grand temps de réverbération. D'autre part, les conditions virtuelles d'écoute et de visualisation peuvent avoir joué un rôle dans le plafonnement à 70% du taux de positivité. En effet, on peut supposer que la question de la cohérence visuo-auditive est intimement liée à celle du réalisme. Or, il a été montré en chapitre 2 que le dispositif de spatialisation ambisonique utilisé pour cette étude présentait un certain nombre

biais par rapport à des conditions d'écoutes réelles. Parallèlement, les types de visualisations proposés ici possèdent tous certaines distorsions par rapport à la réalité, discutées en section 4.2.2 du présent chapitre. Il faut aussi noter que la source n'est pas représentée dans l'interface visuelle. Cette absence d'une référence visuelle a été relevée par certains sujets et pourrait avoir un impact négatif sur la sensation de réalisme durant l'expérience. Enfin, il semble plausible que le niveau d'expertise hétérogène des participants, en termes d'acoustique des salles, ait également eu une influence sur ces résultats.

4.4.2 Influences du stimulus et du type rendu visuel sur la cohérence visuo-auditive

Une analyse par facteur a également été réalisée. Les résultats par facteur peuvent être consultés en annexe C. Pour mesurer l'influence du stimulus sonore et du type de rendu visuel, un test d'indépendance du χ^2 a été réalisé pour ces deux facteurs. Le tableau 4.3 présente les résultats de ces tests.

Facteur	DDL	χ^2	p
Stimulus sonore	24	35.6526	0.056
Type de rendu	48	46.7451	0.5243

TABLE 4.3 – Résultats du test d'indépendance du χ^2, pour les facteurs "Stimulus Sonore" et "Type de rendu".

Pour le facteur "Type de rendu", la p valeur $p = 0.5243$ indique qu'il n'y a pas d'effet du type de rendu visuel. Bien que très différents les uns des autres, les 3 types de visualisation (photo 360, nuage de point, modèle simple) ont conduit a des résultats similaires. Ce constat indique que la mauvaise représentation du volume dans le cas de photos 360 et le manque de détails et d'information de texture dans le cas des modèles simple n'ont pas eu d'influence notable sur le jugement de la cohérence. Toutefois, il est possible que pour les représentations photos 360 et nuages de points, contenant des informations de textures, les participants aient perçu la texture d'un matériau peu absorbant. Le choix d'une texture uni, gris mat, pour les modèles simples pourrait avoir également évoqué un matériau du même type. La question de l'influence de la texture mériterait d'être investiguée davantage en réalisant une expérience à partir des modèles simples dont plusieurs conditions de textures seraient testées. En ce qui concerne le stimulus sonore, la p valeur $p = 0.056$ ne permet pas non plus de conclure sur un effet significatif du stimulus.

4.4.3 Modèle perceptif de la cohérence visuo-auditive d'environnements acoustiques

Comme le suggèrent les premiers résultats, il semble possible d'établir une relation entre les grands attributs acoustiques et visuels permettant d'expliquer le jugement de la cohérence multimodale des environnements. A partir des observations

précédentes, nous formulons en première hypothèse qu'une certaine proportionnalité entre le temps de réverbération RT entendu et le volume physique observé V_{3D} est requise pour obtenir un niveau de cohérence élevé. Au contraire lorsque ces deux grandeurs ne respectent pas cette proportionnalité (temps de réverbération élevé et petit volume et vice versa), le niveau de cohérence visuo-auditive attendu est faible. Toutefois, on peut rappeler que d'après [Maempel et Jentsch, 2013], la taille perçue d'une salle est plutôt représentée par une grandeur à une dimension qu'on appellera V'_{1D}, équivalente à la racine cubique du volume perçu V'_{3D} (calculé à partir des dimensions longueur, largeur et hauteur perçues). Il a aussi été montré dans un grand nombre d'études (résumées dans l'article de [Loomis et Knapp, 2003]), que la perception visuelle de la distance est linéaire. Nous pouvons ainsi en déduire l'existence d'une relation linéaire entre le volume perçu V'_{3D} et le volume physique mesuré V_{3D}.

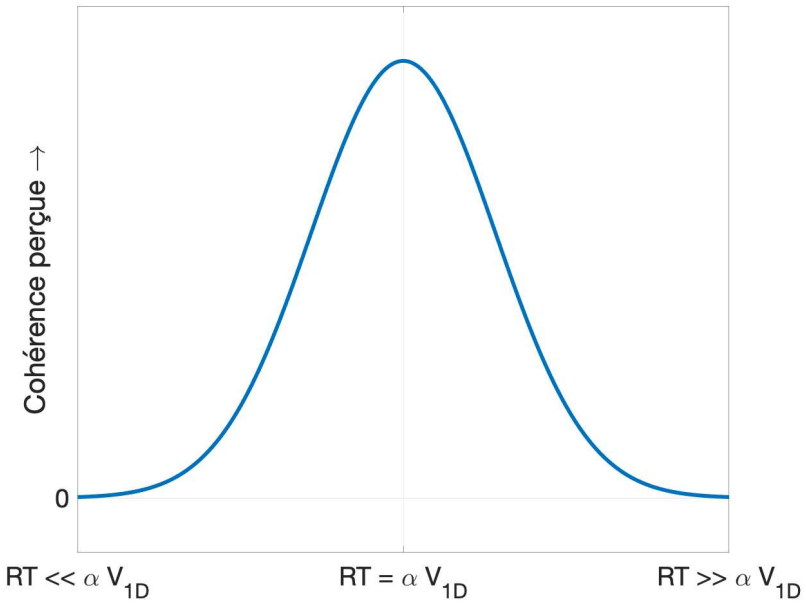


FIGURE 4.6 – Hypothèse d'un modèle de la perception de la cohérence visuo-auditive perçue d'environnements, en fonction du temps de réverbération (RT_{20}) et du volume perçu (V'_{1D}). α représente le coefficient de proportionnalité entre RT_{20} et V'_{1D}.

Le modèle théorique envisagé en prenant en compte ces remarques exprime donc le niveau de cohérence visuo-auditive attendu $\widehat{\tau}_+$ sous la forme d'une fonction gaussienne dépendante d'une relation linéaire entre RT_{20} et V'_{1D}, $x(RT_{20}, V'_{1D})$. Ce modèle est représenté en figure 4.6 et peut être exprimé par la formule suivante :

$$\widehat{\tau}_+ = a \times e^{-\left(\frac{x'(RT_{20}, V'_{1D})}{b}\right)^2} \tag{4.2}$$

avec a, b : paramètres de la fonction gaussienne ; respectivement son amplitude et son écart-type.

$x'(RT_{20}, V'_{1D})$ exprime la relation linéaire entre RT_{20} et V'_{1D}, comme suit :

$$x'(RT_{20}, V'_{1D}) = \alpha' V'_{1D} + RT_{20} + \beta' \tag{4.3}$$

avec α', β' : paramètres de la fonction affine, respectivement le coefficient de proportionnalité et l'ordonnée à l'origine.

En considérant que $V'_{1D} \propto \sqrt[3]{V_{3D}}$ on obtient une nouvelle expression de x, en fonction de RT_{20} et V_{3D} :

$$x(RT_{20}, V_{3D}) = \alpha \cdot \sqrt[3]{V_{3D}} + RT_{20} + \beta \tag{4.4}$$

en intégrant 4.4 dans 4.2, il est possible d'exprimer le taux de positivité attendu, en fonction du temps de réverbération RT_{20} et du volume réel V_{3D} d'un environnement :

$$\widehat{\tau}_+ = a \times e^{-\left(\frac{\alpha \cdot \sqrt[3]{V_{3D}} + RT_{20} + \beta}{b}\right)^2} \tag{4.5}$$

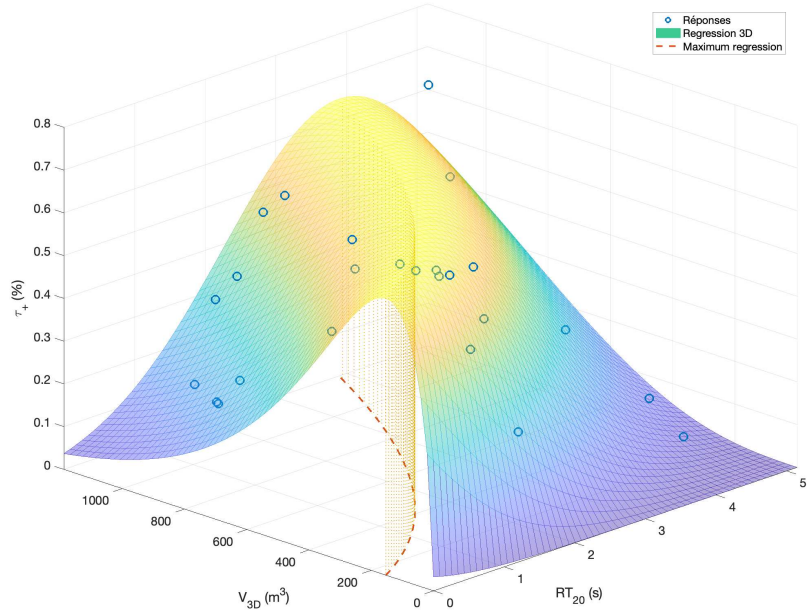
Une régression du modèle donné par 4.5 est réalisée sur les données de l'expérience. Les paramètres a, b, α et β sont estimés ainsi que leur intervalle de confiance à .95. La qualité de la prédiction de la régression est donnée par le coefficient de Pearson R^2. Ces valeurs sont présentées dans le tableau 4.4. La figure 4.7 représente les données du test ainsi que le modèle prédictif.

Paramètre	Valeur	(CI à 0.95)
a	0.6477	(0.5679, 0.7275)
b	0.5364	(0.4108, 0.662)
α	-0.1732	(-0.2165, -0.1298)
β	0.9268	(0.5728, 1.281)
R^2	0. 7773	

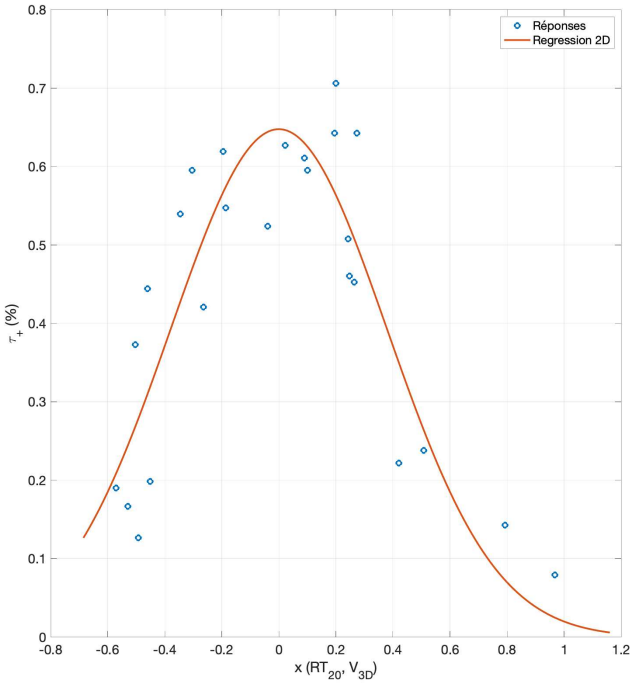
TABLE 4.4 – Résultats de la régression du modèle présenté en équation 4.5 sur les taux de positivité obtenus, toutes conditions de stimulus sonore et de rendu visuel confondues.

Discussions

La régression présentée précédemment donne un coefficient de corrélation de Pearson $R^2 = 0.7773$, ce qui signifie que les résultats de l'expérience sont bien



(a) Représentation 3D de la régression en fonction de RT_{20} et V_{3D}.



(b) Représentation 2D de la régression en fonction de $x(RT_{20}, V_{3D})$.
L'expression de $x(RT_{20}, V_{3D})$ est donnée en équation 4.4.

FIGURE 4.7 – Représentations de la régression du modèle prédicatif du taux de positivité τ_+, en fonction du temps de réverbération RT_{20} et du volume physique V_{3D}.

expliqués par le modèle choisi. Cela semble suggérer que d'autres indices acoustiques, notamment les indices spatiaux, ne sont pas déterminants dans le jugement de la cohérence visuo-auditive. Afin de vérifier cette hypothèse, il conviendrait d'intégrer ces indices dans le modèle et de mesurer l'effet de cette intégration sur la qualité de la prédiction. En ce qui concerne les paramètres de la régression, a et b sont liés au caractère gaussien du modèle. a représente l'amplitude maximum de la fonction. Avec $a = 0.65 \pm 0.08$, le modèle prédit un taux de positivité maximal de 65%. Cette valeur est cohérente avec les premiers résultats de l'expérience donnant des τ_+ n'excédant pas 71%. Le paramètre b quand à lui représente l'écart-type de la fonction gaussienne, plus il est élevé plus la fonction est étendue. Ici, $b = 0.53 \pm 0.126$ donne une indication sur la vitesse à laquelle la cohérence attendue décroît en fonction de la déviation entre RT_{20} et V_{3D}.

La relation utilisée dans le modèle pour exprimer la quantité de déviation entre RT_{20} et V_{3D} est donnée par $x(RT_{20}, V_{3D})$ (*c.f.* equation 4.4). Cet axe est représenté en abscisse de la figure 4.7b. Le niveau de cohérence maximal est attendu pour une déviation nulle, c'est à dire pour $x(RT_{20}, V_{3D}) = 0$, soit :

$$RT_{20} = -\alpha \cdot \sqrt[3]{V_{3D}} - \beta \tag{4.6}$$

La courbe associée à l'équation 4.6 est représentée en figure 4.7a par la ligne rouge pointillée. Cette relation permet donc de trouver le temps de réverbération qui sera perçu comme le plus cohérent à un volume donné et vice-versa. Le paramètre $\beta = 0.927 \pm 0.354$ représente l'ordonnée à l'origine de la courbe. Il indique que d'après notre modèle, un temps de réverbération nul ne serait pas forcément associé à un volume nul, mais plutôt à un volume d'environ $150m^3$. Il est fort probable toutefois qu'aux conditions limites pour des temps de réverbération et volumes proche de zéro, le modèle perceptif décrit ici ne soit pas adapté. En effet, bien qu'il ne soit pas absurde qu'un RT_{20} proche de zéro ne soit pas forcément associé à un volume nul, il y a fort à penser que dans le cas des petits volumes, d'autres indices acoustiques rentre en jeu, tels que les caractéristiques modales de la pièce et le comportement précoce de la réponse acoustique. Il serait intéressant de réaliser une nouvelle expérience du même type afin de préciser notre perception en termes de cohérence dans le cas particulier de petits volumes. Le paramètre $\alpha = -0.173 \pm 0.432$ représente l'allure à laquelle devrait évoluer le volume pour rester cohérent avec le temps de réverbération, lorsqu'on augmente ce dernier.

La validité du modèle peut être questionnée en dehors des limites de notre étude. La figure 4.8 représente une extrapolation du modèle à des volumes 10 fois supérieurs à ceux présentés dans cette étude. On peut notamment voir que le temps de réverbération associé à un volume de $10000m^3$ (ordre de grandeur de la grande chapelle du Palais des Papes d'Avignon) est d'environ 12 secondes, ce qui ne paraît pas aberrant. Il est également intéressant de constater une analogie entre la formule décrite par ce modèle (eq. 4.6) et la formule de Sabine permettant de déterminer un

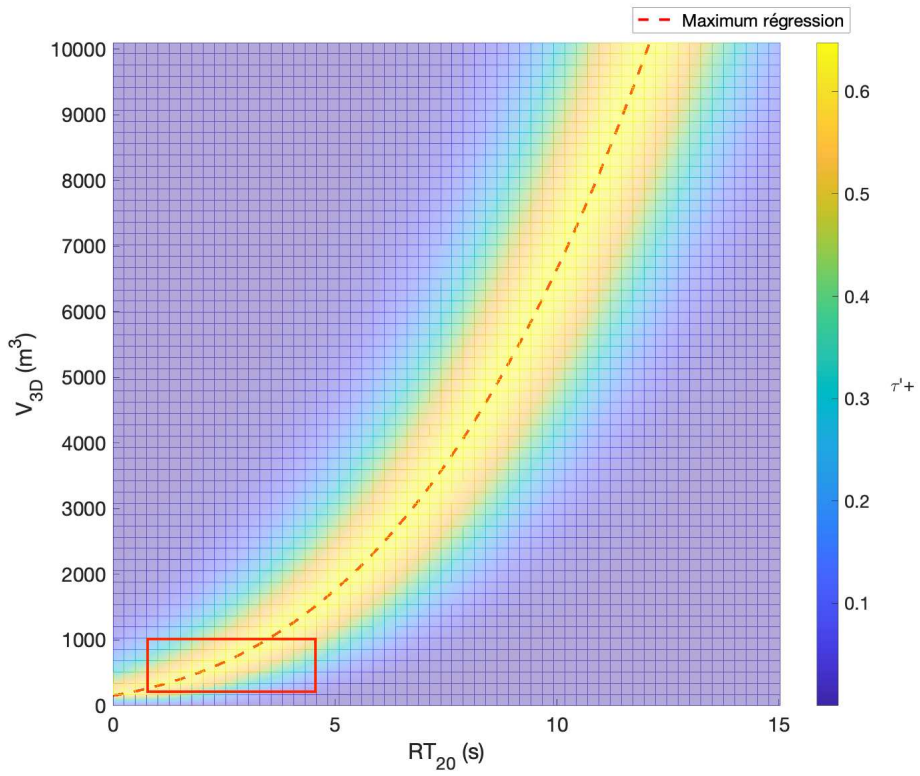


FIGURE 4.8 – Extrapolation du modèle à des volumes et temps de réverbérations plus élevés. Le rectangle rouge représente la zone du corpus utilisé pour cette étude.

temps de réverbération théorique RT_{60} en fonction du volume V d'une salle, donnée par l'équation suivante :

$$RT_{60} = 0.161 \cdot V/A \tag{4.7}$$

avec : $A = \sum_i \alpha_i \times S_i$: surface d'absorption équivalente. α_i est le coefficient d'absorption relatif à la surface S_i. En effet, en faisant une analyse dimensionnelle, on peut constater que la formule de Sabine exprime elle aussi un temps de réverbération théorique en fonction d'un volume (m^3) ramené à une dimension, en divisant par la surface d'absorption équivalente (m^2). Ainsi, en mettant en relation le modèle et la formule de Sabine, il serait par exemple possible, en connaissant la géométrie exacte des lieux, de déterminer un coefficient d'absorption moyen, caractéristique de la perception représentée par le modèle.

4.5 Conclusions et perspectives

Cette expérience nous a permis d'explorer la perception visuo-auditive d'environnements acoustiques. Les participants n'ont pas été capables d'associer l'acoustique

des lieux avec leur représentation visuelle, ce qui est cohérent avec les résultats de [Greif et Ackermann, 2020]. En revanche, il a été montré que le niveau de cohérence perçu semble être fortement lié à l'adéquation entre le volume, jugé visuellement et le temps de réverbération, jugé auditivement, des environnements. Un modèle perceptif a pu être proposé, permettant ainsi de prédire un niveau de cohérence attendu entre une acoustique et un espace visuel à partir des simples valeurs de RT_{20} et de volume. En effet, bien que l'expérience ait été conduite dans des conditions d'écoute spatialisée, le jugement de la cohérence visuo-auditive ne semble pas avoir tenu compte d'indices acoustiques spatiaux. D'autre part, le type de rendu visuel (photos 360, nuages de points et modèles simples) n'a pas montré d'influence statistique sur les réponses des participants, tout comme le stimulus sonore (extrait de guitare, extrait de parole). Bien qu'instructive, cette étude a avant tout une portée prospective et les résultats obtenus encouragent à imaginer de nouvelles expériences sur le sujet. Premièrement il serait judicieux de répéter cette expérience sur un ensemble plus vaste d'acoustiques et d'architectures pour confirmer et étendre les tendances observées ici. Du fait que la représentation simplifiée de l'architecture ne semble pas avoir d'effet majeur sur le jugement des participants, il est possible d'envisager cette expérience avec des salles de géométries simples (type "Shoebox"), dont les acoustiques pourraient être facilement simulées grâce à des logiciels de modélisation tels que CATT ou ODEON. Enfin, afin d'enquêter davantage sur l'adéquation entre acoustique et représentation, une approche plus incarnée peut être envisagée. A l'instar de l'étude de [Valente et Braasch, 2010], une expérience où les participants doivent "accorder" la réponse acoustique d'une salle à une représentation visuelle du lieu permettrait de définir plus précisément les attentes perceptives en termes d'adéquation acoustique/architecture. Une telle étude pourrait également faire émerger un jeu de paramètres acoustiques pertinents pour le contrôle et la synthèse d'environnements acoustiques, comme proposé dans les travaux de thèses de [Salmon, 2021]. Enfin, l'extraction automatisée de modèles géométriques simples à partir de nuages de points, décrite en section 4.2.3, ouvre également de nouvelles perspectives dans l'utilisation de logiciels de simulations acoustiques, puisque de part leur simplicité, ces modèles pourraient être directement exploitables dans ces environnements, comme suggéré par [Aspöck et Vorländer, 2016].

Chapitre 5

Conclusion générale

Comme nous l'avons vu tout au long de cette thèse, la compréhension de la perception des environnements acoustiques représente aujourd'hui un enjeu majeur dans des domaines aussi variés que ceux des sciences cognitives, de l'acoustique architecturale, des sciences du patrimoine ou encore des technologies immersives. Les différents travaux présentés dans cette thèse s'inscrivent dans cette démarche en abordant différents aspects de la perception des salles. Les contributions principales de ces travaux et les perspectives de recherches associées sont résumées ici.

5.1 Une interface VR pour l'étude de la perception auditive spatiale

Afin d'étudier la perception auditive spatiale en environnements réverbérants, une méthode de report de la perception des attributs spatiaux de sources sonores basée sur une interface en réalité virtuelle a été développée. Cette méthode originale permet non seulement le report de la position angulaire perçue des sources, tel qu'il pourrait être réalisé avec des méthodes de pointage classiques, mais intègre également la possibilité de rendre compte de sa perception en termes de distance et de taille apparente de source. Malgré ses atouts incontestables (description globale des attributs spatiaux de sources, simplicité de la tâche, portabilité du dispositif), la méthode présentée est sujette à un certain nombre de biais et limitations que nous avons pu identifier.

D'abord, un biais systématique dans le report de la distance perçue est à noter, résultant en une sur-estimation générale de la distance reportée. Ce biais n'est pas critique lorsqu'on s'intéresse à caractériser des différences relatives entre différentes conditions de test, mais rend notre dispositif inapte à un jugement absolu de la distance. D'après la littérature sur la perception visuelle de la distance en réalité virtuelle, une sous-estimation visuelle de la distance du pointeur dans l'interface VR pourrait être à l'origine de ce biais. Une expérience perceptive visant à quantifier l'effet de ce biais visuel est à considérer. D'autre part, l'intégration d'éléments graphiques tels qu'un avatar représentant le corps du sujet ou bien d'autres objets du quotidien faisant office d'étalon sont des pistes envisageables pour limiter cet effet.

Le report de la taille apparente des sources est également sujet à confusion. Celui-ci est effectué en délimitant une zone plus ou moins large autour de la position perçue de la source sonore. L'ambiguïté réside dans le fait que ce report peut être interprété soit comme le report effectif de la taille apparente de la source soit

comme une difficulté à localiser précisément la source et donc plutôt comme représentatif d'un flou de localisation. Nous avons vu qu'en condition réelle d'écoute, le report semble correspondre à la première interprétation tandis qu'en condition d'auralisation, il correspond davantage à la deuxième, en témoigne une corrélation significative entre l'erreur de localisation en azimut et la taille apparente reportées dans diverses conditions auralisées lors de l'expérience 2 du chapitre 3. Un moyen de lever cette ambiguïté serait d'interroger les sujets sur leur confiance dans leur propre jugement à chaque report. Toutefois, cette confusion soulève des questions fondamentales sur la perception auditive spatiale. Le flou de localisation et la taille apparente des sources sont-ils deux percepts indépendants ? Une source diffuse est-elle nécessairement floue ? Du point de vue de la mesure, les paramètres acoustiques couramment employés pour caractériser l'ASW ($IACC$ et LF_E) ne sont-ils pas également représentatifs de la sensation de flou ?

Pour conclure, cette méthode de report est adaptée à l'étude de la perception auditive spatiale des sources sonores dans différents contextes et situations. Son utilisation peut être envisagée aussi bien pour l'évaluation perceptive d'autres dispositifs de restitution spatialisée et d'auralisation que pour des études psychophysiques dans des situations réelles d'écoute. Les différents biais et limitations de notre dispositif demandent à être mieux compris pour pouvoir en limiter les effets et ouvrent également de nombreuses perspectives de recherche.

5.2 Recours à l'auralisation pour l'étude de la perception des environnements acoustiques

Dans l'ensemble des travaux présentés ici, nous avons choisi de nous tourner vers l'étude de la perception des environnements acoustiques à travers leur mesure et leur auralisation, afin de pallier la difficulté d'étudier la perception dans des conditions réelles. La technique d'auralisation retenue repose sur la technologie HOA particulièrement intéressante pour ses moyens de captation de réponses impulsionnelles spatiales de salles et pour ses qualités d'immersion sonore 3D. Dans une première expérience, nous avons souhaité évaluer la qualité du rendu spatial d'un système d'auralisation HOA 3D à l'ordre 4. Grâce à la méthode de report discutée précédemment, il a été possible de comparer les performances de localisation obtenues dans 3 salles d'acoustiques différentes en conditions réelles d'écoutes et en conditions d'auralisation. Cette étude révèle que les performances de localisations sont nettement dégradées lors d'une écoute auralisée et ce notamment en termes de position angulaire et de taille apparente et dans une moindre mesure de distance de source. Il a été observé que la capacité du système à restituer précisément les propriétés spatiales des sources sonores dépendait des propriétés acoustiques de la salle auralisée. Ce constat a été confirmé par l'expérience 2 consistant à comparer les performances de localisations dans plus de 20 environnements acoustiques auralisés différents. En effet, cette deuxième expérience révèle que la précision de la restitution ambisonique est de manière générale fortement corrélée à la force sonore (globale et latérale) de

la salle que l'on souhaite auraliser. Ce résultat semble indiquer que les caractéristiques spatiales des sources sonores auront tendance à être mieux restituées par le système pour des salles de faible gain que pour des salles présentant un gain élevé. Pour déterminer la nature physique de ces dégradations, nous proposons en perspective de réaliser une campagne de mesures des conditions auralisées dégradées et de comparer sur une base de descripteurs objectifs la qualité spatiale des réponses mesurées in-situ avec celles obtenues lors de l'auralisation.

En conséquence, nous pensons que le recours à l'auralisation ambisonique jusqu'à l'ordre 4 ne permet pas d'étudier la perception spatiale des sources sonores en milieu réverbérant de façon écologique, c'est à dire au plus proche de la perception du phénomène réel. A l'instar de la méthode HO-SIRR, récemment proposée par [McCormack *et al.*, 2020], d'autres méthodes d'auralisation de réponses impulsionnelles spatiales, dites paramétriques, ont été développées pour pallier au manque de précision spatiale des systèmes ambisoniques. Toutefois, une validation perceptive de ces méthodes semble encore nécessaire est envisagée comme travail futur.

Malgré un manque de fidélité spatiale de l'auralisation ambisonique, cette technique présente tout de même des qualités d'immersion incontestables. Le recours à cette méthode d'auralisation reste tout de même intéressant pour investiguer d'autres aspects de la perception des environnements acoustiques notamment dans des situations d'immersion multi-sensorielle.

5.3 Des outils pour la caractérisation des acoustiques du patrimoine

Dans le cadre du projet Sésames, nous avons mis en place différents outils méthodologiques pour la caractérisation d'acoustiques du patrimoine. En collaboration avec nos collègues du laboratoire MAP, un protocole d'acquisitions groupées de données métriques, spatiales, visuelles et sonores a été élaboré et déployé dans une vingtaines d'édifices de la région PACA. Une caractérisation acoustique issue de l'analyse factorielle de 23 paramètres de salles, calculés pour l'ensemble des réponses mesurées dans le cadre du projet Sésames, a permis de révéler 3 grandes dimensions objectives respectivement représentatives de la "précision", du "niveau" et de la "réverbérance" des salles et définissant un espace acoustique caractéristique du corpus à l'étude. D'un autre côté, l'expérience 2 visant à caractériser les individus du corpus selon leur capacité à modifier les attributs spatiaux de sources sonores a abouti à la proposition d'une cartographie perceptive et à la classification des différents environnements acoustiques. Il faut toutefois noter que cette caractérisation perceptive ne représente en aucun cas la perception réelle des lieux mais bien la perception de l'auralisation de ces lieux. La pertinence de ces résultats, dans une démarche de caractérisation perceptive et de compréhension du patrimoine, est certainement discutable et il semble que ceux-ci nous renseignent finalement davantage sur le comportement du système d'auralisation que sur notre perception des lieux. Pour aller plus loin sur le plan de la caractérisation perceptive, un test de

dissemblance entre les différentes conditions acoustiques du corpus est envisagée et permettrait certainement de décrire de façon plus globale la manière dont nous percevons ces acoustiques. D'autre part, ces travaux s'inscrivent dans une démarche de caractérisation plus large, prenant également en compte les données collectées par nos collaborateurs du projet Sésames. Un effort de croisement de ces différents jeux de données reste encore à faire pour prétendre à une meilleure compréhension du patrimoine architectural.

5.4 Vers un modèle de la cohérence audio-visuelle d'environnements acoustiques

Une expérience perceptive visant à étudier de la cohérence perçue entre des environnements acoustiques auralisées et différentes représentations visuelles de ces environnements a été réalisée. Le résultat principal de cette étude est que notre jugement de cette cohérence semble principalement s'appuyer sur une adéquation entre le temps de réverbération de la réponse acoustique et le volume estimé de l'espace représenté visuellement. En testant différents types de rendus visuels (photos panoramiques, nuages de points et modèles simplifiés), l'expérience révèle que les différentes conditions visuelles n'ont pas d'influence significative sur le jugement de la cohérence audio-visuelle. Sur la base de ces résultats, nous avons proposé un modèle simple permettant d'estimer un niveau de cohérence attendue entre une acoustique et un espace visuel, à partir des valeurs du temps de réverbération et du volume des environnements sonores et visuels. Ce modèle déterminé à partir des réponses des participants pour des salles de temps de réverbération relativement faibles et de faibles volumes, semble pouvoir être extrapolé à des valeurs de TR et de V plus élevées. Toutefois, de nouvelles expériences doivent être menées pour valider cette hypothèse et affiner le modèle. La relative indépendance du jugement au type de rendu visuel permet d'envisager la conduite des prochaines études avec des représentations visuelles simples de type "Shoeboxs" et des acoustiques obtenues par simulation. Ces premiers résultats peuvent néanmoins d'ores et déjà trouver des applications dans les domaines du cinéma, du jeu vidéo et de la VR, en donnant aux ingénieurs du son et designers sonores des indications dans la création d'environnements sonores en cohérence avec l'image.

Bibliographie

[IEC, 2003] (2003). IEC 60268-16. Sound System Equipment-Part 16 : Objective rating of speech intelligibility by speech transmission index. Rapport technique. (Cité en page 86.)

[ISO, 2009] (2009). ISO 3382-1. Acoustics – Measurements of room acoustic parameters – Part 1 : Performance spaces. Rapport technique. (Cité en page 82.)

[Ahrens *et al.*, 2019] Ahrens, A., Lund, K. D., Marschall, M. et Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE*, 14(3):e0214603. (Cité en pages 54 et 119.)

[Ahrens *et al.*, 2016] Ahrens, A., Marschall, M. et Dau, T. (2016). Evaluating the auralization of a small room in a virtual sound environment using objective room acoustic measures. *In 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan*, volume 140. (Cité en page 35.)

[Alary *et al.*, 2019] Alary, B., Masse, P., Valimaki, V. et Noisternig, M. (2019). Assessing the anisotropic features of spatial impulse responses. *In EAA Spatial Audio Signal Processing Symposium*, page 6 pages. EAA Spatial Audio Signal Processing Symposium. (Cité en pages 11 et 28.)

[Altmann *et al.*, 2013] Altmann, C. F., Ono, K., Callan, A., Matsuhashi, M., Mima, T. et Fukuyama, H. (2013). Environmental reverberation affects processing of sound intensity in right temporal cortex. *European Journal of Neuroscience*, 38(8):3210–3220. (Cité en pages 18 et 105.)

[Anderson et Zahorik, 2014] Anderson, P. W. et Zahorik, P. (2014). Auditory/visual distance estimation : Accuracy and variability. *Frontiers in psychology*, 5:1097. (Cité en page 20.)

[Ando, 2014] Ando, Y. (2014). Concert hall acoustics based on subjective preference theory. *In Springer Handbook of Acoustics*, pages 367–402. Springer. (Cité en page 82.)

[Aoshima, 1981] Aoshima, N. (1981). Computer-generated pulse signal applied for sound measurement. *The Journal of the Acoustical Society of America*, 69(5): 1484–1488. (Cité en page 25.)

[Armbrüster *et al.*, 2008] Armbrüster, C., Wolter, M., Kuhlen, T., Spijkers, W. et Fimm, B. (2008). Depth perception in virtual reality : Distance estimations in peri-and extrapersonal space. *Cyberpsychology & Behavior*, 11(1):9–15. (Cité en page 67.)

[Ashmead *et al.*, 1995] Ashmead, D. H., Davis, D. L. et Northington, A. (1995). Contribution of listeners' approaching motion to auditory distance perception. *Journal of experimental psychology : Human perception and performance*, 21(2): 239. (Cité en page 19.)

[Ashmead *et al.*, 1990] Ashmead, D. H., Leroy, D. et Odom, R. D. (1990). Perception of the relative distances of nearby sound sources. *Perception & Psychophysics*, 47(4):326–331. (Cité en page 17.)

[Aspöck et Vorländer, 2016] Aspöck, L. et Vorländer, M. (2016). Room geometry acquisition and processing methods for geometrical acoustics simulation models. *Proceedings of the EuroRegio*. (Cité en page 136.)

[Bahu *et al.*, 2016] Bahu, H., Carpentier, T., Noisternig, M. et Warusfel, O. (2016). Comparison of different egocentric pointing methods for 3d sound localization experiments. *Acta acustica united with Acustica*, 102(1):107–118. (Cité en pages 41 et 43.)

[Barazzetti *et al.*, 2018] Barazzetti, L., Previtali, M. et Roncoroni, F. (2018). Can we use low-cost 360 degree cameras to create accurate 3D models ? *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42(2). (Cité en page 123.)

[Barre *et al.*, 2014] Barre, S., Dobler, D. et Meyer, A. (2014). Room impulse response measurement with a spherical microphone array, application to room and building acoustics. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 249, pages 5467–5472. Institute of Noise Control Engineering. (Cité en page 28.)

[Barron, 1971] Barron, M. (1971). The subjective effects of first reflections in concert halls—the need for lateral reflections. *Journal of sound and vibration*, 15(4):475–494. (Cité en page 23.)

[Barron, 2001] Barron, M. (2001). Late lateral energy fractions and the envelopment question in concert halls. *Applied Acoustics*, 62(2):185–202. (Cité en pages 23 et 24.)

[Barron et Marshall, 1981] Barron, M. et Marshall, A. H. (1981). Spatial impression due to early lateral reflections in concert halls : The derivation of a physical measure. *Journal of sound and Vibration*, 77(2):211–232. (Cité en pages 21 et 23.)

[Beerends et De Caluwe, 1999] Beerends, J. G. et De Caluwe, F. E. (1999). The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society*, 47(5):355–362. (Cité en page 116.)

[Begault et Trejo, 2000] Begault, D. R. et Trejo, L. J. (2000). 3-D sound for virtual reality and multimedia. Rapport technique. (Cité en page 29.)

[Beranek, 2003] Beranek, L. L. (2003). Subjective rank-orderings and acoustical measurements for fifty-eight concert halls. *Acta Acustica united with Acustica*, 89(3):494–508. (Cité en page 82.)

[Beranek, 2004] Beranek, L. L. (2004). *Concert Halls and Opera Houses : Music, Acoustics, and Architecture*, volume 2. Springer. (Cité en page 22.)

[Berg et Nyberg, 2008] Berg, J. et Nyberg, D. (2008). Listener Envelopment–What Has Been Done and What Future Research Is Needed ? *In Au-*

dio Engineering Society Convention 124. Audio Engineering Society. (Cité en page 23.)

[Bernschütz *et al.*, 2014] Bernschütz, B., Giner, A. V., Pörschmann, C. et Arend, J. (2014). Binaural reproduction of plane waves with reduced modal order. *Acta Acustica united with Acustica*, 100(5):972–983. (Cité en page 32.)

[Bertet, 2009] Bertet, S. (2009). *Formats Audio 3D Hiérarchiques : Caractérisation Objective et Perceptive Des Systemes Ambisonics d'ordres Supérieurs*. Thèse de doctorat, PhD Thesis. (Cité en pages 32 et 41.)

[Bertet *et al.*, 2009] Bertet, S., Daniel, J., Parizet, E. et Warusfel, O. (2009). Influence of microphone and loudspeaker setup on perceived higher order ambisonics reproduced sound field. *In Proceedings of Ambisonics Symposium*. (Cité en pages 40 et 42.)

[Bertet *et al.*, 2013] Bertet, S., Daniel, J., Parizet, E. et Warusfel, O. (2013). Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica*, 99(4):642–657. (Cité en pages 40 et 63.)

[Berzborn *et al.*, 2017] Berzborn, M., Bomhardt, R., Klein, J., Richter, J.-G. et Vorländer, M. (2017). The ITA-Toolbox : An open source MATLAB toolbox for acoustic measurements and signal processing. *In Proceedings of the 43th Annual German Congress on Acoustics, Kiel, Germany*, volume 2017, pages 6–9. (Cité en page 80.)

[Blaise *et al.*, 2021] Blaise, J.-Y., Dudek, I., Pamart, A., Bergerot, L., Vidal, A., Fargeot, S., Aramaki, M., Ystad, S. et Kronland-Martinet, R. (2021). Acquisition & integration of spatial and acoustic features : A workflow tailored to small-scale heritage architecture. *ACTA IMEKO*, 11(2). (Cité en pages 75, 77 et 123.)

[Blauert, 1997a] Blauert, J. (1997a). Chapitre 3.1. *In Spatial Hearing : The Psychophysics of Human Sound Localization*. MIT press. (Cité en page 85.)

[Blauert, 1997b] Blauert, J. (1997b). *Spatial Hearing : The Psychophysics of Human Sound Localization*. MIT press. (Cité en pages 14, 15, 16, 61 et 105.)

[Braasch et Hartung, 2002] Braasch, J. et Hartung, K. (2002). Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data. *Acta acustica united with Acustica*, 88(6):942–955. (Cité en page 42.)

[Bradley, 2011] Bradley, J. S. (2011). Review of objective room acoustics measures and future needs. *Applied Acoustics*, 72(10):713–720. (Cité en pages 84 et 105.)

[Bradley *et al.*, 2000] Bradley, J. S., Reich, R. D. et Norcross, S. G. (2000). On the combined effects of early-and late-arriving sound on spatial impression in concert halls. *The Journal of the Acoustical Society of America*, 108(2):651–661. (Cité en pages 21 et 23.)

[Bradley *et al.*, 2003] Bradley, J. S., Sato, H. et Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, 113(6):3233–3244. (Cité en page 85.)

[Bradley et Soulodre, 1995a] Bradley, J. S. et Soulodre, G. A. (1995a). The influence of late arriving energy on spatial impression. *The Journal of the Acoustical Society of America*, 97(4):2263–2271. (Cité en page 23.)

[Bradley et Soulodre, 1995b] Bradley, J. S. et Soulodre, G. A. (1995b). Objective measures of listener envelopment. *The Journal of the Acoustical Society of America*, 98(5):2590–2597. (Cité en pages 23 et 87.)

[Braun et Frank, 2011] Braun, S. et Frank, M. (2011). Localization of 3D ambisonic recordings and ambisonic virtual sources. *In 1st International Conference on Spatial Audio,(Detmold)*. (Cité en pages 40 et 63.)

[Bronkhorst et Houtgast, 1999] Bronkhorst, A. W. et Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, 397(6719):517–520. (Cité en pages 17 et 18.)

[Brungart, 1999] Brungart, D. S. (1999). Auditory localization of nearby sources. III. Stimulus effects. *The Journal of the Acoustical Society of America*, 106(6): 3589–3602. (Cité en pages 18 et 19.)

[Brungart *et al.*, 1999] Brungart, D. S., Durlach, N. I. et Rabinowitz, W. M. (1999). Auditory localization of nearby sources. II. Localization of a broadband source. *The Journal of the Acoustical Society of America*, 106(4):1956–1968. (Cité en page 19.)

[Brungart et Scott, 2001] Brungart, D. S. et Scott, K. R. (2001). The effects of production and presentation level on the auditory distance perception of speech. *The Journal of the Acoustical Society of America*, 110(1):425–440. (Cité en page 19.)

[Butler *et al.*, 1980] Butler, R. A., Levy, E. T. et Neff, W. D. (1980). Apparent distance of sounds recorded in echoic and anechoic chambers. *Journal of Experimental Psychology : Human Perception and Performance*, 6(4):745. (Cité en page 19.)

[Cabrera *et al.*, 2005] Cabrera, D., Jeong, C., Kwak, H. J. et Kim, J.-Y. (2005). Auditory room size perception for modeled and measured rooms. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2005, pages 2995–3004. Institute of Noise Control Engineering. (Cité en pages 24 et 35.)

[Carlile *et al.*, 2000] Carlile, S., Jin, C. et Van Raad, V. (2000). Continuous virtual auditory space using HRTF interpolation : Acoustic and psychophysical errors. *In Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pages 220–223. (Cité en page 29.)

[Carpentier, 2015] Carpentier, T. (2015). Récents développements du Spatialisateur. *In Journées d'Informatique Musicale*. (Cité en page 160.)

[Cerdá *et al.*, 2015] Cerdá, S., Giménez, A., Cibrián, R., Girón, S. et Zamarreño, T. (2015). Subjective ranking of concert halls substantiated through orthogonal objective parameters. *The Journal of the Acoustical Society of America*, 137(2):580–584. (Cité en page 82.)

[Cerdá *et al.*, 2012] Cerdá, S., Giménez, A. et Cibrián, R. M. (2012). An objective scheme for ranking halls and obtaining criteria for improvements and design. *Journal of the Audio Engineering Society*, 60(6):419–430. (Cité en page 82.)

[Cerdá *et al.*, 2009] Cerdá, S., Giménez, A., Romero, J., Cibrián, R. et Miralles, J. (2009). Room acoustical parameters : A factor analysis approach. *Applied Acoustics*, 70(1):97–109. (Cité en pages 82 et 83.)

[Chernyak, 1968] Chernyak, R. I. (1968). Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise. *In Proc. 6th Int. Congr. Acoust.*, volume 3. (Cité en pages 15 et 16.)

[Coleman, 1963] Coleman, P. D. (1963). An analysis of cues to auditory depth perception in free space. *Psychological bulletin*, 60(3):302. (Cité en page 17.)

[Coleman, 1968] Coleman, P. D. (1968). Dual role of frequency spectrum in determination of auditory distance. *The Journal of the Acoustical Society of America*, 44(2):631–632. (Cité en page 19.)

[Cook *et al.*, 1955] Cook, R. K., Waterhouse, R. V., Berendt, R. D., Edelman, S. et Thompson Jr, M. C. (1955). Measurement of correlation coefficients in reverberant sound fields. *The Journal of the Acoustical Society of America*, 27(6):1072–1077. (Cité en page 11.)

[Côté *et al.*, 2012] Côté, N., Koehl, V. et Paquier, M. (2012). Ventriloquism effect on distance auditory cues. *In Acoustics 2012 joint congress (11ème Congrès Français d'Acoustique - 2012 Annual IOA Meeting)*. (Cité en page 20.)

[Daniel, 2000] Daniel, J. (2000). *Représentation de Champs Acoustiques, Application à La Transmission et à La Reproduction de Scènes Sonores Complexes Dans Un Contexte Multimédia*. Thèse de doctorat, Université Paris 6, France. (Cité en pages 28, 30 et 32.)

[Daniel, 2003] Daniel, J. (2003). Spatial sound encoding including near field effect : Introducing distance coding filters and a viable, new ambisonic format. *In Audio Engineering Society Conference : 23rd International Conference : Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society. (Cité en page 28.)

[de Vries *et al.*, 2001] de Vries, D., Hulsebos, E. M. et Baan, J. (2001). Spatial fluctuations in measures for spaciousness. *The journal of the Acoustical Society of America*, 110(2):947–954. (Cité en page 22.)

[Dick, 2017] Dick, D. A. (2017). *A New Metric to Predict Listener Envelopment Based on Spherical Microphone Array Measurements and Higher Order Ambisonic Reprodictions*. Thèse de doctorat, Pennsylvania State University. (Cité en page 24.)

[Dick et Vigeant, 2015] Dick, D. A. et Vigeant, M. C. (2015). Investigation of listener envelopment and the late sound field using spherical array microphone impulse response measurements. *Proceedings of the Institute of Acoustics*, 37:9. (Cité en page 24.)

[Djelani *et al.*, 2000] Djelani, T., Pörschmann, C., Sahrhage, J. et Blauert, J. (2000). An interactive virtual-environment generator for psychoacoustic research II : Collection of head-related impulse responses and evaluation of auditory localization. *Acta Acustica united with Acustica*, 86(6):1046–1053. (Cité en page 42.)

[Duda et Martens, 1998] Duda, R. O. et Martens, W. L. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058. (Cité en page 19.)

[Dunn et Hawksford, 1993] Dunn, C. et Hawksford, M. J. (1993). Distortion immunity of MLS-derived impulse response measurements. *Journal of the Audio Engineering Society*, 41(5):314–335. (Cité en page 25.)

[Embrechts, 2015] Embrechts, J.-J. (2015). Measurement of 3D room impulse responses with a spherical microphone array. *In Proceedings of the Euronoise 2015 Congress*, pages 143–148. (Cité en page 28.)

[Evjen *et al.*, 2001] Evjen, P., Bradley, J. S. et Norcross, S. G. (2001). The effect of late reflections from above and behind on listener envelopment. *Applied Acoustics*, 62(2):137–153. (Cité en page 24.)

[Fargeot *et al.*, 2019] Fargeot, S., Derrien, O., Parseihian, G., Aramaki, M. et Kronland-Martinet, R. (2019). Subjective evaluation of spatial distorsions induced by a sound source separation process. *In EAA Spatial Audio Signal Processing Symposium*, pages 67–72. (Cité en page 44.)

[Farina, 2000] Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. *In Audio Engineering Society Convention 108*. Audio Engineering Society. (Cité en pages 26, 27, 48 et 77.)

[Farina, 2007] Farina, A. (2007). Advancements in impulse response measurements by sine sweeps. *In Audio Engineering Society Convention 122*. Audio Engineering Society. (Cité en page 26.)

[Farina *et al.*, 2011] Farina, A., Amendola, A., Capra, A. et Varani, C. (2011). Spatial analysis of room impulse responses captured with a 32-capsule microphone array. *In Audio Engineering Society Convention 130*. Audio Engineering Society. (Cité en page 28.)

[Fazenda *et al.*, 2017] Fazenda, B., Scarre, C., Till, R., Pasalodos, R. J., Guerra, M. R., Tejedor, C., Peredo, R. O., Watson, A., Wyatt, S., Benito, C. G., Drinkall, H. et Foulds, F. (2017). Cave acoustics in prehistory : Exploring the association of Palaeolithic visual motifs and acoustic response. *The Journal of the Acoustical Society of America*, 142(3):1332–1349. (Cité en page 5.)

[Feldstein *et al.*, 2020] Feldstein, I. T., Kölsch, F. M. et Konrad, R. (2020). Egocentric Distance Perception : A Comparative Study Investigating Differences

Between Real and Virtual Environments. *Perception*, 49(9):940–967. (Cité en pages 67 et 119.)

[Fischler et Bolles, 1981] Fischler, M. A. et Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395. (Cité en page 124.)

[Furuya *et al.*, 2001] Furuya, H., Fujimoto, K., Ji, C. Y. et Higa, N. (2001). Arrival direction of late sound and listener envelopment. *Applied Acoustics*, 62(2): 125–136. (Cité en page 24.)

[Gagnon *et al.*, 2013] Gagnon, K. T., Geuss, M. N. et Stefanucci, J. K. (2013). Fear influences perceived reaching to targets in audition, but not vision. *Evolution and Human Behavior*, 34(1):49–54. (Cité en page 19.)

[Gandemer, 2016] Gandemer, L. (2016). *Son et Posture : Le Rôle de La Perception Auditive Spatiale Dans Le Maintien de l'équilibre Postural*. Thèse de doctorat, Aix-Marseille. (Cité en page 160.)

[Gardner, 1994] Gardner, W. G. (1994). Efficient convolution without input/output delay. *In Audio Engineering Society Convention 97*. Audio Engineering Society. (Cité en page 6.)

[Gardner et Martin, 1995] Gardner, W. G. et Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908. (Cité en page 29.)

[Gerzon, 1976] Gerzon, M. A. (1976). Unitary (energy-preserving) multichannel networks with feedback. *Electronics Letters*, 11(12):278–279. (Cité en page 6.)

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871. (Cité en pages 28 et 32.)

[Gibson, 1979] Gibson, J. J. (1979). *The Ecological Approach to Visual Perception : Classic Edition*. Psychology Press. (Cité en pages 13 et 116.)

[Gilkey *et al.*, 1995] Gilkey, R. H., Good, M. D., Ericson, M. A., Brinkman, J. et Stewart, J. M. (1995). A pointing technique for rapidly collecting localization responses in auditory research. *Behavior Research Methods, Instruments, & Computers*, 27(1):1–11. (Cité en page 42.)

[Giménez *et al.*, 2014] Giménez, A., Cibrián, R. M., Cerdá, S., Girón, S. et Zamarreño, T. (2014). Mismatches between objective parameters and measured perception assessment in room acoustics : A holistic approach. *Building and Environment*, 74:119–131. (Cité en page 82.)

[Gorzel *et al.*, 2012] Gorzel, M., Corrigan, D., Kearney, G., Squires, J. et Boland, F. (2012). Distance perception in virtual audio-visual environments. *In 25th UK Conference of the Audio Engineering Society : Spatial Audio In Today's 3D World (2012)*, pages 1–8. (Cité en page 117.)

[Greif et Ackermann, 2020] Greif, J. et Ackermann, D. (2020). *Can you hear the shape of a concert hall ? An audiovisual test in simulated 3D environments*. Thèse de doctorat, M. Sc. thesis, Institut für Sprache und Kommunikation, Technische Universitat Berlin, Berlin, Germany. (Cité en pages 117, 129 et 136.)

[Gröhn *et al.*, 2007] Gröhn, M., Lokki, T. et Takala, T. (2007). Localizing Sound Sources in a CAVE-Like Virtual Environment with Loudspeaker Array Reproduction. *Presence : Teleoperators and Virtual Environments*, 16(2):157–171. (Cité en page 41.)

[Gupta *et al.*, 2018] Gupta, R., Ranjan, R., He, J. et Woon-Seng, G. (2018). Investigation of effect of VR/AR headgear on Head related transfer functions for natural listening. *In Audio Engineering Society Conference : 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society. (Cité en page 43.)

[Guski et Vorländer, 2015] Guski, M. et Vorländer, M. (2015). Impulsive noise detection in sweep measurements. *Acta Acustica united with Acustica*, 101(4):723–730. (Cité en page 26.)

[Haas, 1951] Haas, H. (1951). über den Einfluβ eines Einfachechos auf die Hörsamkeit von Sprache. *Acta Acustica united with Acustica*, 1(2):49–58. (Cité en page 85.)

[Haber *et al.*, 1993] Haber, L., Haber, R. N., Penningroth, S., Novak, K. et Radgowski, H. (1993). Comparison of nine methods of indicating the direction to objects : Data from blind adults. *Perception*, 22(1):35–47. (Cité en page 41.)

[Hamasaki *et al.*, 2004] Hamasaki, K., Hatano, W., Hiyama, K., Komiyama, S. et Okubo, H. (2004). 5.1 and 22.2 multichannel sound productions using an integrated surround sound panning system. *In Audio Engineering Society Convention 117*. Audio Engineering Society. (Cité en page 33.)

[Hanyu et Kimura, 2001] Hanyu, T. et Kimura, S. (2001). A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections. *Applied Acoustics*, 62(2):155–184. (Cité en pages 23 et 24.)

[Hartmann, 1983] Hartmann, W. M. (1983). Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74(5):1380–1391. (Cité en pages 20 et 40.)

[Hidaka et Beranek, 2000] Hidaka, T. et Beranek, L. L. (2000). Objective and subjective evaluations of twenty-three opera houses in Europe, Japan, and the Americas. *The Journal of the Acoustical Society of America*, 107(1):368–383. (Cité en page 82.)

[Hidaka *et al.*, 1995] Hidaka, T., Beranek, L. L. et Okano, T. (1995). 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*, 98(2):988–1007. (Cité en pages 22 et 86.)

[Holt et Thurlow, 1969] Holt, R. E. et Thurlow, W. R. (1969). Subject orientation and judgment of distance of a sound source. *The Journal of the Acoustical Society of America*, 46(6B):1584–1585. (Cité en page 19.)

[Howard et Rogers, 2002] Howard, I. P. et Rogers, B. J. (2002). *Seeing in Depth, Vol. 2 : Depth Perception.* University of Toronto Press. (Cité en page 123.)

[Huisman *et al.*, 2021] Huisman, T., Ahrens, A. et MacDonald, E. (2021). Sound source localization in virtual reality with ambisonics sound reproduction. (Cité en pages 43, 63 et 64.)

[Iannace et Trematerra, 2014] Iannace, G. et Trematerra, A. (2014). The acoustics of the caves. *Applied Acoustics*, 86:42–46. (Cité en page 5.)

[Ihlefeld et Shinn-Cunningham, 2011] Ihlefeld, A. et Shinn-Cunningham, B. G. (2011). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America*, 130(1):324–333. (Cité en pages 20, 40 et 48.)

[Jesteadt *et al.*, 1977] Jesteadt, W., Wier, C. C. et Green, D. M. (1977). Intensity discrimination as a function of frequency and sensation level. *The Journal of the acoustical society of America*, 61(1):169–177. (Cité en page 17.)

[Johnson, 2018] Johnson, D. (2018). *Towards the Perceptual Optimisation of Virtual Room Acoustics.* Thèse de doctorat, University of Huddersfield. (Cité en pages 24 et 35.)

[Jot, 1992] Jot, J.-M. (1992). *Etude et Réalisation d'un Spatialisateur de Sons Par Modèles Physiques et Perceptifs.* Thèse de doctorat, Ecole nationale supérieure des télécommunications, France. (Cité en page 6.)

[Jot *et al.*, 1999] Jot, J.-M., Larcher, V. et Pernaux, J.-M. (1999). A comparative study of 3-D audio encoding and rendering techniques. *In Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction.* Audio Engineering Society. (Cité en page 33.)

[Kahle, 1995] Kahle, E. (1995). Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras. *Unpublished Ph. D. dissertation, Université du Maine, Le Mans.* (Cité en page 82.)

[Käsbach *et al.*, 2014] Käsbach, J., May, T., Le Goff, N. et Dau, T. (2014). The importance of binaural cues for the perception of apparent source width at different sound pressure levels. *In Proceedings of DAGA.* (Cité en page 105.)

[Kasbach *et al.*, 2015] Kasbach, J., Wiinberg, A., May, T., Jepsen, M. L. et Dau, T. (2015). Apparent source width perception in normal-hearing, hearing-impaired and aided listeners. *DAGA, Nürnberg*, page 4. (Cité en page 40.)

[Kelly *et al.*, 2018] Kelly, J. W., Cherep, L. A., Klesel, B., Siegel, Z. D. et George, S. (2018). Comparison of two methods for improving distance perception in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 15(2):1–11. (Cité en page 67.)

[Khaykin et Rafaely, 2012] Khaykin, D. et Rafaely, B. (2012). Acoustic analysis by spherical microphone array processing of room impulse responses. *The Journal of the Acoustical Society of America*, 132(1):261–270. (Cité en page 28.)

[Kim *et al.*, 2001] Kim, H.-Y., Suzuki, Y., Takane, S. et Sone, T. (2001). Control of auditory distance perception based on the auditory parallax model. *Applied Acoustics*, 62(3):245–270. (Cité en page 19.)

[Kirkeby et Nelson, 1999] Kirkeby, O. et Nelson, P. A. (1999). Digital filter design for inversion problems in sound reproduction. *Journal of the Audio Engineering Society*, 47(7/8):583–595. (Cité en page 161.)

[Klockgether et van de Par, 2014] Klockgether, S. et van de Par, S. (2014). A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception. *Acta Acustica united with Acustica*, 100(5):964–971. (Cité en page 22.)

[Kolarik *et al.*, 2016] Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S. et Pardhan, S. (2016). Auditory distance perception in humans : A review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78(2):373–395. (Cité en pages 17, 40, 61, 66 et 104.)

[Kopčo et Shinn-Cunningham, 2011] Kopčo, N. et Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *The Journal of the Acoustical Society of America*, 130(3):1530–1541. (Cité en pages 18 et 19.)

[Kuttruff, 2017] Kuttruff, H. (2017). *Room Acoustics*. CRC Press/Taylor & Francis Group, Boca Raton, sixth edition édition. (Cité en page 10.)

[Kuznetsova *et al.*, 2015] Kuznetsova, A., Christensen, R. H., Bavay, C. et Brockhoff, P. B. (2015). Automated mixed ANOVA modeling of sensory and consumer data. *Food Quality and Preference*, 40:31–38. (Cité en page 54.)

[Langendijk et Bronkhorst, 1997] Langendijk, E. H. et Bronkhorst, A. W. (1997). Collecting localization response with a virtual acoustic pointer. *The Journal of the Acoustical Society of America*, 101(5):3106–3106. (Cité en page 42.)

[Larsson *et al.*, 2001] Larsson, P., Västfjäll, D. et Kleiner, M. (2001). Ecological acoustics and the multi-modal perception of rooms : Real and unreal experiences of auditory-visual virtual environments. *In 2001 International Conference on Auditory Display*, page 5. (Cité en pages 117 et 118.)

[Li *et al.*, 2016a] Li, M., Nan, L., Smith, N. et Wonka, P. (2016a). Reconstructing building mass models from UAV images. *Computers & Graphics*, 54:84–93. (Cité en page 124.)

[Li *et al.*, 2016b] Li, M., Wonka, P. et Nan, L. (2016b). Manhattan-world urban reconstruction from point clouds. *In European Conference on Computer Vision*, pages 54–69. Springer. (Cité en page 124.)

[Lindau *et al.*, 2014a] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F. et Weinzierl, S. (2014a). A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica*, 100(5):984–994. (Cité en page 23.)

[Lindau *et al.*, 2014b] Lindau, A., Maempel, H.-J. et Horn, M. (2014b). SEACEN : P9 - Audio-visual perception of acoustical environments. https ://www.seacen.tu-berlin.de/subprojects/p9/parameter/en/. (Cité en page 118.)

[Litovsky *et al.*, 1999] Litovsky, R. Y., Colburn, H. S., Yost, W. A. et Guzman, S. J. (1999). The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654. (Cité en pages 20 et 21.)

[Little *et al.*, 1992] Little, A. D., Mershon, D. H. et Cox, P. H. (1992). Spectral Content as a Cue to Perceived Auditory Distance. *Perception*, 21(3):405–416. (Cité en pages 19 et 105.)

[Loomis et Knapp, 2003] Loomis, J. M. et Knapp, J. M. (2003). Visual perception of egocentric distance in real and virtual environments. *Virtual and adaptive environments*, 11:21–46. (Cité en pages 67 et 131.)

[Luizard *et al.*, 2018] Luizard, P., Brauer, E., Weinzierl, S. et Bernardoni, N. H. (2018). How singers adapt to room acoustical conditions. *In Auditorium Acoustics 2018*. (Cité en page 35.)

[Lundeby *et al.*, 1995] Lundeby, A., Vigran, T. E., Bietz, H. et Vorländer, M. (1995). Uncertainties of measurements in room acoustics. *Acta Acustica united with Acustica*, 81(4):344–355. (Cité en pages 81 et 82.)

[Macdonald et McGurk, 1978] Macdonald, J. et McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257. (Cité en page 116.)

[Macé *et al.*, 2012] Macé, M. J.-M., Dramas, F. et Jouffrais, C. (2012). Reaching to sound accuracy in the peri-personal space of blind and sighted humans. *In International Conference on Computers for Handicapped Persons*, pages 636–643. Springer. (Cité en page 48.)

[Maempel, 2017] Maempel, H.-J. (2017). Apples and oranges : A methodological framework for basic research into audiovisual perception. Retrieved from `http://dx.doi.org/10.14279/depositonce-6424`. (Cité en page 118.)

[Maempel et Horn, 2017] Maempel, H.-J. et Horn, M. (2017). The virtual concert hall : A research tool for the experimental investigation of audiovisual room perception. *International Journal on Stereo and Immersive Media*. (Cité en page 118.)

[Maempel et Horn, 2018] Maempel, H.-J. et Horn, M. (2018). Audiovisual perception of real and virtual rooms. *Journal of Virtual Reality and Broadcasting*, 14(5). (Cité en pages 118 et 119.)

[Maempel et Jentsch, 2013] Maempel, H.-J. et Jentsch, M. (2013). Auditory and visual contribution to egocentric distance and room size perception. *Building Acoustics*, 20(4):383–401. (Cité en pages 20, 116 et 131.)

[Majdak *et al.*, 2010] Majdak, P., Goupell, M. J. et Laback, B. (2010). 3-D localization of virtual sound sources : Effects of visual environment, pointing

method, and training. *Attention, Perception, & Psychophysics*, 72(2):454–469. (Cité en pages 42 et 43.)

[Majdak *et al.*, 2008] Majdak, P., Laback, B., Goupell, M. et Mihocic, M. (2008). The accuracy of localizing virtual sound sources : Effects of pointing method and visual environment. *In Audio Engineering Society Convention*, volume 124. (Cité en page 42.)

[Mason *et al.*, 2005] Mason, R., Brookes, T. et Rumsey, F. (2005). Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *The Journal of the Acoustical Society of America*, 117(3):1337–1350. (Cité en pages 22 et 105.)

[Mason *et al.*, 2000] Mason, R., Ford, N., Rumsey, F. et De Bruyn, B. (2000). Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. *Audio Engineering Society Preprint*, 5225. (Cité en page 41.)

[Massé *et al.*, 2020] Massé, P., Carpentier, T., Warusfel, O. et Noisternig, M. (2020). Denoising Directional Room Impulse Responses with Spatially Anisotropic Late Reverberation Tails. *Applied Sciences*, 10(3):1033. (Cité en page 28.)

[McCormack *et al.*, 2020] McCormack, L., Pulkki, V., Politis, A., Scheuregger, O. et Marschall, M. (2020). 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*, 68(5):338–354. (Cité en pages 28, 29, 34, 68 et 139.)

[McKenzie *et al.*, 2019] McKenzie, T., Murphy, D. et Kearney, G. (2019). Interaural Level Difference Optimization of Binaural Ambisonic Rendering. *Applied Sciences*, 9(6):1226. (Cité en page 36.)

[Merimaa et Pulkki, 2004] Merimaa, J. et Pulkki, V. (2004). Spatial impulse response rendering. *In Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*, volume 146. (Cité en page 29.)

[Mershon *et al.*, 1989] Mershon, D. H., Ballenger, W. L., Little, A. D., McMurtry, P. L. et Buchanan, J. L. (1989). Effects of Room Reflectance and Background Noise on Perceived Auditory Distance. *Perception*, 18(3):403–416. (Cité en pages 18, 19 et 105.)

[Mershon et Bowers, 1979] Mershon, D. H. et Bowers, J. N. (1979). Absolute and Relative Cues for the Auditory Perception of Egocentric Distance. *Perception*, 8(3):311–322. (Cité en pages 17 et 66.)

[Meyer et Elko, 2002] Meyer, J. et Elko, G. (2002). A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. *In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1781. IEEE. (Cité en pages 31 et 32.)

[Moreau *et al.*, 2006] Moreau, S., Daniel, J. et Bertet, S. (2006). 3D sound field recording with higher order ambisonics–Objective measurements and validation of a 4th order spherical microphone. *In 120th Convention of the AES*, pages 20–23. (Cité en page 31.)

[Morimoto *et al.*, 1993] Morimoto, M., Iida, K. et Furue, Y. (1993). Relation between auditory source width in various sound fields and degree of interaural cross-correlation. *Applied Acoustics*, 38(2-4):291–301. (Cité en pages 21 et 22.)

[Morimoto *et al.*, 2008] Morimoto, M., Nakagawa, K. et Iida, K. (2008). The relation between spatial impression and the law of the first wavefront. *Applied Acoustics*, 69(2):132–140. (Cité en page 21.)

[Morse et Ingard, 1986] Morse, P. M. et Ingard, K. U. (1986). *Theoretical Acoustics*. Princeton university press. (Cité en page 12.)

[Nan et Wonka, 2017] Nan, L. et Wonka, P. (2017). Polyfit : Polygonal surface reconstruction from point clouds. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 2353–2361. (Cité en page 124.)

[Nélisse et Nicolas, 1997] Nélisse, H. et Nicolas, J. (1997). Characterization of a diffuse field in a reverberant room. *The Journal of the Acoustical Society of America*, 101(6):3517–3524. (Cité en page 11.)

[Nielsen, 1992] Nielsen, S. H. (1992). Auditory distance perception in different rooms. *In Audio Engineering Society Convention 92*. Audio Engineering Society. (Cité en page 18.)

[Noisternig *et al.*, 2003] Noisternig, M., Sontacchi, A., Musil, T. et Holdrich, R. (2003). A 3D Ambisonic Based Binaural Sound Reproduction System. *In Audio Engineering Society Conference : 24th International Conference : Multichannel Audio, The New Reality*. Audio Engineering Society. (Cité en pages 32 et 36.)

[Okano *et al.*, 1998] Okano, T., Beranek, L. L. et Hidaka, T. (1998). 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*, 104(1):255–265. (Cité en pages 22 et 105.)

[Otani *et al.*, 2009] Otani, M., Hirahara, T. et Ise, S. (2009). Numerical study on source-distance dependency of head-related transfer functions. *The Journal of the Acoustical Society of America*, 125(5):3253–3261. (Cité en page 19.)

[Paquier *et al.*, 2016] Paquier, M., Côté, N., Devillers, F. et Koehl, V. (2016). Interaction between auditory and visual perceptions on distance estimations in a virtual environment. *Applied Acoustics*, 105:186–199. (Cité en page 20.)

[Parseihian *et al.*, 2014] Parseihian, G., Jouffrais, C. et Katz, B. F. G. (2014). Reaching nearby sources : Comparison between real and virtual sound and visual targets. *Frontiers in Neuroscience*, 8. (Cité en page 17.)

[Parseihian et Katz, 2012] Parseihian, G. et Katz, B. F. (2012). Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America*, 131(4):2948–2957. (Cité en page 29.)

[Perrott et Buell, 1982] Perrott, D. R. et Buell, T. N. (1982). Judgments of sound volume : Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise. *The Journal of the Acoustical Society of America*, 72(5):1413–1417. (Cité en pages 21 et 22.)

[Philbeck et Mershon, 2002] Philbeck, J. W. et Mershon, D. H. (2002). Knowledge about typical source output influences perceived auditory distance. *The Journal of the Acoustical Society of America*, 111(5):1980–1983. (Cité en page 19.)

[Polack, 1992] Polack, J.-D. (1992). Modifying chambers to play billiards : The foundations of reverberation theory. *Acta Acustica united with Acustica*, 76(6): 256–272. (Cité en page 11.)

[Postma, 2017] Postma, B. (2017). *Serious Auralizations*. Thèse de doctorat, Université Paris-Saclay (ComUE). (Cité en page 36.)

[Postma et Katz, 2017a] Postma, B. N. J. et Katz, B. F. G. (2017a). The influence of visual distance on the room-acoustic experience of auralizations. *The Journal of the Acoustical Society of America*, 142(5):3035–3046. (Cité en pages 117 et 119.)

[Postma et Katz, 2017b] Postma, B. N. J. et Katz, B. F. G. (2017b). Influence of visual rendering on the acoustic judgements of a theater auralization. *In 173rd Meeting of Acoustical Society of America and 8th Forum Acusticum*, page 015008, Boston, Massachusetts. (Cité en page 119.)

[Prawda *et al.*, 2022] Prawda, K., Schlecht, S. J. et Välimäki, V. (2022). Robust selection of clean swept-sine measurements in non-stationary noise. *The Journal of the Acoustical Society of America*, 151(3):2117–2126. (Cité en page 26.)

[Pulkki, 1997] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466. (Cité en pages 29 et 33.)

[Pulkki, 2007] Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516. (Cité en pages 29 et 34.)

[Pulkki et Hirvonen, 2004] Pulkki, V. et Hirvonen, T. (2004). Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, 13(1):105–119. (Cité en pages 33 et 42.)

[Rafaely, 2015] Rafaely, B. (2015). *Fundamentals of Spherical Array Processing*, volume 8. Springer. (Cité en page 31.)

[Rakerd et Hartmann, 1986] Rakerd, B. et Hartmann, W. M. (1986). Localization of sound in rooms, III : Onset and duration effects. *The Journal of the Acoustical Society of America*, 80(6):1695–1706. (Cité en page 20.)

[Regan et Spekreijse, 1977] Regan, D. et Spekreijse, H. (1977). Auditory—Visual Interactions and the Correspondence between Perceived Auditory Space and Perceived Visual Space. *Perception*, 6(2):133–138. (Cité en page 116.)

[Reznikoff, 2008] Reznikoff, I. (2008). Sound resonance in prehistoric times : A study of Paleolithic painted caves and rocks. *The Journal of the Acoustical Society of America*, 123(5):3603–3603. (Cité en page 5.)

[Rindel et Christensen, 2003] Rindel, J. H. et Christensen, C. L. (2003). Room acoustic simulation and auralization–how close can we get to the real room. *In Proc. 8th Western Pacific Acoustics Conference, Melbourne*. (Cité en page 24.)

[Romanov *et al.*, 2017] Romanov, M., Berghold, P., Frank, M., Rudrich, D., Zaunschirm, M. et Zotter, F. (2017). Implementation and Evaluation of a Low-Cost Headtracker for Binaural Synthesis. *In Audio Engineering Society Convention 142*. Audio Engineering Society. (Cité en page 69.)

[Ronsse et Wang, 2012] Ronsse, L. M. et Wang, L. M. (2012). Effects of Room Size and Reverberation, Receiver Location, and Source Rotation on Acoustical Metrics Related to Source Localization. *Acta Acustica united with Acustica*, 98(5): 768–775. (Cité en page 18.)

[Routray *et al.*, 2021] Routray, G., Dwivedi, P. et Hegde, R. M. (2021). Binaural Reproduction of HOA Signal using Sparse Multiple Measurement Vector Projections. *In 2021 National Conference on Communications (NCC)*, pages 1–6. IEEE. (Cité en page 32.)

[Rumsey, 2018] Rumsey, F. (2018). Perceptual Evaluation : Listening Strategies, Methods, and VR. *Journal of the Audio Engineering Society*, 66(4):301–305. (Cité en page 43.)

[Rychtáriková *et al.*, 2009] Rychtáriková, M., den Bogaert, T. V., Vermeir, G. et Wouters, J. (2009). Binaural sound source localization in real and virtual rooms. *Journal of the Audio Engineering Society*, 57(4):205–220. (Cité en page 40.)

[Salmon, 2021] Salmon, F. (2021). *Contrôle des impressions spatiales dans un environnement acoustique virtuel*. Thèse de doctorat, Université de Bretagne occidentale - Brest, France. (Cité en page 136.)

[Schiff et Oldak, 1990] Schiff, W. et Oldak, R. (1990). Accuracy of judging time to arrival : Effects of modality, trajectory, and gender. *Journal of Experimental Psychology : Human Perception and Performance*, 16(2):303. (Cité en page 19.)

[Schnabel *et al.*, 2007] Schnabel, R., Wahl, R. et Klein, R. (2007). Efficient RANSAC for point-cloud shape detection. *In Computer Graphics Forum*, volume 26, pages 214–226. Wiley Online Library. (Cité en page 124.)

[Schoeffler *et al.*, 2014] Schoeffler, M., Westphal, S., Adami, A., Bayerlein, H. et Herre, J. (2014). Comparison of a 2D- and 3D-based graphical user interface for localization listening tests. *In Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014*, page 6. (Cité en page 42.)

[Schroeder, 1962] Schroeder, M. R. (1962). Natural sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223. (Cité en page 6.)

[Schroeder, 1979] Schroeder, M. R. (1979). Integrated-impulse method measuring sound decay without using impulses. *The Journal of the Acoustical Society of America*, 66(2):497–500. (Cité en page 25.)

[Schroeder, 1987] Schroeder, M. R. (1987). Statistical parameters of the frequency response curves of large rooms. *Journal of the Audio Engineering Society*, 35(5):299–306. (Cité en page 12.)

[Schroeder *et al.*, 1974] Schroeder, M. R., Gottlob, D. et Siebrasse, K. F. (1974). Comparative study of European concert halls : Correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America*, 56(4):1195–1201. (Cité en page 82.)

[Schroeder et Kuttruff, 1962] Schroeder, M. R. et Kuttruff, K. H. (1962). 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*, 34(1):76–80. (Cité en page 12.)

[Schultz, 1971] Schultz, T. J. (1971). Diffusion in reverberation rooms. *Journal of Sound and Vibration*, 16(1):17–28. (Cité en page 11.)

[Seeber, 2002] Seeber, B. (2002). A new method for localization studies. *In ACUSTProc. Int. Congress on Acoustics ICA/Acta Acustica*, volume 88, pages 446–450. (Cité en page 42.)

[Seeber *et al.*, 2010] Seeber, B. U., Kerber, S. et Hafter, E. R. (2010). A system to simulate and reproduce audio–visual environments for spatial hearing research. *Hearing research*, 260(1-2):1–10. (Cité en page 118.)

[Shinn-Cunningham, 2000] Shinn-Cunningham, B. G. (2000). Learning reverberation : Considerations for spatial auditory displays. *In Proceedings of the 2000 International Conference on Auditory Display, Atlanta, GA, 2000*. Georgia Institute of Technology. (Cité en page 19.)

[Shinn-Cunningham, 2001] Shinn-Cunningham, B. G. (2001). Localizing sound in rooms. *ACM/SIGGRAPH and Eurographics Campfire : Acoustic Rendering for Virtual Environments*, pages 1–6. (Cité en pages 20 et 40.)

[Shinn-Cunningham *et al.*, 2005] Shinn-Cunningham, B. G., Kopco, N. et Martin, T. J. (2005). Localizing nearby sound sources in a classroom : Binaural room impulse responses. *The Journal of the Acoustical Society of America*, 117(5):3100–3115. (Cité en page 19.)

[Stan *et al.*, 2002] Stan, G.-B., Embrechts, J.-J. et Archambeau, D. (2002). Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262. (Cité en page 25.)

[Stecker, 2019] Stecker, G. C. (2019). Using Virtual Reality to Assess Auditory Performance. *The Hearing Journal*, 72(6):20–22. (Cité en page 43.)

[Stitt *et al.*, 2016] Stitt, P., Hendrickx, E., Messonnier, J.-C. et Katz, B. F. (2016). The Role of Head Tracking in Binaural Rendering. *In 29th Tonmeistertagung, International VDT Convention*, page 7. (Cité en page 69.)

[Tervo *et al.*, 2013] Tervo, S., Pätynen, J., Kuusinen, A. et Lokki, T. (2013). Spatial decomposition method for room impulse responses. *Journal of the Audio Engineering Society*, 61(1/2):17–28. (Cité en pages 29, 34 et 68.)

[Thurlow et Jack, 1973] Thurlow, W. R. et Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Perceptual and motor skills*, 36(3_suppl):1171–1184. (Cité en page 116.)

[Valente et Braasch, 2010] Valente, D. L. et Braasch, J. (2010). Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues. *The Journal of the Acoustical Society of America*, 128(4):1952–1964. (Cité en pages 117, 119 et 136.)

[Vennerød, 2014] Vennerød, J. (2014). Binaural Reproduction of Higher Order Ambisonics-A Real-Time Implementation and Perceptual Improvements. Mémoire de D.E.A., NTNU. (Cité en page 36.)

[Vliegen et Van Opstal, 2004] Vliegen, J. et Van Opstal, A. J. (2004). The influence of duration and level on human sound localization. *The Journal of the Acoustical Society of America*, 115(4):1705–1713. (Cité en page 63.)

[von Békésy, 1938] von Békésy, G. (1938). Uber die entstehung der entfernungsempfindung beim horen (On the origin of the sensation of distance in hearing). *Akustische zeitschrift*, 3:21–31. (Cité en page 19.)

[Vorländer, 2020] Vorländer, M. (2020). *Auralization : Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. RWTHedition. Springer International Publishing, Cham. (Cité en pages 24 et 30.)

[Vorländer *et al.*, 2015] Vorländer, M., Schröder, D., Pelzer, S. et Wefers, F. (2015). Virtual reality for architectural acoustics. *Journal of Building Performance Simulation*, 8(1):15–25. (Cité en page 119.)

[Wakuda *et al.*, 2003] Wakuda, A., Furuya, H., Fujimoto, K., Isogai, K. et Anai, K. (2003). Effects of arrival direction of late sound on listener envelopment. *Acoustical science and technology*, 24(4):179–185. (Cité en page 24.)

[Wallach *et al.*, 1949] Wallach, H., Newman, E. B. et Rosenzweig, M. R. (1949). A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4):468–468. (Cité en page 85.)

[Wang *et al.*, 2020] Wang, P., Lin, Z. et Qiu, X. (2020). Influence of interaural cross-correlation coefficient and loudness level on auditory source width at different frequency. *Applied Acoustics*, 162:107198. (Cité en pages 40 et 105.)

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244. (Cité en page 101.)

[Weinzierl *et al.*, 2018] Weinzierl, S., Lepa, S. et Ackermann, D. (2018). A measuring instrument for the auditory perception of rooms : The Room Acoustical Quality Inventory (RAQI). *The Journal of the Acoustical Society of America*, 144(3):1245–1257. (Cité en page 93.)

[Wightman et Kistler, 1989] Wightman, F. L. et Kistler, D. J. (1989). Headphone simulation of free-field listening. II : Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878. (Cité en page 41.)

[Winter *et al.*, 2017] Winter, F., Wierstorf, H. et Spors, S. (2017). Improvement of the reporting method for closed-loop human localization experiments. *In Audio Engineering Society Convention 142*. Audio Engineering Society. (Cité en page 42.)

[Xie, 2018] Xie, B. (2018). SCUT Near-Field HRTFs Database. http ://sofacoustics.org/data/database/scut/. (Cité en page 86.)

[Yadav *et al.*, 2011] Yadav, M., Cabrera, D. et Martens, W. (2011). Auditory Room Size Perceived from a Room Acoustic Simulation with Autophonic Stimuli. *Acoustics Australia*, 39(3). (Cité en page 24.)

[Zacharov *et al.*, 2016] Zacharov, N., Pedersen, T. et Pike, C. (2016). A common lexicon for spatial sound quality assessment - latest developments. *In 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, Lisbon, Portugal. IEEE. (Cité en page 23.)

[Zahorik, 2002a] Zahorik, P. (2002a). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846. (Cité en pages 17 et 18.)

[Zahorik, 2002b] Zahorik, P. (2002b). Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117. (Cité en page 18.)

[Zahorik *et al.*, 2005] Zahorik, P., Brungart, D. S. et Bronkhorst, A. W. (2005). Auditory distance perception in humans : A summary of past and present research. *ACTA Acustica united with Acustica*, 91(3):409–420. (Cité en pages 17, 18, 40 et 66.)

[Zahorik et Wightman, 2001] Zahorik, P. et Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature neuroscience*, 4(1):78–83. (Cité en page 17.)

[Zaunschirm *et al.*, 2018] Zaunschirm, M., Schörkhuber, C. et Höldrich, R. (2018). Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *The Journal of the Acoustical Society of America*, 143(6):3616–3627. (Cité en page 36.)

[Zotter *et al.*, 2012] Zotter, F., Pomberger, H. et Noisternig, M. (2012). Energy-Preserving Ambisonic Decoding. *Acta Acustica united with Acustica*, 98(1):37–47. (Cité en page 31.)

Annexe A

Présentation d'un système multicanal pour la spatialisation 3D

Le laboratoire PRISM dispose d'un système de spatialisation 3D, en chambre semi-anéchoïque. 42 hauts-parleurs (Genelec 8020A), répartis de façon homogène sur une structure géodésique de 4m de diamètre, permettent une spatialisation complète dans les 3 dimensions de l'espace. En effet, à la différence de la plupart des systèmes ambisoniques existants, qui sont des systèmes de spatialisation 2D, permettant une spatialisation dans le plan horizontal, ou 3D hémisphériques, prenant la forme d'un dôme, la qualité quasi-sphérique de notre système permet également la restitution de sources sonores provenant de l'hémisphère inférieure et lui vaut son surnom : la Sphère (cf. figure A.1).

Figure A.1 – Photo du système de spatialisation 42 canaux en chambre semi-anéchoïque.

Le dispositif actuel est calqué sur sa version précédente proposée, conçue et décrite par Lennie Gandemer, lors de ses travaux de thèse [Gandemer, 2016]. Gandemer relevait dans son mémoire de thèse l'importance de l'utilisation d'un tel système dans des conditions acoustiques les plus neutres possibles. Pour cela, il a réalisé un traitement acoustique de la salle dans laquelle se trouvait le système HOA, permettant ainsi d'en abaisser le TR_{60} de 1.2s à 0.61s à 250 Hz et de 0.66s à 0.41s à 8000 Hz. Bien que satisfaisantes dans le cadre des travaux de Gandemer sur l'étude de l'influence du son spatial sur la posture, ces caractéristiques acoustiques ne le sont plus lorsque l'on souhaite utiliser ce système pour l'étude perceptive des environnements acoustiques eux-mêmes. En effet, les perturbations induites par la salle d'expérimentation peuvent altérer la perception de l'acoustique des lieux à l'étude. C'est pourquoi nous avons décidé de démanteler l'ancien système et de le remonter en chambre semi-anéchoïque (de $TR_{60} < 0.1$s, et de fréquence de coupure $f_c = 80$ Hz). Au passage, la nouvelle version a été montée sur une structure géodésique identique à celle de Gandemer, à la différence près de son diamètre : ici 4m (contre 3m pour l'ancienne version).

Calibration du système

Bien que les haut-parleurs soient répartis de manière homogène et quasi-équidistante de la position centrale d'écoute, aussi appelée *sweet spot*, le système doit être calibré pour une utilisation optimale. Une calibration du niveau et du retard pour chaque haut-parleur a été réalisée avec les outils de calibration du Spat5 [Carpentier, 2015] dans Max/MSP. La calibration de la réponse en fréquence du système est également discutée dans cette section.

Calibration en niveau La calibration en niveau consiste à régler le gain appliqué à chacun des haut-parleurs pour qu'ils fournissent le même niveau sonore à la position centrale d'écoute du dispositif. Elle est effectuée en deux étapes. Les hauts-parleur Genelec 8020A sont équipés d'un potentiomètre de volume permettant un réglage grossier du niveau sonore. La première étape consiste à régler à la main et à l'œil chacun des potentiomètres sur une même valeur (représentée par la position angulaire du potentiomètre). La deuxième étape vise à réaliser un réglage fin des niveaux sonores et compenser d'éventuelles différences de niveaux engendrées par le premier réglage manuel et la variabilité dans les propriétés acoustiques des 42 haut-parleurs. Le spat5 propose un outil de calibration du gain, appelé spat5.calibrate.gain~. Le niveau sonore généré par chaque haut-parleur, jouant à tour de rôle un bruit blanc de 3s, est mesuré par un microphone omnidirectionnel, placé au centre du dispositif. Les différences de niveau mesurées (sd = $\pm$ 2.35 dB) sont alors compensées en appliquant en entrée des HP un gain numérique correspondant.

Calibration en temps La librarie spat5 propose également un objet de calibration des retards inter-haut-parleurs appelé spat5.calibrate.delay~. Ce processus permet de compenser les retards temporels induits par des différences de distance entre les haut-parleurs et le centre du système. De la même manière que pour la calibration en niveau, un microphone omnidirectionnel est placé au *sweet spot*. Les retards sont calculés à partir de la mesure des réponses impulsionnelles par sinus-glissant exponentiels de 1s. Le retard moyen mesuré est de 0.11 ± 0.05 ms, correspondant à une erreur moyenne de positionnement des hauts parleurs de 3.7 cm. Ces retards sont compensés numériquement en entrée de chaque HP.

Calibration de la réponse en fréquence La calibration de la réponse en fréquence du système est un problème pouvant être abordé selon plusieurs approches. L'approche que nous avons choisie consiste à égaliser la réponse en fréquence de tous les HP, un par un, par rapport au centre du dispositif en procédant de façon analogue à la calibration en temps. En effet, une réponse en fréquence par haut-parleur peut être calculée à partir de la mesure de réponses impulsionnelles par sinus glissant. Ces réponses peuvent ensuite être compensées en appliquant un filtre inverse à chacun des HP. Kirkeby propose une méthode de filtrage inverse avec régularisation adaptée aux systèmes mutli-canaux [Kirkeby et Nelson, 1999]. Cette méthode est notamment employée pour la restitution sur haut-parleurs d'un flux binaural. Bien qu'efficace, elle peut devenir coûteuse pour des systèmes composés d'un grand nombre de HP (un filtre FIR par HP).

Annexe B

Données de l'expérience 2 : Caractérisation de la perception auditive spatiale dans un corpus d'environnements acoustiques mesurés

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience.

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
Anecho	MA-eg	6.47 (3.93 :10.65)	9.43 (4.56 :19.49)	2.25 (1.55 :3.26)	7.44 (4.17 :13.28)
	MA-ec	3.87 (1.65 :9.08)	8.56 (4.49 :16.31)	1.98 (1.39 :2.82)	7.24 (3.74 :14.00)
	MA-ed	3.82 (1.57 :9.28)	7.69 (3.75 :15.77)	2.09 (1.45 :3.00)	7.42 (4.40 :12.50)
	MC-ec	6.60 (4.82 :9.04)	6.22 (2.23 :17.41)	2.24 (1.61 :3.13)	5.17 (2.46 :10.86)
AviGC	MA-eg	4.59 (2.82 :7.44)	9.90 (4.95 :19.79)	6.04 (4.77 :7.64)	5.61 (3.81 :8.25)
	MA-ec	4.09 (2.54 :6.59)	10.74 (5.18 :22.26)	3.96 (3.10 :5.06)	7.52 (4.75 :11.89)
	MA-ed	3.83 (2.08 :7.06)	9.62 (4.29 :21.58)	6.17 (4.87 :7.81)	6.48 (3.95 :10.62)
	MC-ec	4.35 (2.67 :7.08)	9.31 (4.90 :17.68)	8.17 (6.22 :10.73)	4.33 (2.53 :7.40)

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience. (Suite)

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
AviStMa	MA-eg	3.41 (2.26 :5.14)	7.76 (3.40 :17.69)	8.66 (6.81 :11.01)	7.32 (4.68 :11.44)
	MA-ec	7.62 (5.37 :10.81)	9.70 (4.66 :20.16)	5.62 (4.27 :7.41)	10.85 (6.73 :17.51)
	MA-ed	5.07 (2.96 :8.68)	9.76 (4.37 :21.79)	8.28 (6.46 :10.62)	8.71 (5.26 :14.43)
	MC-ec	11.40 (8.20 :15.85)	8.77 (3.29 :23.36)	9.50 (7.15 :12.64)	11.32 (5.20 :24.68)
AviTri	MA-eg	4.28 (2.58 :7.10)	12.21 (7.05 :21.14)	7.03 (5.68 :8.70)	4.97 (3.19 :7.74)
	MA-ec	3.88 (2.65 :5.70)	12.30 (7.06 :21.42)	5.35 (4.22 :6.79)	5.94 (3.78 :9.32)
	MA-ed	2.63 (1.33 :5.19)	12.73 (8.06 :20.11)	6.61 (5.49 :7.95)	7.03 (4.36 :11.36)
	MC-ec	2.97 (1.82 :4.85)	7.15 (3.19 :16.06)	8.59 (6.92 :10.67)	4.87 (2.86 :8.31)
Bras	MA-eg	10.23 (6.58 :15.92)	8.03 (2.84 :22.68)	8.07 (5.84 :11.15)	8.92 (5.52 :14.40)
	MA-ec	4.76 (1.86 :12.20)	13.30 (7.05 :25.07)	6.32 (4.82 :8.28)	8.72 (4.62 :16.48)
	MA-ed	8.88 (5.66 :13.92)	11.76 (6.17 :22.41)	8.31 (6.30 :10.95)	10.08 (5.33 :19.06)
	MC-ec	8.03 (5.02 :12.85)	11.00 (5.31 :22.77)	9.83 (7.52 :12.85)	13.32 (7.66 :23.15)
Bru	MA-eg	3.56 (1.89 :6.69)	10.40 (5.92 :18.28)	5.76 (4.60 :7.20)	6.10 (3.43 :10.84)
	MA-ec	7.57 (4.76 :12.04)	12.36 (6.10 :25.02)	5.63 (4.66 :6.80)	10.07 (5.30 :19.13)
	MA-ed	8.37 (5.15 :13.59)	6.62 (2.46 :17.79)	5.60 (4.56 :6.89)	8.63 (5.57 :13.39)
	MC-ec	5.74 (2.30 :14.35)	11.28 (6.60 :19.27)	7.43 (6.10 :9.04)	5.99 (3.24 :11.10)

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience. (Suite)

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
Esp	MA-eg	5.49 (3.37 :8.95)	15.69 (8.96 :27.48)	6.30 (4.80 :8.28)	9.94 (6.41 :15.43)
	MA-ec	6.21 (2.92 :13.24)	11.90 (6.17 :22.94)	5.83 (4.63 :7.34)	8.46 (4.42 :16.22)
	MA-ed	6.44 (3.50 :11.83)	11.22 (5.71 :22.06)	6.11 (4.96 :7.52)	10.13 (5.89 :17.44)
	MC-ec	7.29 (4.63 :11.49)	11.72 (6.11 :22.48)	8.61 (6.58 :11.27)	8.63 (5.44 :13.70)
LaVdre	MA-eg	4.14 (2.04 :8.39)	11.14 (5.32 :23.35)	6.40 (5.28 :7.75)	8.94 (5.25 :15.20)
	MA-ec	7.55 (5.37 :10.61)	12.74 (5.84 :27.81)	5.16 (3.88 :6.88)	10.61 (6.35 :17.74)
	MA-ed	8.92 (4.99 :15.92)	14.47 (8.40 :24.93)	5.83 (4.44 :7.66)	14.43 (8.68 :23.97)
	MC-ec	11.05 (6.26 :19.53)	11.58 (5.91 :22.68)	6.47 (5.29 :7.91)	11.21 (5.66 :22.21)
LMSHo	MA-eg	5.54 (2.83 :10.88)	13.40 (7.50 :23.96)	5.46 (4.21 :7.07)	8.80 (4.91 :15.79)
	MA-ec	4.29 (1.72 :10.71)	10.98 (4.89 :24.68)	5.36 (4.37 :6.56)	8.32 (4.96 :13.96)
	MA-ed	8.20 (4.33 :15.52)	12.59 (7.03 :22.55)	5.66 (4.20 :7.64)	7.46 (4.20 :13.23)
	MC-ec	7.83 (5.30 :11.57)	11.14 (5.42 :22.92)	8.25 (6.79 :10.04)	8.10 (4.99 :13.14)
LMSRch	MA-eg	7.10 (3.89 :12.98)	9.47 (5.40 :16.63)	7.27 (5.53 :9.55)	11.59 (6.20 :21.66)
	MA-ec	8.68 (5.41 :13.92)	11.71 (5.13 :26.72)	5.40 (4.16 :7.00)	13.75 (7.88 :24.00)
	MA-ed	3.76 (1.47 :9.58)	8.89 (3.85 :20.51)	9.06 (6.66 :12.34)	8.72 (4.00 :19.02)
	MC-ec	12.10 (7.61 :19.26)	11.88 (6.26 :22.55)	8.30 (6.40 :10.76)	12.27 (6.57 :22.92)

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience. (Suite)

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
Peyn	MA-eg	9.03 (5.44 :14.97)	14.15 (6.85 :29.22)	6.90 (5.53 :8.61)	8.51 (4.65 :15.60)
	MA-ec	4.12 (2.32 :7.30)	12.26 (5.61 :26.78)	5.97 (4.84 :7.38)	7.82 (3.47 :17.61)
	MA-ed	8.68 (6.12 :12.30)	9.24 (4.02 :21.26)	7.26 (5.60 :9.42)	9.74 (5.68 :16.68)
	MC-ec	11.65 (7.31 :18.57)	7.98 (3.71 :17.16)	9.02 (6.66 :12.23)	10.21 (6.28 :16.61)
Peyrl	MA-eg	4.52 (2.46 :8.29)	10.81 (6.13 :19.06)	5.98 (4.85 :7.38)	5.16 (2.81 :9.48)
	MA-ec	6.64 (3.51 :12.58)	8.35 (3.24 :21.46)	4.07 (3.13 :5.31)	9.24 (5.61 :15.22)
	MA-ed	9.30 (6.14 :14.07)	5.73 (2.33 :14.09)	5.05 (4.08 :6.26)	6.72 (3.82 :11.82)
	MC-ec	7.59 (5.18 :11.11)	10.61 (5.34 :21.07)	6.22 (4.93 :7.84)	6.76 (4.36 :10.50)
Prue	MA-eg	6.07 (3.39 :10.89)	11.93 (6.40 :22.22)	5.77 (4.59 :7.26)	9.00 (5.61 :14.45)
	MA-ec	4.78 (1.91 :11.97)	11.23 (5.55 :22.73)	3.99 (3.09 :5.16)	7.86 (4.09 :15.11)
	MA-ed	5.31 (2.11 :13.38)	9.67 (4.51 :20.75)	5.53 (4.16 :7.34)	9.93 (5.27 :18.68)
	MC-ec	6.05 (3.29 :11.14)	12.05 (7.63 :19.04)	8.12 (6.35 :10.37)	7.66 (4.30 :13.63)
PuySt Ma	MA-eg	5.51 (2.74 :11.08)	12.16 (6.82 :21.65)	6.22 (5.03 :7.69)	9.37 (6.09 :14.43)
	MA-ec	3.98 (1.90 :8.33)	10.73 (5.17 :22.27)	4.93 (3.70 :6.57)	9.06 (4.88 :16.82)
	MA-ed	5.12 (3.19 :8.23)	10.34 (5.93 :18.01)	6.42 (4.83 :8.54)	6.78 (2.70 :17.06)
	MC-ec	6.31 (2.97 :13.38)	9.15 (3.76 :22.23)	6.90 (4.92 :9.67)	5.89 (3.50 :9.91)

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience. (Suite)

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
PuySt Pce	MA-eg	4.19 (2.13 :8.21)	9.39 (4.35 :20.23)	6.31 (4.97 :8.00)	7.30 (3.12 :17.10)
	MA-ec	8.42 (5.06 :14.02)	9.27 (3.76 :22.86)	5.03 (3.87 :6.55)	13.26 (8.07 :21.76)
	MA-ed	6.78 (2.71 :17.00)	11.10 (5.90 :20.88)	6.11 (4.49 :8.32)	10.82 (5.39 :21.72)
	MC-ec	9.96 (5.63 :17.63)	9.11 (4.30 :19.31)	6.66 (4.91 :9.05)	8.22 (4.38 :15.43)
Pvrt	MA-eg	4.98 (2.38 :10.42)	10.67 (5.57 :20.43)	5.25 (3.70 :7.45)	9.10 (5.11 :16.24)
	MA-ec	6.51 (3.68 :11.54)	11.41 (5.52 :23.60)	5.05 (3.78 :6.76)	11.21 (5.90 :21.28)
	MA-ed	5.89 (2.81 :12.35)	8.67 (3.95 :19.02)	4.85 (3.62 :6.51)	11.55 (6.15 :21.70)
	MC-ec	6.71 (4.12 :10.93)	6.91 (3.09 :15.47)	7.73 (6.50 :9.18)	6.87 (3.58 :13.17)
StMaPa	MA-eg	7.51 (4.43 :12.73)	11.13 (5.43 :22.81)	5.69 (4.32 :7.49)	10.99 (6.86 :17.59)
	MA-ec	8.69 (4.85 :15.55)	17.11 (9.89 :29.61)	4.18 (3.10 :5.64)	14.72 (9.12 :23.76)
	MA-ed	7.37 (4.70 :11.54)	8.99 (3.92 :20.60)	6.28 (4.77 :8.27)	14.01 (6.83 :28.74)
	MC-ec	7.79 (3.98 :15.27)	12.77 (6.93 :23.53)	5.91 (4.54 :7.67)	14.23 (8.72 :23.21)
TvNDSalt	MA-eg	5.83 (2.59 :13.15)	13.52 (7.36 :24.84)	6.53 (5.09 :8.37)	7.40 (4.47 :12.25)
	MA-ec	6.96 (4.35 :11.15)	12.33 (6.22 :24.43)	4.25 (3.41 :5.29)	8.05 (4.71 :13.75)
	MA-ed	6.33 (2.87 :13.97)	10.30 (5.52 :19.25)	6.24 (4.90 :7.94)	5.50 (3.24 :9.36)
	MC-ec	6.31 (4.30 :9.26)	10.69 (6.03 :18.96)	7.21 (5.79 :8.98)	6.29 (4.11 :9.62)

Table B.1: Données brutes de l'expérience 2 : Valeurs moyennes et écarts–types entre parenthèses pour l'ensemble des conditions acoustiques de l'expérience. (Suite)

Salle	**Mesure**	$\lvert\varepsilon_\theta\rvert$ (°)	$\lvert\varepsilon_\phi\rvert$ (°)	R_p (m)	S_{eq} (10^{-2} sr)
TvStPbce	MA-eg	3.44 (1.47 :8.05)	7.38 (3.48 :15.67)	6.82 (5.49 :8.46)	4.84 (3.09 :7.59)
	MA-ec	5.27 (3.03 :9.16)	14.89 (7.86 :28.22)	4.76 (3.53 :6.43)	7.92 (4.49 :13.96)
	MA-ed	9.07 (5.93 :13.86)	14.67 (9.75 :22.07)	5.67 (4.40 :7.30)	10.40 (6.02 :17.96)
	MC-ec	5.09 (2.64 :9.81)	12.33 (6.58 :23.10)	7.39 (5.86 :9.32)	8.68 (5.57 :13.50)

	AviGC	AviStMa	AviTri	Bras	Bru	Esp	LaVdre	LMSHo	LMSRch	Peyn	Peyrl	Prue	PuyStMa	PuyStPce	Pvrt	StMaPa	TvNDSalt	TvStPbce
Anecho	2.40	-0.49	3.37	-3.54	-1.27	-1.38	-2.58	-1.98	-3.41	-2.73	-1.31	-1.93	-0.14	-3.45	-1.31	-3.28	-1.05	-0.55
AviGC		-2.89	0.97	-5.94	-3.67	-3.78	-4.98	-4.38	-5.81	-5.13	-3.71	-4.33	-2.54	-5.85	-3.71	-5.68	-3.45	-2.96
AviStMa			3.86	-3.05	-0.78	-0.89	-2.09	-1.49	-2.92	-2.24	-0.82	-1.44	0.35	-2.96	-0.82	-2.79	-0.56	-0.07
AviTri				-6.91 (*)	-4.64	-4.76	-5.95 (*)	-5.36	-6.78 (*)	-6.11 (**)	-4.68	-5.31	-3.51	-6.82	-4.68	-6.66 (**)	-4.43	-3.93
Bras					2.27	2.16	0.96	1.56	0.13	0.81	2.23	1.61	3.40	0.09	2.23	0.26	2.49	2.99
Bru						-0.11	-1.31	-0.72	-2.14	-1.46	-0.04	-0.66	1.13	-2.18	-0.04	-2.01	0.21	0.71
Esp							-1.19	-0.60	-2.03	-1.35	0.07	-0.55	1.24	-2.07	0.07	-1.90	0.33	0.83
LaVdre								0.59	-0.83	-0.16	1.27	0.64	2.44	-0.87	1.27	-0.71	1.52	2.02
LMSHo									-1.43	-0.75	0.68	0.05	1.85	-1.47	0.67	-1.30	0.93	1.43
LMSRch										0.68	2.10	1.48	3.27	-0.04	2.10	0.13	2.36	2.86
Peyn											1.42	0.80	2.59	-0.72	1.42	-0.55	1.68	2.18
Peyrl												-0.62	1.17	-2.14	-0.00	-1.97	0.25	0.75
Prue													1.79	-1.52	0.62	-1.35	0.88	1.38
PuyStMa														-3.31	-1.17	-3.14	-0.92	-0.42
PuyStPce															2.14	0.17	2.40	2.90
Pvrt																-1.97	0.26	0.76
StMaPa																	2.23	2.73
TvNDSalt																		0.50

TABLE B.2 – Résultats des tests post-hoc (ajustement de Tukey) pour l'erreur absolue de localisation en azimut $|\varepsilon_\theta|$ (°). EMMs (*Estimated Marginal Means*) et significativité entre parenthèses : $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

	AviGC	AviStMa	AviTri	Bras	Bru	Esp	LaVdre	LMSHo	LMSRch	Peyn	Peyrl	Prue	PuyStMa	PuyStPce	Pvrt	StMaPa	TvNDSalt	TvStPbce
Anecho	-4.06 (***)	-6.28 (***)	-4.75 (***)	-6.53 (***)	-3.87 (***)	-4.73 (***)	-3.88 (***)	-4.10 (***)	-5.88 (***)	-5.42 (***)	-3.12 (***)	-3.79 (***)	-4.27 (***)	-4.12 (***)	-3.74 (***)	-3.57 (***)	-3.94 (***)	-4.14 (***)
AviGC		-2.22 (***)	-0.69	-2.47 (***)	0.19	-0.67	0.18	-0.04	-1.82 (*)	-1.36	0.94	0.27	-0.21	-0.06	0.32	0.49	0.12	-0.09
AviStMa			1.54	-0.24	2.41 (**)	1.55	2.40 (**)	2.18 (**)	0.41	0.86	3.16 (***)	2.50 (***)	2.01 (**)	2.16 (**)	2.54 (***)	2.71 (***)	2.35 (**)	2.14 (**)
AviTri				-1.78	0.88	0.02	0.87	0.65	-1.13	-0.68	1.63 (**)	0.96	0.48	0.62	1.01	1.17 (*)	0.81	0.60
Bras					2.66 (**)	1.80	2.65 (***)	2.43 (**)	0.65	1.11	3.41 (***)	2.74 (***)	2.26 (**)	2.41 (***)	2.79 (***)	2.95 (***)	2.59 (***)	2.38 (**)
Bru						-0.86	-0.01	-0.23	-2.01	-1.55	0.75	0.08	-0.40	-0.25	0.13	0.30	-0.07	-0.28
Esp							0.85	0.63	-1.15	-0.69	1.61 (*)	0.94	0.46	0.61	0.99	1.16	0.79	0.58
LaVdre								-0.22	-2.00	-1.54	0.76	0.09	-0.39	-0.24	0.14	0.31	-0.06	-0.26
LMSHo									-1.78	-1.32	0.98	0.31	-0.17	-0.02	0.36	0.53	0.17	-0.04
LMSRch										0.45	2.76 (***)	2.09 (**)	1.61	1.75	2.14 (**)	2.30 (***)	1.94	1.73
Peyn											2.30 (***)	1.63 (*)	1.15	1.30	1.68 (**)	1.85 (**)	1.49	1.28
Peyrl												-0.67	-1.15	-1.00	-0.62	-0.45	-0.81	-1.02
Prue													-0.48	-0.34	0.05	0.21	-0.15	-0.36
PuyStMa														0.15	0.53	0.70	0.33	0.13
PuyStPce															0.38	0.55	0.19	-0.02
Pvrt																0.17	-0.19	-0.40
StMaPa																	-0.36	-0.57
TvNDSalt																		-0.21

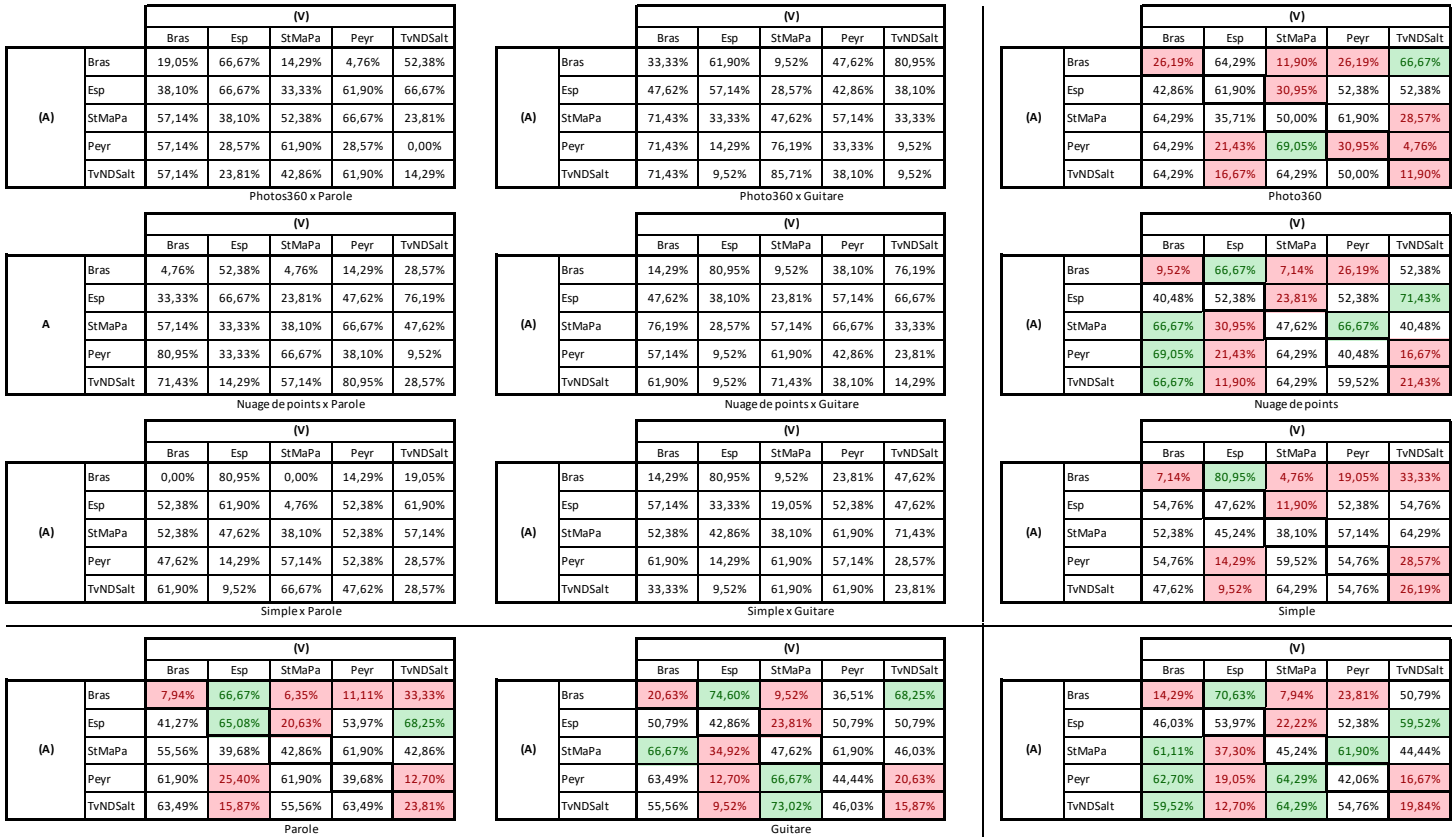
TABLE B.3 – Résultats des tests post-hoc (ajustement de Tukey) pour la distance reportée R_p (m). EMMs (*Estimated Marginal Means*) et significativité entre parenthèses : $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

	AviGC	AviStMa	AviTri	Bras	Bru	Esp	LaVdre	LMSHo	LMSRch	Peyn	Peyrl	Prue	PuyStMa	PuyStPce	Pvrt	StMaPa	TvNDSalt	TvStPbce
Anecho	3.73	-4.87	4.08	-5.77	-0.55	-2.01	-6.01 (*)	-1.54	-8.92 (**)	-3.46	1.55	-1.96	-1.39	-5.04	-4.46	-8.83 (***)	1.89	0.76
AviGC		-8.60 (*)	0.35	-9.49 (**)	-4.28	-5.73 (*)	-9.74 (***)	-5.27	-12.64 (***)	-7.18	-2.17	-5.68	-5.11	-8.77 (*)	-8.19 (*)	-12.56 (***)	-1.84	-2.97
AviStMa			8.94 (**)	-0.90	4.32	2.86	-1.14	3.33	-4.05	1.41	6.42	2.91	3.48	-0.17	0.41	-3.97	6.76	5.63
AviTri				-9.84 (***)	-4.63	-6.08 (*)	-10.08 (***)	-5.62	-12.99 (***)	-7.53 (*)	-2.52	-6.03	-5.46	-9.12 (**)	-8.53 (**)	-12.91 (***)	-2.18	-3.32
Bras					5.22	3.76	-0.24	4.23	-3.15	2.31	7.32	3.81	4.38	0.73	1.31	-3.07	7.66	6.53
Bru						-1.45	-5.46	-0.99	-8.36	-2.91	2.11	-1.41	-0.83	-4.49	-3.91	-8.28 (**)	2.44	1.31
Esp							-4.00	0.46	-6.91	-1.45	3.56	0.05	0.62	-3.04	-2.45	-6.83	3.90	2.77
LaVdre								4.47	-2.91	2.55	7.56 (*)	4.05	4.62	0.97	1.55	-2.82	7.90 (*)	6.77
LMSHo									-7.37	-1.92	3.10	-0.42	0.16	-3.50	-2.92	-7.29 (*)	3.43	2.30
LMSRch										5.46	10.47 (**)	6.96	7.53	3.87	4.46	0.08	10.81 (**)	9.67
Peyn											5.01	1.50	2.07	-1.58	-1.00	-5.38	5.35	4.22
Peyrl												-3.51	-2.94	-6.60	-6.01	-10.39 (***)	0.34	-0.80
Prue													0.57	-3.08	-2.50	-6.88	3.85	2.72
PuyStMa														-3.65	-3.07	-7.45 (**)	3.28	2.15
PuyStPce															0.58	-3.79	6.93	5.80
Pvrt																-4.38	6.35	5.22
StMaPa																	10.73 (***)	9.59 (**)
TvNDSalt																		-1.13

TABLE B.4 – Résultats des tests post-hoc (ajustement de Tukey) pour la taille apparente de source reportée S_{eq} (10^{-2} sr). EMMs (*Estimated Marginal Means*) et significativité entre parenthèses : $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

Annexe C

Résultats bruts de l'expérience 3 : Etude de la cohérence audio-visuelle d'environnements virtuels



(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	19,05%	66,67%	14,29%	4,76%	52,38%
Esp	38,10%	66,67%	33,33%	61,90%	66,67%
StMaPa	57,14%	38,10%	52,38%	66,67%	23,81%
Peyr	57,14%	28,57%	61,90%	28,57%	0,00%
TvNDSalt	57,14%	23,81%	42,86%	61,90%	14,29%

Photos360 x Parole

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	33,33%	61,90%	9,52%	47,62%	80,95%
Esp	47,62%	57,14%	28,57%	42,86%	38,10%
StMaPa	71,43%	33,33%	47,62%	57,14%	33,33%
Peyr	71,43%	14,29%	76,19%	33,33%	9,52%
TvNDSalt	71,43%	9,52%	85,71%	38,10%	9,52%

Photo360 x Guitare

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	26,19%	64,29%	11,90%	26,19%	66,67%
Esp	42,86%	61,90%	30,95%	52,38%	52,38%
StMaPa	64,29%	35,71%	50,00%	61,90%	28,57%
Peyr	64,29%	21,43%	69,05%	30,95%	4,76%
TvNDSalt	64,29%	16,67%	64,29%	50,00%	11,90%

Photo360

A \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	4,76%	52,38%	4,76%	14,29%	28,57%
Esp	33,33%	66,67%	23,81%	47,62%	76,19%
StMaPa	57,14%	33,33%	38,10%	66,67%	47,62%
Peyr	80,95%	33,33%	66,67%	38,10%	9,52%
TvNDSalt	71,43%	14,29%	57,14%	80,95%	28,57%

Nuage de points x Parole

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	14,29%	80,95%	9,52%	38,10%	76,19%
Esp	47,62%	38,10%	23,81%	57,14%	66,67%
StMaPa	76,19%	28,57%	57,14%	66,67%	33,33%
Peyr	57,14%	9,52%	61,90%	42,86%	23,81%
TvNDSalt	61,90%	9,52%	71,43%	38,10%	14,29%

Nuage de points x Guitare

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	9,52%	66,67%	7,14%	26,19%	52,38%
Esp	40,48%	52,38%	23,81%	52,38%	71,43%
StMaPa	66,67%	30,95%	47,62%	66,67%	40,48%
Peyr	69,05%	21,43%	64,29%	40,48%	16,67%
TvNDSalt	66,67%	11,90%	64,29%	59,52%	21,43%

Nuage de points

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	0,00%	80,95%	0,00%	14,29%	19,05%
Esp	52,38%	61,90%	4,76%	52,38%	61,90%
StMaPa	52,38%	47,62%	38,10%	52,38%	57,14%
Peyr	47,62%	14,29%	57,14%	52,38%	28,57%
TvNDSalt	61,90%	9,52%	66,67%	47,62%	28,57%

Simple x Parole

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	14,29%	80,95%	9,52%	23,81%	47,62%
Esp	57,14%	33,33%	19,05%	52,38%	47,62%
StMaPa	52,38%	42,86%	38,10%	61,90%	71,43%
Peyr	61,90%	14,29%	61,90%	57,14%	28,57%
TvNDSalt	33,33%	9,52%	61,90%	61,90%	23,81%

Simple x Guitare

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	7,14%	80,95%	4,76%	19,05%	33,33%
Esp	54,76%	47,62%	11,90%	52,38%	54,76%
StMaPa	52,38%	45,24%	38,10%	57,14%	64,29%
Peyr	54,76%	14,29%	59,52%	54,76%	28,57%
TvNDSalt	47,62%	9,52%	64,29%	54,76%	26,19%

Simple

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	7,94%	66,67%	6,35%	11,11%	33,33%
Esp	41,27%	65,08%	20,63%	53,97%	68,25%
StMaPa	55,56%	39,68%	42,86%	61,90%	42,86%
Peyr	61,90%	25,40%	61,90%	39,68%	12,70%
TvNDSalt	63,49%	15,87%	55,56%	63,49%	23,81%

Parole

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	20,63%	74,60%	9,52%	36,51%	68,25%
Esp	50,79%	42,86%	23,81%	50,79%	50,79%
StMaPa	66,67%	34,92%	47,62%	61,90%	46,03%
Peyr	63,49%	12,70%	66,67%	44,44%	20,63%
TvNDSalt	55,56%	9,52%	73,02%	46,03%	15,87%

Guitare

(A) \ (V)	Bras	Esp	StMaPa	Peyr	TvNDSalt
Bras	14,29%	70,63%	7,94%	23,81%	50,79%
Esp	46,03%	53,97%	22,22%	52,38%	59,52%
StMaPa	61,11%	37,30%	45,24%	61,90%	44,44%
Peyr	62,70%	19,05%	64,29%	42,06%	16,67%
TvNDSalt	59,52%	12,70%	64,29%	54,76%	19,84%

TABLE C.1 – Cohérence audio-visuelle perçue, indiquée par le taux de positivité, calculé pour chaque condition de test. A : environnements acoustiques, V : environnements visuels. Les cases en couleurs indiquent une différence statistique par rapport à la distribution théorique du hasard (intervalle de confiance à 0,95). En rouge : les conditions jugées non cohérentes, en vert : les conditions jugées cohérentes.

Résumé :

L'étude de la perception des environnements acoustiques est un champ de recherche largement exploré depuis le début du XX^{ème} siècle. Ces dernières décennies, l'essor de techniques et outils de captation et d'auralisation de champs acoustiques spatialisés, couplé au développement des technologies de réalité virtuelle permettent d'envisager de nouvelles problématiques relatives à l'immersion et à l'impression spatiale de ces environnements. Les travaux présentés dans cette thèse visent à proposer des nouveaux protocoles expérimentaux, basés sur ces outils et techniques, pour la caractérisation de la perception d'environnements acoustiques en conditions d'immersion 3D.

La première expérience présentée dans cette thèse porte sur l'évaluation perceptive de la qualité spatiale de l'auralisation ambisonique d'acoustiques de salles mesurées, en comparant les performances de localisation de sources sonores en contexte d'auralisation avec celles observées en conditions réelles d'écoute. Cette expérience, basée sur une méthode de report des attributs spatiaux de sources en réalité virtuelle met en évidence des dégradations de la précision spatiale des sources sonores induites par l'auralisation, et révèle également une dépendance de ces attributs saptiaux aux différentes conditions acoustiques à l'étude. Dans un second temps, un travail de caractérisation multi-dimensionnelle (acoustique et perceptive) d'un corpus d'acoustiques mesurées dans une vingtaines d'édifices du patrimoine est présenté. La caractérisation acoustique, basée sur le calcul d'un ensemble de descripteurs acoustiques, permet de proposer une cartographie acoustique des édifices et une catégorisation des paramètres acoustiques en trois grands facteurs. De son côté, la caractérisation perceptive, fondée sur la méthodologie de la première expérience révèle que, dans un contexte d'auralisation HOA, la distance perçue et la précision spatiale des sources semblent affectées par des propriétés acoustiques des salles étudiées. La troisième expérience propose d'étudier l'adéquation audio-visuelle entre des environnements acoustiques et leurs représentations visuelles en contexte d'immersion multi-modale. Cette étude révèle que la cohérence audio-visuelle est principalement évaluée sur une relation simple entre le temps de réverbération de la salle et le volume représenté visuellement. Ces travaux ouvrent des perspectives concrètes autour de l'amélioration des techniques d'auralisation, la caractérisation perceptive des acoustiques du patrimoine et la synthèse d'espaces acoustiques en contexte d'immersion multi-sensorielle.

Mots clés : Perception auditive spatiale, auralisation, réalité virtuelle, acoustique des salles.