Représentation et perception des espaces auditifs virtuels

Rozenn Nicol

REPRESENTATION ET PERCEPTION DES ESPACES AUDITIFS VIRTUELS

Rozenn NICOL

Mémoire d'Habilitation à Diriger des Recherches

Soutenue le 30 juin 2010

JURY

Président	**Prof. Jacques TISSEAU**
Directeur	**Prof. Claude DEPOLLIER**
Rapporteurs	**Prof. Jens BLAUERT**
	Prof. Stéphane NATKIN
	Enrique LOPEZ-POVEDA
Examinateur	**Olivier WARUSFEL**

Résumé

Un espace auditif virtuel (par référence à l'acronyme anglais VAS pour *Virtual Auditory Space*) est une scène sonore virtuelle constituée d'un ensemble de sources sonores qui n'existent que dans l'espace perceptif de l'auditeur. Cet espace est généré au moyen des technologies de spatialisation sonore (telles que : stéréophonie, technologie binaurale, Wave Field Synthesis ou Higher Order Ambisonics) qui reposent sur des modèles de représentation de la scène sonore. La modélisation est le premier aspect à étudier et concerne notamment les étapes de la captation et de la restitution de l'information spatiale. La notion de format audio spatialisé (et par la même les questions de conversion de format et de compression) est implicite. A l'autre extrêmité se situe la perception de l'espace ainsi généré pour évaluer comment l'auditeur perçoit les sources sonores virtuelles. Ce mémoire ouvre une réflexion sur ces différentes problématiques. En complément d'un état des lieux des connaissances actuelles, deux questions sont traitées en détails. La première question porte sur les technologies de spatialisation multi haut-parleurs en se focalisant sur les technologies Wave Field Synthesis (WFS) et Higher Order Ambisonics (HOA). Il est montré quel(s) système(s) concret(s) peuvent être mis en œuvre à partir des équations théoriques. Grâce à un formalisme unifié, les convergences entre les deux technologies sont mises en évidence, pour ouvrir sur une évaluation comparée. La seconde question concerne l'application de la spatialisation sonore à des terminaux individuels (c'est à dire mono auditeur) et portables, impliquant de façon préférentielle un rendu sur casque. Il s'agit du domaine de la technologie binaurale qui consiste à reproduire les signaux acoustiques à l'entrée des conduits auditifs. Cette technologie repose principalement sur la reproduction des indices de localisation qui résultent de l'interaction des ondes acoustiques avec le corps de l'auditeur et sont par la même fortement individuels. Il est décrit comment modéliser ces indices (notamment les informations temporelles correspondant à l'Interaural Time Difference ou ITD et les informations spectrales associées aux Indices Spectraux ou IS) et comment individualiser cette modélisation.

Abstract

A Virtual Auditory Space (VAS) is a virtual sound scene which is composed of several sound sources which only exist in the perceptive space of the listener. This space is created by technologies of sound spatialization (such as : stereophony, binaural technology, Wave Field Synthesis or Higher Order Ambisonics) which relies on models for representing the sound scene. Modelling is the first issue to be investigated : it concerns the steps of recording and rendering the spatial information. The concept of spatial audio format (as well as the related topics concerning format adaptation and spatial audio coding) is implicit. The opposite issue is the perception of the VAS, i.e. how the listener perceives the virtual sound sources. This document provides food for thought about all these issues. In addition to an overview of current knowledge, two questions are examined in details. The first question concerns spatialization technologies for multi loudspeaker array, focussing on Wave Field Synthesis (WFS) and Higher Order Ambisonics (HOA). It is shown how to derive feasable systems from the theoretical equations. A unified description allows one to point out the convergence between the two technologies and opens a comparative study. The second question deals with the adaptation of sound spatialization to individual (i.e. mono listener) and handheld devices, which implyies rendering over headphones. It is based on binaural technology which consists in reproducing the acoustic signals at the entrance of the listener's ear. This technology relies on the reproduction of the localization cues which result from the interaction of the acoustic wave with the listener's body and are therefore strongly individual. It is presented how to model these localization cues, considering the temporal information (i.e. Interaural Time Difference or ITD) and the spectral information (i.e. the Spectral Cues or SC), and how to customize them for one particular individual.

REMERCIEMENTS

Je souhaite tout d'abord exprimer mon immense gratitude à Claude Depollier et Laurent Simon, pour avoir cru en moi et pour m'avoir encouragée, pas à pas, tout au long de mon projet d'Habilitation à Diriger les Recherches. Sans eux je n'aurais sans doute jamais commencé, ni jamais fini...

Je voudrais ensuite remercier la Direction de France Telecom pour son accompagnement dans ce projet. Je tiens en particulier à remercier Jean-Pierre Petit, Responsable du laboratoire TECH/SSTP, et Christine Marcatte, qui lui a succédé à la tête du laboratoire TECH/OPERA. Je souhaite aussi remercier Bruno Lozach, Responsable de l'Unité de Recherche TPS, pour la confiance et le soutien qu'il m'a témoignés.

J'ai rédigé ce document pour présenter mes travaux de recherche sur la spatialisation sonore depuis mon arrivée au sein de la Division R&D de France Telecom. Mais, au delà de mes modestes contributions, ce mémoire est avant tout un témoignage des travaux de l'Equipe Son 3D d'Orange Labs. Je tiens à remercier ici mes "'compagnons de route"' : Marc Emerit dit "'Marco"', Jérôme Daniel dit "'Hieronymus"' et Gregory Pallone dit "'Greg"'.

Mais aussi, et surtout, cette Habilitation, je l'ai construite et acquise au travers des doctorants qui m'ont accompagnée pendant toutes ces années. J'espère leur avoir donné autant qu'ils m'ont appris. Un très grand merci à vous tous : Sylvain, Pierre, Adrien et Romain.

Et une pensée toute particulière à toi, Jean-Marie... Tu m'as tant apporté, aussi bien sur le plan scientifique que personnel, qu'il n'y a pas de semaine où je ne pense à toi... Tu es présent, tel un filigrane, tout au long de ce document... A jamais !...

Je n'oublie pas non plus ma famille dont le soutien et l'affection a largement contribué à ce projet, même si cet acronyme "'HDR"' vous a semblé si mystérieux... Souvenirs de soirées pordicaises à refaire le monde dans la véranda :-)

Et un "'big trugarez"' à tous mes amis dont la présence a su illuminer toute cette aventure : Antoine, Alan, Laetitia, Jean-Fi, Meme, Jean-Luc, Magali, Awena, Yuna, Tonton, Matmot, Ludo, Charly, Anne-Marie, Sabine, Bernard, Mandie, Nono, Pascal... Toutes mes excuses à ceux que j'ai oubliés...

Et, pour finir, une spéciale dédicace à toi... qui te reconnaîtras ;-)

Guide de lecture

Il me semble indispensable, en préambule de ce mémoire, de resituer les travaux qui vont y être présentés dans leur contexte. Il s'agit d'abord d'un contexte industriel (c'est à dire une équipe de recherche et développement à France Telecom R&D) qui impose certaines contraintes exogènes aux travaux menés. Par exemple, j'ai commencé par étudier la spatialisation sonore basée sur des réseaux multi hauts-parleurs, du type *Wave Field Synthesis* ou *Ambisonics*. Cependant, avec la montée en puissance du téléphone mobile, la technologie binaurale est apparue comme une solution privilégiée de spatialisation pour les terminaux mobiles et par suite est devenue mon axe prioritaire de recherche. Une autre spécificité du contexte industriel est l'accent mis sur le dépôt de brevets en amont (et potentiellement au détriment) de la rédaction d'articles.

Un deuxième élément du contexte que je voudrais souligner est que les travaux décrits résultent avant tout d'un travail d'équipe. Au delà de mon mémoire de recherche, une des ambitions de ce document est d'offrir une synthèse des travaux menés sur la spatialisation sonore au sein de l'équipe "Son 3D" du laboratoire TECH/OPERA (anciennement TECH/SSTP) d'*Orange Labs* (Division recherche et développement de France Telecom). C'est sur le fond de ce travail collectif que je tiens à présenter mes contributions personnelles qui, sans ce contexte général, perdraient tout leur sens. Mais c'est aussi parce que, pour moi, les échanges et les collaborations entre chercheurs sont un des ingrédients vitaux de tout travail de recherche, sans lequel il ne peut progresser et prospérer pleinement. Je pense tout particulièrement à l'étroite collaboration qui lie un doctorant et son encadrant. D'un côté l'encadrant apporte son expérience et son expertise du domaine. De l'autre le doctorant apporte, outre ses compétences scientifiques et son énergie, le point de vue critique d'un interlocuteur privilégié. De la bonne synergie de ces deux forces en présence jaillissent les étincelles de l'innovation. Ceci vaut également pour les relations avec des stagiaires même si l'échelle temporelle est plus courte. Pour l'essentiel, les travaux que je vais présenter sont le fruit de ces collaborations et il me semble important de le rappeler ici. A égale mesure, mon travail s'est aussi nourri de multiples collaborations avec les autres membres de l'équipe "Son 3D". Ainsi l'objectif de ce mémoire est d'illustrer mes travaux de recherche en interactivité avec mon équipe de recherche. La paternité de ces travaux est donc bien collective et doit être projetée sur chacun des membres (permanents et temporaires) de l'équipe.

Le document s'organise en deux parties. La première partie retrace les principaux éléments de mon curriculum vitae, en présentant mon parcours professionnel. Y sont précisées mes activités d'encadrement de recherche. La liste des brevets et des publications auxquels j'ai collaboré est donnée. Les partenariats scientifiques auxquels j'ai participé sont aussi mentionnés. La seconde partie du mémoire présente les travaux de recherche auxquels j'ai contribué dans le domaine de la spatialisation sonore, consistant à créer ou recréer des espaces auditifs virtuels. Cette seconde partie se compose des chapitres 1 à 3 :

- Le chapitre 1 vise à planter le décor des travaux en indiquant les concepts généraux relatifs aux espaces auditifs virtuels et en introduisant la terminologie associée.
- Le chapitre 2 présente mes travaux menés sur les technologies *Wave Field Synthesis* et *Higher*

Order Ambisonics. Ces technologies ont pour point commun d'utiliser des réseaux multi haut-parleurs et d'être destinées à une spatialisation multi auditeurs. Les travaux ont cherché à faire le point sur les convergences et divergences des deux méthodes, dans le prolongement de mes travaux de thèse.

– Le chapitre 3 décrit mes travaux sur la synthèse binaurale qui, par opposition aux deux méthodes précédentes, est une technologie dédiée aux terminaux individuels. Un des principaux obstacles auxquels se heurte cette technologie est l'adaptation à l'auditeur des filtres de spatialisation qui dépendent fortement de sa morphologie. Les travaux ont essentiellement porté sur cette question.

Pour chaque étude présentée, les travaux sont replacés dans leur contexte en précisant l'état des lieux des connaissances antérieures, les questions qui ont suscité l'étude en question, les points qui restent non résolus, ainsi que les applications potentielles identifiées ou mises en oeuvre. La perception des sources virtuelles est un axe de recherche qui est abordée de façon transverse au sein de chaque chapitre.

Le dernier chapitre conclut le mémoire en présentant les perspectives et les futurs travaux à mener.

Curriculum Vitae

NICOL Rozenn

Orange Labs

TECH/OPERA/TPS

2 Avenue Pierre Marzin, 22307 Lannion Cedex

Tél : 02 96 05 16 99

E-mail : rozenn.nicol@orange-ftgroup.com

Profession

Ingénieur de recherche en audio 3D

Domaines de compétences : audio 3D (technologies binaurales, Wave Field Synthesis, Ambisonic), réalité virtuelle, acoustique physique, électro-acoustique, acoustique des salles, traitement du signal

Formation

1999 : Docteur ès Acoustique, Université du Maine (Mention Très Honorable avec les Félicitations du Jury)

1996 : D.E.A. d'Acoustique Appliquée, Université du Maine (mention Très Bien)

1995 : Ingénieur C.N.A.M (Paris), spécialité Acoustique (mention Très Bien)

1993 : D.E.S.T. Physique, spécialité Acoustique (C.N.A.M., Paris)

1991 : B.T.S. Cinéma Option Son (Ecole Nationale Louis Lumière, Paris)

Parcours professionnel

Depuis juillet 2000 — **Ingénieur de recherche en audio 3D** (*Expert senior*)
Orange Labs, Division R&D de France Telecom (Lannion)

Chef de Projet : gérer, piloter et animer les travaux de recherche
- Projet de recherche "Conférence Audio Spatialisée" (2009)
- Projet de recherche "Audio Multicanal" (2009)
- Projet de recherche (TC) "Audio" (2009-2010)
- Projet de recherche (Opération) "Technologies de représentation, codage et perception" du Macropôle Interface Sciences (2007-2008)
- Lot Son 3D du projet HOLOS du Programme Vision CyberMonde et du PACR PACTPARSON (2003-2006)

Missions scientifiques : développer, enrichir et acquérir des technologies de spatialisation sonore pour les intégrer dans des applications de télécommunication
- Développement et évaluation des nouvelles technologies de rendu audio 3D (briques de spatialisation sonore pour la librairie logicielle FT : Wave Field Synthesis, multicanal 5.1, Ambisonic)
- Développement et intégration de technologies audio 3D dédiées aux terminaux mobiles (technologies binaurales)
- Développement de nouveaux modèles associant compression audio et spatialisation sonore pour le codage des signaux multicanaux

1999 - 2000 — **Ingénieur de Recherche en Acoustique Sous-Marine** (Stage post-doctoral, 9 mois)
I.F.R.E.M.E.R., Institut Français de Recherche pour l'Exploitation de la Mer (Brest)
Mission : développement de la chaîne de traitement d'un sondeur multi-faisceau à émission frontale (application à la cartographie des fonds marins), librairie logicielle de traitement et d'analyse des enregistrements sonar

1996 - 1999 — **Ingénieur de Recherche en Spatialisation Sonore** (Thèse de Doctorat, 3 ans)
France Telecom, Division R&D (Lannion)
Mission : conception, mise en œuvre et évaluation d'un système de restitution sonore spatialisée par holophonie pour le contexte de visioconférence (concept de téléprésence)

1996 — **Stage de D.E.A. en Acoustique Physique** (4 mois)
L.A.U.M., Laboratoire d'Acoustique de l'Université du Maine (Le Mans)
Mission : étude et modélisation de la propagation des ondes acoustiques dans les espaces à 1 et 2 dimensions (Principe de Huygens)

1994 - 1995 **Stage de Mémoire Ingénieur en Acoustique des Salles** (1 an)
I.R.C.A.M., Institut de Recherche et Coordination Acoustique et Musique (Paris)
Mission : Analyse de l'influence de l'effet de salle sur les performances d'une antenne acoustique par des modèles numériques (acoustique prévisionnelle, auralisation)

Encadrement de recherche

Encadrement de thèses

depuis déc. 2008 Direction de la thèse de Romain Deprez
Adaptation et optimisation de la diffusion d'un contenu multicanal à une configuration hétérogène et peu contrainte du système d'écoute (Université de Marseille, LMA, E. Friot)

depuis déc. 2007 Direction de la thèse d'A. Daniel
Modèle de masquage audio 3D pour les codeurs audio multicanaux (Université McGill, CIRMMT, S. MacAdams)

2005-2008 Direction de la thèse de Pierre Guillon
Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF (Université du Maine, LAUM, L. Simon)

2002-2005 Direction de la thèse de Sylvain Busson
Individualisation d'indices acoustiques pour la synthèse binaurale (IRCAM, O. Warusfel & Université de Marseille, LMA, P.-O. Mattei)

2000-2003 Co-direction de la thèse de Jean-Marie Pernaux
Spatialisation du son par les techniques binaurales : Application aux services de télécommunications (INPG, N. Martin)

Jurys

2009 Membre (*advisor*) du jury de thèse d'Audun Solvang (Norwegian University of Science and Technology, Directeur : P. Svensson)
Representation of High Quality Spatial Audio

2006 Membre invité du jury de thèse de Sébastien Moreau (Université du Maine, LAUM, Directeur : C. Depollier)
Etude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance

2006 Membre du jury de mémoire d'Ingénieur CNAM de Antoine Hurtado-Huyssen
Intensité acoustique appliquée aux métiers du son

2004 Membre (*advisor*) du jury de thèse de Werner de Bruijn (Technological University of Delft, Supervisor : M. Boone)
Application of Wave Field Synthesis in Videoconferencing

Encadrement de stages

2008-2010 Stage d'Apprentissage Ingénieur (Telecom Lille) de Matthieu Berjon
Technologies de spatialisation sonore : optimisation du rendu sur réseau multi haut-parleurs

de type 5.1

2008 Stage Master (Université d'Aix-Marseille, Ecole Centrale Marseille) de Thomas Guignard
Adaptation morphologique de HRTF non individuelles

2006 Stage Licence III (Université du Maine) de Kevin Derval
Etude d'optimisation des performances acoustiques sous contraintes fortes pour terminaux téléphoniques sans fil

2004-2005 Co-direction du post-doctorat de J.Faure
Etude, réalisation et évaluation de systèmes de synthèse binaurale statique et dynamique

2004 Co-direction du mémoire de fin d'études de l'Ecole Nationale Supérieure Louis Lumière de Raphaël Mouterde
Etude perceptive en vue de l'utilisation d'un système WFS au cinéma

Stage Ingénieur UTT de V. Choqueuse
Etude exploratoire de l'application des réseaux de neurones à la prédiction d'une base de données de HRTF

2003 Co-direction du stage DEA de Manuel Briand
Extraction & encodage des informations de spatialisation sonore : le Binaural Cue Coding (BCC)

2002 Stage DEA de Guillaume Lenost
Modélisation de fonctions de transfert acoustiques de têtes humaines (HRTF) et application à l'individualisation de la synthèse binaurale : modèles sphériques et ellipsoïdaux

2001 Stage MST Image et Son de Alan Blum
Mise en place et réalisation de tests de perception de la spatialisation du son à l'aide de techniques binaurales

2000 Stage DEA de Cyril Renard
Analyse objective et subjective d'une technique de rendu sonore 2D sur une zone d'écoute étendue, l'holophonie, en vue de réaliser un mur de téléprésence

1999 Stage DEA de Jean-Marie Pernaux
Restitution sonore spatialisée par antenne de haut-parleurs selon un procédé holophonique : étude des phénomènes de diffraction et mise en œuvre d'une solution basée sur l'annulation du champ diffracté

Activités d'enseignement

1998 - 2010 **Université de Bretagne Occidentale** (Brest)
Licence / Master Image & Son

- Acoustique des Salles
- Audio 3D

2009 - 2010 **ENSSAT** (Lannion)
3ième année, module Multimédia

- Spatialisation sonore

2009 - 2010 **Telecom Lille**

- Spatialisation sonore

2001 - 2010 **Ecole Nationale Supérieure des Télécommunications**
Master Signal Télécommunications Images Radar

- Audio 3D

1998 - 1999 **C.N.A.M.** (Lannion)
Cycle B

- Traitement Numérique du Signal

1997 - 1998 **A.B.R.E.T.** (Association Bretonne pour la Recherche et la Technologie, Lannion)
Intervention dans une classe de C.M.2 pour réaliser un atelier scientifique sur le Bruit dans la Ville

Partenariats

Initiation et pilotage de **contrats de recherche externes** :

- **Université de Bretagne Occidentale / ENIB (LYSiC)**
  Evaluation de la technologie H.O.A. (Higher Order Ambisonics) pour la création de contenus audio multicanaux (captation, post-production, restitution) en vue d'un contexte de production professionnelle (2008-2010)
- **University of York** (Department of Electronics / The Intelligent Systems Group/ Audio Lab) :
  Frontal externalization of binaural synthesis (2006 - 2009)
- **IRCAM** (Equipe d'Acoustique des Salles) :
  Technologies binaurales (2002 - 2005)
- **Delft University of Technology** (Laboratory of Acoustical Imaging and Sound Control) :
  Application of Wave Field Synthesis in Videoconferencing (2001 - 2004)

Participation à des **projets collaboratifs**

- Projet **CARROUSO** (IST 1999 20993 - 5th framework) : Creating, Assessing and Rendering

in Real Time of High Quality Audio-Visual Environments in MPEG-4 Context (2001 - 2003)
Ce projet était dédié à l'enregistrement, la transmission et la restitution d'une scène sonore réelle ou virtuelle préservant ses propriétés perceptives, notamment spatiales, et autorisant leur manipulation interactive. Ce projet s'appuyait d'une part, sur le format de codage MPEG4 qui privilégie, sur le plan de la spatialisation, une approche descriptive et paramétrique de la scène sonore, et d'autre part, sur la technologie de Wave Field Synthesis (WFS) pour la spatialisation sonore.

Co-direction de thèses
- **LMA** (Laboratoire de Mécanique et d'Acoustique, Université d'Aix-Marseille)
  co-direction de la thèse de Romain Deprez (2008-2011)
- **CIRMMT** (Université McGill)
  co-direction de la thèse d'Adrien Daniel (2007-2010)
- **LAUM** (Laboratoire d'Acoustique de l'Université du Maine)
  direction de la thèse de Pierre Guillon (2005 - 2008)
- **LMA** (Laboratoire de Mécanique et d'Acoustique, Université d'Aix-Marseille)
  direction de la thèse de Sylvain Busson (2002 - 2005)

Brevets

02/2010	Compression des flux audio multicanaux	FR1051420
10/2009	Traitement de données sonores encodées dans un domaine de sous-bandes	FR0957118
02/2008	Procédé et dispositif pour la détermination de fonctions de transfert de type HRTF	WO2009106783
03/2006	Procédé de synthèse binaurale prenant en compte un effet de salle	FR2899424
10/2005	Individualisation de HRTFs utilisant une modélisation par éléments finis couplée à un modèle correctif	EP1946612
01/2005	Procédé de modélisation de HRTF pour l'interpolation et l'individualisation des HRTF	FR2880755
12/2003	Procédé de spatialisation de sons synthétiques	EP1695335
02/2003	Procédé et système d'obtention automatisée de fonctions de transfert acoustiques associées à la morphologie d'un individu	FR2851878

Publications

Thèse

1999 Restitution sonore spatialisée sur une zone étendue : Application à la téléprésence
R. Nicol
Université du Maine, 1999.

Livre

2010 Binaural Technologies
R. Nicol

En cours de publication dans la Collection *AES Monograph* (Audio Engineering Society Inc., New York, U.S.A.)

2006 An anthology of articles on Spatial Sound Techniques, Part 2 : Multichannel Audio Technologies - Francis Rumsey Editor (pp. 136-153)
Audio Engineering Society, Inc. Library of Congress - New York, USA

Articles

2010 Sound spatialization by Higher Order Ambisonics : Encoding and decoding a sound scene in practice from a theoretical point view...
R. Nicol
2nd International Symposium on Ambisonics and Spherical Acoustics, Paris, May 6-7, 2010 (invited paper)

Lateralization threshold in binaural synthesis
S. Busson, R. Nicol , O. Warusfel & L. Gros
En cours de soumission à *IEEE Transactions on Multimedia*

Modelling interaural time difference for virtual auditory space : Assessment of various models in the light of auditory perception
S. Busson, R. Nicol , O. Warusfel & L. Gros
Article en préparation pour soumission à Acta Acustica

An initial validation of individual crosstalk cancellation filters for binaural perceptual experiments
A.H. Moore, A.I. Tew & R. Nicol
J. Audio Eng. Soc., Vol. 58, No. 1/2, 2010 January/February

2008 L'Acoustique dans les Télécommunications
R. Nicol, K. Bartkova, D. Charlet, O. Rosec, D. Virette, J.-L. Garcia, A. Guérin & L. Gros
Acoustique & Technique n°53 (article invité pour le Livre Blanc de l'Acoustique de la Société Française d'Acoustique)

Le son 3D dans toutes ses dimensions
R. Nicol, J. Daniel, M. Emerit, G. Pallone, D. Virette, N. Chetry, P. Guillon & S. Bertet
Acoustique & Technique n°52 (article invité)

Head-Related Transfer Functions reconstruction from sparse measurements considering a priori knowledge from database analysis : a pattern recognition approach
P. Guillon, R. Nicol & L. Simon
125ème Convention de l'A.E.S. (Audio Engineering Society), San Fransisco, 2-5 octobre 2008

Head-related Transfer Function customization by frequency scaling and rotation shifts based on a new morphological matching method
P. Guillon, Th. Guignard, R. Nicol & L. Simon
125ème Convention de l'A.E.S. (Audio Engineering Society), San Fransisco, 2-5 octobre 2008

An initial validation of individualised crosstalk cancellation filters for binaural perceptual experiments
A.H. Moore, A.I. Tew & R. Nicol
125ème Convention de l'A.E.S. (Audio Engineering Society), San Fransisco, 2-5 octobre 2008

2007 Headphone transparification : A novel method for investigating the externalisation of binaural sounds
A.H. Moore, A.I. Tew & R. Nicol
123ème Convention de l'A.E.S. (Audio Engineering Society), New York, 5-8 octobre 2007

Efficient binaural filtering in QMF domain for BRIR
D. Virette, P. Philippe, G. Pallone, R. Nicol, J. Faure, M. Emerit & A. Guérin
122ème Convention de l'A.E.S. (Audio Engineering Society), Vienne, 5-8 mai 2007

2006 Looking for a relevant similarity criterion for HRTF clustering : a comparative study
R. Nicol, V. Lemaire, A. Bondu & S. Busson
120ème Convention de l'A.E.S. (Audio Engineering Society), Paris, 20-23 mai 2006

2005 Subjective investigations of the interaural time difference in the horizontal plane
S. Busson, R. Nicol & B. Katz
118ème Convention de l'A.E.S. (Audio Engineering Society), Barcelone, 28-31 mai 2005

2003 Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging
J. Daniel, R. Nicol & S. Moreau
114ème Convention de l'A.E.S. (Audio Engineering Society), Amsterdam, 22-25 mars 2003

Perceptual Evaluation of Binaural Sound Synthesis : The Problem of Reporting Localization Judgments
JM. Pernaux, M. Emerit & R. Nicol
114ème Convention de l'A.E.S. (Audio Engineering Society), Amsterdam, 22-25 mars 2003

1998 Propagation dans les espaces de dimensions paires
J. Hardy, C. Depollier & R. Nicol
ESAIM, Vol. 5, pp. 111-116, 1998

Conférences invitées

2010 Sound spatialization by Higher Order Ambisonics : Encoding and decoding a sound scene in practice from a theoretical point view...
R. Nicol
2nd International Symposium on Ambisonics and Spherical Acoustics, Paris, May 6-7, 2010 (Keynote talk)

2007 Traitement du signal basé sur la perception : I. Audio 3D, réalité virtuelle, compression audio, II. Son 3D binaural

R. Nicol
Journées Fondatrices du Groupe Perception Sonore de la Société Française d'Acoustique (SFA) - 18-19 janvier 2007

2006 Technologies audio 3D pour les services de télécommunication
R. Nicol, M. Emerit, J. Daniel & G. Pallone
Journées d'Etude de la Spatialisation sonore (Télécom Paris, IRCAM et SFA), Paris, 24-25/01/2006

2003 Comparaison des approches Wave Field Synthesis et (Higher Order) Ambisonics pour l'encodage et la restitution de scènes sonores
J. Daniel & R. Nicol
Forum International du Son Multicanal, Paris, 23-24 octobre 2003

(Re)création de scènes sonores 3D par des réseaux multi haut-parleurs : Holophonie et concept de Wave Field Synthesis
R. Nicol
Journée transversale "Méthodes d'enregistrement et de restitution sonore pour l'évaluation de la perception des bruits de transports" (S.F.A.), Lyon, 15 septembre 2003

2002 Création, transmission et reproduction de champs sonores. Application de réseaux de microphones et de haut-parleurs à la spatialisation
R. Nicol
S.F.A., Paris, 2-3 décembre 2002 : Journées transversales "Réseaux de sources et de capteurs"

2001 Réseau de haut-parleurs & Spatialisation sonore : Concept de WaveField Synthesis, Application à la visioconférence
R. Nicol
Journée d'étude S.F.A./A.E.S. sur les diverses applications des réseaux de transducteurs en acoustique, Paris, 27 avril 2001

2000 3D-Sound Reproduction by Soundfield Reconstruction : An Holophonic Approach
R. Nicol
108ème Convention de l'A.E.S., Paris, 19-22 février 2000

Congrès

2010 Validation théorique de la correction des réflexions par l'ajout de sources virtuelles, sur la base d'une représentation en harmoniques sphériques
R. Deprez, R. Nicol & E. Friot
CFA10, 10ème Congrès Français d'Acoustique (Société Française d'Acoustique), Lyon, 12-16 avril 2010

2006 Non-linear interpolation of Head Related Transfer Function
S. Busson, R. Nicol, V. Choqueuse & V. Lemaire
CFA06, 8ème Congrès Français d'Acoustique (Société Française d'Acoustique), Tours, 24-27

avril 2006

2005 Individualized HRTFs from few measurements : a statistical learning approach
V. Lemaire, F. Clérot, S. Busson, R. Nicol & V. Choqueuse
International Joint Conference on Neural Networks IJCNN 2005, Montreal, Canada, 31 Juillet - 4 Août 2005

2004 Spatial perception of virtual 3D sound scene : how to assess the quality of 3D audio rendering by WFS ?
R. Nicol & L. Gros
CFADAGA 2004 (sociétés françaises et allemandes d'Acoustique), Strasbourg, 22-25 mars 2004

Influence of the ear canal location on spherical model for the individualized interaural time difference
S. Busson, R. Nicol & O. Warusfel
CFADAGA 2004 (sociétés françaises et allemandes d'Acoustique), Strasbourg, 22-25 mars 2004

2003 Listening room compensation for Wave Field Synthesis. What can be done ?
E. Corteel & R. Nicol
23ème Conférence Internationale de l'A.E.S. on Signal Processing in Audio Recording and Reproduction, Helsingor, 23-25 mai 2003

2002 Perceptual Evaluation of Static Binaural Sound Synthesis
JM. Pernaux, M. Emerit, J. Daniel & R. Nicol
22ème Conférence Internationale de l'A.E.S. on "Virtual, Synthetic and Entertainment Audio", Espoo, 15-17 juin 2002

1999 3D-Sound Reproduction over an Extensive Listening Area : a Hybrid Method derived from Holophony and Ambisonic
R. Nicol & M. Emerit
16ème Conférence Internationale de l'A.E.S. on Spatial Sound Reproduction, Rovaniemi, 10-12 avril 1999

Holophony versus Ambisonic : Deriving a Hybrid Method for 3D-Sound Reproduction in Videoconferencing
R. Nicol & M. Emerit
Forum Acusticum 99 (European Acoustics Association & Acoustical Society of America), Berlin, 14-19 mars 1999

1998 Reproducing 3D-Sound for Videoconferencing : a Comparison between Holophony and Ambisonic
R. Nicol & M. Emerit
D.A.F.X. 98 (Digital Audio Effects), Barcelone, 19-21 novembre 1998

Lending an Ear to the Dimensionality of Space
R. Nicol, C.Depollier & J. Hardy
16ème I.C.A. (International Congress on Acoustics), Seattle, 20-26 juin 1998

Mur de téléprésence pour la visioconférence : une approche holophonique
R. Nicol & M. Emerit
C.O.R.E.S.A. 98 (4èmes Journées d'Etudes et d'Echanges "Compression et Représentation des Signaux Audiovisuels"), Lannion, 9-10 juin 1998

Comités d'organisation

2010 2nd International Symposium on Ambisonics and Spherical Acoustics (IRCAM, Paris, France, 6-7 Mai 2009)
comité d'organisation

CFA10, 10ème Congrès Français d'Acoustique (Société Française d'Acoustique, Lyon, 12-16 Avril 2010)
co-organisation de la Session *"Réalité virtuelle et spatialisation sonore"*

2008 ITU-T Workshop on "From Speech to Audio Bandwidth extension, binaural perception" (Lannion, France, 10-12 Septembre 2008)
comité d'organisation

Ears Wide Open (Rennes, France, 11-13 Mars 2008)
organisation (Orange Labs) de Journées d'échanges sur la spatialisation sonore, en partenariat avec la SFA et l'AES

Comités de Review

2010 AES 40th International Conference on Spatial Audio (Tokyo, 8-10 Octobre 2010)

2009 International Congress on Auditory Display 09 (Copenhagen, 18-22 Mai 2009)

2008 Seventh Research Framework Programme (European Commission)

2007 AES 30th International Conference on Intelligent Audio Environments (Saariselkä, Finland, 15-17 mars 2007)

Associations

Membre de l'Audio Engineering Society (A.E.S.) depuis 1999

Membre de la Société Française d'Acoustique (S.F.A.) depuis 1995 & membre du Bureau du Groupe Perception Sonore de la SFA depuis 2009

Table des matières

Liste des acronymes

BEM Boundary Element Method
CPA Covert Peak Area
DF Diffuse Field
FEM Finite Element Method
FF Free Field
HOA Higher Order Ambisonics
HpTF Headphone Transfer Function
HRIR Head Related Impulse Response
HRTF Head Related Transfer Function
ICA Independent Component Analysis
IHL In-Head-Locatedness
ILD Interaural Level Difference
IPD Interaural Phase Difference
IS Indices Spectraux
ITD Interaural Time Difference
PCA Principal Component Analysis
SC Spectral Cues
SFRS Spectrum Frequency Response Surface
SRF Spatial Receptive Field
VAS Virtual Auditory Space
WFS Wave Field Synthesis

Chapitre 1

Les espaces auditifs virtuels

Un espace auditif virtuel (*Virtual Auditory Space* ou *VAS* en anglais) se définit comme une scène sonore perçue en tant que telle par l'auditeur, mais qui n'a pas de support tangible dans le monde physique. Du moins, à l'instant où l'auditeur perçoit cette scène, les sources perçues n'existent pas dans le monde physique. Cependant elles ont pu existé à des instants antérieurs s'il s'agit d'un enregistrement. En d'autres termes un espace auditif virtuel n'existe que dans la *perception* de l'auditeur : c'est une image mentale *suggérée* à l'auditeur. La suggestion s'effectue par le biais de signaux acoustiques appliqués aux tympans de l'auditeur et convenablement contrôlés de façon à produire l'illusion auditive souhaitée. A l'origine de l'espace auditif virtuel se trouvent certes des *sources réelles* : il s'agit des hauts-parleurs du système de reproduction sonore ou des transducteurs du casque d'écoute, mais dans ce contexte particulier les sources réelles ne sont pas perçues en tant que telles par l'auditeur[1] : elles "s'évanouissent" au profit des *sources virtuelles* autour desquelles se structure la scène sonore perçue.

Même si l'espace auditif virtuel est avant tout un phénomène perceptif, il prend sa source dans le monde physique : il repose en effet sur la stimulation de l'appareil auditif par des signaux acoustiques. Il semble raisonnable de considérer que ces signaux sont traités à l'instar de ceux créés par une scène sonore *naturelle*. Cependant les signaux issus d'une scène naturelle respectent certaines règles (par exemple les relations de cohérence entre les indices temporels et spectraux de localisation auditive) qui, lorsqu'elles sont violées, aboutissent à des distorsions de l'espace auditif. La localisation intracrânienne[2] des sources virtuelles dans certaines situations d'écoute est un exemple de telles distorsions.

Dans la mise en œuvre d'un espace auditif virtuel, deux principales étapes interviennent :

- d'une part la **synthèse des signaux acoustiques** appliqués aux tympans de l'auditeur, cette synthèse étant le résultat de la mise en œuvre d'un modèle de représentation d'une scène audio 3D,
- d'autre part les **processus de traitement de ces signaux par l'appareil auditif** opérant aussi bien au niveau périphérique (au niveau de l'oreille interne : codage de l'information fréquentielle, extraction des indices temporels de localisation auditive par exemple) qu'au niveau central (au niveau du cortex auditif : exploitation des indices spectraux, analyse de scène auditive par exemple).

La synthèse d'un espace auditif virtuel suppose donc de contrôler les signaux acoustiques soumis aux tympans de l'auditeur sur la base de la connaissance des mécanismes d'analyse du système auditif

[1] Du moins le cas où l'auditeur identifie les transducteurs du système de reproduction comme des sources est jugé comme un dysfonctionnement qui doit être à tout prix évité.

[2] La localisation intracrânienne correspond à la situation où les sources virtuelles sont localisées à l'intérieur de la tête de l'auditeur.

pour interpréter et construire la scène sonore perçue. Quelle que soit la méthode de synthèse de l'espace auditif virtuel, il est impossible d'occulter le système auditif et la perception de l'auditeur. Même si certaines méthodes visent la reconstruction parfaite de la pression acoustique à l'identique de celle induite par des sources réelles (comme par exemple l'holophonie qui sera décrite plus loin), il importe toujours de tenir compte du système auditif et de la façon dont il va traiter ces informations, ne serait-ce que parce qu'aucune méthode n'est parfaite et introduit des artéfacts qu'il convient d'évaluer du point de vue de la perception. Mes travaux s'organisent ainsi autour de deux thématiques suivantes :

- les **modèles de représentation d'une scène audio 3D** permettant de synthétiser un espace audtif virtuel,
- la **perception de ces modèles**, en incluant les processus intervenant à la fois au niveau périphérique et au niveau central.

L'objectif de ce chapitre est de donner les concepts fondamentaux relatifs aux espaces auditifs virtuels afin d'introduire le cadre et la terminologie des études présentées aux chapitres suivants. La première partie montre comment un espace auditif virtuel est en fait un *espace pluriel* constitué de la concaténation de plusieurs espaces représentant les différentes étapes entre l'espace physique de la scène sonore et l'image mentale de l'auditeur. La seconde partie introduit le schéma général d'un modèle de représentation d'une scène audio 3D dans lequel on retrouve une démarche d'*analyse et synthèse*. Il existe aujourd'hui une grande variété de modèles de scènes audio 3D. La troisième partie présente un panorama des principaux modèles disponibles. Au delà du seul problème de représentation de la scène sonore, les modèles sont amenés à prendre en charge des fonctionnalités complémentaires, telles que la possibilité de manipuler la scène sonore ou le souci d'une représentation efficace à des fins de compression. Ces aspects sont traités dans la quatrième partie. Enfin la cinquième partie propose un état des lieux des travaux de recherche sur les espaces auditifs virtuels en indiquant les questions qu'il reste à résoudre.

1.1 Un espace pluriel

Entre la création[3] originelle de la scène sonore et sa re-création dans l'espace perceptif de l'auditeur, il existe plusieurs étapes intermédiaires qui correspondent à autant d'espaces[4] différents de représentation de la scène sonore. Ces espaces définissent les multiples composantes qui vont concourir à l'espace auditif virtuel.

Espace physique A l'origine de la scène sonore se trouve d'abord l'espace *physique* des sources acoustiques. Cet espace est décrit par :

- des paramètres géométriques et acoustiques des sources : position, directivité, orientation et signal émis,
- des paramètres géométriques et acoustiques de l'environnement des sources : géométrie de la salle, propriétés d'absorption et de réflexion des ondes acoustiques par les parois, présence d'éléments diffractants ou diffusants etc...

Espace acoustique primaire Les sources acoustiques rayonnent dans l'espace physique et génèrent ainsi une onde acoustique. La représentation de cette onde acoustique en fonction des coordonnées d'espace et du temps définit l'espace (ou représentation) *acoustique primaire* de la scène sonore. Il est qualifié de primaire au sens où il résulte directement des sources

[3]Cette création n'existe au sens propre que dans le cas d'une scène sonore naturelle enregistrée au moyen d'un système de captation sonore donné.

[4]Le concept de ces espaces est issu de discussions avec Jérôme Daniel.

acoustiques *primaires* originelles par opposition aux sources *secondaires* (hauts-parleurs par exemple) qui vont permettre de le simuler.

Espace de captation En vue de sa reproduction, la scène sonore est enregistrée au moyen d'un système de captation donné constitué d'un ensemble de microphones. Les signaux microphoniques définissent la représentation de la scène sonore dans l'espace de *captation*.

Espace de restitution Les signaux microphoniques sont destinés à alimenter, soit directement (par exemple dans le cas de la Wave Field Synthesis), soit après une transformation (qui peut consister par exemple en un matriçage dans le cas d'Ambisonic), un système de restitution (dispositif de plusieurs hauts-parleurs ou casque d'écoute). Les signaux alimentant le système de restitution définissent une nouvelle représentation, cette fois dans l'espace de restitution.

Espace acoustique secondaire Le système de restitution est composé de sources acoustiques (telles que des hauts-parleurs ou les transducteurs d'un casque) qui peuvent être considérées comme des sources *secondaires* par opposition aux sources *primaires* présentes dans l'espace acoustique primaire. Ces sources secondaires donnent lieu à une nouvelle onde acoustique dont la finalité est, une fois interprêtée dans l'espace perceptif de l'auditeur, de susciter une image de la scène sonore la plus proche de celle qu'aurait suscitée l'onde acoustique primaire. Cette onde acoustique secondaire définit l'espace acoustique virtuel.

Espace binaural L'onde acoustique secondaire induite par le système de restitution se propage jusqu'à l'auditeur. Elle va alors interagir avec le corps de l'auditeur par un jeu de diffractions et de réflexions sur les différents éléments de sa morphologie (principalement la tête, les épaules, le haut du torse sans oublier le rôle particulier du pavillon). L'onde acoustique résultante qui vient exciter les tympans de l'auditeur définit l'espace *binaural* de représentation de la scène audio 3D. La transformation de l'espace acoustique à l'espace binaural traduit l'opération d'encodage acoustique de la direction des sources, dans laquelle sont introduits les indices de localisation auditive destinés à être interprêtés au niveau périphérique et central.

Espace perceptif périphérique L'onde acoustique qui met en vibration le tympan est encodé au niveau de l'oreille interne sous la forme d'impulsions électriques qui sont transmises par le nerf auditif au système central. Les informations véhiculées dans l'impulsion nerveuse concernent la fréquence et l'intensité des sons, ainsi que les différences interaurales de temps (ITD) et d'intensité (ILD) qui sont des indices de localisation[5]. Ces informations résultent des premières analyses de la scène sonore par le système auditif périphérique et définissent l'espace perceptif périphérique.

Espace perceptif central En complément de l'analyse du système périphérique, de nouvelles informations sur la scène sonore sont extraites par le système central. Par exemple, les olives supérieures médiane (MSO) et latérale (LSO), et les noyaux cochléaires dorsaux (DCN) ont été identifiés comme participant au traitement de l'ITD, l'ILD et des IS, respectivement [Wanrooij & Opstal, 2006]. Interviennent ici également des processus cognitifs qui participent notamment à l'analyse de scène auditive [Bregman, 1990]. Le système central construit la représentation mentale de la scène sonore correspondant à l'espace perceptif central. Cette représentation définit l'illusion auditive perçue par l'auditeur.

L'enchaînement de transformations qui vient d'être décrit s'applique en fait au cas d'une **scène naturelle**, c'est à dire d'une scène créée à partir de sources primaires réelles et *naturellement* captée par un ensemble de microphones. On peut aussi rencontrer le cas d'une **scène artificielle** pour laquelle les deux premiers espaces (espace physique et espace acoustique réel) n'existent pas. A ces

[5]Plus précisément l'ILD et l'ITD sont des indices de *latéralisation*, c'est à dire qu'ils permettent de localiser les sources sonores sur l'axe gauche-droite traversant les oreilles de l'auditeur.

espaces se substitue un espace dit de **synthèse** spécifique aux scènes artificielles et correspondant aux paramètres décrivant (ces paramètres définissent en fait un format donné de description) les propriétés de la scène artificielle en termes de sources (nombre, position, directivité, spectre etc..) et de leur environnement acoustique (géométrie, comportement acoustique des parois...). A partir de cette représentation de synthèse sont générés les *pseudo* signaux microphoniques définissant l'espace de captation *virtuel*. Si l'on désire associer des éléments d'une scène naturelle et d'une scène artificielle, il est possible de les combiner à partir de l'espace de captation.

1.2 Modèle de représentation d'une scène audio 3D

Dans ce qui précède, seuls les espaces de captation, de restitution et l'espace acoustique secondaire relèvent à proprement parler du domaine des technologies de spatialisation sonore. Il convient de les distinguer des espaces intervenant en amont et en aval et qui se rattachent respectivement au domaine de l'acoustique des salles et du domaine de la perception auditive spatialisée. Dans cette partie nous allons nous focaliser sur le domaine des technologies audio 3D afin d'introduire les concepts généraux qui les définissent.

Avant tout une technologie de spatialisation sonore se fonde sur un modèle de représentation d'une scène audio 3D. Ce modèle emprunte un schéma classique d'analyse/synthèse reposant sur les deux étapes suivantes :

- **analyse** : Il s'agit d'extraire de l'espace acoustique primaire les informations à la fois temporelles et spatiales caractérisant la scène audio 3D et qui serviront à la reproduire. Cette extraction est réalisée par une distribution donnée de microphones (cf. espace de captation) qui est spécifique à chaque technologie et qui présente ses avantages et ses défauts. Cette étape est souvent référencée comme un *encodage* de la scène sonore (à savoir dans l'espace de captation). Les signaux microphoniques qui en résultent définissent un *format audio 3D*.
- **synthèse** : Il s'agit de synthétiser une onde acoustique qui est destinée à exciter les tympans de l'auditeur pour lui donner l'illusion de la scène sonore. Cette opération comporte deux étapes : d'abord la conversion des signaux microphoniques en signaux destinés à alimenter le système de reproduction (distribution de hauts-parleurs ou casque), puis la génération par les transducteurs de ce système d'ondes acoustiques élémentaires dont la superposition donne lieu à l'onde acoustique secondaire que va percevoir l'auditeur. Faisant suite à l'opération d'encodage, ce processus correspond au *décodage*[6] des signaux microphoniques.

Dans la plupart des cas, le type d'encodage de la scène audio 3D détermine de façon univoque le décodage qui doit être appliqué aux signaux microphoniques pour la restituer. Par exemple un enregistrement stéréophonique est dédié à être écouté sur une paire stéréophonique de hauts-parleurs disposés selon une configuration standardisée (norme ITU). Il en est ainsi de la plupart des technologies de spatialisation sonore. Cependant, avec la multiplication des formats audio 3D, la question de la conversion des formats est devenue un axe actif de recherche. Les premiers travaux se portent sur les technologies de *upmix*, pour convertir un flux stéréo en un flux multicanal 5.1, et de *downmix*, pour la conversion inverse. L'alternative au *downmix stéréo* est le *downmix binaural* qui s'applique à n'importe quel flux audio 3D destinés à un dispositif donné de haut-parleurs et qui consiste à simuler en synthèse binaurale un dispositif virtuel de hauts-parleurs. Il en résulte un signal binaural destiné à une écoute au casque. Ces outils de conversion de format audio 3D introduisent une certaine flexibilité entre l'encodage et le décodage en offrant des passerelles entre

[6]Selon certains auteurs, le décodage se limite à la première étape de conversion des signaux microphoniques en signaux destinés à alimenter les transducteurs de restitution. Dans tout ce document on considère que le décodage englobe à la fois l'ensemble des processus (traitement du signal, émission et propagation de l'onde acoustique secondaire) qui conduisent à l'onde acoustqiue excitant les tympans de l'auditeur.

formats. Il ne faut pas non plus perdre de vue que l'équipement de rendu audio spatialisé qu'on puisse espérer rencontrer le plus couramment chez le particulier reste encore pour de nombreuses années le système 5.1 (cf. Recommandation ITU-R BS. 775-1). Si l'on veut pouvoir diffuser de nouveaux formats audio 3D, un atout pour leur succès est leur compatibilité avec une diffusion sur un système 5.1. Cependant la conversion de formats n'est qu'un pis-aller : la qualité de restitution est toujours dégradée en comparaison du décodage dédié qui offre évidemment la qualité optimale.

Un modèle de représentation d'une scène audio 3D est donc caractérisé principalement par la nature des informations qui sont extraites de la scène sonore, par les paramètres qui vont être chargés de représenter ces informations et par la méthode par laquelle ces dernières sont exploitées pour créer l'illusion de la scène à l'auditeur. La construction d'un modèle se base à la fois sur des connaissances de la propagation des ondes acoustiques et de la perception auditive. Les deux questions fondamentales posées concernent la **description** d'une scène audio 3D et sa **(re)création**.

1.3 Principaux modèles audio 3D

Cette partie propose un panorama des principales technologies de spatialisation sonore disponibles aujourd'hui [Nicol et al., 2008]. Chaque technologie est présentée sous la forme d'un tableau synthétique décrivant notamment son principe d'analyse/synthèse, le nombre de dimensions[7] effectivement spatialisées, les systèmes de prise et restitution sonore associés, les atouts et les défauts observés.

[7]Le nombre de dimensions définit le nombre de dimensions géométriques de la scène audio 3D selon lesquelles une technologie de spatialisation sonore est effectivement capable de faire évoluer les sources virtuelles. Un rendu monophonique correspond à une prise de son par un seul microphone associé à un restitution sur un haut-parleur unique. Même s'il est traditionnellement opposé aux technologies de spatialisation sonore comme le degré 0 de spatialisation, il n'est pas totalement dénué d'informations spatiales. L'information de distance des sources sonores, principalement à travers la perception du rapport entre les énergies de l'onde directe et de la réverbération, est en effet préservée dans un enregistrement monophonique. Un rendu monophonique possède donc une dimension spatiale : la distance entre l'auditeur et la source, c'est à dire le rayon dans un système de coordonnées sphériques. Par suite on qualifiera la monophonie de spatialisation 1D. A fortiori l'information de distance est aussi présente dans n'importe quel rendu audio 3D. Ainsi la spatialisation minimale attendue d'une technologie "audio 3D" est supérieure à 1D. A l'information de distance s'ajoute alors l'information en direction correspondant à deux dimensions et décrite par les angles d'azimut et d'élévation dans un système de coordonnées sphériques. Au passage il convient de remarquer que dans le présent document on rassemble sous l'appellation générique "audio 3D" des technologies qui n'offrent pas forcément une spatialisation 3D complète.

Technologie	**Stéréophonie**
Modèle	Modèle essentiellement perceptif basé sur les indices de latéralisation que sont les différences interaurales de temps (ITD) et de niveau (ILD) et qui pilotent la localisation des sources sonores selon l'axe gauche-droite de l'auditeur
Analyse	Extraction d'une différence de temps et/ou d'intensité entre deux points de l'espace acoustique primaire
Synthèse	Reproduction d'une différence de temps et/ou d'intensité entre les oreilles de l'auditeur permettant de localiser une source virtuelle ou *fantôme* entre les deux hauts-parleurs : La reproduction stéréophonique bénéficie d'un *artéfact* de la perception par lequel en présence de deux sources réelles (c'est à dire les hauts-parleurs) l'auditeur perçoit une source virtuelle unique localisée entre les deux sources réelles. Il s'agit d'un phénomène de *fusion* résultant de processus d'analyse de scène auditive [Bregman, 1990].
Nombre de dimensions	1D 1/6 : L'espace des sources virtuelles est limité à la portion du plan horizontal comprise entre les deux hauts-parleurs stéréophoniques.
Prise de son	Couple de microphones coïncidents (stéréophonie par différence d'intensité) ou distants (stéréophonie de temps) ou une combinaison des deux (stéréophonie mixte, par exemple couple AB). De nombreux systèmes sont disponibles
Reproduction sonore	Paire de hauts-parleurs disposés selon le triangle équilatéral stéréophonique (Recommandation ITU-R BS. 775-1)
Encodage virtuel	Panoramique d'intensité (loi des sinus, loi des tangentes)
Format associé	Stéréophonie (2 signaux)
Compatibilité	Stéréo → multicanal 5.1 (upmix : conversion d'un flux stéréo en flux 5.1) Multicanal 5.1 → stéréo (downmix : conversion d'un flux 5.1 en flux stéréo)
Domaines d'application	Prise de son musicale, cinéma, radio
Atouts	Simplicité de mise en oeuvre Spatialisation robuste Format compact (2 signaux) Le dispositif d'écoute stéréophonique tend à devenir l'équipement de base (PC multimédia par exemple).
Défauts	Spatialisation restreinte (zone horizontale frontale) Pas de spatialisation en élévation Un seul point d'écoute optimale (*sweet spot*)

Technologie	**Multicanal 5.1** (ainsi que ses futures déclinaisons : 6.1, 7.1, 10.2, 22.2...)
Modèle	Extension de la stéréophonie : ajout d'un canal central pour stabiliser les sources frontales, et de canaux arrières pour les ambiances et l'effet de salle
Analyse	Extraction de différences d'intensité et/ou de temps entre plusieurs points de l'espace acoustique primaire (de façon à couvrir l'ensemble du plan horizontal, en distinguant la zone frontale privilègiée de la zone arrière)
Synthèse	reproduction d'une différence de temps et/ou d'intensité entre les oreilles de l'auditeur permettant de localiser des sources virtuelles dans le plan horizontal.
Nombre de dimensions	2D
Prise de son	Arbres multicanaux : INA 5, Fukada-Tree, OCT-Surround, IRT-Cross, Hamasaki-Square [Theile, 2001]
Reproduction sonore	Configuration de 5 hauts-parleurs et un caisson de graves selon la Recommandation ITU-R BS. 775-1
Encodage virtuel	Panoramique d'intensité
Format associé	Multicanal 5.1
Compatibilité	Multicanal 5.1 ↔ stéréo (downmix, upmix) Multicanal 5.1 → binaural (downmix binaural : conversion d'un flux 5.1 en flux binaural) Ambisonic/HOA → multicanal 5.1 (adaptation d'un flux Ambisonic/HOA pour un système d'écoute 5.1) Multicanal 5.1 → WFS (adaptation d'un flux 5.1 pour un système d'écoute WFS, en synthétisant 5 ondes planes dans les directions des hauts-parleurs du système 5.1)
Domaines d'application	Prise de son musicale, cinéma, radio
Atouts	Format compact (6 canaux) Standard
Défauts	Un seul point d'écoute optimal (sweet spot) Pas de spatialisation en élévation Spatialisation horizontale hétérogène en fonction de l'azimut : zone frontale privilégiée, zones latérales défavorisées à cause de l'écart angulaire des haut-parleurs

Technologie	**Ambisonic & Higher Order Ambisonic (HOA)** [Gerzon, 1980] [Gerzon, 1985] [Gerzon, 1992a] [Gerzon, 1992b] [Bamford, 1995] [Daniel, 2000]
Modèle	Modèle basé sur une décomposition *mathématique* de l'onde acoustique (espace acoustique primaire) utilisant la base des harmoniques sphériques (fonctions propres de l'équation des ondes acoustiques en géométrie sphérique) Ambisonic correspond à la décomposition limité à l'ordre 1, tandis que HOA en est la généralisation aux ordres supérieurs [Daniel, 2000].
Analyse	Extraction des coefficients de la décomposition en harmoniques sphériques (analyse de la distribution spatiale des ondes acoustiques, analyse de plus en plus fine au fur et à mesure que l'ordre des harmoniques augmente)
Synthèse	Projection de la distribution spatiale de l'onde acoustique primaire sur le dispositif de hauts-parleurs, la projection pouvant être optimisée au sens d'un ou plusieurs critères de décodage, ce qui donne lieu à différentes lois de décodage (par exemple : décodage *basique*, décodage *max* r_E, décodage *in phase*)
Nombre de dimensions	2D ou 3D
Prise de son	Microphone Soundfield (Ambisonic à l'ordre 1) [Farrar, 1979a] [Farrar, 1979b] [Craven & Gerzon, 1977] Sphère de microphones (HOA) (problèmes d'échantillonnage spatial et de troncature de la décomposition en harmoniques sphériques) [Moreau, 2006]
Reproduction sonore	Dispositif de N (N $\geq$ 4) hauts-parleurs (distribution régulière ou non)
Encodage virtuel	Loi de panoramique agissant sur l'intensité et la phase des signaux (espace de captation virtuel) simulant un encodage sur la base des harmoniques sphériques
Format associé	B-format (Ambisonic à l'ordre 1) HOA ($[2M + 1]^2$ signaux correspondant à une décomposition à l'ordre M)
Compatibilité	Ambisonic/HOA $\rightarrow$ multicanal 5.1 (adaptation d'un flux Ambisonic/HOA pour un système d'écoute 5.1) Ambisonic/HOA $\rightarrow$ binaural (downmix binaural : conversion d'un flux 5.1 en flux binaural)
Domaines d'application	Utilisation marginale d'Ambisonic à l'ordre 1 dans le monde audio professionnel Technologie majoritairement expérimentale aujourd'hui

Technologie	**Ambisonic & Higher Order Ambisonic (HOA)** (suite)
Atouts	Format audio 3D *hiérarchique* (chaque nouvelle composante vient seulement compléter l'information contenue dans les harmoniques inférieures) et *flexible* (le décodage s'adapte au nombre et à la disposition des hauts-parleurs du dispostif d'écoute) Spatialisation 3D complète Extension de la zone d'écoute avec les ordres supérieurs Possibilité de manipuler la scène sonore à l'issue de l'enregistrement
Défauts	Nombre élevé de signaux Qualité audio qui reste à améliorer pour convaincre les ingénieurs du son

Technologie	**Holophonie & Wave Field Synthesis (WFS)** [Berkhout, 1988] [Vogel, 1993] [Start, 1997] [Verheijen, 1998] [de Bruijn, 2004] [Nicol, 1999]
Modèle	Modèle basé sur le Principe de Huygens : recomposition d'une onde acoustique (espace acoustique primaire) par superposition d'ondelettes (décomposition *physique* de l'onde acoustique) [Jessel, 1973]
Analyse	Extraction des différences de temps et d'intensité sur une distribution dense et étendue de points de l'espace acoustique primaire
Synthèse	Chaque haut-parleur émet une ondelette convenablement paramétrée en temps et en intensité et qui en se superposant aux ondelettes générées par les autres hauts-parleurs va reconstituer une copie de l'onde acoustique primaire
Nombre de dimensions	2D ou 3D
Prise de son	Réseau étendu de microphones (en théorie), pas de système utilisé en pratique (encodage virtuel)
Reproduction sonore	Réseau étendu de hauts-parleurs (problème d'échantillonnage spatial et de troncature)
Encodage virtuel	Contrôle en amplitude et en temps des signaux (espace de captation virtuel) pour simuler une prise de son par un réseau microphonique (concept de *source notionnelle* [Berkhout et al., 1993])
Format associé	Aucun
Compatibilité	Holophonie/WFS → binaural (downmix binaural en théorie, mais non évalué) Stéréophonie, multicanal 5.1 → holophonie/WFS par synthèse d'ondes planes
Domaines d'application	Technologie expérimentale, quelques exemples de mise oeuvre au cinéma [IOSONO Sound, 2010]
Atouts	Spatialisation 2D (voire 3D, mais encore aujourd'hui on dispose de peu de recul sur le rendu 3D) complète et naturelle : les sources virtuelles sont perçues à la fois présentes et naturelles. Zone d'écoute étendue
Défauts	Absence de système de prise de son associé Nombre élevé de signaux

Technologie	**Binaural** [Møller, 1992]
Modèle	Modèle basé sur l'imitation de la perception auditive et visant à reproduire au niveau des tympans de l'auditeur (espace binaural) les indices de localisation perçus en situation d'écoute naturelle
Analyse	Extraction la plus exhaustive possible des indices de localisation sonore (notamment les différences interaurales de temps et d'intensité, ainsi que les indices spectraux)
Synthèse	Reproduction des indices de localisation sonore
Nombre de dimensions	3D
Prise de son	Paire de microphones binauraux placés sur une tête naturelle ou artificielle
Reproduction sonore	Casque Paire de hauts-parleurs [Gardner, 1997] : par exemple système transaural [Atal & Schroeder, 1966] [Cooper & Bauck, 1989] [Bauck & Cooper, 1996], stéréo dipôle [Kirkeby et al., 1997] Système de 4 hauts-parleurs [Guastavino et al., 2007]
Encodage virtuel	Synthèse binaurale : mise en œuvre de filtres binauraux reproduisant les fonctions de transfert acoustiques entre la source sonore et les tympans de l'auditeur (fonctions de transfert dites HRTF pour Head Related Transfer Function)
Format associé	Signal binaural (2 canaux)
Compatibilité	Multicanal 5.1, Ambisonic/HOA → binaural (downmix binaural)
Domaines d'application	Réalité virtuelle Outil d'expérimentation pour la perception auditive spatiale
Atouts	Spatialisation 3D complète et naturelle Format compact Restitution au casque (compatible avec les terminaux mobiles)
Défauts	Spatialisation individuelle Introduction de colorations spectrales qui nuisent à la transparence des timbres

Technologie	**Vector Base Amplitude Panning (VBAP) & Vector Base Intensity Panning (VBIP)** [Pulkki & Lokki, 1998] [Pernaux et al., 1998]
Modèle	Projection de la position de la source virtuelle sur une base vectorielle constituée par une paire (2D) ou un triplet (3D) de hauts-parleurs (généralisation de la loi des tangentes utilisée en stéréophonie)
Analyse	Spécification de la position de la source virtuelle dans l'espace physique
Synthèse	Combinaison linéaire des contributions des 2 (2D) ou 3 (3D) hauts-parleurs les plus proches de la position cible de la source virtuelle (stéréophonie locale par panoramique d'intensité selon une loi des tangentes)
Nombre de dimensions	2D ou 3D
Prise de son	Encodage virtuel seul
Reproduction sonore	Réseau sphérique de hauts-parleurs
Encodage virtuel	Loi de panoramique locale (limité à 2 ou 3 hauts-parleurs), agissant en amplitude (VBAP) ou en énergie (VBIP) selon la loi des tangentes, généralisée à une reproduction 3D le cas échéant
Format associé	Aucun
Compatibilité	VBAP/VBIP → binaural (downmix binaural)
Domaines d'application	Technologie expérimentale
Atouts	Simplicité de mise en œuvre
Défauts	Zone d'écoute restreinte (voisinage du centre de la sphère de hauts-parleurs) Hétérogénité du rendu quand la source virtuelle se déplace [Pernaux et al., 1998]

1.4 Fonctionnalités avancées des modèles audio 3D

La première finalité d'un modèle audio 3D est de représenter le plus fidèlement possible une scène sonore, notamment au travers d'un format audio 3D. Au delà de la représentation de la scène sonore, un modèle audio 3D est susceptible d'assumer des fonctionnalités complémentaires [Nicol et al., 2008] :

- **manipulation** de la scène sonore,
- **compatibilité** avec les autres modèles,
- **flexibilité** en termes de prise et de restitution du son,
- **compression** de la représentation.

1.4.1 Manipulation de la scène sonore

Il ne faut pas perdre de vue qu'à l'étape d'analyse de la scène sonore (prise de son), un point de vue particulier[8] est choisi pour représenter la scène. La perception de l'auditeur est par la suite assujettie à ce point de vue. Or l'auditeur peut désirer changer de point de vue et se déplacer à l'intérieur de la scène sonore. Il est intéressant d'examiner quels modèles lui offrent cette possibilité, à l'issue de l'étape de captation de la scène sonore, une fois que la scène est encodée dans un format

[8]Tout le talent de l'ingénieur du son est justement de choisir le point de vue optimal au sens de sa création artistique.

donné (espace de captation). Dans les formats stéréo et multicanal 5.1 la scène est représentée par un nombre limité de canaux dans lesquels les composantes de la scène sont mélangées de façon non séparable a posteriori. Dans ces conditions il est évident qu'il est très difficile de manipuler la scène, à moins d'avoir recours à des méthodes de traitement du signal permettant d'extraire certaines composantes (sources localisées, réverbération). A l'opposé, le format Ambisonic/HOA se fonde sur une structuration spatiale de la scène avec une décomposition des ondes acoustiques en fonction des directions. Une telle structure se prête bien à des manipulations de l'espace sonore au moyen de simples rotations du repère de coordonnées. Le format Ambisonic/HOA est ainsi le format par excellence qui offre à l'auditeur, par simple manipulation des signaux d'encodage, la possibilité de faire tourner la scène sonore autour de lui et de distordre la perspective en faisant une focalisation sur une zone donnée de l'espace. A un degré moindre, le format WFS offre une relative interactivité avec la scène dans la mesure où le rendu WFS s'étend sur une large zone d'écoute et permet à l'auditeur de se déplacer au milieu des hauts-parleurs (espace acoustique secondaire) et d'expérimenter ainsi différents points d'écoute et perspectives. En revanche le cas de la synthèse binaurale dynamique qui consiste à modifier la direction de la source virtuelle en fonction des mouvements de tête de l'auditeur ne peut être considéré comme une interactivité du format binaural, car la manipulation s'effectue lors de la synthèse des signaux binauraux. L'auditeur ne peut en bénéficier que si ces signaux sont générés en temps réel. Aucune manipulation ne peut être appliquée a posteriori sur les signaux binauraux. A l'instar des formats stéréo et multicanal, le format binaural est totalement dénué d'interactivité puisqu'un signal binaural est non seulement assujetti à un point de vue, mais aussi à un individu et une perspective d'écoute.

1.4.2 Compatibilité avec les autres modèles

Un atout déterminant pour la diffusion et le développement d'un modèle audio 3D est sa compatibilité avec d'autres modèles. En effet un premier frein à l'innovation est l'inertie des équipements techniques : on ne change pas du jour au lendemain son système d'écoute. Par ailleurs des équipements complexes impliquant de nombreux hauts-parleurs et le coût de calcul associé seront toujours une exception. Il est donc essentiel pour la viabilité d'une technologie audio 3D qu'elle soit compatible avec des technologies standardisées et/ou largement répandues. L'enjeu est double : il s'agit d'abord de disposer d'un potentiel d'écoute, mais il s'agit aussi pour une technologie audio 3D de faire la preuve de ses performances supérieures de spatialisation sur un système d'écoute non optimal. Dans la section précédente, les compatibilités entre les différentes technologies ont été identifiées. Par une sorte d'héritage naturel, stéréophonie et multicanal sont fortement compatibles. La technologie binaurale offre à toutes les technologies multi hauts-parleurs une opportunité séduisante d'écoute au casque. Enfin HOA propose des passerelles intéressantes avec le multicanal.

1.4.3 Flexibilité en termes de prise et de restitution du son

La flexibilité dénote la capacité d'une technologie audio 3D à varier ses espaces de captation et de restitution. L'exemple type de la technologie flexible est sans aucun doute HOA, puisqu'un enregistrement HOA peut être restitué sur une grande variété de systèmes d'écoute, à la fois en nombre et en dispostion de hauts-parleurs. Une caractéristique spécifique de cette technologie est justement la matrice de décodage HOA qui vient adapter l'espace de captation à l'espace de restitution. La flexibilité ne joue pas seulement entre captation et restitution, elle intervient déjà au niveau de la captation dans la mesure où la décomposition en harmoniques sphériques peut être limitée à n'importe quel ordre arbitraire, ce qui donne lieu en pratique à une infinie variété de systèmes de captation. A l'opposé les technologies stéréophoniques et multicanales sont assujetties à un système de restitution standardisé. De même la variété des systèmes de prise de son associés ne peut être

assimilé à une flexibilité. La technologie binaurale possède une flexibilité limitée dans la mesure où la restitution peut s'adapter à un rendu sur un dispositif de hauts-parleurs. Dans son principe, l'holophonie est flexible puisque la géométrie du réseau de transducteurs n'est pas contrainte. En revanche la géométrie du réseau de microphones en théorie détermine celle du réseau de hauts-parleurs même si des méthodes de compensation ont été proposées [Berkhout et al., 1993].

1.4.4 Compression de la représentation

La représentation d'un scène audio 3D constitue une quantité importante de données. En dépit des capacités croissantes de stockage et de transmission, la compression de ces informations reste une nécessité, ainsi qu'une gageure. La solution la plus simple consiste à utiliser les méthodes de compression développées pour les signaux monophoniques et à les appliquer sur chaque signal pris séparément. Pour aller plus loin il faut déjà examiner en quoi chaque modèle audio 3D intègre ou non cette question de compression. Là encore HOA se démarque des autres technologies par sa propriété de *représentation hiérarchique*. Dans la décomposition en harmoniques sphériques, la composante à l'ordre 0 constitue déjà une représentation complète de la scène audio 3D et se suffit à elle-même. En fait elle correspond à un enregistrement monophonique de la scène sonore : toutes les sources sonores sont bien présentes, mais sans aucune spatialisation. Les composantes de l'ordre 1 ne vont alors que compléter l'information de la composante à l'ordre 0 pour en enrichir la spatialisation, et ainsi de suite pour les composantes des ordres supérieurs. L'information est ainsi répartie de façon hiérarchique sur les différents ordres. On saisit tout l'avantage d'une telle représentation du point de vue de la compression : en tronquant la décomposition à un ordre donné on est certain de conserver une information utile et efficace. HOA est le seul modèle audio 3D à intégrer ainsi une propriété de compression.

1.5 Défis à relever

On vient de voir que le domaine de la spatialisation sonore dispose aujourd'hui d'un large éventail de technologies : stéréophonie, multicanal, ambisonic, holophonie, VBAP, VBIP. Ces différentes technologies n'ont pas la même maturité : certaines sont relativement bien maîtrisées (stéréophonie, et dans une certaine mesure multicanal 5.1), alors que d'autres soulèvent encore un certain nombre de questions (notamment HOA, pour laquelle la mise en oeuvre d'un système de prise de son et l'optimisation du rendu sont d'actuels sujets d'étude). Encore aujourd'hui la majorité de ces technologies reste inconnue du grand public. Le monde audio professionnel reste attaché aux technologies maîtrisées et adopte une certaine frilosité à l'égard de technologies plus "innovantes". L'acceptation de ces dernières est d'ailleurs un frein réel à leur développement. Il est à noter cependant qu'il n'existe pas véritablement de rivalité entre les technologies disponibles, ni qu'il existe une technologie meilleure qu'une autre. L'ensemble des technologies offre plutôt des propriétés complémentaires, en sorte que pour un problème donné une technologie sera mieux adaptée que les autres. Ainsi, par exemple, pour doter un téléphone mobile d'un rendu audio spatialisé, la technologie binaurale est la solution la plus pertinente compte tenu des contraintes d'encombrement et de mobilité. A l'inverse la mise en œuvre d'un système holophonique ne prend véritablement son sens que dans le cas d'une diffusion pour un grand nombre d'auditeurs ou des auditeurs se déplaçant dans un large espace. Pour l'enregistrement de scènes naturelles, les systèmes de captation binaurale ou ambisonic assurent une qualité incomparable de réalisme et de naturel. Pour une audioconférence spatialisée impliquant un faible nombre de locuteurs, la stéréophonie apporte une solution robuste et efficace à moindre coût. On pourrait ainsi multiplier les exemples où les contraintes applicatives amènent à privilégier une technologie en particulier.

La pluralité des technologies de spatialisation sonore indique une certaine saturation du domaine et il est peu probable que de nouvelles technologies de spatialisation sonore soient mises à jour dans les années à venir. Les enjeux actuels concernent plutôt les questions suivantes [Nicol et al., 2008] :

- La **conversion de formats** : La multiplicité des formats audio 3D proposés aujourd'hui impose de chercher des outils permettant de convertir un flux audio 3D à un autre format donné en un autre format, afin par exemple de s'adapter au système de restitution. La compatibilité avec les autres formats audio 3D est un atout déterminant pour une technologie audio 3D.
- La **compression de données** : A la fin des années 90, le monde du codage a vu apparaître une nouvelle discipline : le *codage audio 3D* avec principalement le codage multicanal qui adresse le problème de la compression d'un flux stéréophonique ou multicanal 5.1 en élaborant des schémas de codage spécifiques aux flux audio 3D [Faller & Baumgarte, 2002] [Breebaart et al., 2005b] [Breebaart et al., 2005a] [Engdegard et al., 2008].
- L'évaluation de la **perception des modèles audio 3D** : On se place ici au niveau de l'espace perceptif de l'auditeur. L'objectif est d'évaluer comment la scène audio 3D synthétisée par les sources secondaires est perçue, analysée et interprétée par l'auditeur. Une méthode d'évaluation consiste à mener des tests de localisation. On évalue comment l'auditeur localise les sources virtuelles. Les performances de localisation, à la fois en exactitude et en précision du jugement de localisation, permettent d'évaluer la capacité d'un modèle audio 3D à spatialiser les sons. Une autre méthode d'évaluation est de demander à l'auditeur de noter sa perception sur la base de critères donnés (naturel, impression d'espace, enveloppement, préférence etc...). Le travail de recherche ici est donc double : il s'agit d'une part de mettre au point une méthododologie d'évaluation adaptée à la perception des modèles audio 3D et d'autre part d'évaluer cette perception afin de caractériser et d'optimiser les modèles.
- Les **interactions avec les autres modalités sensorielles** : Dans notre expérience quotidienne, nous percevons des stimuli multi-sensoriels où sont sollicités simultanément nos cinq sens. La réalité virtuelle vise l'intégration de toutes ces modalités dans un moteur de synthèse pour une illusion convaincante. Pour les modèles audio 3D, il s'agit alors d'étudier et de prendre en compte les interactions notamment entre l'audition et la vision, ainsi que l'audition et le toucher (perception des vibrations).
- L'**analyse de scène auditive** : Il semble qu'au niveau des analyses effectuées par le système central, l'information de spatialisation soit reléguée au second rang derrière les informations spectrales [Bregman, 1990]. Ainsi l'*illusion de Diana Deutsch* [Deutsch, 2009] suggère que la perception spatialisée des sons est influencée par leur contenu spectral. Des sons distincts spatialement peuvent être perçus à la même position sous l'influence de similitudes spectrales. Ces résultats apportent un éclairage nouveau au domaine de la spatialisation sonore. Il est probable que les progrès sur la connaissance des mécanismes de l'analyse de scène auditive vont conduire à revisiter les modèles audio 3D dans les années à venir.

L'ensemble de ces thèmes de recherche n'est pas lié à une technologie en particulier : ces questions se posent quelle que soit la technologie considérée. Il existe aussi des questions spécifiques à chaque technologie, notamment :

- la mise au point de systèmes de prise de son HOA,
- la convergence entre HOA et WFS,
- l'individualisation des filtres binauraux (synthèse binaurale).

1.6 Ma contribution personnelle

Mes travaux de recherche se focalisent sur trois technologies :

- WFS,

- HOA,
- binaural.

Les principales questions que j'ai traités sont les suivantes :

- évaluation de la perception d'un rendu WFS,
- convergence entre les technologies WFS et HOA,
- individualisation des filtres de spatialisation de la synthèse binaurale,
- évaluation de la perception de la synthèse binaurale,
- externalisation du rendu binaural,
- compression des flux audio 3D.

Ces questions témoignent de mon souci d'aborder l'ensemble des problèmes soulevés par les espaces auditifs virtuels qui ont été évoqués dans la section précédente. Dans ce document seront détaillées les deux questions suivantes :

- convergence entre les technologies WFS et HOA (Chapitre 2),
- modèle de synthèse binaurale (Chapitre 3).

Dans chaque chapitre, l'état des lieux des technologies est présenté avant de décrire ma contribution personnelle. La question de l'évaluation (objective et subjective), et par suite de la perception des modèles audio 3D est un thème transvserse qui est abordé de façon récurrente tout au long du document.

Chapitre 2

Convergence(s) entre WFS et HOA

Les technologies WFS et HOA proposent deux approches fortement similaires de la captation et la restitution d'une scène audio 3D. Ces technologies sont d'abord similaires dans leur objectif : elles visent à la *reconstruction physique* de l'onde acoustique (primaire) sous la forme d'une onde secondaire qui se propage dans un espace étendu autour de l'auditeur et est perçue par ce dernier d'une façon idéalement identique à la façon dont se serait propagée et aurait été perçue l'onde primaire. Les technologies WFS et HOA sont aussi similaires dans leurs moyens : la captation s'effectue par un réseau de microphones (réseau étendu pour WFS, réseau compact pour HOA), tandis que la restitution met en œuvre un réseau de haut-parleurs. Un autre point de convergence entre les technologies WFS et HOA porte sur leur modèle audio 3D : elles sont basées en effet sur **deux représentations équivalentes** de l'onde acoustique dans un espace 3D, *l'intégrale de Kirchhoff* pour WFS et la *décomposition sur la base des harmoniques sphériques* pour HOA [Nicol, 1999].

Les technologies WFS et HOA se ressemblent aussi dans le fait qu'elles restent des concepts généraux qui laissent beaucoup de libertés quant à leur interprétation et leur mise en œuvre. Les équations théoriques ouvrent de nombreuses possibilités pratiques, ce qui peut apporter quelques confusions. Souvent il faut bien maîtriser les aspects théoriques pour faire la part des choses. Par exemple, il n'est pas forcément évident de répondre à des questions élémentaires, telles que : quel est le nombre optimal de microphones ou de haut-parleurs, comment les disposer, quelles approximations puis-je commettre ?... La difficulté est qu'en théorie beaucoup de configurations sont possibles, mais qu'il faut en pratique un peu de savoir-faire pour ajuster les paramètres d'une configuration donnée afin d'en optimiser les performances, compte tenu des contraintes spécifiques au problème considéré. C'est là sans doute un des freins au développement de ces technologies, notamment dans le monde audio professionnel. Il s'avère néanmoins qu'il existe des configurations privilégiées qui conviennent à la plupart des cas. Un de mes premiers objectifs a été d'identifier ces configurations, d'en expliciter les paramètres et d'en préciser les limites. Il s'agit d'abord de clarifier les esprits sur les systèmes pratiques de captation et restitution auxquels correspondent les technologies WFS et HOA. Il s'agit aussi de fixer les idées en proposant une configuration "optimale" au sens d'un compromis entre efficacité, complexité et faisabilité, ce qui permet d'associer à chaque technologie un système concret. Par la suite ces configurations seront désignées sous le terme de *pragmatiques*, car ces configurations sont issues d'une démarche pragmatique cherchant à décanter les équations théoriques pour en extraire le squelette essentiel, dans la perspective de les mettre en œuvre dans des systèmes réels de captation et restitution d'une scène sonore. C'est l'objet de la première partie de ce chapitre. A partir de ces configurations, mon second objectif a consisté à rassembler les technologies WFS et HOA sous un formalisme unifié permettant une comparaison directe des deux technologies. Cette évaluation comparée est menée dans la troisième partie. Des

outils et des critères basés sur la perception auditive sont proposés. L'ensemble de ce chapitre est le fruit d'une étroite collaboration avec Jérôme Daniel et se nourrit des travaux de thèse de Sébastien Moreau et Stéphanie Bertet [Moreau, 2006] [Bertet, 2009].

2.1 Des équations théoriques à la mise en œuvre de systèmes de captation et de restitution audio 3D

Dans cette partie, chaque technologie est analysée de l'étape de captation jusqu'à la restitution, en partant des fondements théoriques pour dégager les règles essentielles de sa mise en œuvre en ayant le souci de la faisabilité pratique. Au final une configuration optimale est proposée.

2.1.1 Technologie WFS

Intégrale de Kirchhoff

La technologie WFS [Berkhout, 1988] [Vogel, 1993] [Start, 1997] [Verheijen, 1998] [de Bruijn, 2004] est un exemple particulier de mise en oeuvre de l'holophonie. Dans son principe fondamental l'holophonie se définit comme l'équivalent acoustique de l'holographie : à partir d'un enregistrement sur une surface on cherche à reproduire une onde acoustique à l'intérieur d'un volume [Jessel, 1973]. Physiquement il s'agit d'un problème aux limites. L'espace est décomposé en deux sous-espaces : un sous-espace Ω_1 contenant les sources acoustiques (sources dites primaires) et un sous-espace Ω_2 ne contenant aucune source et constituant la zone d'écoute. On montre alors que la pression acoustique $p(\vec{r},\omega)$ en tout point $\vec{r}$ à l'intérieur de Ω_2 et à la pulsation ω peut être exprimée sous la forme de l'équation intégrale suivante dite *Intégrale de Kirchhoff* [Bruneau, 1983] :

$$p(\vec{r},\omega) = \iint_{\partial\Omega_0} \left[g(\vec{r}-\vec{r}_0,\omega)\nabla_0 p_0(\vec{r}_0,\omega) - p_0(\vec{r}_0,\omega)\nabla_0 g(\vec{r}-\vec{r}_0,\omega) \right] .\vec{n} dS_0 \tag{2.1}$$

Dans l'intégrale, le vecteur $\vec{r}_0$ désigne la variable d'intégration et pointe donc sur un élément de surface de $\partial\Omega_0$. Le vecteur $\vec{n}$ représente la normale unitaire à la surface $\partial\Omega_0$ et extérieure au domaine Ω_2. La fonction $g(\vec{r}-\vec{r}_0,\omega)$ est la fonction de Green associée au problème, c'est à dire l'opérateur de propagation qui traduit la propagation acoustique entre un monopôle situé en un point $\vec{r}_0$ et un point d'observation $\vec{r}$. Ainsi l'intégrale 2.1 s'interprête de la façon suivante : aux sources acoustiques primaires est substituée une distribution de sources secondaires dont les propriétés de propagation sont décrites par les termes $g(\vec{r}-\vec{r}_0,\omega)$ et $\nabla_0 g(\vec{r}-\vec{r}_0,\omega)$ et dont les amplitudes sont définies par la pression et le gradient de pression (c'est à dire la vitesse particulaire à un facteur multiplicatif près) induits par les sources primaires sur la surface $\partial\Omega_0$. Chaque source secondaire comporte deux composantes. On retrouve l'idée contenue dans le *Principe de Huygens* selon lequel, dans le processus de propagation d'une onde, un front d'onde se comporte comme une distribution de sources secondaires qui émettent des ondelettes dont la superposition est capable de reconstruire l'onde primaire. La principale différence entre le Principe de Huygens et l'intégrale 2.1 est que la surface $\partial\Omega_0$ n'est pas nécessairement un front d'onde : si tel est le cas, les sources secondaires sont contrôlées uniquement en amplitude, mais, dans le cas général d'une surface quelconque, les sources sont pilotées en amplitude et en phase.

L'intégrale 2.1 est d'abord un modèle de représentation d'une onde acoustique, mais, si on l'examine en termes de captation et de restitution d'une scène sonore, cette intégrale propose aussi un modélisation d'une scène audio 3D, intégrant un modèle d'encodage (captation) et de décodage (restitution) de cette scène. L'encodage est réalisé par un réseau de capteurs (pression et gradient de pression) répartis en périphérie de la zone d'écoute (surface $\partial\Omega_0$). Le décodage est effectué par

un réseau identique de sources secondaires dont les propriétés sont déterminées par la fonction de Green.

L'intégrale 2.1 définit l'holophonie dans son principe **le plus générique**. On se rend compte que cette formulation laisse deux degrés de libertés sur :

- le choix de la fonction de Green : N'importe quel choix arbitraire de fonction de Green est possible. Par exemple on peut choisir une fonction de Green qui s'annule ou dont le gradient s'annule sur la surface $\partial\Omega_0$, ce qui permet d'éliminer une des composantes des sources secondaires. Cependant cette simplification peut n'être qu'illusoire : la suppression d'une des composantes est probablement compensée par un accroissement de la complexité de la composante restante. En pratique on choisit le plus souvent la fonction de Green définie en espace libre qui est donnée par :

$$g(\vec{r}-\vec{r}_0,\omega) = \frac{e^{ik|\vec{r}-\vec{r}_0|}}{4\pi|\vec{r}-\vec{r}_0|} \tag{2.2}$$

  où k désigne le nombre d'onde et i l'imaginaire pur. Ce choix conduit à une forme bien particulière des sources secondaires correspondant à l'association d'un monopôle et d'un dipôle.
- le choix de la géométrie (taille et forme) de la surface $\partial\Omega_0$.

En conséquence, lorsqu'on veut mettre en œuvre un système holophonique, une infinie variété de réalisations est possible, à la fois dans les spécificités des sources secondaires (fonction de Green) et la géométrie du réseau de transducteurs pour la captation et la restitution (surface $\partial\Omega_0$).

Le concept WFS

Le concept WFS se fonde sur l'intégrale 2.1, à partir de laquelle il propose des hypothèses ou des approximations [Nicol, 1999] au niveau de :

- La captation : Une captation holophonique *naturelle* n'est pas considérée. Le concept WFS se base sur un encodage *artificiel* où les signaux d'encodage sont synthétisés en simulant la propagation des sources primaires jusqu'au réseau de capteurs. L'absence de système de captation naturelle est d'ailleurs une des limitations de WFS.
- La nature des sources secondaires : On montre que les deux composantes des sources secondaires sont redondantes[1]. Par conséquent le concept WFS ne prend en compte qu'une des composantes, à condition d'appliquer une pondération spatiale au réseau.
- La géométrie du réseau de transducteurs : Le réseau de sources secondaires peut être vu comme une fenêtre ouverte sur la scène sonore à restituer. Par suite la taille et la position de cette fenêtre peuvent être adaptées en fonction du contenu de la scène sonore à restituer. Ainsi, si des zones de l'espace sont exemptes de sources primaires, il n'est pas nécessaire d'y disposer des sources secondaires. Par exemple, si la scène ne comporte que des sources frontales, le réseau de sources peut être limité à la zone frontale. De même, souvent les sources sont principalement situées dans le plan horizontal, ce qui permet de mettre en œuvre un réseau de sources secondaires undimensionnel au lieu d'une surface (*Approximation de la phase stationnaire* [Nicol, 1999]). Dans ce cas, on parle alors d'un **décodage 2D** (restriction du réseau de sources secondaires au plan horizontal), par opposition au **décodage 3D** où l'auditeur est complètement enveloppé par le réseau de sources secondaires, ce qui permet de contrôler les sources virtuelles en élévation.

[1] La présence des deux composantes est principalement utile pour discriminer les sources primaires internes (présentes dans Ω_2) des sources externes (présentes dans Ω_1). On retrouve un comportement similaire dans la décomposition en harmoniques sphériques [Daniel et al., 2003].

– La discrétisation du réseau de transducteurs : L'intégrale décrit une distribution continue de transducteurs. Le concept WFS lui substitue un réseau discret ce qui pose le problème de l'échantillonnage spatial et du repliement spectral associé.

L'holophonie en question

Partant de l'intégrale de Kirchhoff (Equ. 2.1) avec l'éclairage du concept WFS, quelles règles peut-on retenir pour mettre en œuvre un système holophonique ?

La première idée à retenir du concept WFS est l'utilisation de sources secondaires d'un seul type (monopôle ou dipôle). Un haut-parleur monté sur une enceinte close peut être considéré en bonne approximation (et sur une gamme limitée de fréquences) comme un monopôle. Pour cette raison, il semble plus judicieux d'opter pour des sources secondaires monopolaires. L'encodage spatial s'en trouve également simplifié : il suffit de capter le gradient de pression à l'emplacement de la source secondaire. Cependant l'élimination d'une des composantes des sources secondaires ne se fait pas sans prendre quelques précautions, car le travail de reconstruction de l'onde acoustique par les sources secondaires opère par un subtil jeu d'interférences constructives et destructrices. Si on identifie la source acoustique à une source lumineuse, on peut séparer le réseau de sources secondaires en deux parties : une partie éclairée et une partie dans l'ombre. Cette dernière doit alors être désactivée, car elle correspond à des sources qui, étant donné qu'elles sont situées à l'opposé de la source primaire par rapport à la zone d'écoute, devraient émettre avant d'avoir été excitées [Nicol, 1999]. Dans le cas où les composantes monopolaires et dipolaires des sources secondaires sont présentes, la contribution de la partie non éclairée est nulle, car monopôles et dipôles s'annulent les uns les autres. Si seuls les monopôles sont présents, il faut donc les désactiver sur cette portion du réseau de sources secondaires, ce qui revient à pondérer spatialement le réseau. A noter que cette pondération dépend de la position de la source primaire. On montre qu'au lieu d'appliquer cette pondération spatiale au décodage, on peut l'appliquer à l'encodage [Nicol, 1999], en utilisant des microphones unidirectifs, par exemple cardioïdes, afin d'éliminer la contribution des sources situées à l'opposé de la zone d'écoute. C'est cette solution que nous adopterons dans la suite.

La seconde idée qu'on retiendra du concept WFS est la restriction à un décodage 2D focalisé sur un plan horizontal. Cette restriction est acceptable dans de nombreuses situations, du fait que les sources acoustiques primaires sont majoritairement situées dans un plan horizontal. De plus, il faut aussi considérer qu'en termes de performances de localisation, le système auditif privilégie le plan horizontal avec une discrimination de quelques degrés contre quelques dizaines de degrés en élévation [Blauert, 1983]. Pour ces deux raisons, le décodage 2D est une configuration privilégiée à laquelle nous allons nous borner dans ce qui suit. La géométrie circulaire qui permet de couvrir de façon homogène toutes les directions du plan horizontal est la géométrie retenue. C'est d'ailleurs la géométrie adoptée pour les systèmes de restitution multicanal. Le choix d'un réseau undimensionnel circulaire de sources secondaires préserve donc la compatibilité avec le décodage des formats multicanal. L'utilisation d'un réseau 2D suppose cependant quelques précautions. Le décodage ne permettant de ne restituer que des sources dans le plan horizontal, il conviendrait, lors de l'encodage, de restreindre la captation aux seules sources primaires qui sont situées dans le plan horizontal. A un décodage 2D il faut donc associer un encodage 2D, correspondant à imposer une directivité verticale (en élévation) au système de captation. Si cette précaution n'est pas prise, les sources n'appartenant pas au plan horizontal sont "repliées" sur le plan horizontal et viennent polluer la restitution des sources horizontales. Ce problème n'est pas forcément critique, car on peut préférer maintenir l'ensemble des sources présentes, même si la localisation de certaines est erronée, mais il faut en avoir conscience. La seconde précaution porte sur le décodage 2D qui implique un facteur correctif lié à l'*Approximation de la Phase Stationnaire* permettant de rem-

placer une surface de sources secondaires par un réseau undimensionnel (en l'occurrence un cercle) [Berkhout et al., 1993] [Start, 1997]. Mais ce facteur correctif ne peut pas être appliqué en pratique, car il dépend de la position de la source primaire et du point d'écoute [Start, 1997]. D'un point de vue physique, le rôle de cette correction vise à compenser, d'une part, la diminution du nombre de sources secondaires et, d'autre part, le rayonnement d'un réseau linéaire de sources qui implique notamment une décroissance de l'amplitude de l'onde acoustique en $\frac{1}{\sqrt{r}}$ au lieu d'une décroissance en $\frac{1}{r}$ pour un monopôle acoustique. Compte tenu de la difficulté à mettre en œuvre cette correction, il convient de se poser la question de son utilité. Des écoutes informelles suggérant que l'absence de correction de la phase stationnaire n'entraînait pas d'artefacts notables, en termes de rendu et de spatialisation des sources virtulles, nous ont conduit à renoncer à l'appliquer. Ce choix est appliqué pour la solution retenue.

Ce qui fait défaut au concept WFS, c'est un système de captation pour l'encodage de scènes sonores naturelles. Des choix précédents, il découle que la solution la mieux adaptée est un réseau de microphones cardioïdes épousant la géométrie du réseau de haut-parleurs utilisé pour le décodage. Les microphones sont pointés vers les sources primaires. A noter que cette solution n'est fiable qu'à condition que la scène sonore ne comporte que des sources dans le plan horizontal. Sinon il faut opter pour un autre système de captation permettant d'éliminer les sources en dehors du plan horizontal, ou se résigner au repliement de sources non horizontales sur la plan horizontal.

La dernière question à se poser dans la mise en œuvre d'un système holophonique concerne la valeur de l'espacement entre les transducteurs (haut-parleurs ou microphones). On sait que le repliement spatial (*spatial aliasing* en anglais) est inévitable, tant que l'on ne disposera pas de haut-parleurs infiniment petits. On sait aussi que l'incidence du repliement spatial n'est pas critique tant qu'il se produit à des fréquences supérieures à 1.5 kHz [Start, 1997]. Sous cette condition, la localisation des sources sonores est préservée, mais le timbre des sources est susceptible d'être dégradé [Start, 1997]. En pratique un espacement de l'ordre de 10 cm est recommandé. Le repliement spatial apparaît alors à partir de la fréquence 1.7 kHz. Afin d'évaluer la tolérance sur l'espacement entre les haut-parleurs, un test de localisation[2] dont l'objectif était de comparer deux valeurs d'espacements (15 et 30 cm) a été mené à Orange Labs [Renard, 2000]. Les résultats montrent qu'un espacement de 30 cm n'altère pas la localisation des sources virtuelles.

Solution WFS retenue

L'ensemble des choix qui viennent d'être discutés est résumé par l'équation suivante qui peut être considérée comme la déclinaison de l'intégrale de Kirchhoff (Equ. 2.1) qui sera retenue pour la suite :

$$\hat{p}_{WFS}(\vec{r},\omega) = \sum_{l=1}^{N_L} c_{WFS}(l,\omega)\frac{e^{jk\vec{\rho}_l}}{4\pi\rho_l} \tag{2.3}$$

[2]Pour ce test, le système holophonique était constitué d'un réseau linéaire de 16 haut-parleurs. Les signaux alimentant les haut-parleurs étaient synthétisés par contrôle d'un gain et d'un retard en accord avec la position de la source virtuelle. Un total de 12 positions de source virtuelle a été considéré, couvrant l'ensemble de la zone située derrière le réseau de haut-parleurs de 50 cm à 6 m. La tâche du sujet consistait à localiser la source virtuelle en pointant sa tête dans la direction perçue. Un head-tracker placé sur la tête du sujet permettait de relever l'orientation de sa tête. Chaque source était localisée pour un ensemble de 6 positions d'écoute réparties sur l'ensemble de la zone d'écoute. A partir des différentes positions, il était possible de croiser les différentes directions pointées par le sujet : le point d'intersection ainsi obtenu donnait une estimation de la position perçue de la source perçue, non seulement en azimut, mais aussi en distance. Les résultats du test ont montré que le système holophonique (en comparaison d'un contrôle de la position de la source virtuelle par un gain seul, du type "panoramique d'intensité") offrait un rendu effectif de la profondeur des sources virtuelles, ce qui le démarquait de tous les autres systèmes de spatialisation.

avec :

$$\vec{\rho_l} = \vec{r} - \vec{r}_L(l),\ \rho_l = |\vec{\rho_l}|$$

Dans ces expressions, $\hat{p}_{WFS}$ désigne l'onde reconstruite selon le procédé WFS, N_L est le nombre de haut-parleurs et chaque vecteur $\vec{r}_L(l)$ repère la position du lème haut-parleur. Les N_L haut-parleurs sont équirépartis sur un cercle de rayon r_L. Le signal $c_{WFS}(l, \omega)$ correspond au signal de sortie du lème microphone (de type cardioïde) associé au lème haut-parleur. Chaque microphone est dirigé selon la direction radiale, en pointant vers l'extérieur. Si le cercle des haut-parleurs est centré sur l'origine du repère, la position de chaque haut-parleur est donnée par :

$$\vec{r}_L(l) \left| \begin{array}{l} x = r_L \cos\phi_l \\ y = r_L \sin\phi_l \\ z = 0 \end{array} \right.$$

L'espacement Δ entre les haut-parleurs est alors donné par :

$$\Delta = \frac{2\pi r_L}{N_L},$$

Le repliement spatial se produit aux fréquences supérieures à la valeur suivante qui définit la *fréquence d'aliasing spatial* :

$$f_{al} = \frac{c}{2\Delta} = \frac{cN}{4\pi r_L},$$

où c est la célérité des ondes acoustiques.

L'équation 2.3 se base sur les hypothèses suivantes :

- réseau discret de transducteurs,
- réseau circulaire horizontal,
- captation par un réseau de microphones cardioïdes,
- restitution par un réseau de monopôles selon une géométrie identique au réseau de capteurs.

2.1.2 Système HOA

Un modèle de représentation d'une scène audio 3D

Fondamentalement, la technologie HOA est bâtie autour d'un modèle de représentation d'une onde acoustique : ce modèle est le **développement de l'onde sur la base des fonctions propres** de l'équation des ondes acoustiques en coordonnées sphériques (r : rayon, ϕ : angle d'azimuth, θ : angle d'élévation). Ces fonctions propres combinent des *fonctions de Bessel sphériques* $j_m(kr)$[3] et $n_m(kr)$[4] et/ou des *fonctions de Hankel sphériques* $h_m^+(kr)$[5] et $h_m^-(kr)$[6] qui décrivent les dépendances **radiales**, et des *harmoniques sphériques* $Y_{mn}^{\sigma}(\phi, \theta)$ qui décrivent les dépendances **angulaires** de l'onde acoustique [Bruneau, 1983]. Comme pour l'holophonie, l'espace est décomposé en deux sous-espaces : un sous-espace Ω_1 contenant les sources primaires et un sous-espace Ω_2 ne contenant aucune source acoustique et constituant le domaine d'écoute. Etant donné la géométrie sphérique du problème, l'espace est structuré sur la base de sphères concentriques, ce qui ne nuit en rien à la généralité du modèle. Le domaine Ω_2 est ainsi délimité par deux sphères de rayon R_1 et R_2 situées de part et d'autre du point d'écoute $\vec{r}$ de façon à exclure toute source primaire ($R_1 < |\vec{r}| = r < R_2$). Le domaine Ω_1 correspond à l'espace restant et se compose du domaine intérieur à la sphère de

[3] Fonctions de Bessel sphériques de première espèce.
[4] Fonctions de Bessel sphériques de seconde espèce ou fonction de Neumann.
[5] Fonctions de Hankel sphériques de première espèce : onde progressive convergente.
[6] Fonctions de Hankel sphériques de seconde espèce : onde progressive divergente.

rayon R_1 et du domaine extérieur à la sphère de rayon R_2. La pression $p(\vec{r},\omega)$ en tout point $\vec{r}$ situé à l'intérieur de Ω_2 s'exprime sous la forme d'une combinaison linéaire des fonctions propres du problème, comme suit :

$$\begin{aligned} p(\vec{r},\omega) = &\sum_{m=0}^{+\infty} i^m h_m^-(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} A_{mn}^{\sigma}(\omega) Y_{mn}^{\sigma}(\phi,\theta) \\ &+ \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma}(\omega) Y_{mn}^{\sigma}(\phi,\theta) \end{aligned} \tag{2.4}$$

Les harmoniques sphériques sont donnés par :

$$Y_{mn}^{\sigma}(\phi,\theta) = \sqrt{(2m+1)\epsilon_n \frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\theta) \times \begin{cases} \cos(n\phi) \ si \ \sigma = +1 \\ \sin(n\phi) \ si \ \sigma = -1 \end{cases} \tag{2.5}$$

où le coefficient ϵ_n vaut 1 si $n = 0$ et 2 si $n > 0$. Les fonctions $P_{mn}(\sin\theta)$ sont les polynômes associés de Legendre définis par :

$$P_{mn}(\sin\theta) = \frac{d^n P_m(\sin\theta)}{d(\sin\theta)^n} \tag{2.6}$$

La fonction P_m est le polynôme de Legendre de première espèce d'ordre m. Les harmoniques sphériques Y_{mn}^{σ} définissent une base orthonormée au sens du produit scalaire appliqué sur la sphère de rayon $r = 1$:

$$\frac{1}{4\pi} \int_{\phi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} Y_{mn}^{\sigma}(\phi,\theta) Y_{m'n'}^{\sigma'}(\phi,\theta) \cos\theta \ d\theta d\phi = \delta_{mm'}\delta_{nn'}\delta_{\sigma\sigma'} \tag{2.7}$$

Les coefficients A_{mn}^{σ} et B_{mn}^{σ} du développement (Equ. 2.4) sont des signaux qui dépendent de la pulsation ω, c'est à dire de la fréquence (temporelle). Ces signaux constituent la **représentation HOA** de l'onde acoustique. Ils sont calculés en exploitant la propriété d'orthonormalité des harmoniques sphériques (Equ. 2.7). La pression est considérée connue sur une sphère de rayon R ($R_1 < R < R_2$) centrée sur l'origine du repère. Conformément à l'équation 2.4, cette pression s'exprime :

$$\begin{aligned} p(R,\phi,\theta,\omega) = &\sum_{m'=0}^{+\infty} i^{m'} h_{m'}^-(kR) \sum_{n'=0}^{m'} \sum_{\sigma'=\pm 1} A_{m'n'}^{\sigma'}(\omega) Y_{m'n'}^{\sigma'}(\phi,\theta) \\ &+ \sum_{m'=0}^{+\infty} i^{m'} j_{m'}(kR) \sum_{n=0}^{m'} \sum_{\sigma'=\pm 1} B_{m'n'}^{\sigma'}(\omega) Y_{m'n'}^{\sigma'}(\phi,\theta) \end{aligned} \tag{2.8}$$

Définissons la quantité U_{mn}^{σ} comme le résultat de la projection (produit scalaire) de cette pression sur les harmoniques sphériques, conformément à l'Equ. 2.7 :

$$U_{mn}^{\sigma}(\omega) = \frac{1}{4\pi R^2} \int_{\phi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} p(R,\phi,\theta,\omega) Y_{mn}^{\sigma}(\phi,\theta) \cos\theta \ d\theta d\phi \tag{2.9}$$

Du fait de l'orthonormalité des harmoniques sphériques (cf. Equ. 2.7), le second membre de l'équation précédente devient :

$$U_{mn}^{\sigma}(\omega) = i^m h_m^-(kR) A_{mn}^{\sigma}(\omega) + i^m j_m(kR) B_{mn}^{\sigma}(\omega) \tag{2.10}$$

Cette équation est insuffisante pour déterminer le couple de coefficients (A^{σ}_{mn}, B^{σ}_{mn}). En revanche, si, à la connaissance de la pression acoustique sur la sphère de rayon R, on rajoute la connaissance de la vitesse particulaire normale à la surface à la sphère, on obtient une seconde équation qui va nous permettre de résoudre le problème [Daniel et al., 2003]. La vitesse particulaire normale à la sphère, v_n, se déduit de la pression par la relation d'Euler :

$$\begin{aligned} v_n(R,\phi,\theta,\omega) =& \frac{1}{i\omega R}\frac{\partial p}{\partial R}(R,\phi,\theta) \\ =& \sum_{m'=0}^{+\infty} \frac{i^{m'-1}}{cR} h'^{-}_{m'}(kR) \sum_{n'=0}^{m'} \sum_{\sigma'=\pm 1} A^{\sigma'}_{m'n'}(\omega) Y^{\sigma'}_{m'n'}(\phi,\theta) \\ &+ \sum_{m'=0}^{+\infty} \frac{i^{m'-1}}{cR} j'_{m'}(kR) \sum_{n'=0}^{m'} \sum_{\sigma'=\pm 1} B^{\sigma'}_{m'n'}(\omega) Y^{\sigma'}_{m'n'}(\phi,\theta) \end{aligned} \tag{2.11}$$

où :

$$\begin{aligned} h' &\equiv \tfrac{\partial h}{\partial R} \\ j' &\equiv \tfrac{\partial j}{\partial R} \end{aligned}$$

On définit la quantité V^{σ}_{mn} comme le résultat de la projection de la vitesse particulaire v_n sur les harmoniques sphériques :

$$V^{\sigma}_{mn}(\omega) = \frac{1}{4\pi R^2} \int_{\phi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} v_n(R,\phi,\theta,\omega) Y^{\sigma}_{mn}(\phi,\theta) \cos\theta \; d\theta d\phi \tag{2.12}$$

Comme pour la pression, il vient :

$$V^{\sigma}_{mn}(\omega) = \frac{i^{m-1}}{cR} h'^{-}_{m}(kR) A^{\sigma}_{mn}(\omega) + \frac{i^{m-1}}{cR} j'_{m}(kR) B^{\sigma}_{mn}(\omega) \tag{2.13}$$

Les équations 2.10 et 2.13 permettent ainsi de calculer les coefficients A^{σ}_{mn} et B^{σ}_{mn} à partir des quantités U^{σ}_{mn} et V^{σ}_{mn}, c'est à dire de la connaissance de la **pression** et de la **vitesse particulaire** de l'onde acoustique sur la sphère de rayon R :

$$\begin{aligned} A^{\sigma}_{mn}(\omega) &= i^{-m} \frac{j'_{m}(kR) U^{\sigma}_{mn}(\omega) - icR j^{-}_{m}(kR) V^{\sigma}_{mn}(\omega)}{j'_{m}(kR) h^{-}_{m}(kR) - j_m(kR) h'^{-}_{m}(kR)} \\ B^{\sigma}_{mn}(\omega) &= i^{-m} \frac{h'_{m}(kR) U^{\sigma}_{mn}(\omega) - icR h^{-}_{m}(kR) V^{\sigma}_{mn}(\omega)}{j_m(kR) h'^{-}_{m}(kR) - j'_{m}(kR) h^{-}_{m}(kR)} \end{aligned} \tag{2.14}$$

Ce résultat présente deux points communs avec l'holophonie. On remarque d'abord que, comme dans l'intégrale de Kirchhoff (Equ. 2.1), l'onde acoustique est représentée à partir des informations de pression et de vitesse particulaire. Deuxièmement il suffit de collecter cette information sur une surface (sphère de rayon R) pour décrire parfaitement les propriétés de l'onde acoustique à l'intérieur d'un volume (domaine Ω_2), ce qui est un aspect fondamental de l'holophonie. De plus, il faut noter que le choix du rayon R de la sphère sur laquelle pression et vitesse sont enregistrées est, d'un point de vue théorique[7], indifférent. On a donc, comme pour l'holophonie, une représentation exhaustive de l'onde acoustique à partir d'une mesure sur une fraction du domaine considéré (une surface), fraction qu'on peut choisir arbitrairement.

[7]D'un point de vue pratique, en revanche, on verra dans la suite que les propriétés des fonctions de Bessel et de Hankel sphériques imposent des contraintes sur le choix de R.

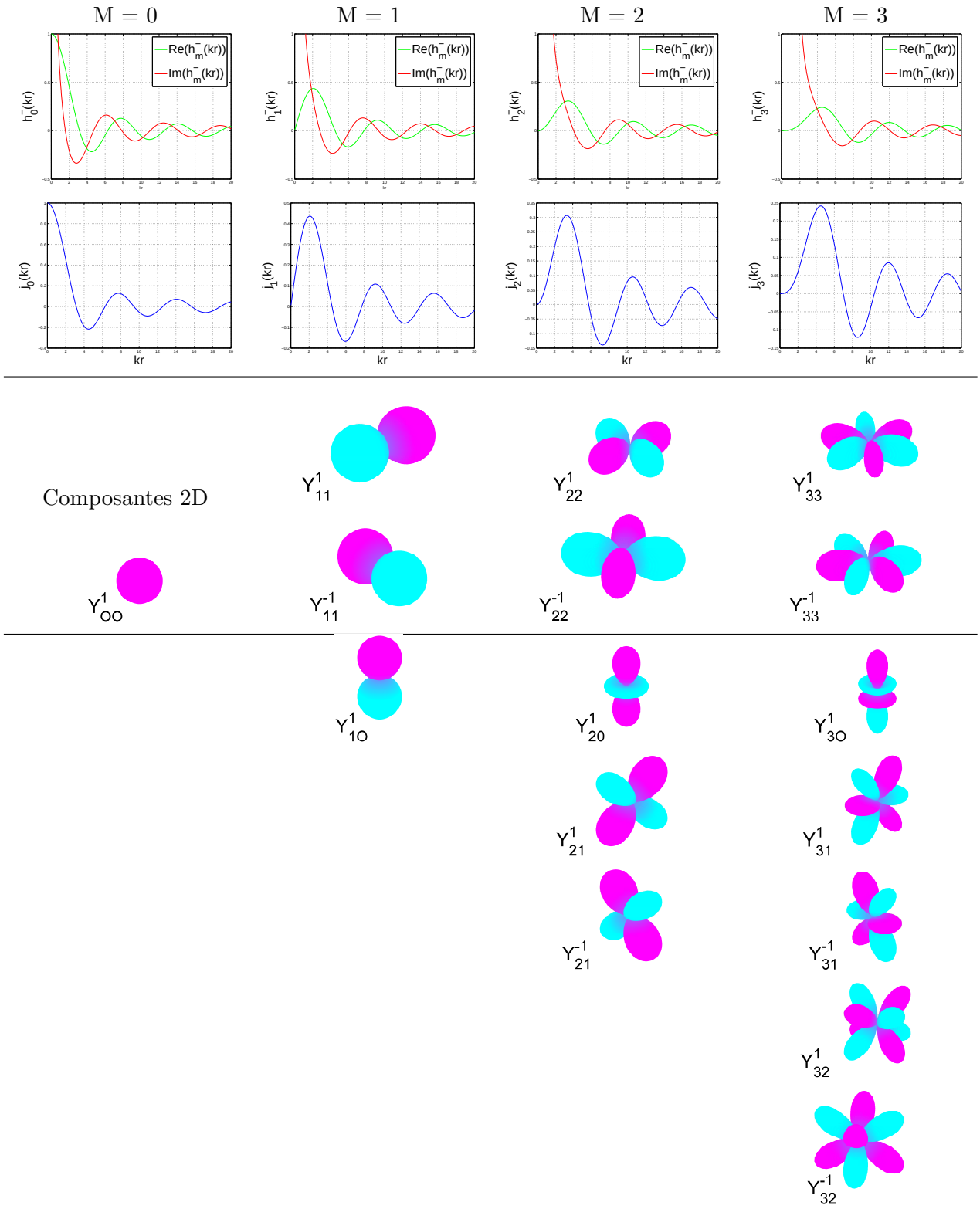


FIG. 2.1 – Combinaison des fonctions de Bessel et de Hankel sphériques avec les harmoniques sphériques pour donner les ondelettes élémentaires de la représentation HOA

Les "ondelettes" HOA

L'intégrale de Kirchhoff (Equ. 2.1) représente l'onde acoustique sous la forme d'une superposition d'*ondelettes* émises par une infinité de sources secondaires réparties en périphérie de la zone d'écoute. On retrouve le même principe dans l'équation 2.4 : l'onde acoustique primaire s'obtient comme la superposition d'une infinité d'ondelettes de la forme $j_m(kr)Y^\sigma_{mn}(\phi,\theta)$ et $h^-_m(kr)Y^\sigma_{mn}(\phi,\theta)$. Contrairement à l'holophonie, les ondelettes possèdent des propriétés spatiales différentes les unes des autres (Fig. 2.1). Ces propriétés dépendent en effet de l'ordre m de l'ondelette. Les variations spatiales selon les angles ϕ et θ tendent à devenir de plus en plus complexes et rapides au fur et à mesure où l'indice m croît. En ce qui concerne les variations radiales, on distingue deux catégories d'ondelettes [Daniel et al., 2003] :

- Les ondelettes dont la dépendance radiale obéit à une *fonction de Hankel sphérique de seconde espèce* $h^-_m(kr)$: elles correspondent à la propagation d'**ondes divergentes** et représentent la contribution d'ondes émises à partir de sources situées à l'**intérieur** de la sphère de rayon R_1.
- Les ondelettes dont la dépendance radiale est définie par une *fonction de Bessel sphérique de première espèce* $j^-_m(kr)$: elles correspondent à la contribution d'ondes émises à partir de sources situées à l'**extérieur** de la sphère de rayon R_2.

Ainsi le double jeu de coefficients (A^σ_{mn}, B^σ_{mn}) offre une représentation discriminée des sources internes ($r < R_1$) et externes ($r > R_2$). On retrouve une dualité similaire à celle des capteurs monopolaires et dipolaires présents dans l'intégrale de Kirchhoff (Equ. 2.1). Dans les deux cas, cette dualité permet de discriminer les sources externes des sources internes.

Par suite, si la sphère intérieure (de rayon R_1) ne contient aucune source, les ondelettes de la première catégorie n'ont pas lieu d'être. Le premier terme de l'équation 2.4 disparaît alors, tous les coefficients A^σ_{mn} étant nuls. Les coefficients B^σ_{mn} suffisent donc à représenter l'onde acoustique résultant des sources extérieures à la sphère de rayon R_2. Pour simplifier, on supprime la sphère de rayon R_1, en sorte que le domaine d'écoute Ω_2 s'étend à tout l'intérieur de la sphère de rayon R_2. La géométrie du problème devient ainsi très proche de celle de l'holophonie, au détail près que le domaine Ω_2 est ici une sphère et non un volume quelconque. Cette configuration convient à la plupart des cas : il semble raisonnable de considérer que le domaine d'écoute est exempt de sources primaires. Aussi adopterons-nous cette hypothèse pour la suite. L'expression de la pression (Equ. 2.4) se simplifie pour devenir :

$$p(\vec{r},\omega) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm1} B^\sigma_{mn}(\omega) Y^\sigma_{mn}(\phi,\theta) \tag{2.15}$$

Le format Ambisonic proposé par Gerzon [Gerzon, 1992a] est un cas particulier de cette représentation, dans lequel le développement de l'équation 2.15 est limité à l'ordre 1 et ne comporte donc que les 4 premières composantes B^1_{00}, B^1_{10}, B^1_{11}, B^1_{11}. Le format HOA est alors la généralisation aux ordres supérieurs du format Ambisonic. Il reste qu'en pratique la représentation HOA doit aussi être tronquée à un ordre M donné, ce qui conduit à représenter la scène audio 3D par $(M+1)^2$ composantes B^σ_{mn} (m=0, 1, ..., M ; n=0, 1, ..., m ; $\sigma = \pm 1$).

Une représentation véritablement universelle

Il faut noter que, comme l'intégrale de Kirchhoff (Equ. 2.1), la représentation HOA est parfaitement universelle et permet de décrire n'importe quelle onde acoustique sur un domaine dépourvu de sources. A titre d'exemple, le développement donne :

– pour une **onde sphérique** :

$$p_{OS}(\vec{r},\omega) = \frac{O_s}{4\pi}\left[\sum_{m=0}^{+\infty} i^m j_m(kr) i^{-(m+1)} \frac{h_m^-(kr_s)}{k} \sum_{n=0}^{m}\sum_{\sigma=\pm 1} Y_{mn}^{\sigma}(\phi_s,\theta_s) Y_{mn}^{\sigma}(\phi,\theta)\right] \tag{2.16}$$

– pour une **onde plane** :

$$p_{OP}(\vec{r},\omega) = \frac{O_p}{4\pi}\left[\sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m}\sum_{\sigma=\pm 1} Y_{mn}^{\sigma}(\phi_p,\theta_p) Y_{mn}^{\sigma}(\phi,\theta)\right] \tag{2.17}$$

Dans ces expressions, les coefficients O_s et O_p désignent respectivement les amplitudes de l'onde sphérique et de l'onde plane. Le vecteur $\vec{r}_s(r_s,\phi_s,\theta_s)$ repère la position de la source de l'onde sphérique ($|\vec{r}_s| > R_2$). La direction de provenance de l'onde plane est définie par (ϕ_p,θ_p). On retiendra donc que le format HOA est apte à représenter aussi bien une onde sphérique qu'une onde plane. La restriction souvent retenue contre Ambisonic de format limité aux ondes planes n'est plus valide pour la représentation HOA.

Les atouts de la représentation HOA

Jusqu'ici nous avons observé de nombreuses analogies entre WFS et HOA. L'intégrale de Kirchhoff (Equ. 2.1) et le développement en fonctions propres (Equ. 2.4) sont certes deux représentations alternatives d'une onde acoustique, il existe cependant des différences notables entre ces représentations. Il faut d'abord remarquer que la représentation HOA est *intrinsèquement* une représentation sur la base d'une **série discrète** d'ondelettes, alors que l'holophonie repose sur un **continuum** d'ondelettes qui doit être discrétisé en pratique, ce qui implique les artéfacts de l'échantillonnage spatial. Mais surtout la représentation HOA est une description **hiérarchique**[8], c'est à dire que les composantes des premiers ordres (m) suffisent à représenter l'onde acoustique, les composantes des ordres supérieurs ne venant que préciser l'information spatiale. Cette propriété est très utile, car elle permet de faire évoluer la représentation, a posteriori de la captation, afin de s'adapter aux contraintes notamment de débit du réseau de transmission ou du système d'écoute disponible. C'est l'avantage majeur de la représentation HOA. Un autre atout est sa **lisibilité**, dans la mesure où cette représentation peut se lire directement en termes de structure spatiale de la scène sonore, offrant une analyse séparée des informations selon 2 axes : la distance (r) et la direction (ϕ, θ). Cette séparation des dépendances radiales et directionnelles est très pertinente du point de vue de la représentation d'une scène audio 3D. Historiquement le format Ambisonic dérive d'ailleurs du format stéréophonique M-S (*Mitte-Seite*) qui repose sur l'association d'un microphone omnidirectif et d'un microphone bidirectif, visant à capter d'une part l'information ominidirectionnelle et d'autre part à séparer les informations latérales gauche et droite. La représentation HOA n'est qu'une généralisation de cette idée, dans laquelle la résolution spatiale est raffinée. Ce lien positionne la représentation HOA dans la filiation d'une démarche *intuitive* de la spatialisation sonore liée au monde des ingénieurs du son. Enfin le développement sur la base des fonctions propres n'est autre qu'une transformation du domaine des coordonnées d'espace (r, ϕ, θ) dans un domaine dual correspondant aux fréquences spatiales. Les coefficients B_{mn}^{σ} définissent ainsi le **spectre spatial** associé à l'onde acoustique, ce qui apporte une lisibilité supplémentaire à la représentation HOA en termes de fréquences (ou de variations) spatiales. Les coefficients B_{mn}^{σ} prennent ainsi le sens d'une **représentation duale** de la scène sonore, les deux étant liés par une transformation et sa

[8] On peut aussi faire le lien entre la décomposition HOA et un développement limité : plus l'ordre des composantes augmente, plus la zone de validité de la décomposition s'élargit autour du centre.

réciproque. Jusqu'à présent nous n'avons parlé que du modèle de représentation HOA sans aborder les aspects liés à la captation et la restitution de la scène sonore. Ces deux points vont être examinés maintenant.

Captation et encodage HOA

L'opération de captation correspond à l'enregistrement de la scène sonore par un dispositif de microphones qui délivrent un ensemble de signaux microphoniques *représentatifs* de cette scène. L'objectif final est d'obtenir les signaux B_{mn}^{σ} qui définissent les composantes du format de représentation HOA. L'idéal serait de capter *directement* ces signaux, c'est à dire que le dispositif de microphones fournissent d'emblée ces signaux B_{mn}^{σ}. Théoriquement ce n'est pas impossible. Ainsi, pour une onde plane, les signaux B_{mn}^{σ} sont donnés par (cf. Equ. 2.17) :

$$B_{mn}^{\sigma}(\omega) = O_p Y_{mn}^{\sigma}(\phi_p, \theta_p) \tag{2.18}$$

Dans ce cas, les signaux B_{mn}^{σ} peuvent s'interpréter comme les signaux de sortie de microphones directifs dont les fonctions de directivité imitent les directivités des harmoniques sphériques Y_{mn}^{σ}. Ces directivités sont illustrées sur la Figure 2.1. Pour les premières composantes, on se rend compte que l'harmonique Y_{00}^{1} correspond à un microphone omnidirectif (soit un capteur de pression) et que les harmoniques Y_{10}^{1}, Y_{11}^{1} et Y_{11}^{-1} correspondent à des microphones à directivité dipolaire (soit des capteurs à gradient de pression), respectivement orientés selon les axes $o\vec{z}$, $o\vec{x}$ et $o\vec{y}$. Cette équivalence entre les harmoniques sphériques et les directivités des microphones monopolaires et dipolaires est d'ailleurs exploitée par la technologie Ambisonic. Les difficultés viennent avec les composantes d'ordre supérieur ($m > 1$), pour lesquelles les figures de directivité deviennent de plus en plus complexes et ne peuvent pas être obtenues par des capteurs simples. Un autre problème est que tous les capteurs sont censés être positionnés au même point.

Il en résulte qu'en pratique, l'extraction directe des signaux B_{mn}^{σ} n'est pas mise en œuvre. Une solution privilégiée consiste à estimer ces signaux à partir de la mesure de la pression sur la surface d'une sphère de rayon r_C, conformément à l'équation 2.14 [Moreau, 2006]. On remarque que dans cette équation, la pression et la vitesse particulaire sont nécessaires pour déterminer les signaux A_{mn}^{σ} et B_{mn}^{σ}. A présent qu'on a fait l'hypothèse que la sphère de rayon R_2 ne contient aucune source acoustique, les signaux B_{mn}^{σ} suffisent à décrire la pression (cf. Equ. 2.15). Par suite, lorsque l'on projète la pression mesurée sur les harmoniques sphériques :

$$U_{mn}^{\sigma}(\omega) = \frac{1}{4\pi r_C^2} \int_{\phi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} p(r_C, \phi, \theta, \omega) Y_{mn}^{\sigma}(\phi, \theta) \cos\theta \; d\theta d\phi \tag{2.19}$$

on obtient une relation *univoque* entre les termes de la projection U_{mn}^{σ} et les signaux B_{mn}^{σ}, relation qui, a priori, ne nécessite pas la connaissance de la vitesse particulaire :

$$U_{mn}^{\sigma}(\omega) = i^m j_m(kr_C) B_{mn}^{\sigma}(\omega) \tag{2.20}$$

Par suite, les signaux B_{mn}^{σ} peuvent être obtenus selon une procédure en trois étapes :

1. mesure de la pression $p(r_C, \phi, \theta)$ sur la surface d'une sphère de rayon r_C,
2. calcul des termes U_{mn}^{σ} par projection de la pression $p(r_C, \phi, \theta)$ sur les harmoniques sphériques (Equ. 2.19),
3. les signaux B_{mn}^{σ} sont obtenus par égalisation des termes U_{mn}^{σ} (cf. Equ. 2.20) :

$$B_{mn}^{\sigma}(\omega) = EQ(kr_C) U_{mn}^{\sigma}(\omega) \tag{2.21}$$

où le terme d'égalisation est défini par :

$$EQ(kr_C) = \frac{1}{i^m j_m(kr_C)} \tag{2.22}$$

Cette procédure met en évidence une des spécificités de la technologie HOA : on se rend compte en effet qu'ici l'opération d'encodage ne se limite pas à une seule étape de captation, comme c'est le cas pour WFS. Les signaux microphoniques (signaux de pression enregistrés sur la surface de la sphère) ne correspondent pas directement au format de représentation HOA (signaux B^{σ}_{mn}). Une étape de traitement est nécessaire pour obtenir les composantes HOA. Les signaux microphoniques définissent une sorte de format intermédiaire "préalable" au format HOA. Par suite l'étape d'égalisation (Equ. 2.21) s'apparente à un **transcodage**. Il s'agit de notions fondamentales à la technologie HOA.

Cette procédure soulève cependant deux problèmes. Le premier problème est posé par les zéros de la fonction de Bessel sphérique (Fig. 2.1) qui intervient au dénominateur du terme d'égalisation (Equ. 2.22). Il en résulte qu'il existe des fréquences pour lesquelles ce terme d'égalisation n'est pas défini. De plus, lorsque la fonction $j_m(kr_C)$ prend des valeurs faibles, l'égalisation introduit une amplification extrême, non seulement difficile à réaliser, mais aussi préjudiciable à la qualité des signaux du fait qu'elle contribue à amplifier les bruits microphoniques. Une solution consiste à compléter la connaissance de la pression par la mesure de la vitesse particulaire. Par exemple, au lieu de placer des microphones de pression sur la surface de la sphère, on peut utiliser des microphones **cardioïdes** dont le signal de sortie $c_{HOA}(r_C, \phi, \theta)$ est proportionnel à une combinaison linéaire de la pression et de son gradient (c'est à dire, à un facteur multiplicatif près, la vitesse particulaire) [Jouhaneau, 1994] :

$$c_{HOA}(r_C, \phi, \theta) = p(r_C, \phi, \theta) - \frac{\vec{\nabla} p(v, \phi, \theta).\vec{n}}{ik} \tag{2.23}$$

C'est alors le signal $c_{HOA}(r_C, \phi, \theta)$ que l'on projète sur les harmoniques sphériques :

$$C^{\sigma}_{mn}(\omega) = \frac{1}{4\pi r_C^2} \int_{\phi=0}^{2\pi} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} c_{HOA}(r_C, \phi, \theta, \omega) Y^{\sigma}_{mn}(\phi, \theta) \cos\theta \; d\theta d\phi \tag{2.24}$$

Ainsi les signaux B^{σ}_{mn} se déduisent des coefficients C^{σ}_{mn} :

$$B^{\sigma}_{mn}(\omega) = EQ(kr_C) C^{\sigma}_{mn}(\omega) \tag{2.25}$$

en appliquant le terme d'égalisation suivant :

$$EQ(kr_C) = \frac{1}{i^m [j_m(kr_C) + k j'_m(kr_C)]} \tag{2.26}$$

Cette fois le dénominateur ne s'annule jamais [Moreau, 2006] : le terme d'égalisation est défini pour toutes les fréquences. Il existe d'autres solutions similaires[9], mais pour la suite nous nous limiterons à cette solution qui a l'avantage d'être simple à mettre en œuvre. On note que, contrairement à une première intuition, la pression ne suffit pas à décrire l'onde acoustique, mais qu'il faut la compléter par l'information de la vitesse particulaire pour avoir une description sans ambiguïté. Ainsi le système de captation associé à la technologie HOA dans la configuration retenue pour la suite est une **sphère de microphone cardioïdes**. On retrouve un système de captation identique

[9]Une autre solution consiste à placer les microphones sur la surface d'une sphère rigide, ce qui conduit à une nouvelle expression du terme d'égalisation qui présente aussi l'avantage d'être définie quelle que soit la fréquence [Moreau, 2006].

à celui de la technologie WFS, à la différence près que, dans le cas de WFS, les rayons de la sphère de microphones (captation) et de la sphère de haut-parleurs (restitution) sont égaux, tandis que, dans le cas de HOA, le rayon de la sphère de microphones est inférieur à celui de la sphère de haut-parleurs.

Le second problème posé par cette solution de captation est qu'idéalement la pression et le gradient de la pression devraient être mesurés continument sur toute la surface de la sphère. En pratique les signaux sont acquis en un nombre fini de points, ce qui soulève le problème de l'**échantillonnage spatial**. Considérons que le système de captation est constitué d'un réseau discret de N_C microphones. La position du qème microphone est définie en coordonnées sphériques par $\vec{r}_C(q)(r_C, \phi_C(q), \theta_C(q))$, où r_C désigne le rayon de la sphère de microphones. Le problème à résoudre est de positionner *au mieux* les points $\vec{r}_C(q)$ sur la surface de la sphère pour capter les informations de la scène sonore de façon optimale au sens des signaux B^{σ}_{mn}. Les contraintes du problème sont les suivantes :

- minimiser l'erreur d'estimation des signaux B^{σ}_{mn},
- minimiser le nombre total N_C de capteurs,
- utiliser une géométrie réaliste de réseau microphonique.

Ce problème possède une *solution exacte* dès lors que les signaux enregistrés ont un spectre spatial à bande limitée, c'est à dire que les signaux B^{σ}_{mn} sont nuls au delà d'un ordre m_{max} qui définit la borne supérieure du spectre spatial. Il s'agit simplement d'une généralisation du **théorème de Shannon** aux fonctions définies sur une sphère. Driscoll et Healy [Driscoll & Healy, 1994] ont en effet montré qu'il suffit d'échantillonner régulièrement et indépendamment les angles d'azimut et d'élévation. Les signaux B^{σ}_{mn} peuvent alors être reconstruits exactement à partir des N_C signaux microphoniques $c_q(\omega) = c_{HOA}(r_C, \phi_C(q), \theta_C(q), \omega) \equiv c_{HOA}(q, \omega)$ [Moreau, 2006]. L'inconvénient de cette solution est qu'elle est sous-optimale du point de vue du nombre total de capteurs, car elle requiert un nombre très important de microphones ($N_C = 4(M+1)^2$ pour un enregistrement jusqu'à l'ordre M) avec une répartition plus dense aux pôles que dans la zone équatoriale [Moreau, 2006].

Afin de réduire N_C, on préfère opter pour une *solution approchée* $\hat{B}^{\sigma}_{mn}$. Chaque signal microphonique $c_{HOA}(q, \omega)$ (cf. Equ. 2.23) peut être développé conformément à l'équation 2.15 :

$$c_{HOA}(q,\omega) = \sum_{m=0}^{M} i^m [j_m(kr_C) + kj'_m(kr_C)] \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B^{\sigma}_{mn}(\omega) Y^{\sigma}_{mn}(\phi_C(q), \theta_C(q)), \quad \text{pour } q = 1, ..., N_C \tag{2.27}$$

ce qui définit N_C équations à $(M+1)^2$ inconnues, les inconnues étant les signaux B^{σ}_{mn}. Le nombre d'inconnues croît donc avec l'ordre maximal M qu'on se fixe pour représenter la scène sonore. L'équation 2.27 définit un système d'équations linéaires qui peut être reformulé sous forme matricielle :

$$\mathbf{c}_{HOA} = \mathbf{Y}_C \ \mathbf{W}_C \ \mathbf{b} \tag{2.28}$$

où le vecteur $\mathbf{c}_{HOA}$ représente le vecteur des signaux microphoniques $c_{HOA}(q)$. Le vecteur $\mathbf{b}$ correspond au vecteur des signaux B^{σ}_{mn} :

$$\mathbf{b} = \begin{bmatrix} B^1_{00} \\ B^1_{10} \\ B^1_{11} \\ B^{-1}_{11} \\ \vdots \\ B^{-1}_{MM} \end{bmatrix} \tag{2.29}$$

Les matrices $\mathbf{Y}_C$ et $\mathbf{W}_C$ sont données par :

$$\mathbf{Y}_C = \begin{bmatrix} Y_{00}^1[\phi_C(1),\theta_C(1)] & Y_{10}^1[\phi_C(1),\theta_C(1)] & \dots & Y_{MM}^{-1}[\phi_C(1),\theta_C(1)] \\ Y_{00}^1[\phi_C(2),\theta_C(2)] & Y_{10}^1[\phi_C(2),\theta_C(2)] & \dots & Y_{MM}^{-1}[\phi_C(2),\theta_C(2)] \\ \vdots & \vdots & \vdots & \vdots \\ Y_{00}^1[\phi_M(N_C),\theta_M(N_C)] & Y_{10}^1[\phi_M(N_C),\theta_M(N_C)] & \dots & Y_{MM}^{-1}[\phi_M(N_C),\theta_M(N_C)] \end{bmatrix}$$

$$\mathbf{W}_C = \begin{bmatrix} [j_0(kr_C) + kj'_0(kr_C)] & 0 & 0 & \dots & 0 \\ 0 & [j_1(kr_C) + kj'_1(kr_C)] & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & i^M[j_M(kr_C) + kj'_M(kr_C)] \end{bmatrix}$$

La première condition pour résoudre ce système est que le nombre d'équations (N_C) soit supérieur ou égal au nombre d'inconnues $(M+1)^2$. En d'autres termes, le nombre requis de microphones vaut au minimum $(M+1)^2$, ce qui représente un gain d'un facteur 4 par rapport à la solution précédente (Théorème de Shannon). Le nombre minimum de microphones dépend de l'ordre M maximum fixé de la représentation HOA. En général, la solution du problème 2.28 est obtenue par une méthode de *minimisation de résidu quadratique* [Moreau, 2006]. Les signaux B_{mn}^{σ} sont estimés en fonction des signaux microphoniques grâce à la *matrice pseudo-inverse de Moore-Penrose* associée à $\mathbf{Y}_C$:

$$\hat{\mathbf{b}} = \mathbf{E}_C \ (\mathbf{Y}_C^t \ \mathbf{Y}_C)^{-1} \ \mathbf{Y}_C^t \ \mathbf{c}_{HOA} \tag{2.30}$$

où la matrice $\mathbf{E}_C$ est définie par :

$$\mathbf{E}_C = \begin{bmatrix} \frac{1}{[j_0(kr_C)+kj'_0(kr_C)]} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{i[j_1(kr_C)+kj'_1(kr_C)]} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{i^M[j_M(kr_C)+kj'_M(kr_C)]} \end{bmatrix}$$

Dans cette expression, $\mathbf{Y}_C^t$ désigne la matrice transposée conjuguée de $\mathbf{Y}_C$. Si cette solution garantit la minimisation du résidu quadratique, elle n'apporte aucune garantie sur la fiabilité de l'estimation des signaux B_{mn}^{σ}. Cette dernière dépend principalement de la sensibilité du système aux erreurs introduites, notamment au niveau du vecteur $\mathbf{c}_{HOA}$ (par exemple : bruit des capteurs, erreur de positionnement, etc...). Le terme d'égalisation présent dans la matrice $\mathbf{E}_C$ est ainsi un facteur d'instabilité du système, car il est susceptible d'amplifier ces erreurs. Afin de minimiser les risques d'instabilité, il est recommandé d'appliquer une méthode de *régularisation* qui revient à modifier la matrice $\mathbf{E}_C$ en remplaçant ses termes diagonaux par des termes de la forme [Moreau, 2006] :

$$F_m(kr_C) = \frac{|i^m[j_m(kr_C) + kj'_m(kr_C)]|^2}{|i^m[j_m(kr_C) + kj'_m(kr_C)]|^2 + \lambda^2} \tag{2.31}$$

où λ définit le paramètre de régularisation à ajuster. Le terme $F_m(kr_C)$ s'interprête comme un filtre de régularisation.

Dans l'expression de la solution de minimisation du résidu quadratique (Equ. 2.30), on note la présence du terme :

$$\mathbf{Y}_C^t \ \mathbf{Y}_C.$$

Si le choix de la distribution des capteurs sur la sphère conserve la **propriété d'orthonormalité** des harmoniques sphériques (Equ. 2.7), ce terme se simplifie :

$$\mathbf{Y}_C^t \ \mathbf{Y}_C = \mathbf{1} \tag{2.32}$$

où **1** est la matrice identité. L'expression des signaux $\hat{B}^{\sigma}_{mn}$ devient :

$$\hat{B}^{\sigma}_{mn}(\omega) = \frac{1}{j_m(kr_C) + kj'_m(kr_C)} \frac{1}{N_C} \sum_{q=1}^{N_C} c_{HOA}(q,\omega) Y^{\sigma}_{mn}[\phi_C(q), \theta_C(q)] \tag{2.33}$$

Ce résultat rappelle l'équation 2.25, en tenant compte des équations 2.24 et 2.26. L'équation 2.33 n'est autre que la transposition discrète de l'équation 2.25. La projection des signaux microphoniques sur les harmoniques sphériques (Equ. 2.24) est réalisée par la sommation discrète des produits :

$$c_{HOA}(q,\omega) Y^{\sigma}_{mn}[\phi_C(q), \theta_C(q))$$

évalués au niveau de chaque microphone. Si les positions des microphones sont choisies sans précaution, la propriété d'orthonormalité des harmoniques sphériques (Equ. 2.32) n'est en général pas conservée par l'échantillonnage spatial. L'expression des signaux $\hat{B}^{\sigma}_{mn}$ comporte alors le terme $\mathbf{Y}^t_C \, \mathbf{Y}_C$ qui traduit le repliement des harmoniques sphériques les uns sur les autres dans l'estimation des signaux B^{σ}_{mn}. La matrice ϵ mesure l'erreur de "non-orthonormalité" de l'échantillonnage :

$$\epsilon = \mathbf{1} - \frac{1}{N_C} \mathbf{Y}^t_C \, \mathbf{Y}_C \tag{2.34}$$

On se rend compte que le problème de l'échantillonnage spatial concerne non seulement l'évaluation des signaux microphoniques $c_{HOA}(r_C, \phi, \theta)$, mais aussi les harmoniques sphériques $Y^{\sigma}_{mn}(\phi, \theta)$. Le choix de la géométrie du réseau microphonique est donc dicté par une *seconde contrainte* liée à l'échantillonnage spatial "implicite" des harmoniques sphériques, c'est à dire le respect de la propriété d'orthonormalité (Equ. 2.32). En pratique cette condition d'orthonormalité est très difficile à satisfaire. Les sommets de certains polyèdres réguliers proposent des géométries préservant l'orthonormalité des harmoniques sphériques jusqu'à un ordre limité (par exemple le tétraèdre jusqu'à l'ordre 1 et l'isocaèdre jusqu'à l'ordre 2). Les polyèdres semi-réguliers offrent des solutions satisfaisantes pour les ordres supérieurs à 2 [Moreau, 2006]. Pour un ordre maximal M d'encodage donné, le choix des positions des microphones sur la sphère est optimisé en observant la matrice ϵ et en recherchant la géométrie qui la minimise pour les ordres $m \leq M$.

Le dispositif de captation HOA est caractérisé par trois principaux paramètres :

- le **nombre** de microphones N_C,
- la **position** de chaque microphone $[\phi_C(q), \theta_C(q)]$,
- le **rayon** de la sphère r_C.

Dans la réalisation d'un système de captation HOA, la première question porte sur le choix de l'ordre maximum M de la représentation que l'on souhaite. La valeur de M impose alors le nombre minimum de capteurs :

$$N_C \geq (1 + M)^2.$$

L'étape suivante consiste à chercher le polyèdre régulier ou semi-régulier proposant au moins $(M + 1)^2$ sommets et minimisant l'erreur de non-orthonormalité (Equ. 2.34) pour les ordres $m \leq M$. Le polyèdre choisi détermine le nombre de microphones et leur position. La dernière question porte sur le choix du rayon r_C de la sphère microphonique. La valeur du rayon joue sur la valeur des fonctions de Bessel $j_m(kr_C)$ et $j'_m(kr_C)$ qui viennent pondérer les composantes de la représentation HOA (Equ. 2.27). On observe que, pour un rayon donné r_C, les valeurs des fonctions $j_m(kr_C)$ et $j'_m(kr_C)$ sont très faibles au delà d'un ordre M_{max} [Moreau, 2006]. On peut ainsi considérer que les composantes d'ordre $m \geq M_{max}$ sont tellement atténuées qu'elles sont quasi-inexistantes. Les pondérations introduites par les fonctions de Bessel réalisent une sorte de **filtre fréquenciel spatial de type Passe-Bas** qui vient limiter les composantes HOA aux ordres inférieurs à M_{max}.

L'ordre M_{max} dépend du rayon r_C : lorsqu'on augmente r_C, on accroît la bande fréquencielle spatiale des signaux captés. Le choix du rayon r_C contrôle ainsi le spectre spatial et joue ainsi le rôle d'un filtre anti-repliement du point de vue de l'échantillonnage spatial. Plus exactement, M_{max} dépend du produit kr_C [Moreau, 2006] :

$$M_{max} = 2kr_C \ .$$

Une fois le rayon r_C fixé, on connaît alors la fréquence limite en dessous de laquelle les ordres supérieurs à M_{max} peuvent être négligés. On cherchera donc à réduire le rayon de captation afin de minimiser les effets de l'aliasing spatial[10], en relation avec l'ordre maximum M de la représentation que l'on s'est fixé. Cependant les fonctions de Bessel $j_m(kr_C)$ et ${j'}_m(kr_C)$ interviennent aussi dans les termes d'égalisation (Equ. 2.26) utilisés pour obtenir les composantes HOA (B^{σ}_{mn}) à partir des signaux microphoniques C^{σ}_{mn}. Or, on remarque que plus le rayon r_C est faible, plus les termes d'égalisation varient fortement [Moreau, 2006]), ce qui conduit à un mauvais conditionnement du problème (Equ. 2.30). Ce phénomène est surtout critique pour les basses fréquences [Moreau, 2006]. En diminuant le rayon de captation, on compromet donc la précision de l'estimation des composantes HOA aux basses fréquences. Par suite, le choix du rayon r_C fait l'objet d'un difficile compromis entre deux aspects :

- la minimisation de l'aliasing spatial,
- la fiabilité de l'estimation des composantes aux basses fréquences.

Le choix est d'autant plus délicat qu'il est impossible de satisfaire tous les critères sur une plage étendue de fréquences. Pour fixer les idées, considérons le cas d'un système de captation HOA visant une représentation jusqu'à l'ordre M = 4 [Moreau, 2006]. Il faut au minimum $N_C = (M+1)^2 = 25$ microphones. Le polyèdre minimisant l'erreur de non-orthonormalité est le pentaki-dodécaèdre qui est constitué de 32 sommets. On fixe le rayon de la sphère microphonique à $r_C = 3.5$ cm. Pour cette valeur de rayon, le nombre de conditionnement du problème reste inférieur à 3.2 pour les fréquences supérieures à 100 Hz. La géométrie du réseau comporte deux valeurs d'écart angulaire entre les capteurs : 0.67 et 0.75 radians. Les fréquences d'apparition du repliement associées sont 7500 et 6700 Hz. Les composantes des ordres supérieurs à M = 4 sont absentes pour les fréquences inférieures à 3 kHz selon la règle empirique $M_{max} = 2kr_C$ [Moreau, 2006]. Ces données illustrent un compromis satisfaisant pour l'ensemble des contraintes du problème de conception d'un système de captation HOA à l'ordre 4. Pour réduire davantage l'aliasing spatial, il est possible de diminuer le rayon r_C de la sphère, ce qui aura pour effet à la fois de réduire les distances entre les capteurs et de réduire la contribution des composantes d'ordre supérieur à M=4 au delà de 3 kHz.

Bien que la technologie HOA soit, en théorie, exempte du problème de l'échantillonnage spatial, on vient de voir qu'elle y est confrontée dès l'étape de captation. C'est un nouveau point de convergence entre les technologies WFS et HOA. Cependant les conséquences de l'échantillonnage spatial présentent des aspects qui sont spécifiques à la technologie HOA : les effets se répercutent non seulement sur l'onde acoustique, mais aussi indirectement sur les harmoniques sphériques (propriété d'orthonormalité). On retiendra qu'un système de captation HOA ne permet pas d'extraire exactement les signaux B^{σ}_{mn}, mais uniquement de les estimer à travers les signaux $\hat{B}^{\sigma}_{mn}$. La mise au point du réseau microphonique (sphère de capteurs cardioïdes) résulte d'un compromis entre de nombreuses contraintes de façon à maximiser la fiabilité de cette estimation.

[10] Il faut notamment limiter la présence des composantes des ordres supérieurs à M qui sont susceptibles de se replier sur des composantes d'ordre inférieur dans le produit $\mathbf{Y}_C^t \ \mathbf{Y}_C$.

Restitution et décodage HOA

L'étape de décodage vise à reconstruire l'onde acoustique originale par un dispositif de haut-parleurs à partir des signaux B_{mn}^{σ}. Considérons un ensemble de N_L haut-parleurs. La position du lème haut-parleur est repérée par le vecteur $\vec{r}_L(l)(r_L(l), \phi_L(l), \theta_L(l))$. Soit $s_{HOA}(l, \omega)$ le signal alimentant le lème haut-parleur, l'onde acoustique *secondaire* $\hat{p}$, c'est à dire synthétisée par le dispositif de haut-parleurs, s'exprime :

$$\hat{p}(\vec{r}, \omega) = \sum_{l=1}^{N_L} s_{HOA}(l, \omega) p_l(\vec{r}, \omega) \tag{2.35}$$

où le terme p_l représente la pression induite par le lème haut-parleur au point $\vec{r}$. La contribution de chaque haut-parleur peut être développée sur la base des fonctions propres de l'équation des ondes en coordonnées sphériques (Equ. 2.15) :

$$p_l(\vec{r}, \omega) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} L_{mn}^{\sigma}(l, \omega) Y_{mn}^{\sigma}(\phi, \theta) \tag{2.36}$$

Les signaux $L_{mn}^{\sigma}(l, \omega)$ sont les composantes HOA décrivant l'onde acoustique rayonnée par le lème haut-parleur. A ce niveau, aucune hypothèse n'est formulée sur la nature de cette onde : il peut s'agir aussi bien d'une onde plane, que d'une onde sphérique ou d'une forme d'onde plus complexe [Solvang, 2009]. De plus chaque source secondaire (ou haut-parleur) peut avoir une forme d'onde spécifique et différente des autres sources secondaires. On note également que les sources secondaires ne sont a priori pas réparties sur la surface d'une sphère, mais occupent des positions quelconques $\vec{r}_l$ situées à des distances $r_L(l)$ de l'origine du repère potentiellement différentes pour chaque source. Ainsi la configuration des sources secondaires de la technologie HOA n'a rien à envier en flexibilité à la technologie WFS. La géométrie du réseau est libre comme pour la WFS. En revanche, les caractéristiques de rayonnement des sources secondaires ne sont pas contraintes, ce qui n'est pas le cas de la WFS, et constitue donc un avantage de HOA sur WFS.

L'onde primaire que l'on cherche à reconstruire est décrite par les signaux B_{mn}^{σ} et se développe sous la forme de l'équation 2.15. Pour déterminer les signaux $s_{HOA}(l)$ à appliquer aux sources secondaires pour synthétiser l'onde primaire, il suffit d'identifier terme à terme les développements de l'onde primaire cible (Equ. 2.15) et de l'onde générée par les sources secondaires (Equ. 2.35 & 2.36) :

$$B_{mn}^{\sigma} = \sum_{l=1}^{N_L} s_{HOA}(l, \omega) L_{mn}^{\sigma}(l, \omega) \tag{2.37}$$

Les représentations HOA étant tronquées à l'ordre m=M, il en résulte $(M+1)^2$ composantes HOA (B_{mn}^{σ} et $L_{mn}^{\sigma}(l)$). L'équation 2.37 se décline pour ces $(M+1)^2$ composantes. Nous obtenons donc un système de $(M+1)^2$ équations linéaires à N_L inconnues. Les inconnues sont les signaux $s_{HOA}(l)$ alimentant les sources secondaires. Le système d'équations peut être reformulé sous une forme matricielle :

$$\mathbf{b} = \mathbf{L}\mathbf{s}_{HOA} \tag{2.38}$$

Dans cette expression, le vecteur $\mathbf{s}_{HOA}$ définit le vecteur des signaux alimentant les haut-parleurs $s_{HOA}(l)$ et la matrice $\mathbf{L}$ décrit les signaux $L_{mn}^{\sigma}(l)$ pour chaque source secondaire :

$$\mathbf{L} = \begin{bmatrix} L_{00}^{1}[\phi_L(1), \theta_L(1)] & L_{00}^{1}[\phi_L(2), \theta_L(2)] & \dots & L_{00}^{1}[\phi_L(N_L), \theta_L(N_L)] \\ L_{10}^{1}[\phi_L(1), \theta_L(1)] & L_{10}^{1}[\phi_L(2), \theta_L(2)] & \dots & L_{10}^{1}[\phi_L(N_L), \theta_L(N_L)] \\ \vdots & \vdots & \vdots & \vdots \\ L_{MM}^{-1}[\phi_L(1), \theta_L(1)] & L_{MM}^{-1}[\phi_L(2), \theta_L(2)] & \dots & L_{MM}^{-1}[\phi_L(N_L), \theta_L(N_L)] \end{bmatrix}$$

Comme précédemment, le vecteur $\mathbf{b}$ représente les signaux B_{mn}^{σ} (Equ. 2.29) décrivant l'onde primaire. L'équation 2.38 rappelle celle obtenue lors de l'étape de captation (Equ. 2.28). On observe ici de nombreuses similarités avec le problème de captation. La résolution du problème dépend du nombre N_L de sources secondaires (nombre d'inconnues) en relation avec l'ordre M des représentations HOA qui détermine le nombre d'équations $(M+1)^2$. Trois cas sont distingués [Poletti, 2005] :

- $N_L < (M+1)^2$: Le système est surdéterminé : il ne possède en général aucune solution exacte. On recherche alors une solution approchée par minimisation du résidu quadratique.
- $N_L = (M+1)^2$: La matrice $\mathbf{L}$ est carrée. Si elle est inversible, la solution est donnée par :

$$\mathbf{s}_{HOA} = \mathbf{L}^{-1}\mathbf{b} \tag{2.39}$$

- $N_L > (M+1)^2$: Le système est sous-déterminé : il possède une infinité de solutions. La solution minimisant l'énergie des signaux des haut-parleurs est obtenue en utilisant la matrice pseudo-inverse associée à $\mathbf{L}$:

$$\mathbf{s} = \mathbf{L}^t(\mathbf{L}\mathbf{L}^t)^{-1}\mathbf{b} \tag{2.40}$$

Il faut ici se poser la question du *nombre optimal N_L de haut-parleurs*. Le cas où $N_L = (M+1)^2$ peut être considéré comme optimal au sens où il minimise l'erreur de reconstruction [Poletti, 2005]. En pratique, le dispositif et le nombre de haut-parleurs sont souvent imposés. Afin de se ramener à la configuration optimale ($N_L = (M+1)^2$), il est alors judicieux :

- si $N_L < (M+1)^2$: d'abaisser l'ordre M de la représentation HOA en n'exploitant pas les composantes d'ordre supérieur jusqu'à ce que $(M+1)^2$ soit le plus proche de N_L,
- si $N_L > (M+1)^2$: de désactiver un partie des haut-parleurs jusqu'à ce que N_L soit le plus proche de $(M+1)^2$.

Cependant Daniel [Daniel, 2000] a remis en cause ce choix optimal, pour la raison que, dans le cas où $N_L = (M+1)^2$, lorsque la direction de la source primaire coïncide avec celle d'un haut-parleur, seul ce haut-parleur est activé, ce qui peut provoquer des artéfacts auditifs en termes de "discrétion" des haut-parleurs et d'homogénéité de la scène sonore quand les sources virtuelles se déplacent. Daniel préconise donc un nombre de haut-parleurs supérieur $N_L > (M+1)^2$. Ces recommandations se fondent uniquement sur l'analyse des équations du problème. En complément, il conviendrait d'examiner cette question à la lumière de tests d'écoute. Des premières évaluations subjectives suggèrent qu'un nombre trop important de haut-parleurs ($N_L > (M+1)^2$) est préjudiciable à la qualité du rendu HOA, car il est facteur d'instabilités, notamment lorsque l'auditeur bouge la tête [Bertet, 2009]. Par suite, nous considérerons que le nombre optimal de haut-parleurs est : $N_L = (M+1)^2$.

Nous venons de présenter le problème du décodage HOA dans sa forme la plus générale. De même que l'encodage ne se réduit pas à une seule opération de captation, le décodage HOA ne se réduit pas à l'opération de restitution acoustique par les sources secondaires, comme c'est le cas pour la technologie WFS. Au préalable, les signaux HOA sont transformés via la matrice de **matrice de décodage D** qui est définie par :

$$\mathbf{D} = \mathbf{L}^{-1} \text{ ou } \mathbf{L}^t(\mathbf{L}\mathbf{L}^t)^{-1} \tag{2.41}$$

pour donner les signaux $s_{HOA}(l)$ destinés à alimenter les haut-parleurs. Cette matrice de décodage opère un **transcodage** (ou **réencodage** [Daniel et al., 2003]) des signaux B_{mn}^{σ} pour les adapter à l'espace des haut-parleurs. Le format B_{mn}^{σ} est donc bien un **format intermédiaire complètement indépendant des formats de captation** ($\mathbf{c}_{HOA}$) et de **restitution** ($\mathbf{s}_{HOA}$). C'est une des principales spécificités de la technologie HOA, ainsi que son atout majeur. Le décodage HOA se compose ainsi de deux étapes : un transcodage (matrice de décodage) suivi de la restitution par les sources secondaires (étape de reconstruction acoustique par les haut-parleurs). La matrice de

décodage est déterminée par la configuration (nombre et géométrie) du dispositif de sources secondaires. En théorie, elle est capable de rendre compte de n'importe quelle configuration. Cependant le "bon sens" recommande d'utiliser plutôt des dispositifs "réguliers" : distribution sur la surface d'une sphère ($r_L(l) = r_L \ \forall l$), répartition homogène sur la sphère... La matrice de décodage en sera d'autant plus simple et stable, au sens du problème mathématique. La matrice de décodage donnée par l'équation 2.41 obéit à une stratégie particulière de décodage correspondant au **décodage basique** qui vise la reconstruction physique à l'identique de l'onde sonore, et est en cela conforme à la philosophie de la technologie WFS. Il existe d'autres stratégies [Daniel, 2000] dans lesquelles le problème du décodage est résolu en prenant en compte des contraintes supplémentaires. On peut chercher par exemple la solution de décodage qui maximise la contribution des sources secondaires proches de la position de la source virtuelle ou qui minimise la contribution des sources secondaires opposées à cette position. Ces stratégies s'appliquent dans le cas où $N_L > (M+1)^2$.

Il est temps de fixer le dispositif de haut-parleurs. Dans ce qui précède, on a vu que ce dispositif est en théorie très peu contraint. Il s'agit ici de considérer une configuration simple à mettre en œuvre, tout en garantissant une qualité optimale. Le dispositif suivant est choisi :

- La géométrie est définie par un **réseau sphérique** de rayon r_L.
- Les sources secondaires sont distribuées selon une configuration **régulière** sur la sphère de rayon r_L. Par analogie avec le réseau de microphones pour la captation, la spécification *régulière* signifie que le positionnement des haut-parleurs respecte la propriété d'orthonormalité des harmoniques sphériques (Equ. 2.7 ou 2.32), c'est à dire que la matrice $\mathbf{L}$ vérifie :

$$\mathbf{L}^t\mathbf{L} = \mathbf{1} \tag{2.42}$$

  Il en résulte que la matrice de décodage prend une forme particulièrement simple : elle est égale à la transposée conjuguée[11] de la matrice $\mathbf{L}$:

$$\mathbf{D} = \mathbf{L}^t \tag{2.43}$$

  Comme on l'a vu pour le réseau microphonique de captation, une configuration régulière de haut-parleurs est définie par les sommets des polyèdres réguliers ou semi-réguliers. Dans le cas d'un décodage 2D limité au plan horizontal, la contrainte de régularité se borne à distribuer les haut-parleurs de façon équirépartie sur le cercle de rayon r_L tous les $\frac{2\pi}{N_L}$ radians.
- Les sources secondaires émettent des **ondes sphériques** (hypothèse raisonnable pour des haut-parleurs). Dans ces conditions,la matrice $\mathbf{L}$ s'écrit :

$$\mathbf{L} = \mathbf{W}_L\mathbf{Y}_L \tag{2.44}$$

  avec :

$$\mathbf{Y}_L = \begin{bmatrix} Y_{00}^1[\phi_L(1),\theta_L(1)] & Y_{00}^1[\phi_L(2),\theta_L(2)] & \dots & Y_{00}^1[\phi_L(N_L),\theta_L(N_L)] \\ Y_{10}^1[\phi_L(1),\theta_L(1)] & Y_{10}^1[\phi_L(2),\theta_L(2)] & \dots & Y_{10}^1[\phi_L(N_L),\theta_L(N_L)] \\ \vdots & \vdots & \vdots & \vdots \\ Y_{MM}^{-1}[\phi_L(1),\theta_L(1)] & Y_{MM}^{-1}[\phi_L(2),\theta_L(2)] & \dots & Y_{MM}^{-1}[\phi_L(N_L),\theta_L(N_L)] \end{bmatrix}$$

[11] Poletti a montré que cette solution a une portée plus générale [Poletti, 2005]. Une solution alternative à l'intégrale de Kirchhoff (Equ. 2.1) consiste à représenter la pression $p(\vec{r},\omega)$ sous la forme d'un potentiel simple couche, c'est à dire sous la forme d'une distribution continue de sources monopolaires dont l'amplitude est définie par le saut de vitesse (ou gradient de pression) au niveau de la frontière $\partial\Omega_0$. En utilisant la représentation HOA de la pression acoustique et une fois discrétisée la distribution de sources secondaires, on retrouve la solution de la matrice transposée conjuguée [Poletti, 2005]. Ce résultat signifie qu'en recherchant une solution de type WFS (c'est à dire une reconstruction de la scène sonore par une distribution continue de sources secondaires), mais en imposant une distribution de monopôles, on retombe naturellement sur une solution HOA, ce qui est une preuve supplémentaire du lien très fort entre les deux technologies et de leur équivalence.

$$\mathbf{W}_L = \begin{bmatrix} (-i) & 0 & 0 & \dots & 0 \\ 0 & -\frac{h_1^-(kr_L)}{k} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{h_M^-(kr_L)}{k} i^{-(M+1)} \end{bmatrix}$$

On se rend compte que le dispositif ainsi défini est identique à un système de rendu WFS. Le choix du nombre de haut-parleurs N_L est déterminé par l'ordre M considéré de la représentation HOA. Dans l'idéal, N_L doit être égal à $(M+1)^2$, mais il est quasiment impossible de trouver une géométrie vérifiant la condition de *régularité* (Equ. 2.42) et constituée exactement de $(M+1)^2$. En pratique, on cherchera donc un polyèdre régulier ou semi-régulier comportant au moins $(M+1)^2$ sommets et vérifiant la propriété d'orthonormalité jusqu'à l'ordre M. On choisira alors la géométrie dont le nombre de sommets est le plus proche de $(M+1)^2$. Dans le cas d'un décodage 2D, le problème est plus simple : quel que soit N_L, il est toujours possible de positionner de façon équirépartie N_L haut-parleurs sur un cercle. De plus, dans la restriction au plan horizontal, la représentation HOA ne comporte plus $(M+1)^2$, mais seulement $(2M+1)$ composantes. Le nombre optimal de haut-parleurs devient donc $N_L = 2M+1$ haut-parleurs.

Dans cette configuration, la qualité du décodage est principalement déterminée par l'ordre M d'encodage et la qualité d'estimation des composantes $\hat{B}^{\sigma}_{mn}(\omega)$. Il est possible d'appliquer une fenêtre de pondération aux composantes HOA afin de réduire les effets de la troncature de la décomposition [Poletti, 2005]. Bien qu'à première vue, l'étape de décodage ne semble pas concernée par les artefacts de l'échantillonnage spatial introduit par l'utilisation d'un réseau discret de sources secondaires pour la raison que le spectre spatial du signal s_{HOA} alimentant les haut-parleurs présente un support limité à l'ordre M, des phénomènes de repliement spatial existent à cause des directivités des haut-parleurs dont les spectres spatiaux ne sont malheureusement pas à support limité . Cependant les effets du repliement spatial diffèrent entre HOA et WFS : alors que pour HOA, la dégradation apparaît progressivement en préservant la zone d'écoute centrale (*sweet spot*), pour WFS le repliement survient brutalement et affecte de façon homogène la zone d'écoute [Spors, 2009].

Solution HOA retenue

Comme pour WFS, on se rend compte que le concept HOA recouvre une infinité de déclinaisons possibles. Dans ce qui précède, une solution préférée a été proposée et discutée pour chaque étape du traitement. La concaténation de ces choix conduit à l'équation globale suivante qui définit la solution HOA retenue dans le cadre d'un décodage 2D en accord avec la solution WFS proposée précédemment :

$$\hat{p}_{HOA}(\vec{r},\omega) = \sum_{l=1}^{N_L} s_{HOA}(l,\omega) \frac{e^{jk\vec{\rho}_l}}{4\pi\rho_l} \tag{2.45}$$

Les signaux $s_{HOA,l}$ alimentant chaque haut-parleur sont donnés par le vecteur $\mathbf{s}_{HOA}$:

$$\mathbf{s}_{HOA} = (\mathbf{W}_L\mathbf{Y}_L)^t \mathbf{W}_C \mathbf{Y}_C^t \mathbf{c}_{HOA} \ .$$

Dans le cas d'un décodage 2D, la représentation HOA est limitée aux seules composantes horizontales (cf. Fig. 2.1), soit (2M+1) composantes pour une représentation à l'ordre M. Il convient donc d'éliminer les termes associées aux composantes verticales dans les matrices $\mathbf{Y}_C$, $\mathbf{W}_C$ $\mathbf{Y}_L$ et $\mathbf{W}_L$.

Cette équation est à rapprocher de l'équation 2.3 obtenue pour WFS. Dans les deux cas, les vecteurs $\mathbf{c}_{WFS}$ et $\mathbf{c}_{HOA}$ représentent les signaux microphoniques captés par des microphones cardioïdes distribués sur la surface d'une sphère :

– pour WFS, les microphones sont situés sur la sphère de rayon $r_C = r_L$ aux emplacements des haut-parleurs,
– pour HOA, les microphones sont disposés sur une sphère de rayon $r_C < r_L$.

Malgré une forte similitude entre les équations 2.3 et 2.45, la solution HOA se distingue par le traitement appliqué aux signaux microphoniques faisant intervenir les matrices $\mathbf{Y}_C$, $\mathbf{W}_C$, $\mathbf{Y}_L$ et $\mathbf{W}_L$ qui sont définies par les propriétés des réseaux de captation et de restitution. La technologie WFS séduit par sa simplicité. Il reste à déterminer les bénéfices des traitements HOA et en quoi ils permettent d'améliorer la qualité du rendu de la scène sonore.

2.2 Formalisme unifié des équations d'encodage et de décodage audio 3D

Les équations décrivant l'onde secondaire reconstruite selon les concepts WFS (Equ. 2.3) et HOA (Equ. 2.45) peuvent se mettre sous la forme générique suivante :

$$\hat{p}(\vec{r},\omega) = \sum_{l=1}^{N_L} s(l,\omega)\frac{e^{jk\vec{\rho}_l}}{4\pi\rho_l} \tag{2.46}$$

avec :

– dans le cas WFS :

$$\mathbf{s}_{WFS} = \mathbf{c}_{WFS}$$

– dans le cas HOA :

$$\mathbf{s}_{HOA} = (\mathbf{W}_L\mathbf{Y}_L)^t\mathbf{W}_C\mathbf{Y}_C^t\mathbf{c}_{HOA}$$

Ce formalisme conduit à comparer les signaux s_l alimentant les haut-parleurs pour chaque technologie. Il est judicieux de mener cette comparaison dans le domaine spectral spatial, c'est à dire dans le domaine des harmoniques sphériques qui peuvent servir de base commune de représentation des signaux. Dans le cas WFS, chaque signal s_l correspond au signal capté par un microphone cardioïde placé à l'emplacement du lième haut-parleur au moment de la captation. Si l'on exprime ce signal sur la base des harmoniques sphériques (représentation HOA), il s'écrit :

$$\begin{array}{rcl} s_{WFS}(l,\omega) & = & \frac{1}{2}\left[p(r_L,\phi_L(l),\theta_L(l),\omega) + \frac{\partial p}{\partial r}(r_L,\phi_L(l),\theta_L(l),\omega)\right] \\ & = & \sum_{m=0}^{+\infty}\frac{i^m}{2}\left[j_m(kr_L) + kj'_m(kr_L)\right]\sum_{n=0}^{m}\sum_{\sigma=\pm1}B_{mn}^{\sigma}(\omega)Y_{mn}^{\sigma}[\phi_L(l),\theta_L(l)] \end{array} \tag{2.47}$$

Les signaux B_{mn}^{σ} contiennent l'information représentative de l'onde acoustique primaire. Dans le cas HOA, si l'on admet que la captation est parfaite, les composantes B_{mn}^{σ} se déduisent exactement des signaux microphoniques $c_{HOA}(q)$, de telle sorte que :

$$\mathbf{s}_{HOA} = (\mathbf{W}_L\mathbf{Y}_L)^t\mathbf{b}.$$

Par suite, si l'on considère que chaque haut-parleur émet une onde sphérique (Equ. 2.16), les signaux alimentant les haut-parleurs s'expriment :

$$s_{HOA}(l,\omega) = \sum_{m=0}^{M}\frac{h_m^-(kr_L)}{h_0^-(kr_L)}\sum_{n=0}^{m}\sum_{\sigma=\pm1}B_{mn}^{\sigma}(\omega)Y_{mn}^{\sigma}[\phi_L(l),\theta_L(l)] \tag{2.48}$$

La similitude entre les équations 2.47 et 2.48 est frappante, d'autant que l'expression 2.45 ne le laissait pas présager. Mais ce résultat n'est pas surprenant, car on se rappelle que les concepts

WFS et HOA sont basés sur deux modes de représentation *équivalents* de l'onde acoustique. Cette équivalence conduit *naturellement* à une convergence si l'on se base sur une représentation commune (développement sur la base des harmoniques sphériques). La différence majeure entre les équations 2.47 et 2.48 portent sur le nombre de termes de la somme : nombre infini dans le cas WFS, nombre fini égal à $(M+1)^2$ dans le cas HOA. La troncature à l'ordre M pour HOA présente l'avantage d'un contrôle sur le nombre de composantes spectrales encodées et décodées. Ce paramètre n'est absolument pas maîtrisé avec WFS.

2.3 Evaluation comparée des systèmes WFS et HOA

2.3.1 Motivations

Depuis la fin des années 90, les technologies WFS et HOA ont fait l'objet de nombreuses études. Dès lors on pourrait croire qu'elles sont parfaitement connues dans leur moindres aspects et qu'elles ne soulèvent plus de questions, ni de débat. C'est d'ailleurs grâce à l'ensemble de ces travaux qu'il est possible de décrire les 2 procédés avec un formalisme unifié, comme nous venons de le voir. De nouvelles études ont confirmé leurs similitudes en leur apportant un nouvel éclairage [Poletti, 2005] [Fazi et al., 2008] [Fazi et al., 2009]. Leurs différences restent en revanche moins bien comprises, ce qui est dommage car c'est vraisemblablement dans cette compréhension que résident des solutions potentielles pour dépasser les limites actuels de la synthèse sonore par un réseau multi haut-parleurs qu'il soit de type WFS ou HOA, voire hybride. Plus généralement, un certain nombre de questions relatives à la mise en œuvre des procédés WFS et HOA restent encore aujourd'hui sans réponse. La première question qui vient à l'esprit concerne les éléments de décision permettant d'opter pour l'une ou l'autre technologie (WFS ou HOA) pour un problème donné. Nous manquons d'études comparant les qualités des rendus WFS et HOA à configuration égale[12] et qui permettraient de choisir la technologie optimale en fonction des contraintes de la situation. En l'état actuel des connaissances, ce choix relève le plus souvent d'une option arbitraire. Les seuls éléments permettant d'opter pour l'une des technologies sont :

- s'il faut effectuer une captation naturelle : seule la technologie HOA propose un système de captation audio 3D (sphère de microphones),
- si l'espacement entre les haut-parleurs est trop important, la solution WFS est à écarter en raison du repliement spatial dont la fréquence d'apparition devient si basse que la qualité de rendu est dégradée sur l'essentiel de la bande audible.

Malgré le nombre conséquent d'études consacrées aux 2 technologies ces dernières années, force est de constater aussi qu'elles se répartissent en deux catégories :

- des études pratiques mettant en œuvre des systèmes HOA limités aux premiers ordres, dans lesquels on ne tire pas parti de tout le potentiel de la technologie HOA [Guastavino & Katz, 2004] [Pulkki & Hirvonen, 2005],
- des explorations théoriques des équations fondamentales, prenant en compte l'apport des ordres supérieurs mais souffrant de l'absence d'un ancrage concret [Poletti, 2005] [Solvang, 2009] [Fazi et al., 2009].

Les travaux de thèse de S. Bertet constituent un premier exemple d'une étude intermédiaire considérant la mise en œuvre des ordres supérieurs avec l'évaluation par des tests d'écoute de prototypes de microphone HOA jusqu'à l'ordre 4 [Bertet, 2009].

Il en résulte que, au delà du choix initial de la technologie, lorqu'on a opté pour un système WFS ou HOA et qu'on souhaite le mettre en œuvre, se posent les questions suivantes qui n'ont pas encore véritablement de réponse :

[12] A ma connaissance une seule étude de ce type a été publiée [Spors, 2009].

- Jusqu'à quel ordre M est-il pertinent de monter ? Existe-t-il un ordre limite au delà duquel aucun bénéfice n'est obtenu ? Pour un ordre M donné qu'apportent les composantes d'ordre M+1 ? De quelle manière cet apport dépend-il de l'ordre M ?
- Pour une qualité attendue de rendu (notamment en termes de taille de zone d'écoute, de précision spatiale des sources virtuelles...), quel est l'ordre M optimal (au sens du meilleur compromis entre résultat et coût) ?
- Quelle est l'origine des "effets de phase" perçus sur un rendu HOA ?
- Comment expliquer que le rendu WFS soit si apprécié (en termes de spatialisation) lors de tests d'écoute, en dépit d'approximations relativement grossières des équations théoriques et du repliement spatial parfois très prononcé ?
- Quel est l'espacement maximal entre les haut-parleurs à partir duquel la qualité du rendu WFS devient inacceptable en raison des artefacts associés au repliement spatial ?
- Pour un dispositif donné de réseau de haut-parleurs, que donne une spatialisation WFS en comparaison de la solution HOA (non seulement en termes de localisation des sources virtuelles, mais aussi de respect de leur timbre) ? Comment les différences entre les deux systèmes évoluent-elles en fonction du nombre N_L de haut-parleurs ? En quoi dépendent-elles de la nature et de la position de la source virtuelle (onde sphérique, onde plane, source intérieure) à synthétiser ?
- En quoi l'observation des signaux alimentant les haut-parleurs (à la fois en amplitude et en phase) permet de comprendre les différences de qualité perçue entre les systèmes WFS et HOA ?
- Comment la qualité des rendus WFS et HOA évolue-t-elle sur la zone d'écoute ? Comment évolue-t-elle lorsque l'auditeur tourne la tête ?

Pour répondre à ces questions, il convient de se doter d'outils et de critères d'évaluation de la qualité de synthèse de la scène sonore. Les différentes méthodes sont illustrées dans la littérature :

- Observation de l'onde synthétisée (obtenue soit par la mesure de systèmes réels, soit par la simulation numérique) [Nicol, 1999] [Daniel, 2000] [Spors, 2009] : La comparaison de l'onde cible avec l'onde synthétique donne une première évaluation macroscopique qui permet de juger de la qualité de reconstruction de la forme spatiale globale de l'onde (front d'onde, propagation...). Mais elle ne permet pas de l'évaluer en détail. Il n'est d'ailleurs pas forcément utile de pousser plus finement l'analyse sur l'onde acoustique, sans chercher à se focaliser sur les seules informations pertinentes du point de vue de la perception, c'est à dire celles qui sont exploitées par le système auditif.
- Evaluation de l'onde telle qu'elle est perçue et analysée par le système auditif : Au final la scène sonore virtuelle est destinée à être écoutée. La méthode d'évaluation qui s'en rapproche le plus et qui semble donc la plus appropriée pour cette raison, consiste à s'intéresser aux signaux à l'entrée des oreilles de l'auditeur. Mais ces signaux ne signifient rien en eux-mêmes, ce qui compte ce sont les informations qu'en extrait le système auditif. Pour avoir accès à ces informations, il faut soit passer par un test d'écoute où l'on donne à entendre et juger l'onde synthétique par des sujets (on demande au sujet de localiser les sources virtuelles - test de localisation - [Sontacchi et al., 2002] [Pulkki & Hirvonen, 2005] [Capra et al., 2007] [Bertet, 2009] ou de juger la scène sonore sur une grille d'attributs perceptifs [Farina & Ugolotti, 1999] [Rumsey, 2001] [Guastavino & Katz, 2004]), soit extraire des informations perceptives des signaux en entrée des conduits auditifs grâce à un modèle du système auditif, auquel cas on s'affranchit du coût et des difficultés pratiques d'un test subjectif [Pulkki & Hirvonen, 2005] [Bertet, 2009] [Solvang, 2009] [Baskind, 2003]. Par exemple, il est possible de calculer l'ITD

et l'ILD. Des modèles[13] de localisation permettent d'en déduire une estimation de la direction perçue de la source virtuelle afin de la comparer à la direction cible [Pulkki & Hirvonen, 2005] [Bertet, 2009]. Il est aussi possible d'évaluer l'écart entre le timbre reproduit en entrée des 2 oreilles et celui qu'aurait produit la source, en vue de quantifier les distorsions spectrales [Solvang, 2009].

2.3.2 Ambition et méthodologie

A ce niveau, ma contribution a pour principale ambition de proposer une panoplie d'outils pour une évaluation comparée des technologies WFS et HOA, en donnant certes des premiers résultats, mais elle ne prétend pas apporter toutes les réponses aux questions que je viens de soulever. Le premier outil est le **formalisme générique** décrivant les procédés de synthèse WFS et HOA.

L'onde qu'on cherche à synthétiser peut être :

– une *onde sphérique* (source localisée au point $\vec{r}_S$, dite *extérieure* si elle est située à l'extérieur du réseau de haut-parleurs, et *intérieure* sinon) :

$$p_S(\omega, \vec{r}) = \frac{e^{-ik\rho_S}}{4\pi\rho_S} \text{ où } \rho_S = |\vec{r} - \vec{r_S}| \tag{2.49}$$

– une *onde plane* (caractérisée par le vecteur d'onde $\vec{k_P}$ de norme k) :

$$p_P(\omega, \vec{r}) = \frac{e^{-i\vec{k_P}.\vec{r}}}{4\pi} \tag{2.50}$$

L'onde synthétisée par le réseau de haut-parleurs s'exprime, en reprenant le formalisme matriciel introduit dans la première section :

– Système WFS[14] :

$$\begin{aligned} \hat{\mathbf{p}}_{WFS} &= \mathbf{p}_L \mathbf{s}_{WFS} \\ &= \mathbf{p}_L \mathbf{c}_{WFS} \end{aligned} \tag{2.51}$$

– Système HOA :

$$\begin{aligned} \hat{\mathbf{p}}_{HOA} &= \mathbf{p}_L \mathbf{s}_{HOA} \\ &= \mathbf{p}_L (\mathbf{W}_L \mathbf{Y}_L)^t \mathbf{b} \end{aligned} \tag{2.52}$$

Dans ces expressions, les vecteurs $\hat{\mathbf{p}}_{WFS}$ et $\hat{\mathbf{p}}_{HOA}$ désignent l'onde synthétisée respectivement par le procédé WFS et HOA aux N_E points d'écoute sélectionnés $(\vec{r_n}, n = 1, ..., N_E)$. La matrice $\mathbf{p}_L$ représente l'ensemble des ondes induites par les N_L haut-parleurs aux N_E points d'écoute. Si l'on considère qu'un haut-parleur émet une onde sphérique de type p_S :

$$p_L(n, l) = \frac{e^{-ik\rho_{(n,l)}}}{4\pi\rho_{(n,l)}} \text{ avec } \rho_{(n,l)} = |\vec{r_n} - \vec{r_l}| \ .$$

[13] Les vecteurs *Vélocité* et *Energie* proposés comme critères de localisation par Gerzon [Gerzon, 1992a] sont un exemple de tels modèles.

[14] Dans le cas d'une source intérieure, le procédé subit une lègère modification. On identifie le haut-parleur le plus éloigné de la source et ce haut-parleur définit la référence de phase, c'est à dire que pour que le système génère des fronts d'onde convergents (et non plus divergents comme pour une source extérieure), ce haut-parleur doit émettre en premier. Il devient effectivement la référence de phase si on normalise la phase des signaux alimentant les haut-parleurs par sa phase.

On peut de façon équivalente considérer que les haut-parleurs rayonnent des ondes planes de type p_P. En ce cas il convient de modifier la matrice de décodage en conséquence. Il suffit d'omettre la matrice $\mathbf{W}_L$. Pour HOA, l'équation mentionne le vecteur $\mathbf{b}$ des coefficients théoriques, mais il peut être remplacé par les coefficients estimés par un réseau de microphones :

$$\hat{\mathbf{b}} = \mathbf{E}_C(\mathbf{Y}_C^t\mathbf{Y}_C)^{-1}\mathbf{Y}_C^t\mathbf{c}_{HOA} \ .$$

Les signaux issus des microphones cardioïdes $\mathbf{c}_{WFS}$ ou $\mathbf{c}_{HOA}$ sont donnés par :

$$c_q = p(\omega, \vec{r_q}) - \frac{\vec{\nabla}p(\omega, \vec{r_q}).\vec{n_i}}{ik} \text{ où } p \equiv p_S \text{ ou } p \equiv p_P \ .$$

Pour obtenir les signaux induits à l'entrée des conduits auditifs, en prenant notamment en compte l'interaction des ondes acoustiques avec la morphologie de l'auditeur (au moyen de ses HRTF), il suffit de modifier la matrice $\mathbf{p}_L$ en la multipliant[15] d'une part par la matrice $\mathbf{H}_{el}$ contenant les HRTF décrivant les fonctions de transfert entre les N_L haut-parleurs et l'oreille gauche de l'auditeur pour chaque point d'écoute et d'autre part par la matrice $\mathbf{H}_{er}$ qui représente l'équivalent de $\mathbf{H}_{el}$ pour l'oreille droite. On obtient ainsi la paire des signaux $\hat{p}_{WFS}(el)$ et $\hat{p}_{WFS}(er)$, ou $\hat{p}_{HOA}(el)$ et $\hat{p}_{HOA}(er)$. Pour chaque position d'écoute, l'incidence de chaque haut-parleur est identifiée pour sélectionner la fonction de transfert associée. L'orientation de la tête de l'auditeur est prise en compte. Par défaut l'auditeur regarde dans la direction $\vec{v}(1, 0, 0)$, mais d'autres pointages peuvent être considérés. Les HRTF utilisées appartiennent à la base *Jean-Marie Pernaux*[16] (base privée d'Orange Labs).

Par suite, cette évaluation comparée des technologies WFS et HOA s'appuie uniquement sur la simulation de la chaîne d'encodage et de décodage d'une source sonore. Les ondes synthétiques $\hat{\mathbf{p}}_{WFS}$ et $\hat{\mathbf{p}}_{HOA}$ qui sont observées sont obtenues par le jeu des calculs matriciels qui viennent d'être décrits. Ce choix offre une totale flexibilité sur la configuration des systèmes pour une étude systématique des phénomènes. Les principaux paramètres[17] des systèmes sont le nombre de haut-parleurs N_L et l'ordre M d'encodage HOA. Deux déclinaisons du système HOA sont considérées : synthèse par des ondes planes ou par des ondes sphériques. Les autres paramètres de l'étude concernent la source sonore à synthétiser : position ($\vec{r}_S$ pour l'onde sphérique ou vecteur d'onde $\vec{k}_P$ pour l'onde plane) et fréquence. Dans cette première étude, on considère une captation HOA idéale (signaux $\mathbf{b}$). La restitution se fait par un réseau circulaire de haut-parleurs de rayon $r_L = 1.5$ m. Les signaux alimentant les haut-parleurs (s_{HOA} ou s_{WFS}) sont normalisés par l'énergie totale des N_L sources.

On s'intéresse aux signaux suivants :

- signaux alimentant les haut-parleurs s_{WFS} et s_{HOA} (amplitude et phase),
- onde synthétique $\hat{p}_{WFS}$ et $\hat{p}_{HOA}$ observée sur la zone d'écoute comprise à l'intérieur du réseau de haut-parleurs,
- ondes synthétiques $\hat{p}_{WFS}(el)$ et $\hat{p}_{WFS}(er)$ (respectivement $\hat{p}_{HOA}(el)$ et $\hat{p}_{HOA}(er)$ pour HOA) produites au niveau des oreilles de l'auditeur.

Les signaux présents à l'entrée des conduits auditifs ne sont pas considérés en tant que tels, mais sont utilisés pour estimer des indices représentatifs des informations dont dispose le système auditif pour construire une "image" perceptive de la scène sonore (notamment pour identifier la localisation des sources sonores). Trois indices sont extraits (cf. Chap. 3) :

- l'ITD (*Interaural Time Difference*) estimée comme le retard de phase des signaux gauche et droite moyenné sur la bande [0-2kHz] [Kulkarni et al., 1999],

[15] Produit terme à terme : $\mathbf{p}_L.\mathbf{H}_{el}$ et $\mathbf{p}_L.\mathbf{H}_{er}$

[16] Les résultats présentés par la suite sont obtenus avec les HRTF du sujet RN de cette base.

[17] D'autres configurations pourraient être étudiées par la suite : décodages HOA non basiques optimisés, configurations irrégulières de haut-parleurs par exemple...

– l'ILD (*Interaural Level Difference*) estimée comme le rapport en dB des énergies des signaux gauche et droite sur la bande [1-5kHz] [Larcher, 2001],
– l'ISSD (*Inter Subject Spectrum Difference*) estimée comme la variance de la différence entre le spectre reproduit et le spectre cible (prenant en compte les HRTF qui auraient filtré le spectre de la source sonore située à la position de la source virtuelle) considérés sur la bande [4-13kHz] [Guillon, 2009].

L'ITD et l'ILD sont des **indices de localisation**, c'est à dire qu'ils font partie des informations potentiellement utilisées par le système auditif pour juger la localisation en azimut des sources sonores. En l'occurrence ils servent à la perception de la latéralisation des sources sonores. Le processus de localisation auditive reste encore aujourd'hui complexe, du moins dans la faculté qu'a le cerveau de s'adapter et de prendre en compte des informations connexes (multimodalité). On sait par exemple que même en présence d'indices de localisation dégradés, le système auditif est capable d'identifier la position de la source, notamment en recoupant les différentes sources d'information. Dans ces conditions, il nous semble délicat d'utiliser un modèle de localisation qui risque de se baser sur des hypothèses peut-être trop limitatives des processus psycho-acoustiques mis en jeu. Nous préférons nous en tenir à la matière brute des indices de localisation que sont l'ITD et l'ILD. Nous cherchons à évaluer dans quelle mesure le procédé de synthèse les reconstruit correctement, sans présumer de la capacité du système auditif à les exploiter correctement.

L'ISSD n'est pas un indice de localisation. Dans sa définition originelle, elle a été proposée pour quantifier la dissimilarité entre les IS de deux HRTF [Middlebrooks, 1999b]. Dans notre cas, elle est appliquée pour évaluer la dissimilarité entre les modules spectraux des signaux induits par la source sonore et l'onde synthétique. Comme il s'agit d'une restitution 2D limitée au plan horizontal, il pourrait ne pas sembler utile d'observer la reconstruction des IS qui sont principalement utilisés pour la localisation en élévation. En effet, avec l'ISSD, le premier objectif est d'évaluer les éventuelles colorations spectrales introduites par la synthèse, en référence au spectre naturellement perçu pour une source sonore localisée à la position de la source virtuelle. A un second niveau, étant donné le flou de la frontière entre détimbrage et IS[18], il peut arriver que les altérations de spectre introduites par le procédé de synthèse soient perçues comme des modifications de l'élévation. La source sonore ne serait alors plus localisée dans le plan horizontal, ce qui peut être un artéfact gênant. l'ISSD reprend alors son sens originel. Les 3 critères ITD, ILD et ISSD sont évalués sur l'ensemble de la zone d'écoute. Leur évolution en fonction de l'orientation de la tête de l'auditeur est aussi étudiée.

2.3.3 Influence du nombre de haut-parleurs

Les figures 2.3 & 2.4 reproduisent les ondes synthétisées par les systèmes WFS et HOA pour des réseaux comportant un nombre croissant de haut-parleurs. L'onde qu'on veut synthétiser est une onde plane (cf. Fig. 2.2). On remarque que l'onde synthétisée par HOA semble très proche de l'onde cible sur l'ensemble de la zone d'écoute, quel que soit le nombre de haut-parleurs. Lorsque N_L augmente, l'onde synthétique gagne seulement en précision, ce qui se traduit par la disparition d'oscillations rapides qui viennent perturber la forme des fronts d'onde. L'onde synthétisée par WFS est très dégradée par le repliement spatial qui se manifeste sur l'ensemble de la zone d'écoute avec un réseau de $N_L = 10$ haut-parleurs ($f_{al} \simeq 180Hz$) . Lorsque N_L croît, l'effet s'estompe progressivement jusqu'à devenir quasi inexistant à $N_L = 80$ ($f_{al} \simeq 1.44kHz$). On vérifie sur la Figure 2.5 que pour le réseau de 10 haut-parleurs l'impact du repliement spectral s'atténue lorsque la fréquence de l'onde décroît.

Les signaux alimentant les haut-parleurs sont représentés sur la Figure 2.6 & 2.7 en fonction

[18]Fondamentalement (c'est à dire du point de vue de l'onde physique), un IS n'est qu'un détimbrage qui selon le cas est interprété comme tel ou comme une élévation de la source sonore.

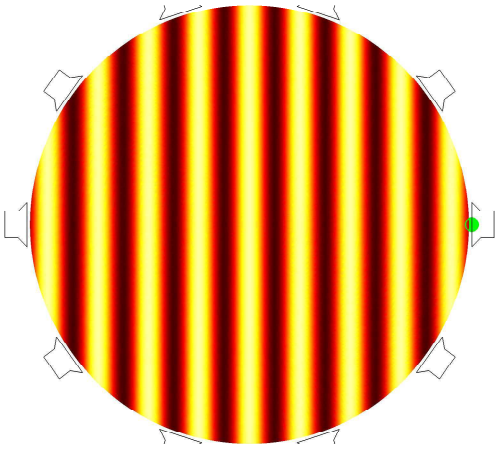
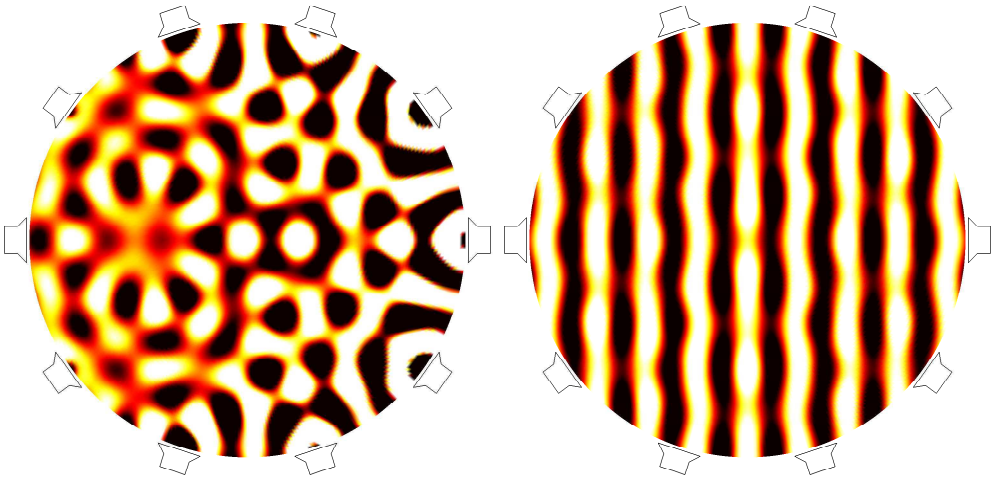


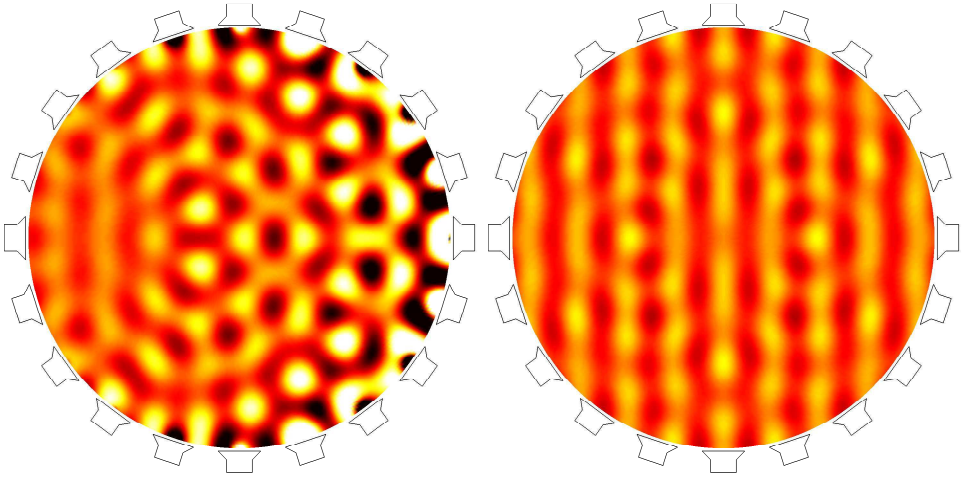
FIG. 2.2 – Onde plane cible à synthétiser (azimut $\phi = 0°$, fréquence : f = 1 kHz). Le point vert repère la direction de l'onde qu'on veut reproduire.

de l'azimut des haut-parleurs, ce qui définit la fonction de *panning* qui traduit la répartition du travail de synthèse sur le réseau. Comme il s'agit d'une onde plane située dans la direction d'un des haut-parleurs du réseau, cette fonction est très sélective pour HOA : l'effort de reconstruction est quasiment porté par un seul haut-parleur. Pour WFS, la fonction de panning présente un maximum pour ce haut-parleur, mais elle s'avère nettement moins sélective. On peut considérer que l'ensemble du réseau contribue à l'effort de synthèse, à l'exception du haut-parleur situé dans la direction opposée à celle de l'onde, où la fonction de panning présente un zéro en raison de la directivité cardioïdes des microphones de la captation. Pour la phase des signaux, on observe que les signaux s_{HOA} sont en phase, c'est à dire que le contrôle des haut-parleurs ne s'applique qu'à leur amplitude. En revanche, pour WFS, il agit à la fois sur leur amplitude et leur phase.

L'ITD évaluée sur l'ensemble de la zone d'écoute pour les ondes synthétisées par WFS et HOA est illustrée sur les Figures 2.10 & 2.11. Comme l'auditeur fait face à l'onde plane, l'ITD attendue (c'est à dire celle qu'aurait induite l'onde plane qu'on cherche à reconstruire) vaut 0 μs quelle que soit la position d'écoute (cf. Fig. 2.9). Pour HOA, l'ITD n'est correcte qu'au centre de la zone d'écoute. Il présente de plus de très fortes variations (de l'ordre de $\pm 700 \mu s$) sur l'ensemble de la zone d'écoute, principalement sur la zone frontale ($\phi \in$ [-90, 90°]). Cette ITD sous-tendrait une localisation potentielle de la source dans les directions $\phi = \pm 90°$ au lieu de 0°. On note que l'évolution de l'ITD reste continue et qu'elle ne présente pas des variations brusques qui risqueraient de provoquer des modifications de la direction perçue au moindre mouvement de l'auditeur. La cartographie de l'ITD ne dépend presque pas du nombre de haut-parleurs. Elle tend seulement à se stabiliser (disparition totale des variations brusques) avec un nombre élevé de haut-parleurs. Les observations pour WFS sont très différentes. Pour $N_L = 10$, l'ITD n'est correct qu'au centre de la zone d'écoute, mais présente une évolution très instable et incohérente sur le reste de la zone d'écoute. La cause est sans aucun doute le repliement spectral. Cependant lorsque le nombre de haut-parleurs augmente, l'ITD se stabilise progressivement jusqu'à valoir 0 μs de façon homogène sur l'ensemble de la zone d'écoute lorsque le réseau comporte 80 haut-parleurs. Déjà avec 40 haut-parleurs, l'ITD est correcte sur une large portion de la zone d'écoute. De plus, pour le réseau de 20 haut-parleurs, on observe que si l'ITD estimée sur la bande [0-2 kHz] est erronée, son estimation limitée aux fréquences inférieures à 500 Hz donne des valeurs beaucoup plus proches de la valeur attendue (cf. Fig. 2.12), ce qui confirme que l'origine de la dégradation de l'ITD est le repliement

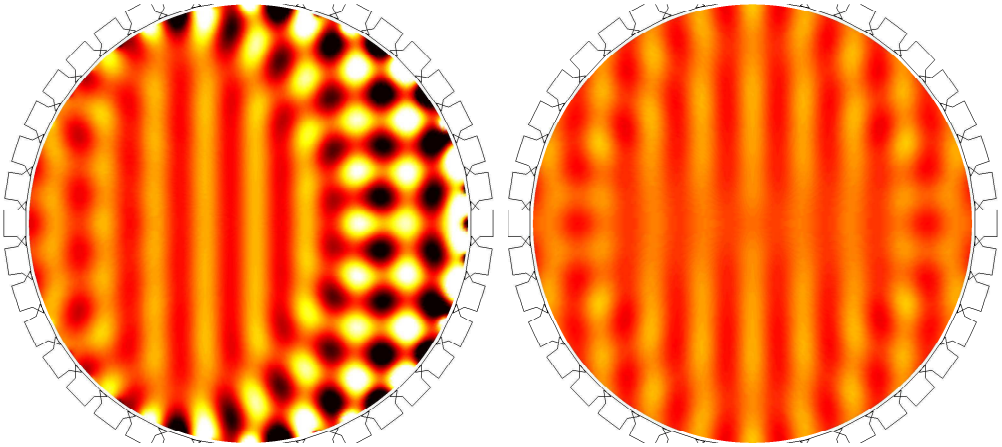


(a) WFS ($N_L = 10$) (b) HOA ($N_L = 10$, M = 4)

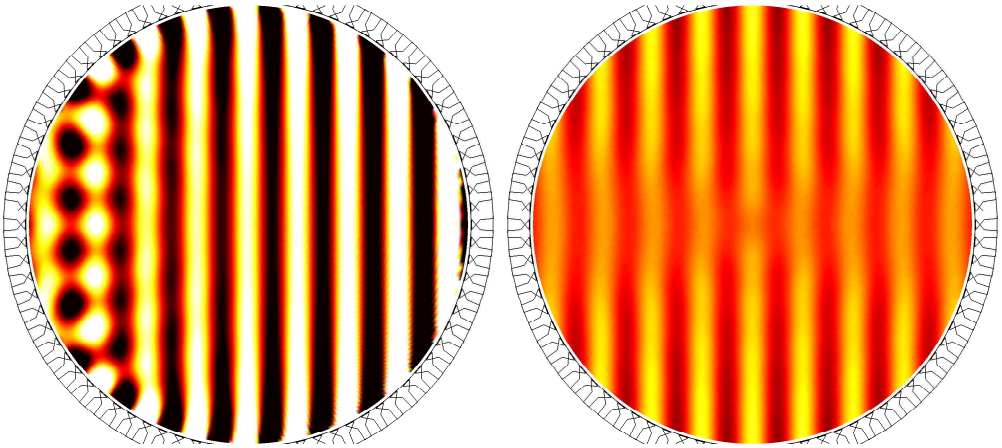


(c) WFS ($N_L = 20$) (d) HOA ($N_L = 20$, M = 9)

FIG. 2.3 – Illustration des ondes synthétisées par les systèmes WFS (synthèse par ondes sphériques) et HOA (synthèse par ondes planes) : Evolution en fonction du nombre de haut-parleurs N_L de 10 à 20 (onde plane d'azimut $\phi = 0°$, fréquence : f = 1 kHz). L'amplitude des ondes est représentée par une échelle de couleurs qui est identique pour les 4 configurations, mais est légèrement différente de celle utilisée pour l'affichage de la Figure 2.2. Cette remarque s'applique à l'ensemble des figures présentées dans cette étude. On observe que l'amplitude de l'onde synthètique varie sensiblement en fonction des paramètres de la synthèse, même si dans chaque une normalisation par l'énergie totale des signaux des haut-parleurs est appliquée pour minimiser les écarts potentiels d'amplitude.

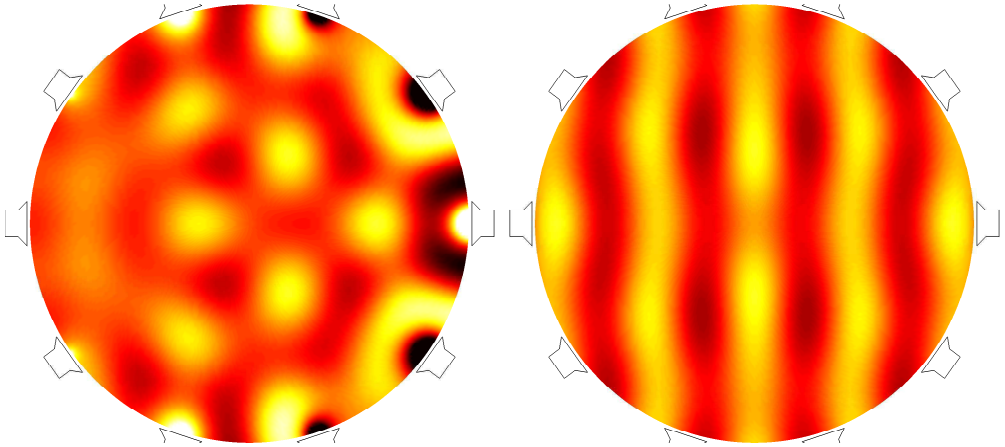


(a) WFS ($N_L = 40$) (b) HOA ($N_L = 40$, M = 19)

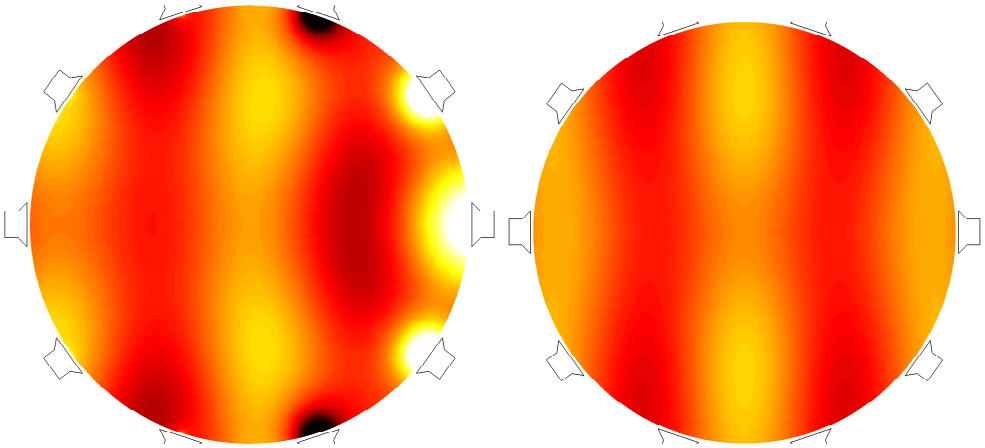


(c) WFS ($N_L = 80$) (d) HOA ($N_L = 80$, M = 39)

FIG. 2.4 – Illustration des ondes synthétisées par les systèmes WFS (synthèse par ondes sphériques) et HOA (synthèse par ondes planes) : Evolution en fonction du nombre de haut-parleurs N_L de 20 à 40 (onde plane d'azimut $\phi = 0°$, fréquence : f = 1 kHz).

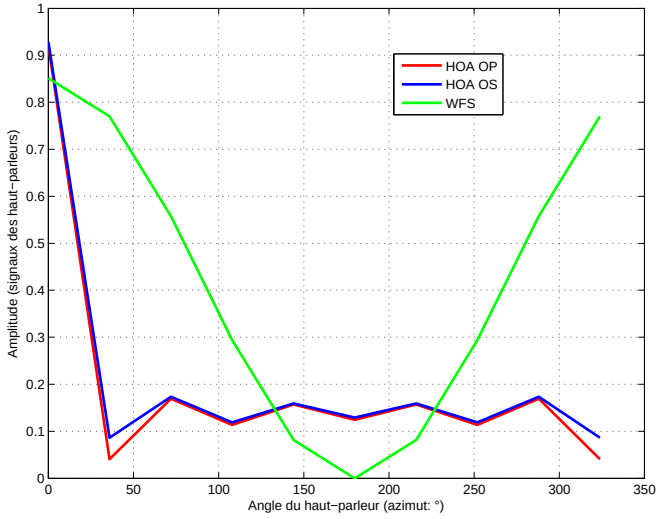


(a) WFS ($N_L = 10$) : f = 500 Hz (b) HOA ($N_L = 10$, M = 4) : f = 500 Hz

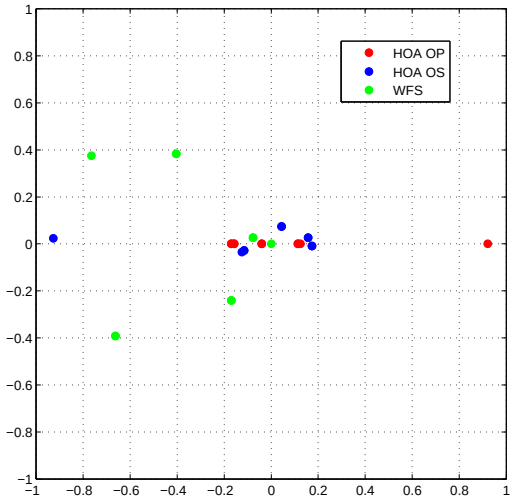


(c) WFS ($N_L = 10$) : f = 250 Hz (d) HOA ($N_L = 10$, M = 4) : f = 250 Hz

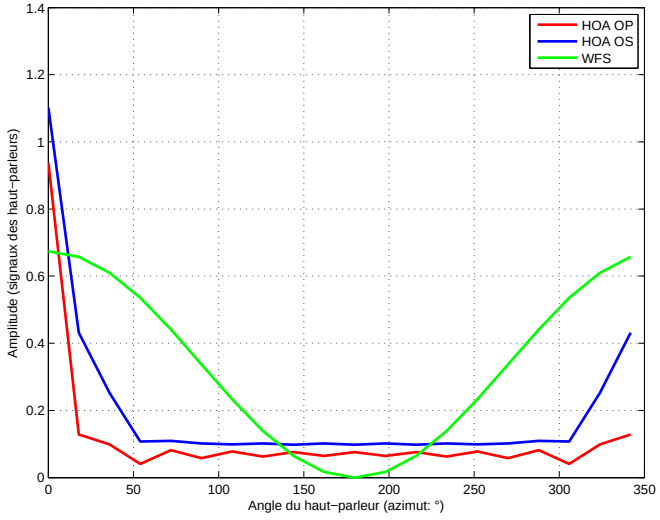
FIG. 2.5 – Illustration des ondes synthétisées par les systèmes WFS (synthèse par ondes sphériques) et HOA (synthèse par ondes planes) : ondes synthétisées aux fréquences f = 500 et 250 Hz (onde plane d'azimut $\phi = 0°$).



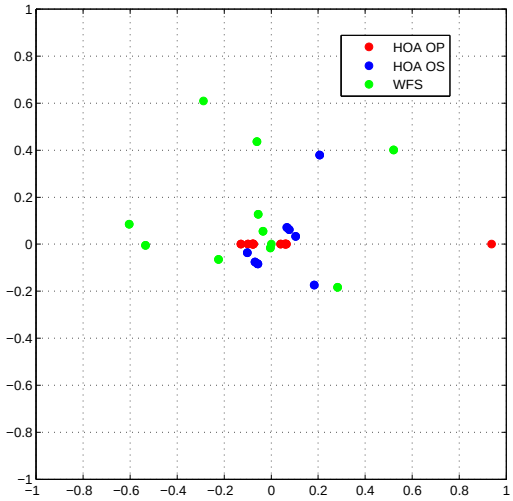
(a) Amplitude ($N_L = 10$, M = 4)



(b) Phase ($N_L = 10$, M = 4)

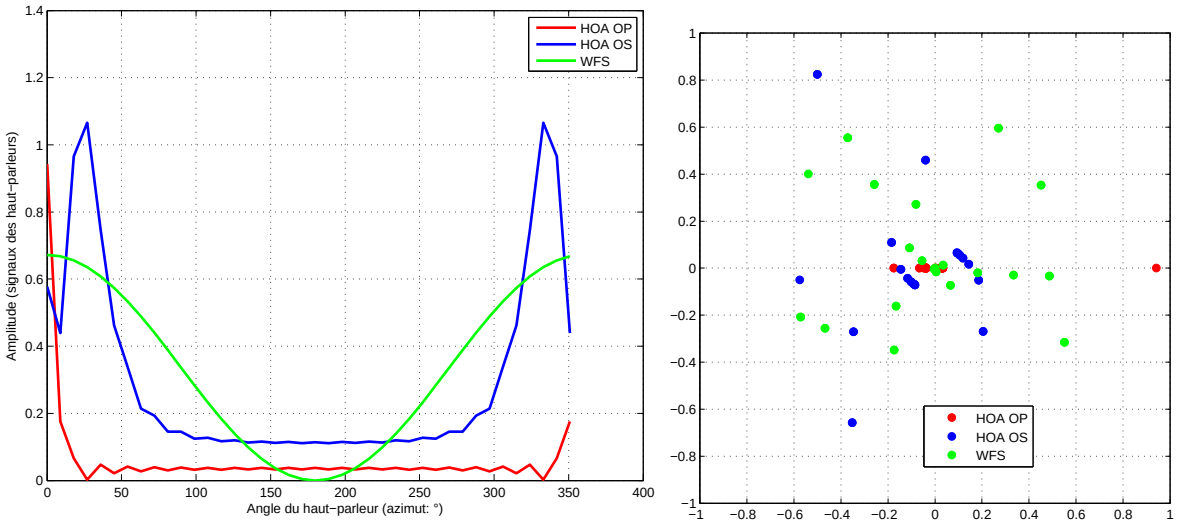


(c) Amplitude ($N_L = 20$, M = 9)



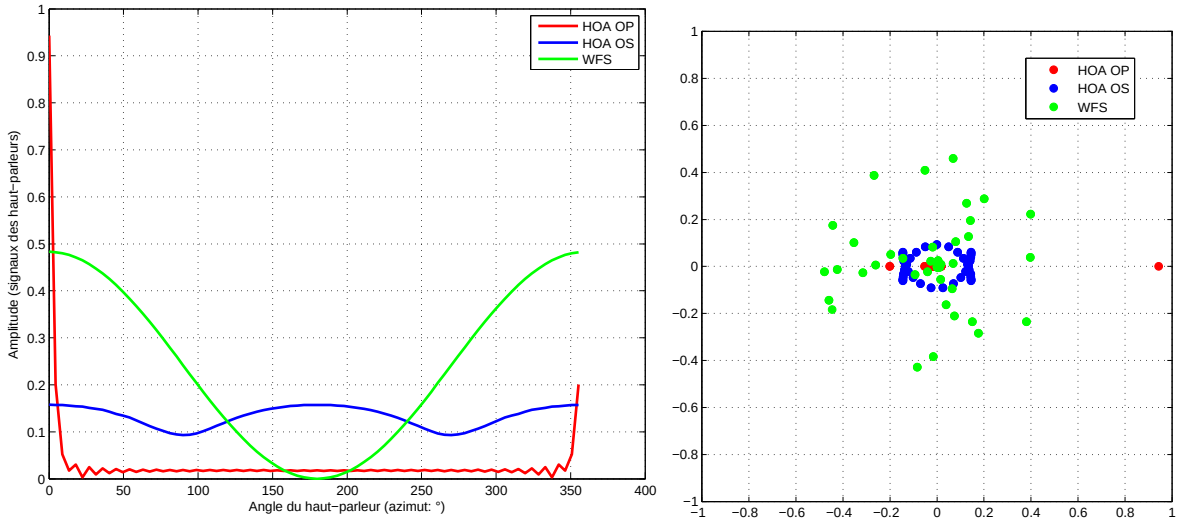
(d) Phase ($N_L = 20$, M = 9)

FIG. 2.6 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde plane (azimut $\phi = 0°$, fréquence : f = 1 kHz) : évolution en fonction du nombre de haut-parleurs. La phase des signaux est illustrée en représentant les signaux dans le plan complexe. Pour la synthèse HOA, deux cas sont considérés : soit les haut-parleurs émettent des ondes planes (HOA OP), soit des ondes sphériques (HOA OS), tandis que pour le rendu WFS ils émettent des ondes sphériques. On remarque qu'en raison de la normalisation des signaux des haut-parleurs par leur énergie totale, l'amplitude maximale de s_{WFS} décroît lorsque le nombre de haut-parleurs augmente.



(a) Amplitude ($N_L = 40$, M = 19)

(b) Phase ($N_L = 40$, M = 19)



(c) Amplitude ($N_L = 80$, M = 39)

(d) Phase ($N_L = 80$, M = 39)

FIG. 2.7 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde plane (azimut $\phi = 0°$, fréquence : f = 1 kHz) : évolution en fonction du nombre de haut-parleurs. La phase des signaux est illustrée en représentant les signaux dans le plan complexe. Pour la synthèse HOA, deux cas sont considérés : soit les haut-parleurs émettent des ondes planes (HOA OP), soit des ondes sphériques (HOA OS), tandis que pour le rendu WFS ils émettent des ondes sphériques. On note le comportement marginal de la synthèse HOA OS, sur lequel on reviendra ultérieurement (cf. Section 2.3.8).

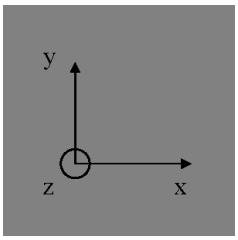
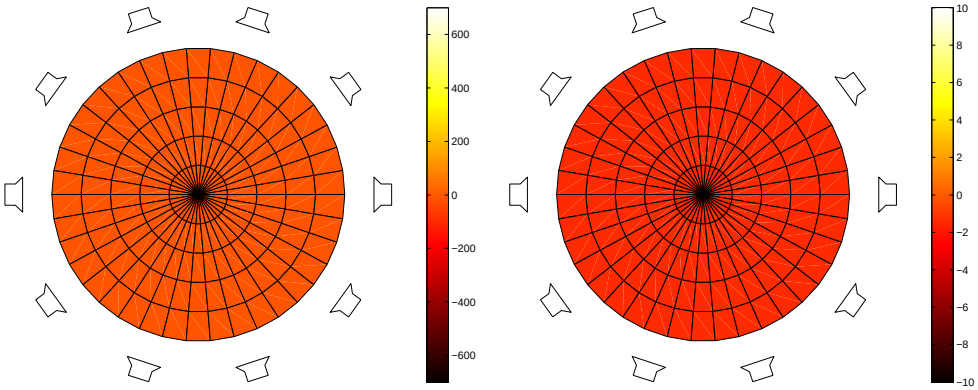


FIG. 2.8 – Système de coordonnées associé à l'affichage des ITD, ILD et ISSD.



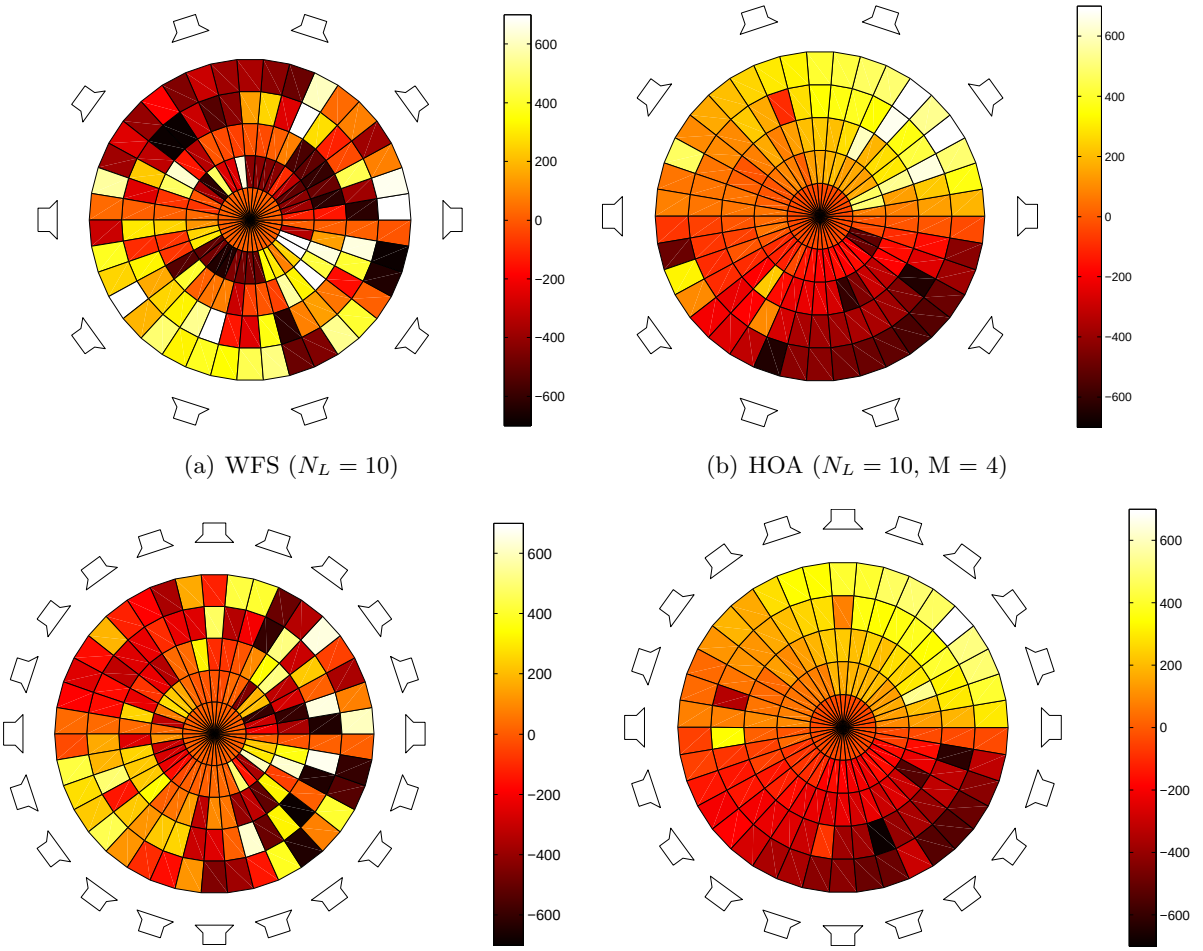
(a) ITD (μs) (b) ILD (dB)

FIG. 2.9 – ITD et ILD évaluées sur la zone d'écoute pour une onde plane d'azimut $\phi = 0°$ (pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$).

spectral. En dessous de la fréquence d'aliasing f_{al}, l'ITD est donc correctement restituée.

Comme pour l'ITD, on s'attend à une ILD nulle (ILD = 0 dB) sur l'ensemble de la zone d'écoute (cf. Fig. 2.9). A l'instar de l'ITD, l'ILD évaluée sur l'onde synthétisée par HOA n'est correcte qu'au centre de la zone. Sur le reste de la zone, sa cartographie est proche de celle de l'ITD avec des variations très fortes (de l'ordre de ±10 dB). En chaque point, l'ILD est cohérente avec l'ITD pour induire une latéralisation de la direction perçue. Lorsque le nombre de haut-parleurs est élevé, on observe des instabilités localisées. Pour WFS, l'ILD semble plus robuste que l'ITD au repliement spatial. En effet, d'emblée avec 10 haut-parleurs, on observe une ITD très proche de 0 dB sur l'ensemble de la zone d'écoute, avec une homogénéité et une stabilité frappantes. L'amélioration apportée en augmentant le nombre de haut-parleurs est très peu sensible. La qualité inattendue de restitution des indices de localisation, à la fois de temps et d'énergie, est potentiellement une raison expliquant le confort de localisation ressenti par les sujets des tests d'écoute avec des systèmes WFS.

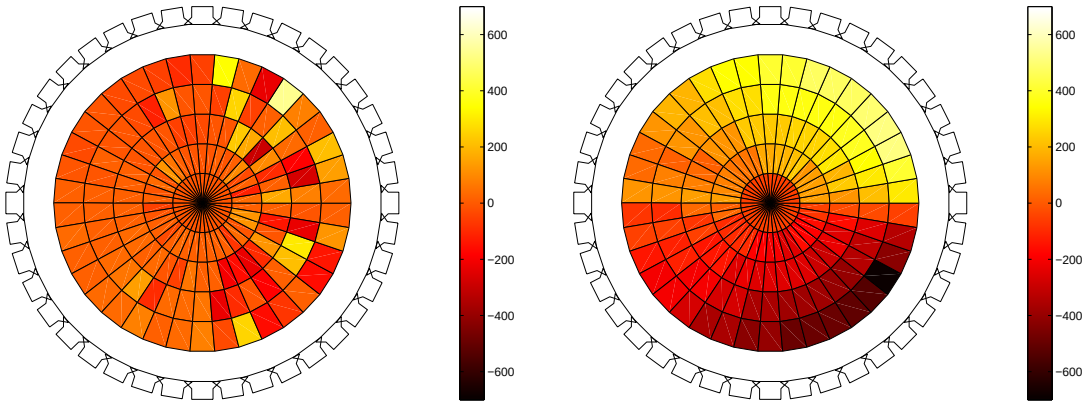
Les Figures 2.15 & 2.16 représentent l'ISSD évaluée pour les ondes synthétisées par les systèmes WFS et HOA. On n'observe pas d'évolution en fonction du nombre de haut-parleurs. HOA se caractérise par des fortes distorsions spectrales dans la partie frontale au voisinage des haut-parleurs. Au centre et dans la partie arrière, les détimbrages sont très faibles en revanche. Pour WFS, les distorsions spectrales restent à un niveau homogène relativement faible sur l'ensemble de la zone d'écoute. La particularité du système WFS semble résider dans sa qualité d'homogénité et de stabilité sur la zone d'écoute.



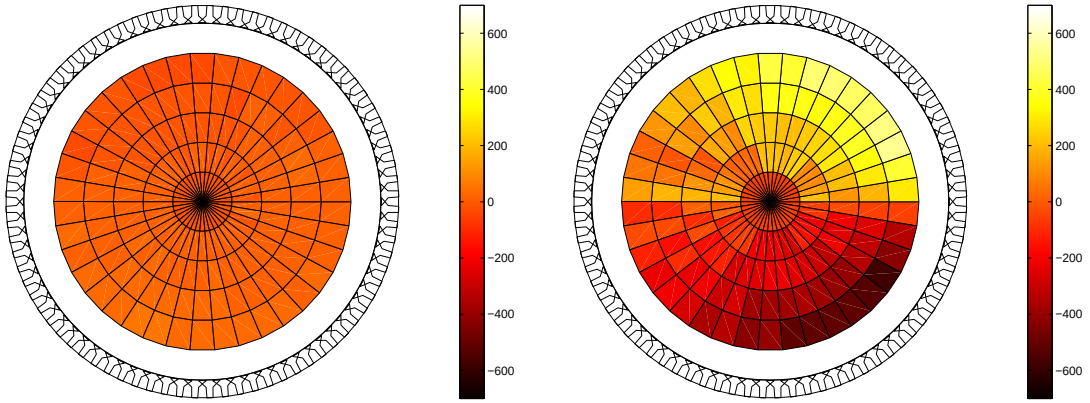
(a) WFS ($N_L = 10$) (b) HOA ($N_L = 10$, M = 4)

(c) WFS ($N_L = 20$) (d) HOA ($N_L = 20$, M = 9)

FIG. 2.10 – ITD (μs) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA(OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut ϕ = 0°). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



(a) WFS ($N_L = 40$) (b) HOA ($N_L = 40$, M = 19)



(c) WFS ($N_L = 80$) (d) HOA ($N_L = 80$, M = 39)

FIG. 2.11 – ITD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.

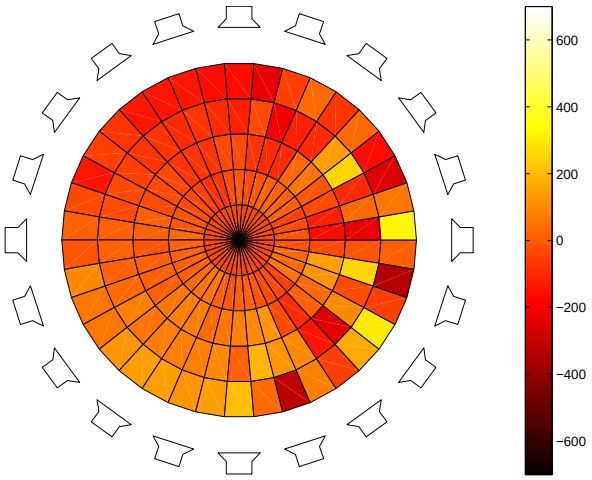
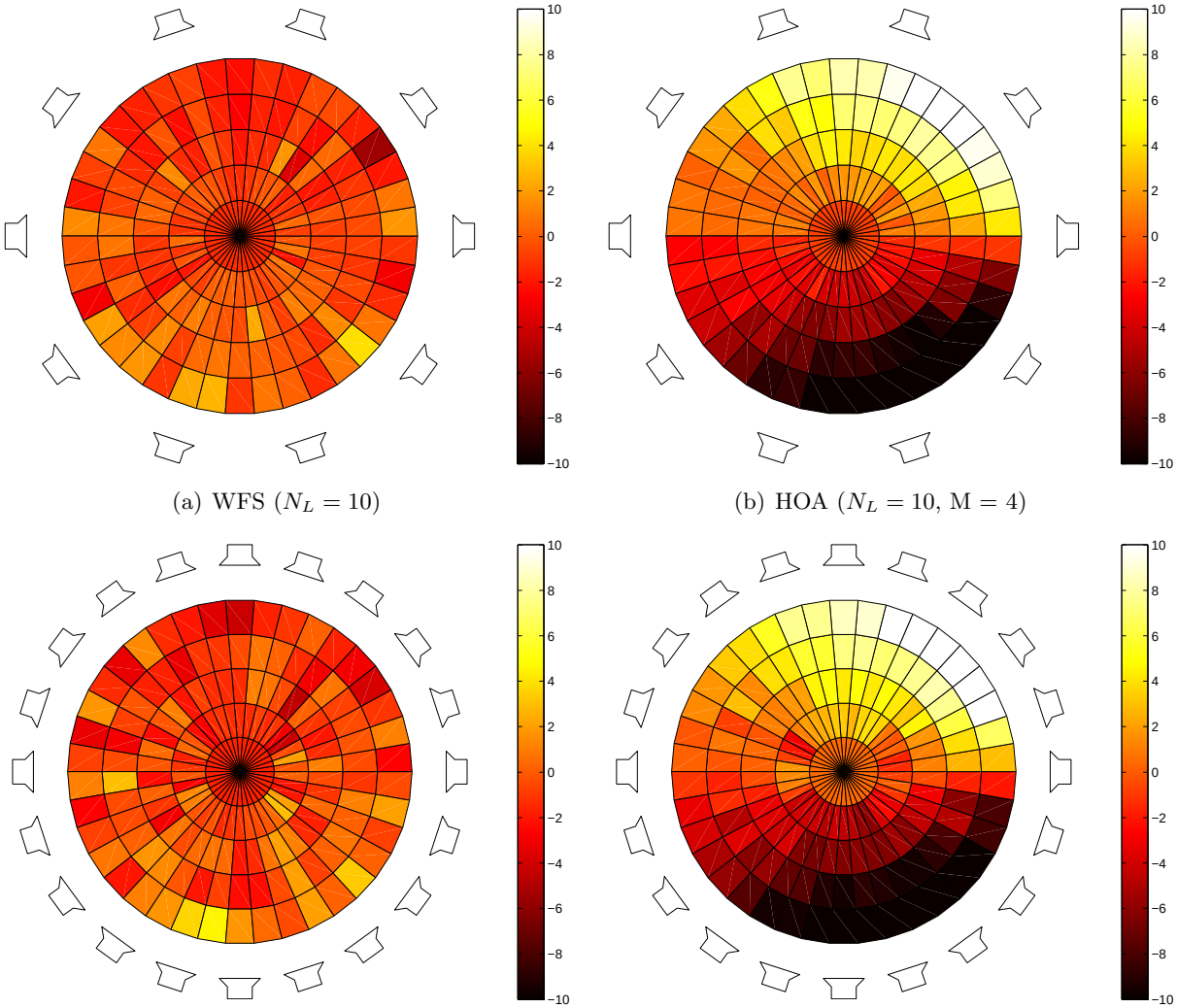


FIG. 2.12 – ITD (μs) évaluée pour la bande [0-500Hz] sur la zone d'écoute pour l'onde synthétisée par le système WFS (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



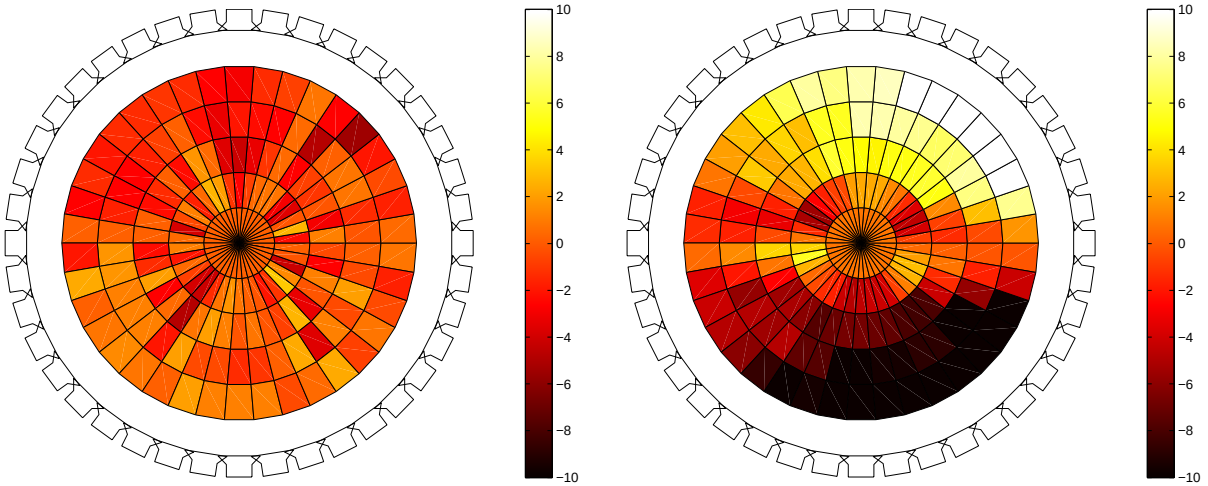
(a) WFS ($N_L = 10$)

(b) HOA ($N_L = 10$, M = 4)

(c) WFS ($N_L = 20$)

(d) HOA ($N_L = 20$, M = 9)

FIG. 2.13 – ILD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



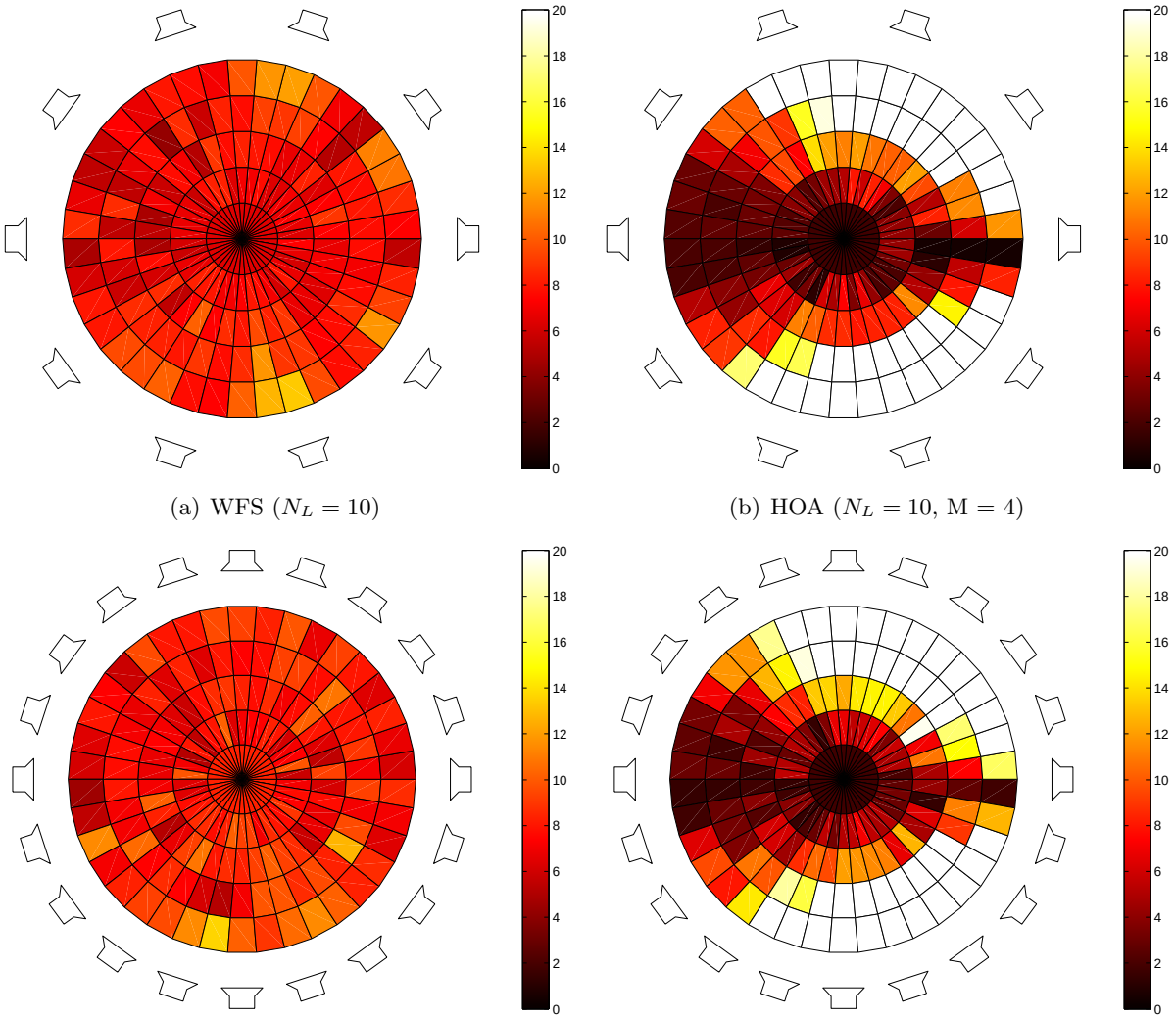
(a) WFS ($N_L = 40$)

(b) HOA ($N_L = 40$, M = 19)

(c) WFS ($N_L = 80$)

(d) HOA ($N_L = 80$, M = 39)

FIG. 2.14 – ILD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



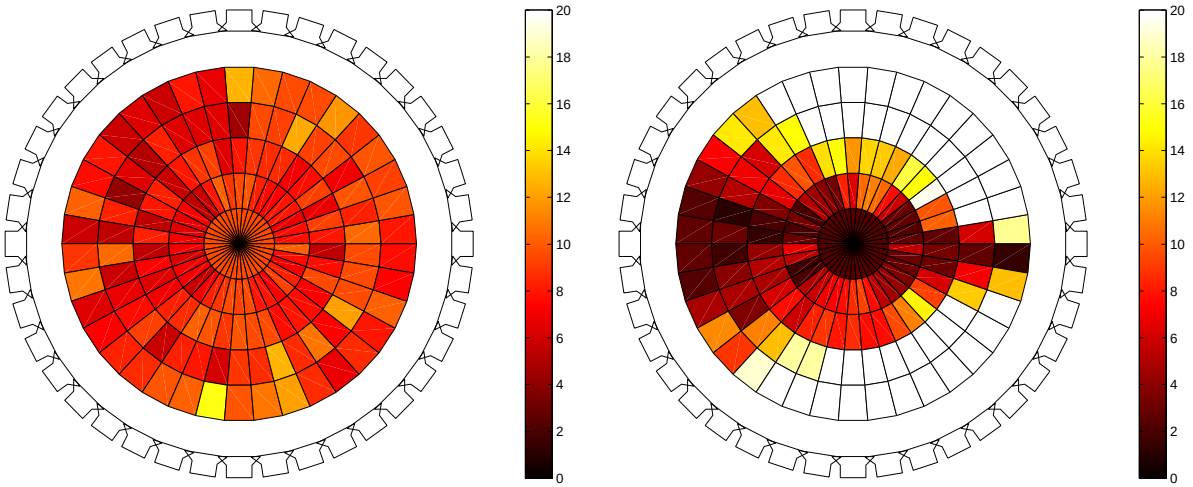
(a) WFS ($N_L = 10$)

(b) HOA ($N_L = 10$, M = 4)

(c) WFS ($N_L = 20$)

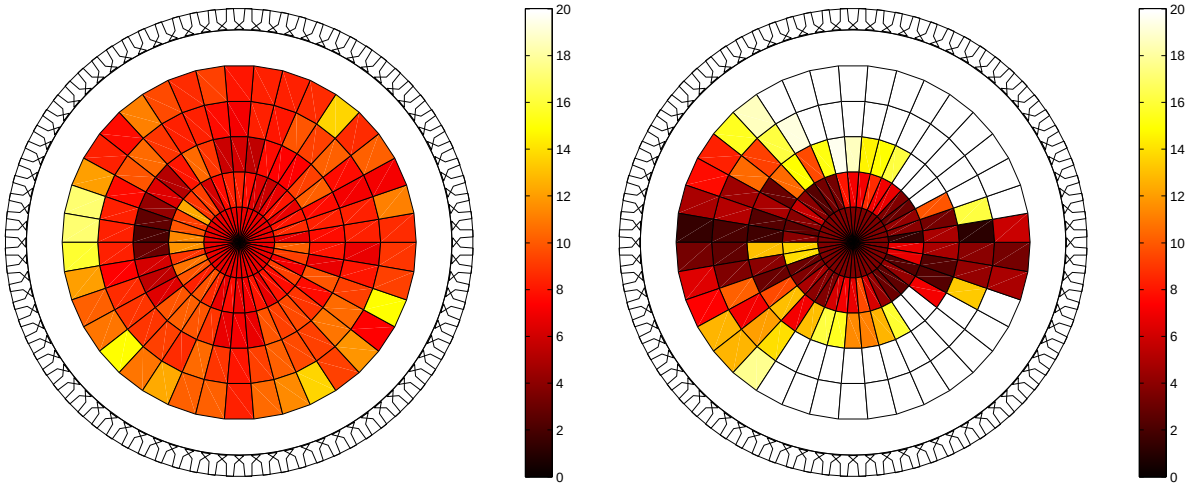
(d) HOA ($N_L = 20$, M = 9)

FIG. 2.15 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



(a) WFS ($N_L = 40$)

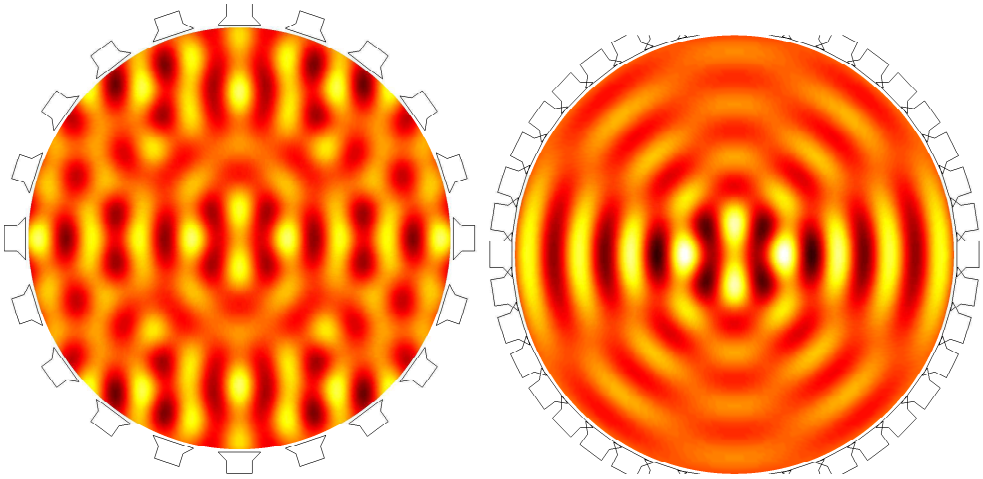
(b) HOA ($N_L = 40$, M = 19)



(c) WFS ($N_L = 80$)

(d) HOA ($N_L = 80$, M = 39)

FIG. 2.16 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Evolution en fonction du nombre de haut-parleurs N_L (onde plane d'azimut $\phi = 0°$). Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



(a) HOA ($N_L = 20$, M = 4) (b) HOA ($N_L = 40$, M = 4)

FIG. 2.17 – Illustration des ondes synthétisées par le système HOA : Augmentation du nombre N_L de haut-parleurs à ordre M constant d'encodage (onde plane d'azimut $\phi = 0°$, fréquence : f = 1 kHz, synthèse HOA OP).

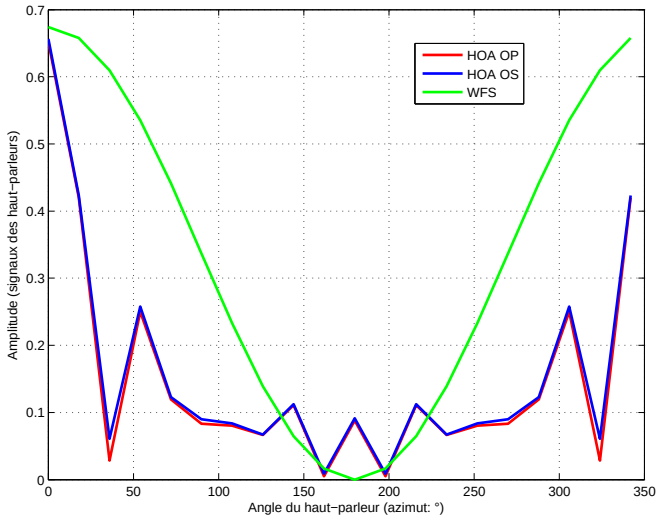
2.3.4 Relation optimale entre l'ordre M et le nombre de haut-parleurs

Pour un ordre donné M d'encodage, le nombre optimal de haut-parleurs à utiliser pour le décodage est souvent considéré comme égal à 2M+1 (décodage 2D). Si l'on dispose d'un nombre inférieur de haut-parleurs, il est impératif d'éliminer des composantes HOA, c'est à dire d'abaisser l'ordre M d'encodage, jusqu'à obtenir que $N_L \geq 2M + 1$. Si l'on ne prend pas cette précaution, on induit le repliement spatial des composantes des ordres supérieurs du fait que l'échantillonnage spatial réalisé par le réseau de haut-parleurs ne satisfait pas le critère de Nyquist. Etant admis que pour un ordre M donné, il faut au moins 2M+1 haut-parleurs, que se se passe-t-il si le réseau comporte plus de de 2M+1 haut-parleurs ? C'est à cette question que nous allons nous intéresser dans cette section.

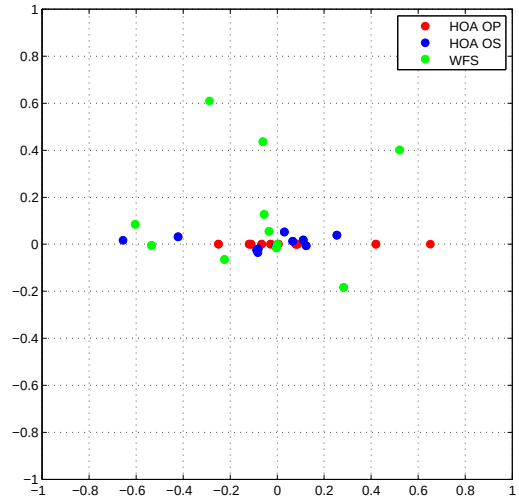
La Figure 2.17 illustre les ondes synthétisées par HOA pour un ordre M=4 d'encodage lorsque le réseau de décodage est composé de 20 ou 40 haut-parleurs au lieu des 9 préconisés. Il est clair que l'onde obtenue est très dégradée par rapport au résultat obtenu avec 10 haut-parleurs (cf. Fig. 2.3). La forme spatiale de l'onde synthétique ne ressemble plus guère à celle d'une onde plane. En termes d'indices de localisation (cf. Fig. 2.19), l'ITD et l'ILD présentent davantage d'instabilités au fur et à mesure que N_L croît. Toutefois on remarque que l'ITD atteint la valeur cible de 0 μs sur une zone relativement étendue correspondant à deux cônes s'élargissant à partir du centre de la zone d'écoute vers l'avant et l'arrière. Cette tendance se confirme pour $N_L = 40$. L'ISSD en revanche augmente sensiblement, ce qui dénote une aggravation des distorsions spectrales. Ces observations corroborent des résultats rapportés dans [Solvang, 2009] [Bertet, 2009].

2.3.5 Azimut de la source virtuelle

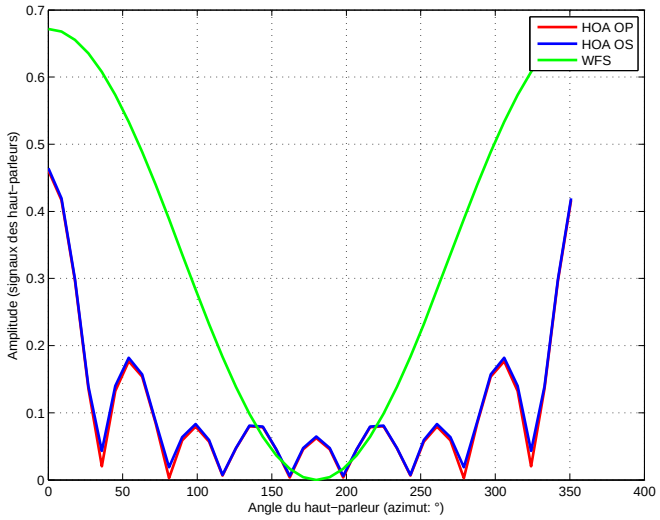
Nous avons considéré jusqu'à présent le cas de la synthèse d'une onde plane d'azimut $\phi = 0°$ qui se situe dans la direction d'un des haut-parleurs du réseau de décodage. Nous allons maintenant étudier le cas d'une onde plane d'incidence $\phi = 60°$ (cf. Fig. 2.20). La Figure 2.21 montre les ondes synthétisées par les systèmes WFS et HOA. On note que la synthèse HOA avec 10 haut-parleurs est moins performante que pour l'onde d'incidence $\phi = 0°$ (cf. Fig. 2.3). Pour WFS, les résultats



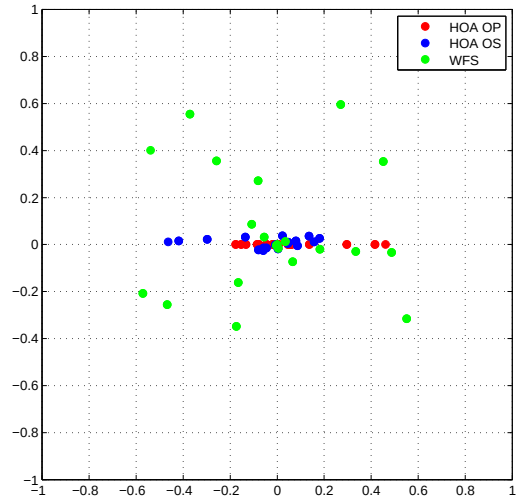
(a) Amplitude ($N_L = 20$, M = 4)



(b) Phase ($N_L = 20$, M = 4)

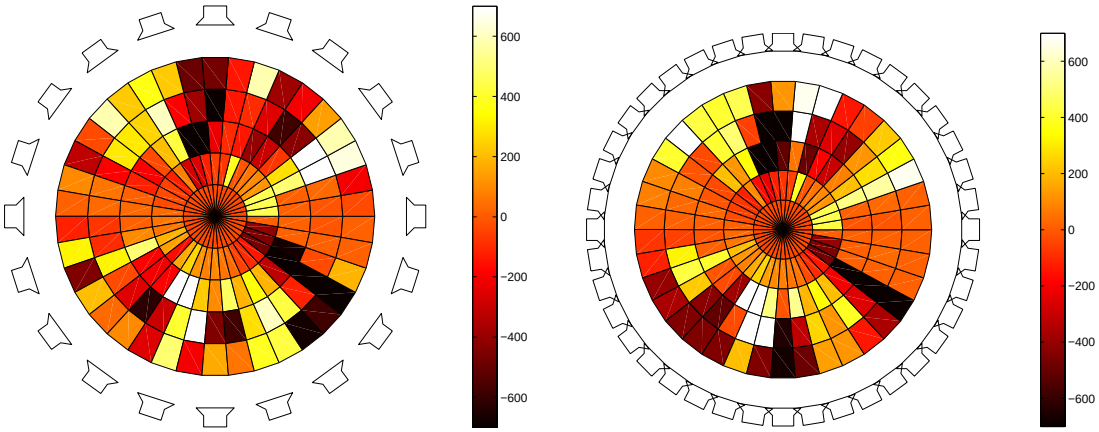


(c) Amplitude ($N_L = 40$, M = 4)



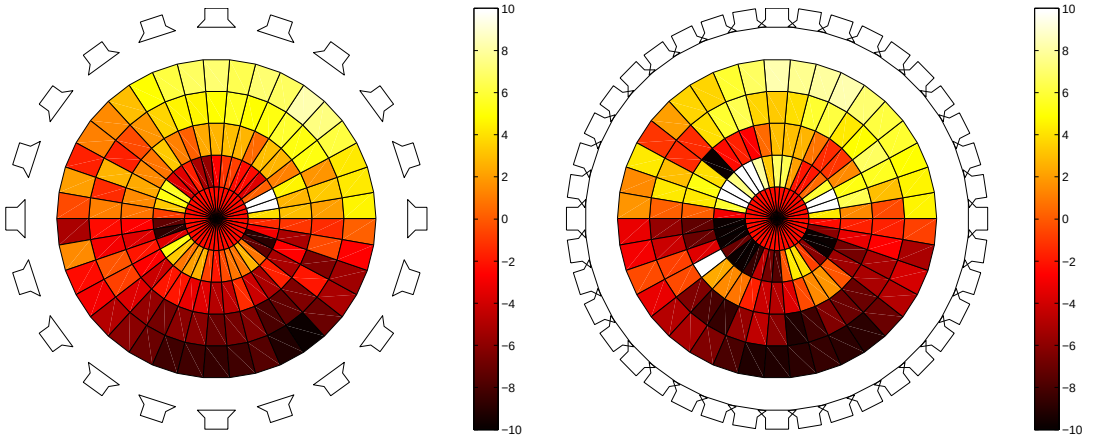
(d) Phase ($N_L = 40$, M = 4)

FIG. 2.18 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde plane (azimut $\phi = 0°$, fréquence : f = 1 kHz) : évolution en fonction du nombre de haut-parleurs à ordre M constant d'encodage.



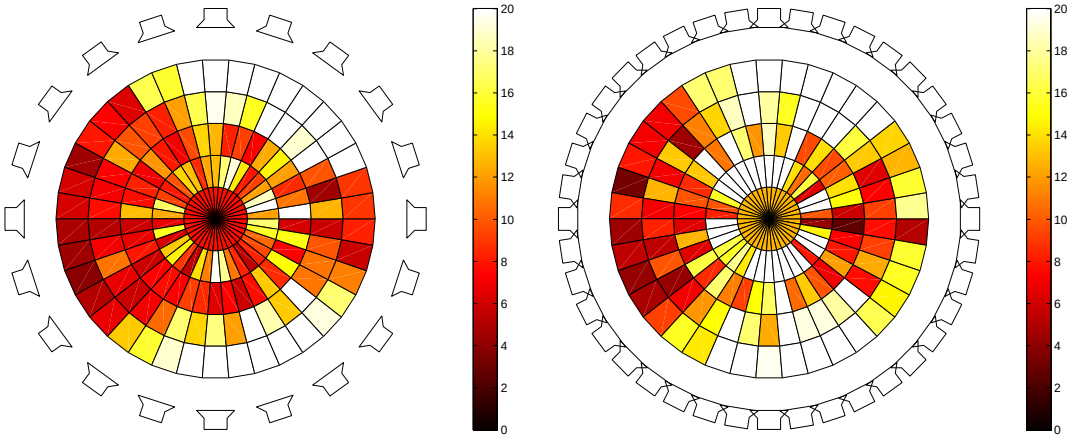
(a) ITD ($N_L = 20$, M = 4)

(b) ITD ($N_L = 40$, M = 4)



(c) ILD ($N_L = 20$, M = 4)

(d) ILD ($N_L = 40$, M = 4)



(e) ISSD ($N_L = 20$, M = 4)

(f) ISSD ($N_L = 40$, M = 4)

FIG. 2.19 – ITD(μs), ILD (dB) et ISSD évaluées sur la zone d'écoute pour les ondes synthétisées par le système HOA (onde plane d'azimut $\phi = 0°$, synthèse HOA OP) : Evolution en fonction du nombre de haut-parleurs N_L à ordre M constant d'encodage. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.

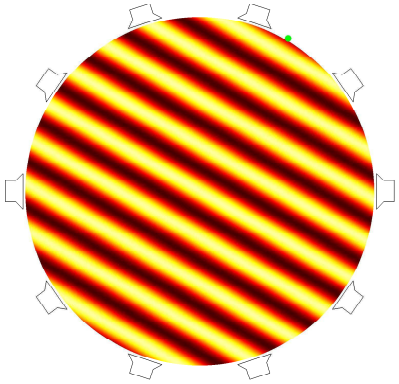
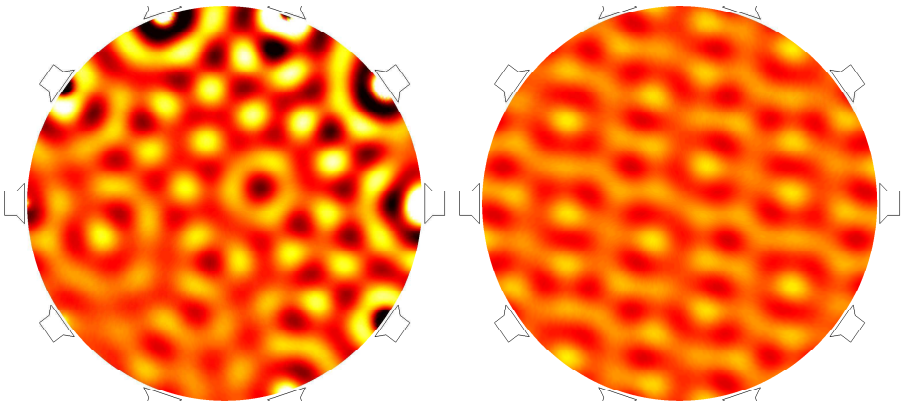
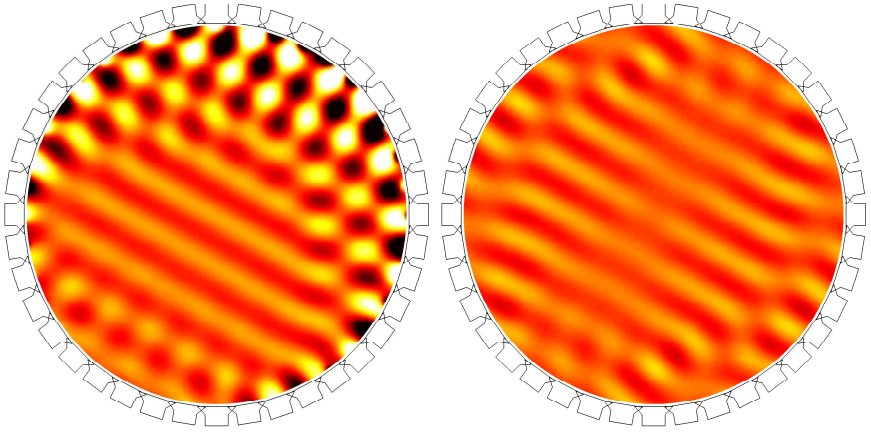


FIG. 2.20 – Onde plane cible à synthétiser (azimut $\phi = 60°$, fréquence : f = 1 kHz). Le point vert repère la direction de l'onde qu'on veut reproduire.



(a) WFS ($N_L = 10$)

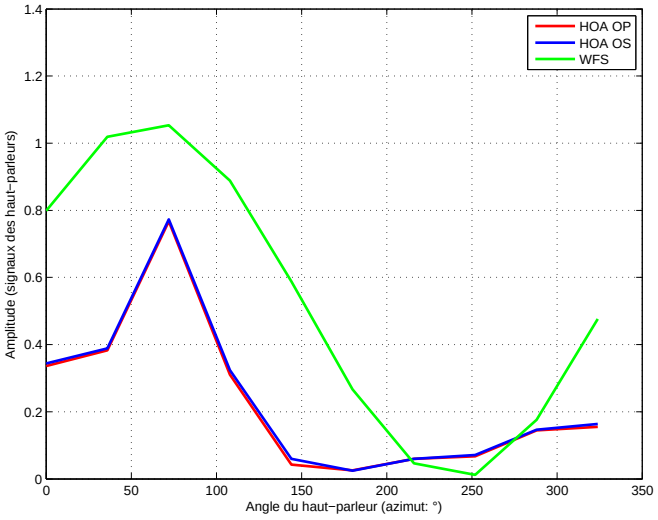
(b) HOA ($N_L = 10$, M = 4)



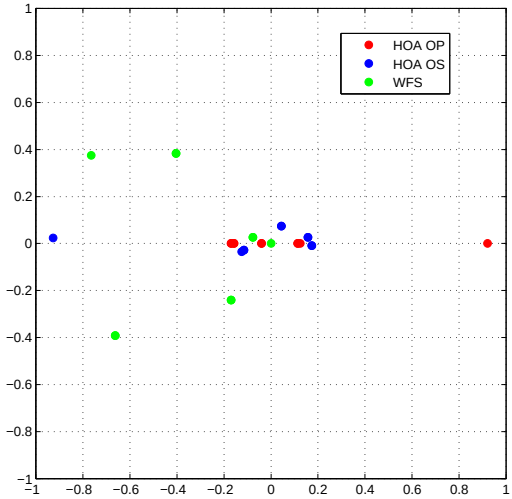
(c) WFS ($N_L = 40$)

(d) HOA ($N_L = 40$, M = 19)

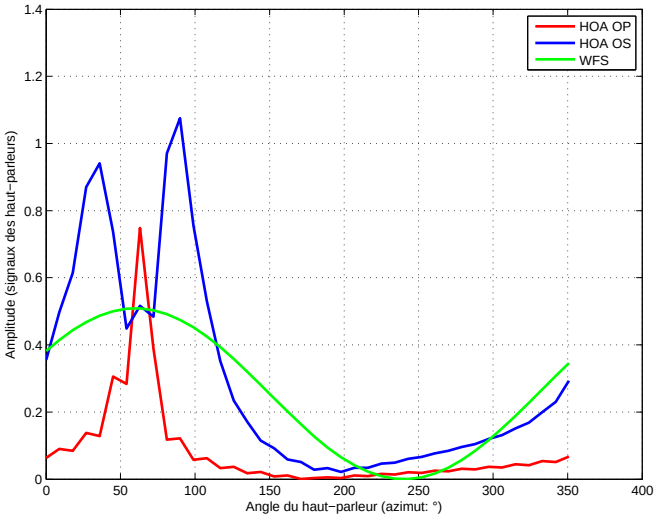
FIG. 2.21 – Illustration des ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde plane d'azimut $\phi = 60°$ (fréquence : f = 1 kHz).



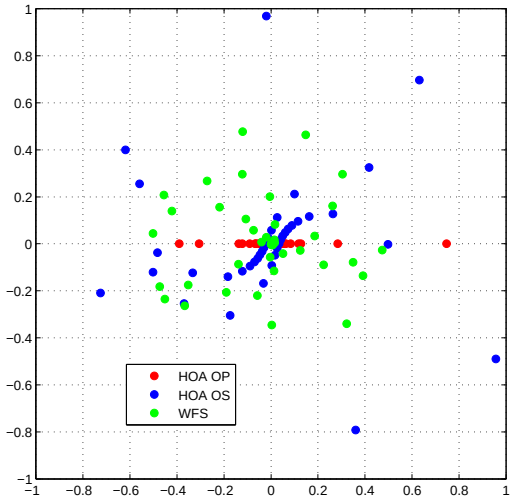
(a) Amplitude ($N_L = 10$, M = 4)



(b) Phase ($N_L = 10$, M = 4)



(c) Amplitude ($N_L = 40$, M = 19)



(d) Phase ($N_L = 40$, M = 19)

FIG. 2.22 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde plane d'azimut $\phi = 60°$ (fréquence : f = 1 kHz).

sont comparables. L'ITD est reproduit sur la Figure 2.23. Alors que pour WFS, la restitution de l'ITD est très proche des résultats obtenus pour l'onde d'azimut 0°, les performances du système HOA sont meilleures avec une ITD à la fois plus homogène sur la zone d'écoute et plus proche de la valeur attendue. La moitié gauche de la zone (azimut $\phi \in$ [60-240°]) présente cependant une sous-latéralisation. L'ILD s'améliore de la même façon pour HOA, tandis que pour WFS l'ILD révèle une sous-latéralisation (cf. Fig. 2.24). On note que pour les deux systèmes, l'ILD reste homogène et stable sur la zone d'écoute. En termes de rendu des timbres (cf. Fig. 2.25), les distorsions spectrales sont considérablement aggravées en comparaison de l'onde d'azimut 0° : l'ISSD atteint à présent des valeurs jusqu'à 80 dB au lieu de 20 dB. Une autre différence est que c'est le système WFS qui présente les détimbrages les plus prononcés avec des valeurs d'ISSD qui restent supérieures à 20 dB, tandis que, pour HOA, l'ISSD est de l'ordre de 10 à 20 dB sur une portion étendue de la zone d'écoute (en général la région opposée à la direction de l'onde).

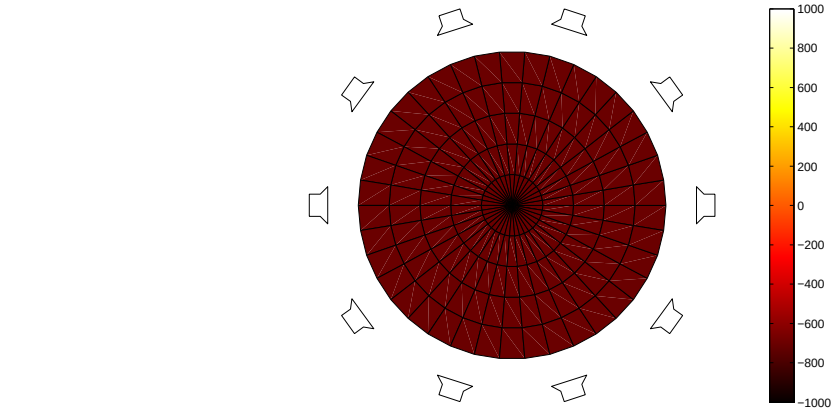
2.3.6 Synthèse d'une onde sphérique (source extérieure)

Après la synthèse d'une onde plane, passons au cas d'une onde sphérique. On considère d'abord une source située à l'extérieur de la zone d'écoute au point $\vec{r}_S(r_S = 3, \phi_S = 0, \theta_S = 0)$ (cf. Fig. 2.26). Les ondes synthétisées par WFS et HOA sont illustrées sur la Figure 2.27. Pour WFS on n'observe pas de différence avec le cas de l'onde plane : avec 10 haut-parleurs l'onde synthétique est très fortement altérée par le repliement spatial, il faut 40 haut-parleurs pour que la forme spatiale de l'onde soit conforme à une onde sphérique (cf. Fig. 2.26) sur une région étendue de la zone d'écoute. Dans le cas de HOA, il apparaît que la synthèse d'une onde sphérique n'est pas "naturelle" pour ce procédé : avec 10 haut-parleurs la forme spatiale de l'onde synthétique est plus proche d'une onde plane que d'une onde sphérique. Il faut un nombre important de haut-parleurs pour obtenir des fronts sphériques sur une zone étendue. Sur la Figure 2.28, on constate que les fonctions de panning pour WFS et HOA présentent une sélectivité comparable. L'effort de synthèse d'une onde sphérique par HOA requiert de mettre en œuvre l'ensemble du réseau. On relève une seconde différence avec le cas de l'onde plane : les haut-parleurs sont contrôlés à la fois en amplitude et en phase, comme pour la WFS.

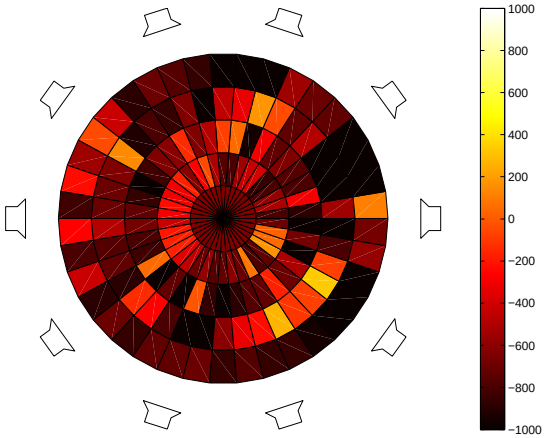
En ce qui concerne les indices de localisation (cf. Fig. 2.29, 2.30 & 2.31), l'ITD présente pour les deux systèmes davantage d'instabilités sur la zone d'écoute, notamment avec un faible nombre de haut-parleurs ($N_L = 10$). Avec un nombre élevé de haut-parleurs, WFS offre une ITD assez proche des valeurs attenues sur l'ensemble de la zone d'écoute. HOA se distingue par une forte instabilité et une amplification des valeurs d'ITD (en valeur absolue), amplification qui est susceptible d'induire une sur-latéralisation de la source virtuelle. Pour l'ILD, les résultats sont similaires à ceux de l'onde plane. L'ILD est correctement restituée par WFS même avec un faible nombre de haut-parleurs, bien qu'avec 40 haut-parleurs, des instabilités marginales soient relevées. En revanche, HOA se caractérise par une sur-latéralisation. Pour HOA, les distorsions spectrales sont nettement plus faibles que pour l'onde plane. Elles sont en général d'un niveau inférieur à celui observé pour WFS qui ne présente pas de différence notable avec le cas de l'onde plane, hormis une rlégère diminution des détimbrages avec un faible nombre de haut-parleurs.

2.3.7 Synthèse d'une onde sphérique (source intérieure)

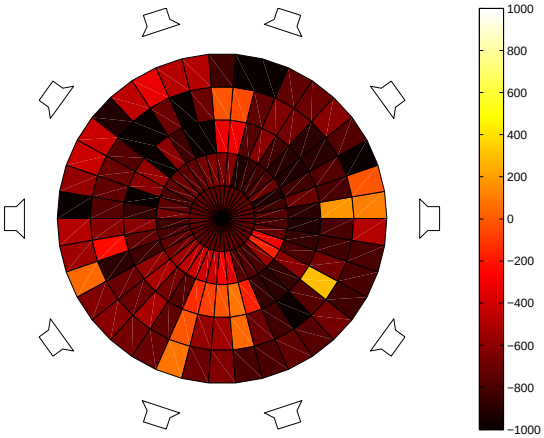
Nous allons maintenant nous intéresser au cas d'une source intérieure à la zone d'écoute située au point $\vec{r}_S(r_S = 1.25, \phi_S = 0, \theta_S = 0)$ (cf. Fig. 2.32). La Figure 2.33 illustre comment les systèmes WFS et HOA réussissent à reproduire une source virtuelle intérieure, apparemment aussi bien qu'une source extérieure. Sur la Figure 2.34, on remarque la forme particulière que prend la fonction de panning HOA : elle n'est plus maximale en 0°, mais présente deux maxima à $\pm$ 27° de part et



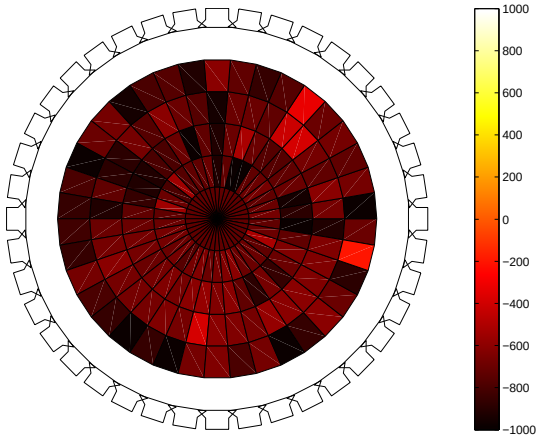
(a) Onde plane cible



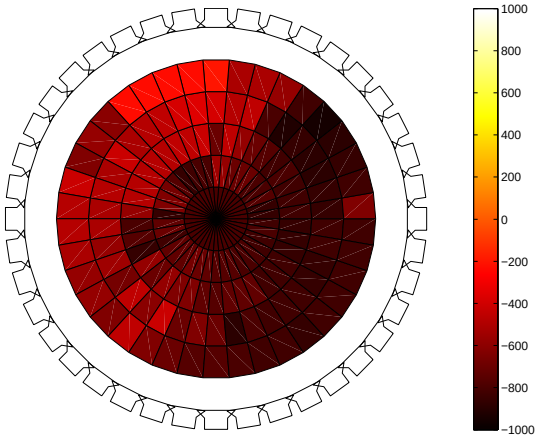
(b) WFS ($N_L = 10$)



(c) HOA ($N_L = 10$, M = 4)

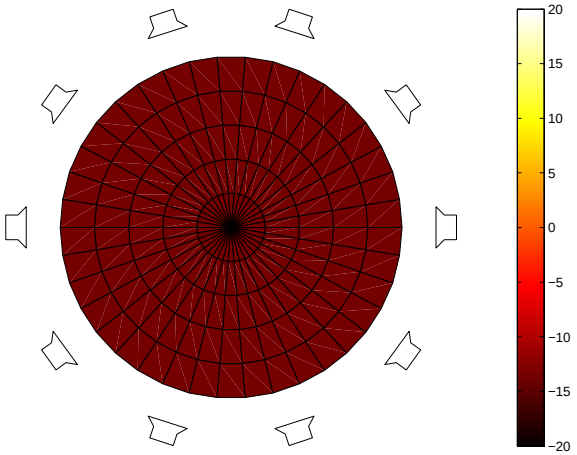


(d) WFS ($N_L = 40$)

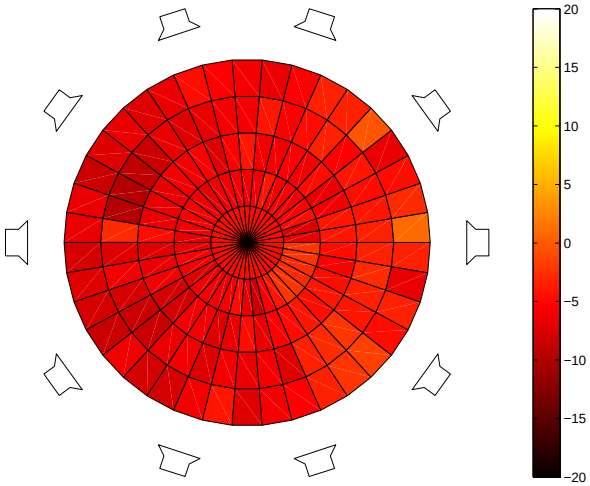


(e) HOA ($N_L = 40$, M = 19)

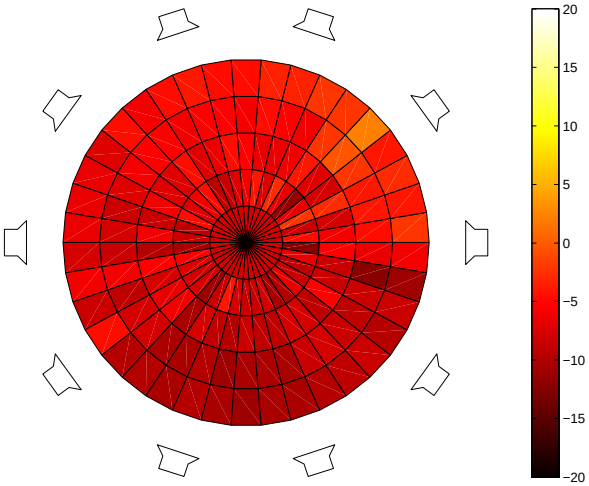
FIG. 2.23 – ITD (μs) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde plane d'azimut $\phi = 60°$. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



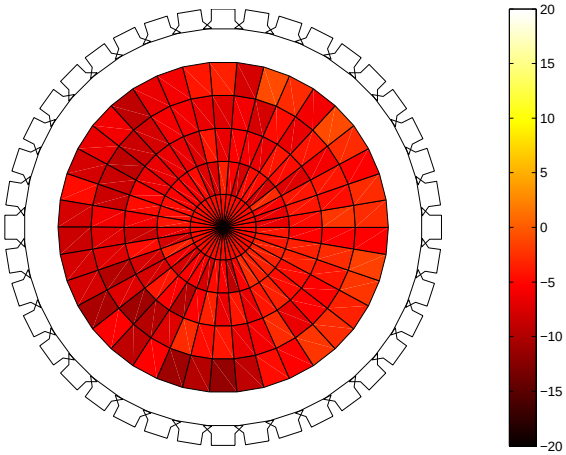
(a) Onde plane cible



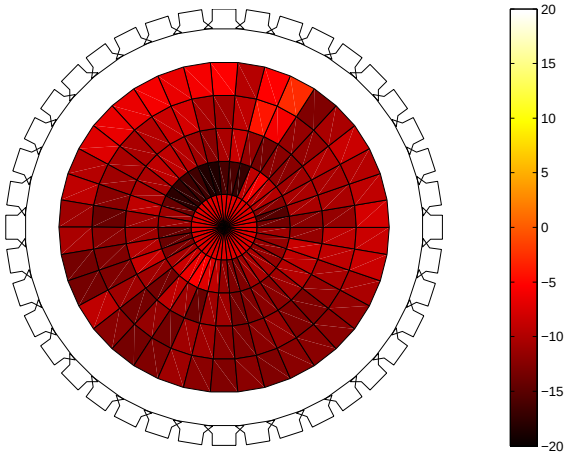
(b) WFS ($N_L = 10$, M = 4)



(c) HOA ($N_L = 10$, M = 4)

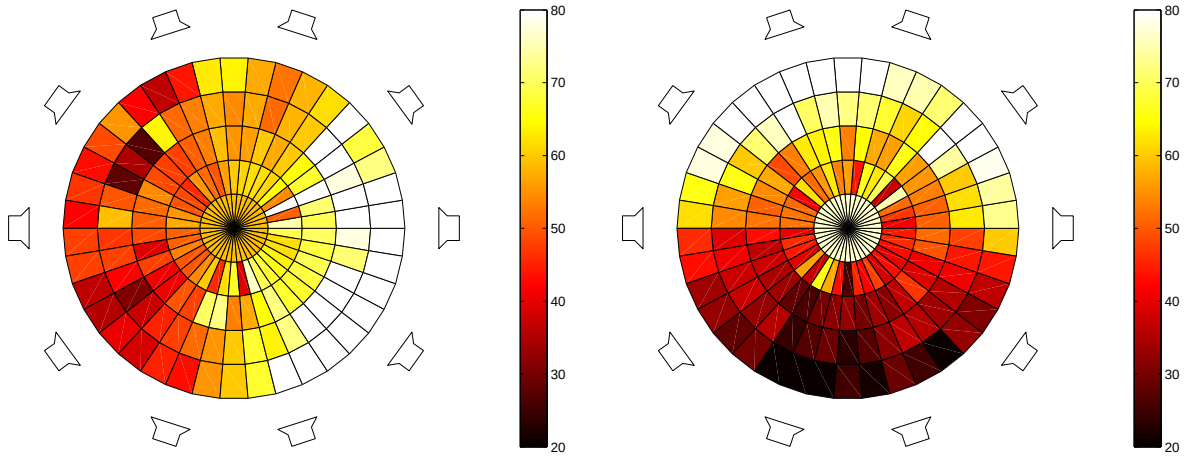


(d) WFS ($N_L = 40$, M = 19)

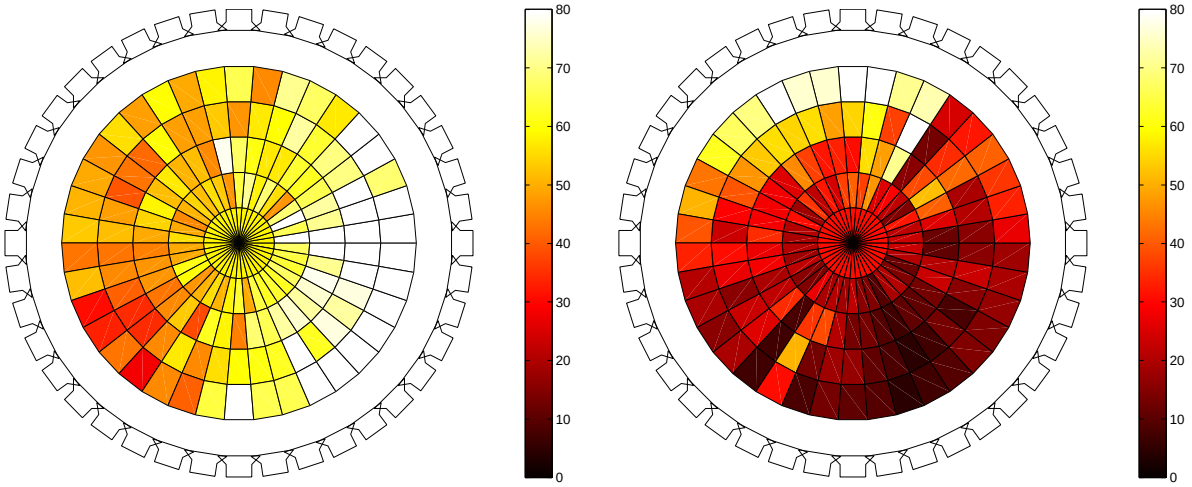


(e) HOA ($N_L = 40$, M = 19)

FIG. 2.24 – ILD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde plane d'azimut $\phi = 60°$. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



(a) WFS ($N_L = 10$, M = 4) (b) HOA ($N_L = 10$, M = 4)



(c) WFS ($N_L = 40$, M = 4) (d) HOA ($N_L = 40$, M = 4)

FIG. 2.25 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde plane d'azimut $\phi = 60°$. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.

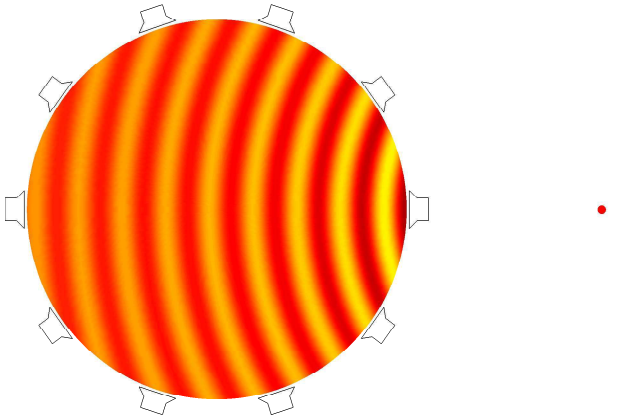
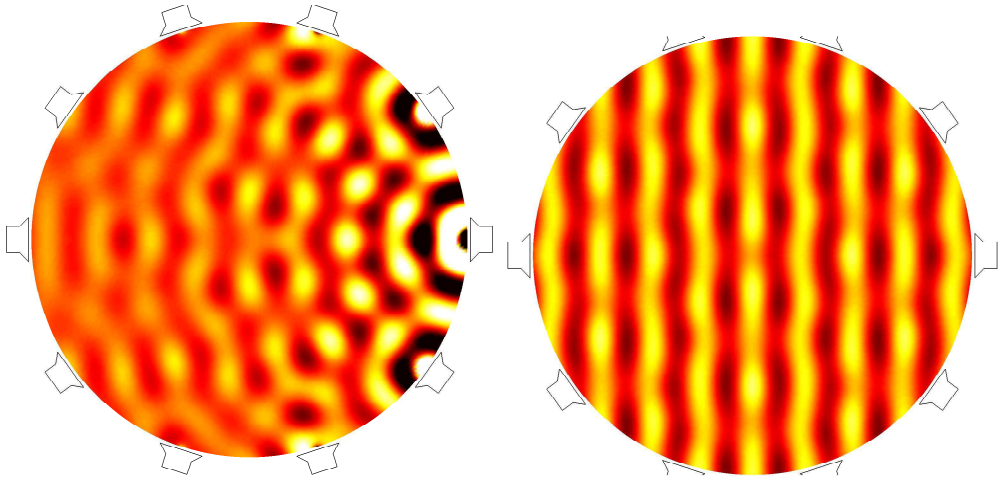


FIG. 2.26 – Onde sphérique cible à synthétiser (rayon $r_S = 3$ m., azimut $\phi = 0°$, fréquence : f = 1 kHz). Le point rouge repère la position de la source virtuelle.

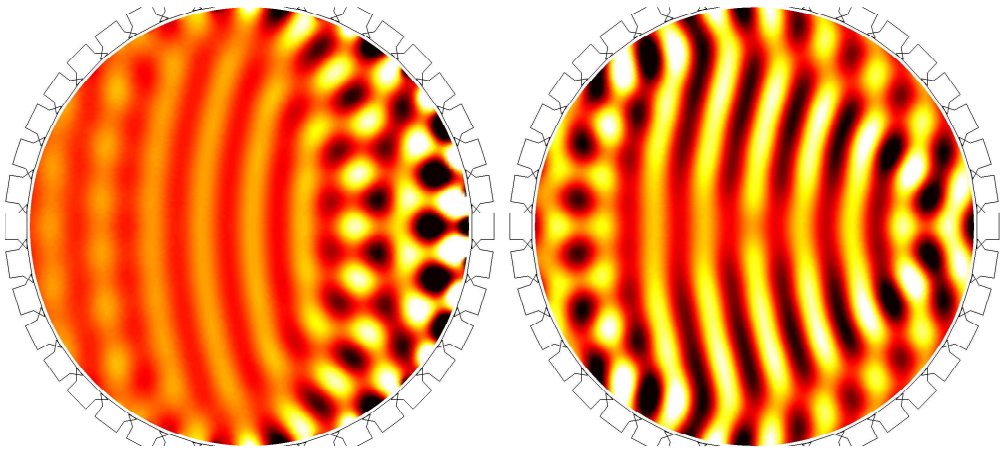
d'autre, ce qui traduit l'effort de focalisation pour la synthèse d'une source intérieure. La fonction de panning WFS présente d'ailleurs aussi deux remontées de part et d'autre de son pic à O°. Les indices de localisation sont décrits sur les Figures 2.35, 2.36 & 2.37. Pour WFS et HOA, l'ITD est assez bien restitué avec seulement 10 haut-parleurs, ce qui n'était pas le cas pour la source extérieure. Cependant, lorsqu'on passe à un réseau de 40 haut-parleurs, on observe la même évolution que dans le cas extérieur : l'ITD s'améliore pour WFS, mais se dégrade pour HOA. De façon un peu surprenante, l'ILD restitué par WFS est erronée puisqu'il reste proche de 0 dB sur l'ensemble de la zone d'écoute quel que soit le nombre de haut-parleurs. En revanche l'ILD obtenue avec HOA est très proche des valeurs attendues sur toute la zone d'écoute avec seulement 10 haut-parleurs. Lorsque N_L s'élève à 40, des instabilités localisées apparaissent. Concernant les distorsions spectrales, on est frappé par leur aggravation pour WFS en comparaison de la source extérieure, alors que HOA garde des valeurs similaires d'ISSD.

2.3.8 Synthèse HOA par des ondes sphériques (HOA OS)

Dans sa définition traditionnelle, HOA implique une reconstruction par des ondes planes. Cependant il est possible de substituer aux ondes planes des ondes sphériques, ce qui permet de se rapprocher des caractéristiques réelles de haut-parleur. La matrice de décodage doit être modifiée en conséquence pour prendre en compte des termes additionnels associés aux fonction de Hankel sphériques (cf. Equ. 2.44). Théoriquement la matrice de décodage est censée adapter les signaux des haut-parleurs aux spécificités du réseau de décodage et par la même rendre aussi transparente que possible l'opération de décodage. La Figure 2.38 illustre les ondes synthétisées en considérant une reconstruction par ondes sphériques avec un nombre croissant de haut-parleurs. Lorsque N_L vaut 10 ou 20, la forme de l'onde synthétique est sphérique, alors que l'onde à reproduire est plane, de la même façon que dans le cas d'une reconstruction par ondes planes, pour la synthèse d'une onde sphérique, le front d'onde reste plan si le nombre de haut-parleurs est faible (cf. Fig. 2.27 & 2.33). Il faut 40 haut-parleurs pour que l'onde synthétisée prenne la forme attendue. Cependant, pour un réseau de 80 haut-parleurs associé à un encodage à l'ordre M = 39, on observe que l'amplitude de l'onde est très atténuée même si sa forme spatiale est juste. L'examen de la matrice de décodage (cf. Fig. 2.39) indique un filtrage passe-haut (lié aux pondérations par la fonction de Hankel sphérique) privilégiant les composantes des ordres supérieures au détriment des premiers

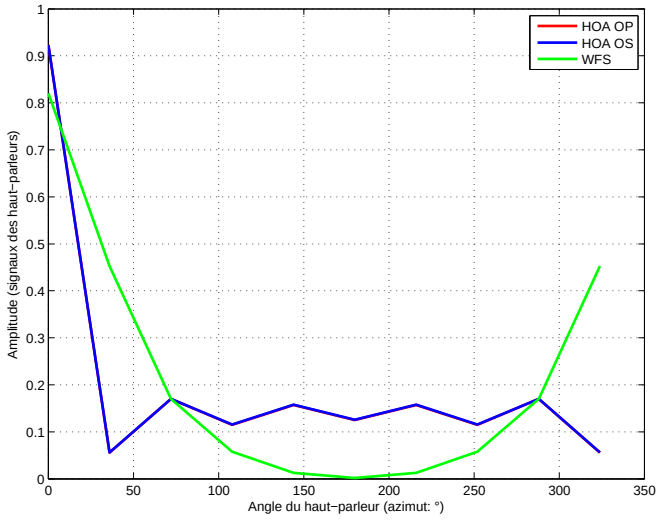


(a) WFS ($N_L = 10$) (b) HOA ($N_L = 10$, M = 4)

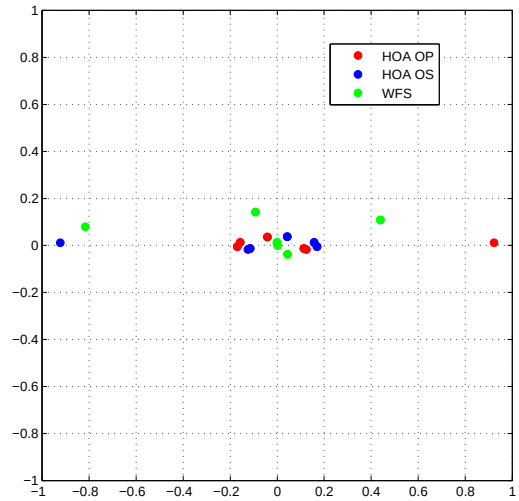


(c) WFS ($N_L = 40$) (d) HOA ($N_L = 40$, M = 19)

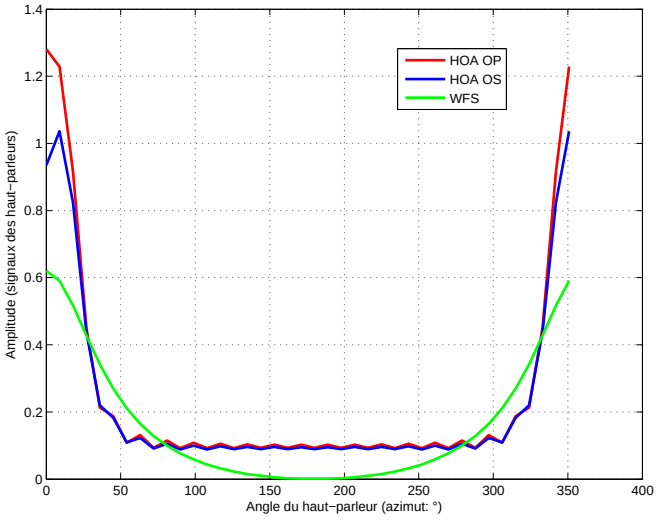
FIG. 2.27 – Illustration des ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 3$ m. (fréquence : f = 1 kHz).



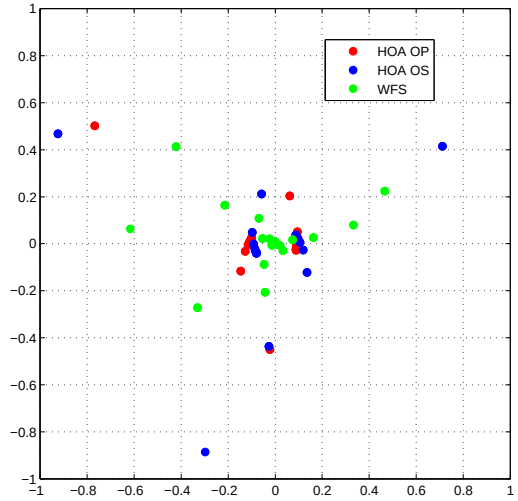
(a) Amplitude ($N_L = 10$, M = 4)



(b) Phase ($N_L = 10$, M = 4)

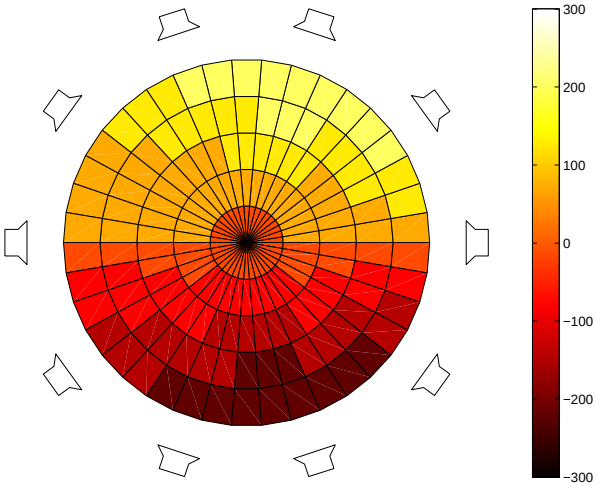


(c) Amplitude ($N_L = 40$, M = 19)

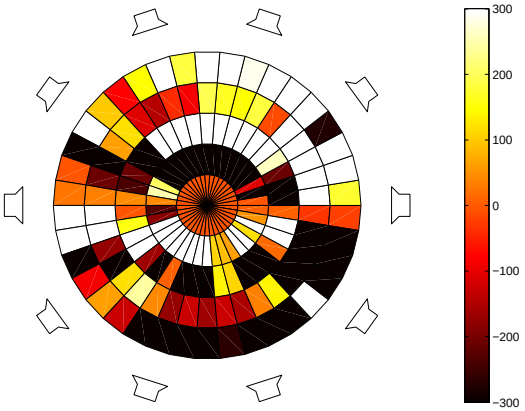


(d) Phase ($N_L = 40$, M = 19)

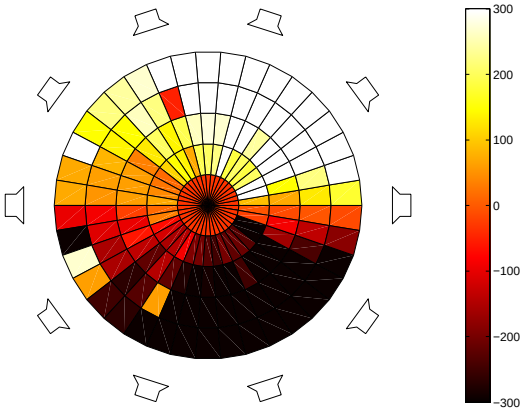
FIG. 2.28 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 3$ m. (fréquence : f = 1 kHz).



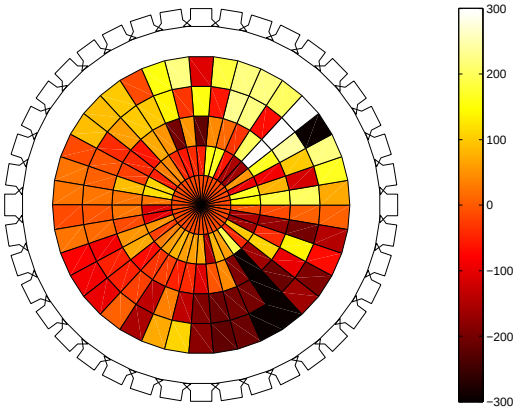
(a) Onde sphérique cible



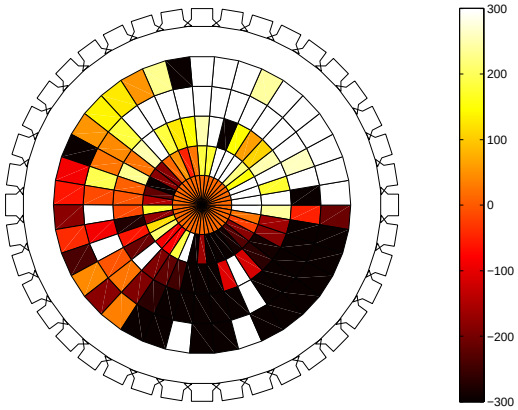
(b) WFS ($N_L = 10$, M = 4)



(c) HOA ($N_L = 10$, M = 4)

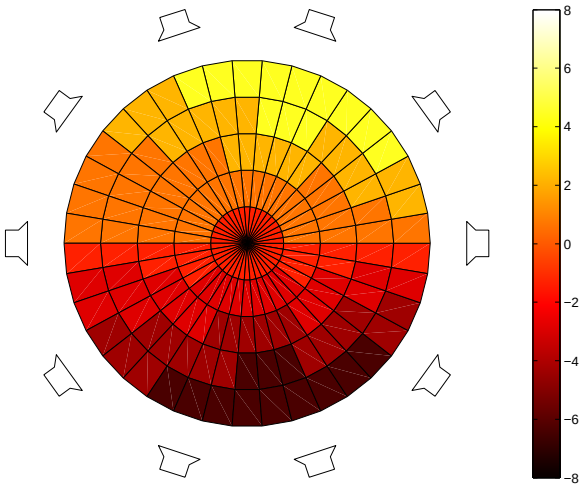


(d) WFS ($N_L = 40$, M = 19)

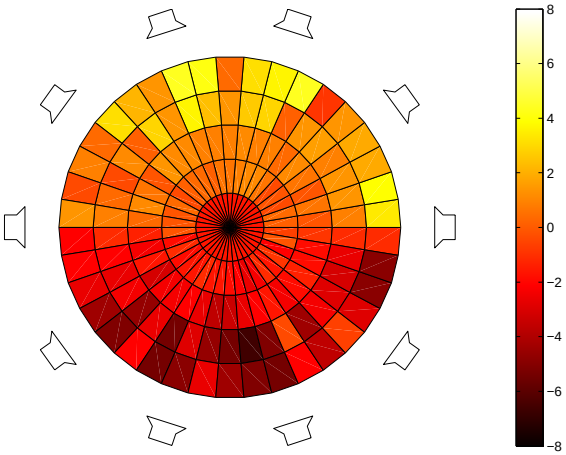


(e) HOA ($N_L = 40$, M = 19)

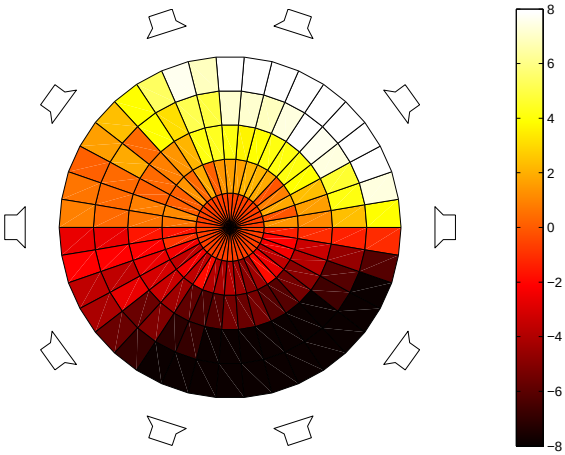
FIG. 2.29 – ITD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 3$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



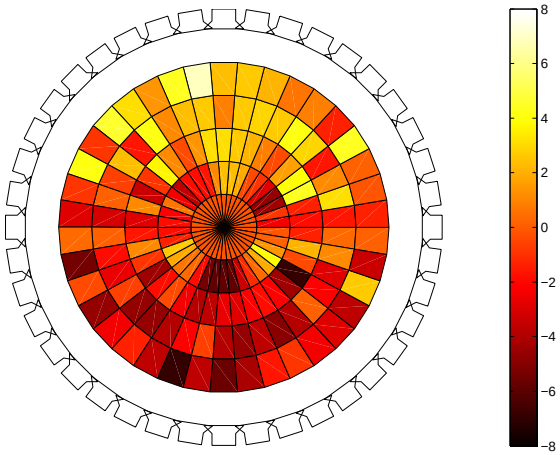
(a) Onde sphérique cible



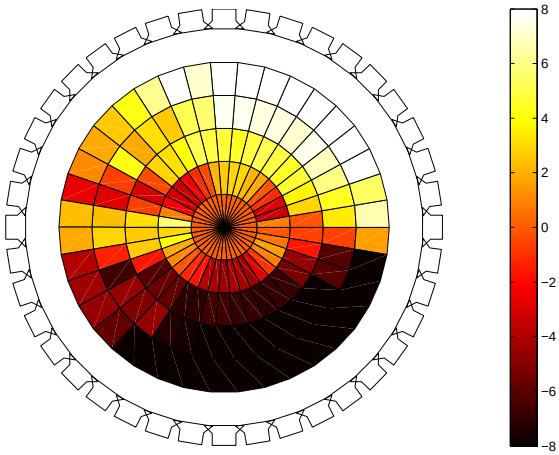
(b) WFS ($N_L = 10$, M = 4)



(c) HOA ($N_L = 10$, M = 4)

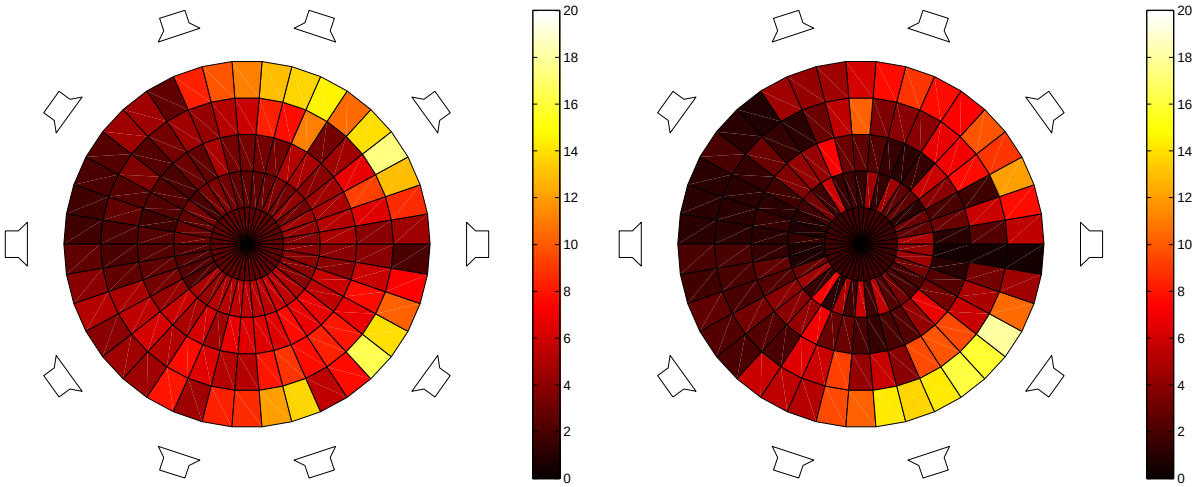


(d) WFS ($N_L = 40$, M = 4)



(e) HOA ($N_L = 40$, M = 19)

FIG. 2.30 – ILD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 3$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



(a) WFS ($N_L = 10$, M = 4)

(b) HOA ($N_L = 10$, M = 4)

(c) WFS ($N_L = 40$, M = 19)

(d) HOA ($N_L = 40$, M = 19)

FIG. 2.31 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 3$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.

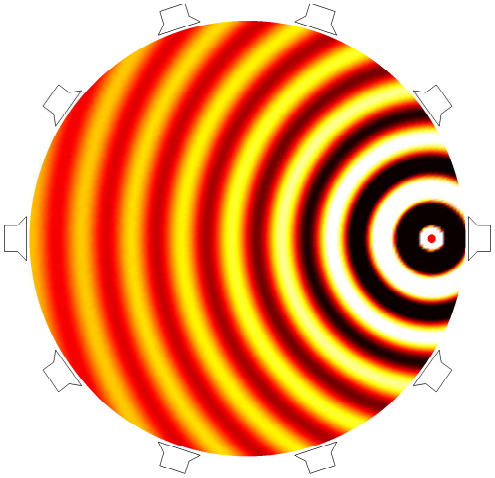


FIG. 2.32 – Onde sphérique cible à synthétiser (rayon $r_S = 1.25$ m., azimut $\phi = 0°$, fréquence : f = 1 kHz). Le point rouge repère la position de la source virtuelle (point de convergence des ondes).

ordres. C'est un artefact de la synthèse par ondes sphériques lorsqu'on augmente l'ordre d'encodage M. L'effet est d'autant plus prononcé aux basses fréquences (k petit). La fonction de panning HOA OS dénote un comportement particulier (cf. Fig. 2.6 & 2.7) : contrairement à la synthèse HOA OP, elle ne présente plus aucune sélectivité. Le décodage s'apparentant à une Transformée de Fourier Circulaire Discrète inverse [Daniel, 2009], on observe ici la conséquence du repliement de la fonction de panning. Dans le cas d'une synthèse HOA OP, le support angulaire de cette fonction est limité, ce qui évite le problème de repliement, alors que pour une synthèse HOA OS le support n'est plus limité (cf. propriétés de diffusion angulaire des ondes sphériques).

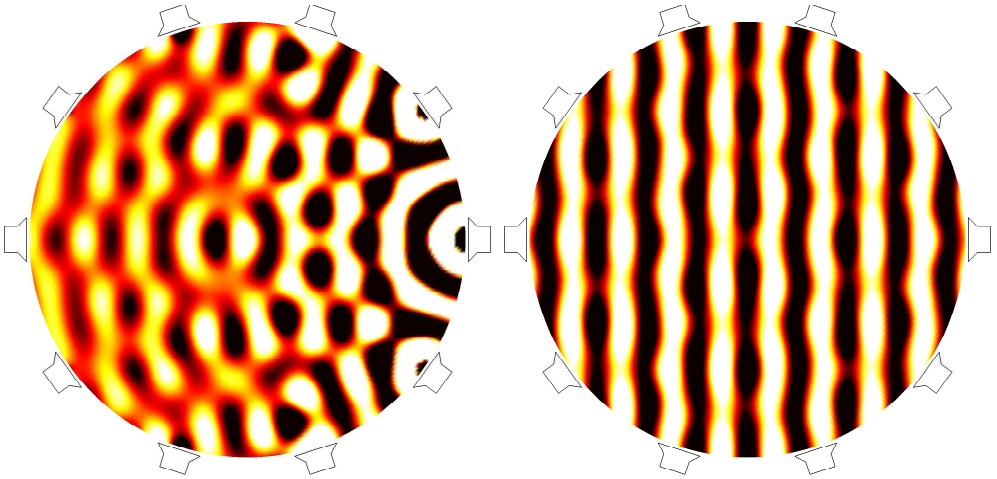
Les indices de localisation sont illustrés sur la Figure 2.40. On observe que l'ITD est beaucoup plus instable que dans le cas de la synthèse HOA OP. Cette instabilité empire lorsque le nombre de haut-parleurs augmente. En revanche l'ILD est très similaire au cas de synthèse HOA OP, de même que les distorsions spectrales (cf. Fig. 2.41) ont un niveau très proche, hormis pour le réseau de 80 haut-parleurs où les détimbrages s'aggravent sensiblement.

2.3.9 Impact des rotations de la tête de l'auditeur sur les indices de localisation

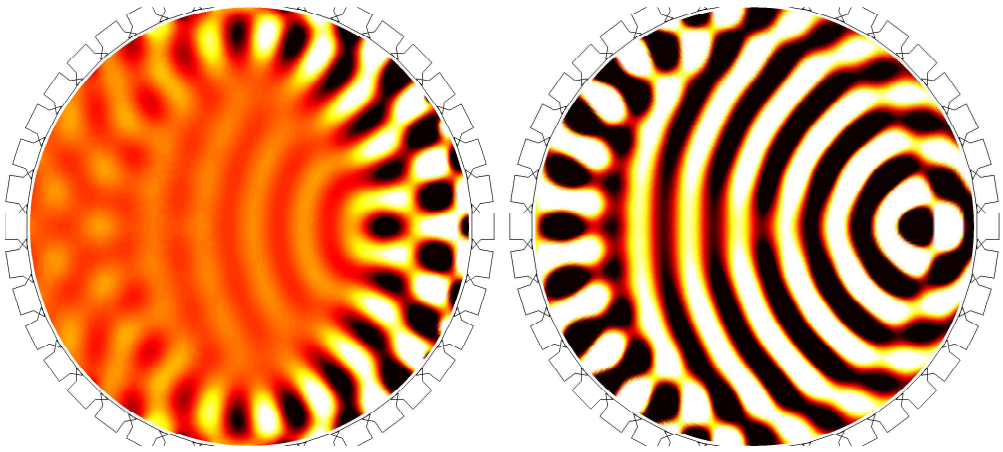
Les Figures 2.42 à 2.55 représentent l'évolution de l'ITD, l'ILD et l'ISSD en fonction de l'orientation de la tête de l'auditeur. Quatre orientations : $\vec{v}(1,0,0)$, $\vec{v}(0,1,0)$, $\vec{v}(-1,0,0)$, $\vec{v}(0,-1,0)$, ont été considérées. Pour l'ITD, on observe que, bien que l'indice soit plus ou moins correctement restitué sur tout ou partie de la zone d'écoute, il tend à évoluer de façon cohérente avec une situation naturelle d'écoute (cf. Fig. 2.42). La qualité de spatialisation des rendus WFS et HOA ne peut que bénéficier de l'apport de cet indice dynamique. Malgré tout on note certains échecs de la synthèse :

- pour WFS lorsque le nombre de haut-parleurs est insuffisant et que l'onde synthétique est trop dégradée par le repliement spatial,
- pour HOA lorsque l'onde plane se propage perpendiculairement à l'axe interaural.

L'ILD est très bien restituée par WFS. Pour HOA, l'indice est moins bien reproduit en valeur absolue (effet de sous-latéralisation) mais il évolue avec l'orientation de la tête de façon cohérente avec une situation naturelle. On retrouve aussi le problème d'une cartographie aberrante lorsque l'onde se propage perpendiculairement à l'axe interaural. Concernant les distorsions spectrales, on

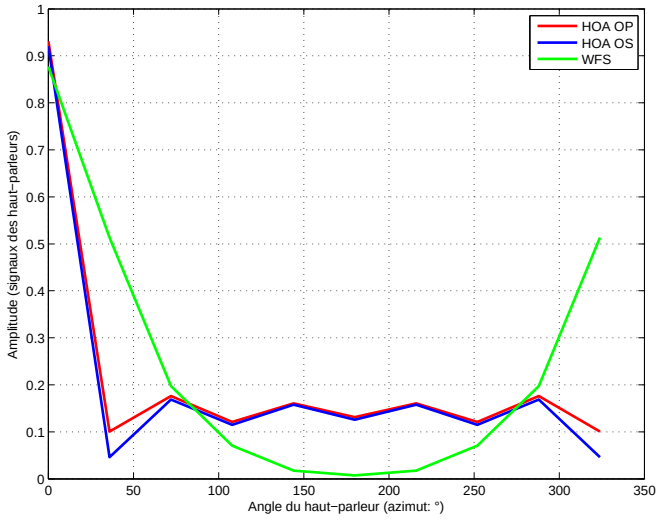


(a) WFS ($N_L = 10$) (b) HOA ($N_L = 10$, M = 4)

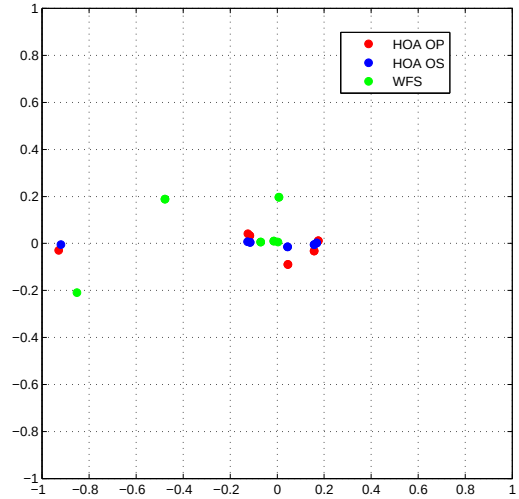


(c) WFS ($N_L = 40$) (d) HOA ($N_L = 40$, M = 19)

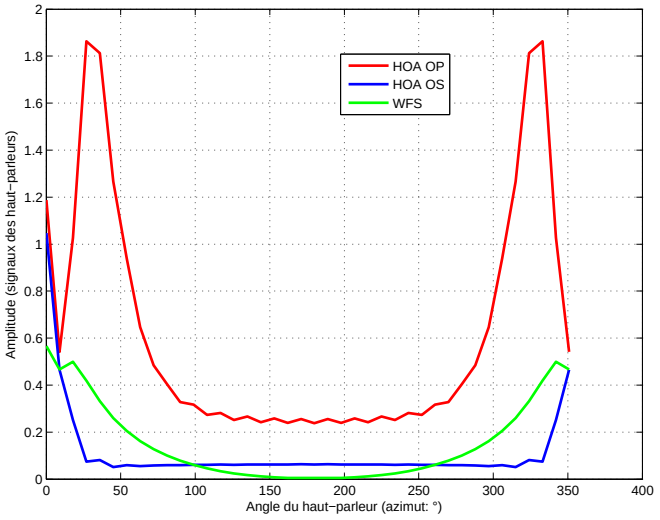
FIG. 2.33 – Illustration des ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 1.25$ m. (fréquence : f = 1 kHz).



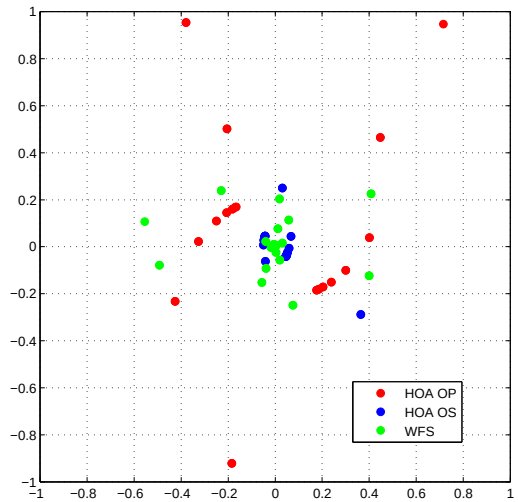
(a) Amplitude ($N_L = 10$, M = 4)



(b) Phase ($N_L = 10$, M = 4)

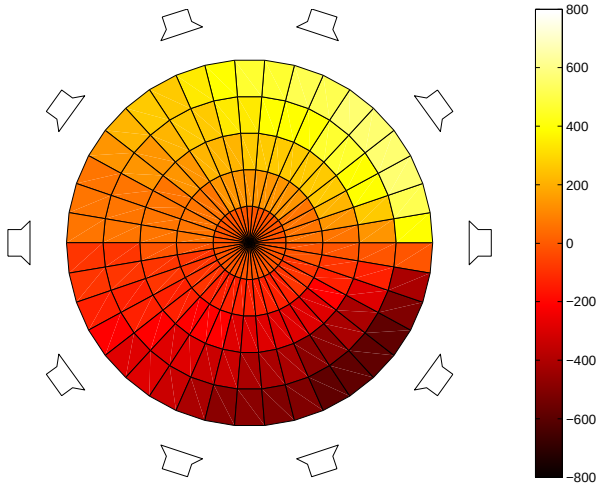


(c) Amplitude ($N_L = 40$, M = 19)

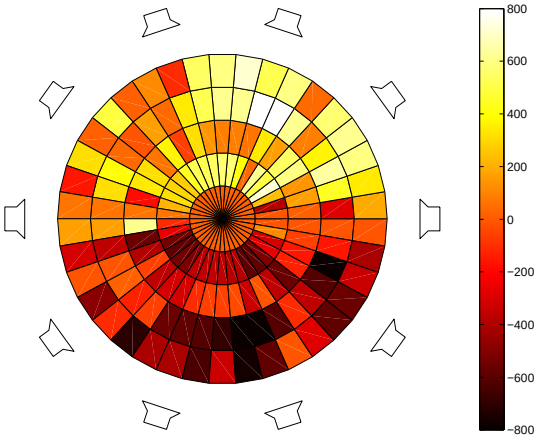


(d) Phase ($N_L = 40$, M = 19)

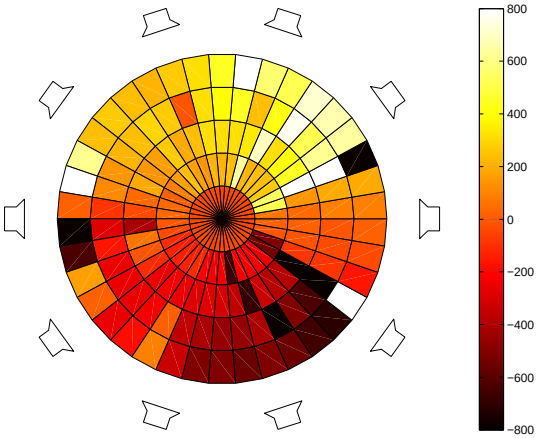
FIG. 2.34 – Amplitude et phase des signaux alimentant les haut-parleurs (s_{WFS} et s_{HOA}) pour synthétiser l'onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 1.25$ m. (fréquence : f = 1 kHz).



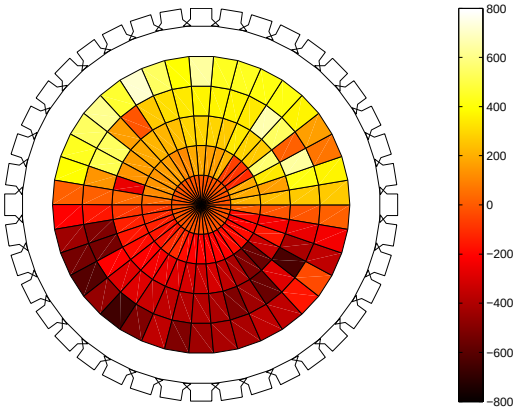
(a) Onde sphérique cible



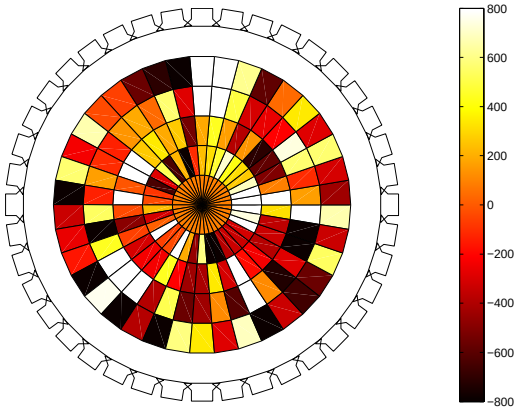
(b) WFS ($N_L = 10$, M = 4)



(c) HOA ($N_L = 10$, M = 4)

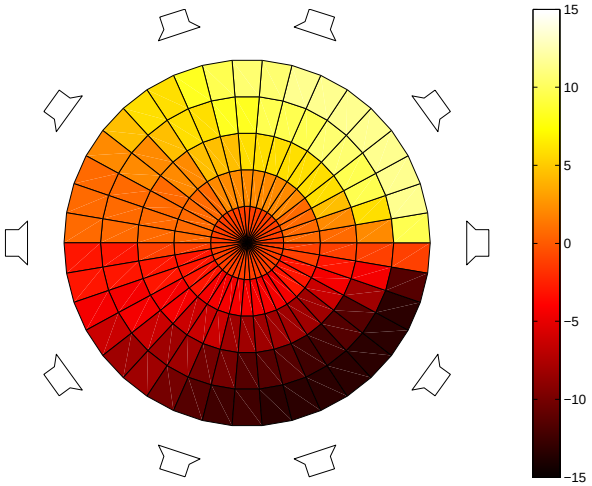


(d) WFS ($N_L = 40$, M = 19)

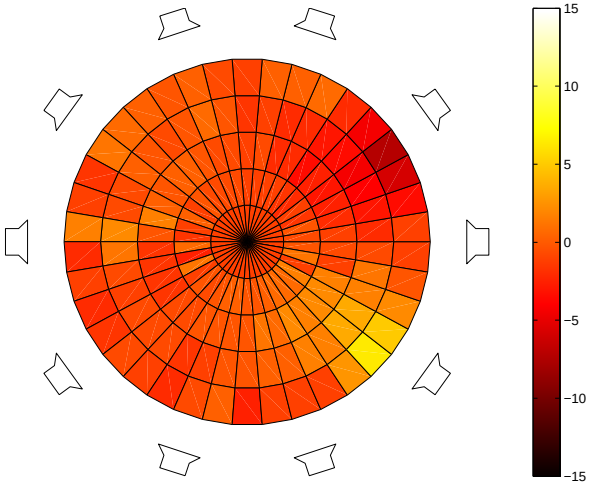


(e) HOA ($N_L = 40$, M = 19)

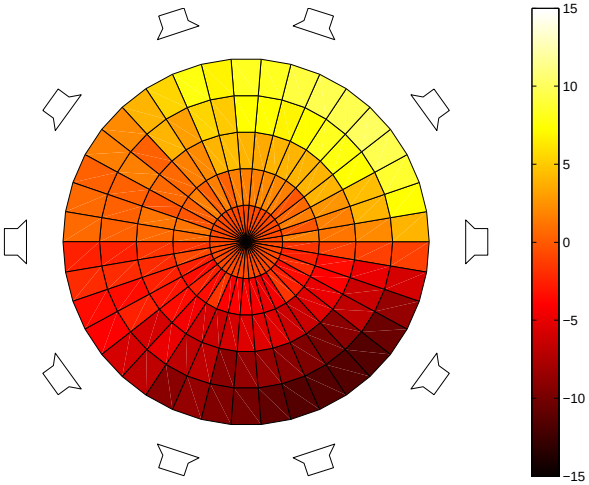
FIG. 2.35 – ITD (μs) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 1.25$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



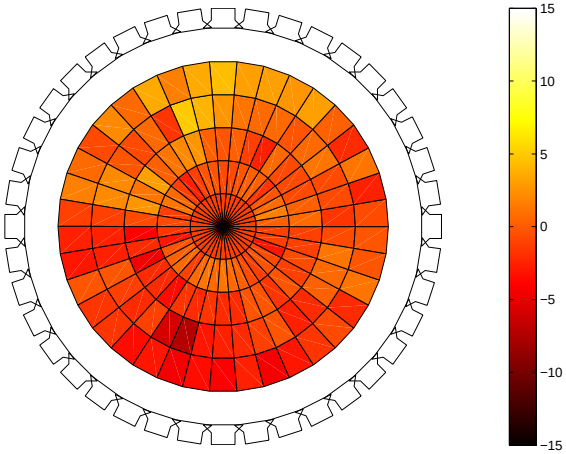
(a) Onde sphérique cible



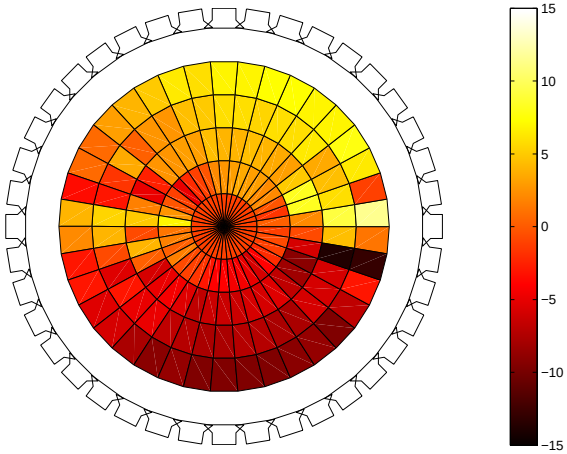
(b) WFS ($N_L = 10$, M = 4)



(c) HOA ($N_L = 10$, M = 4)

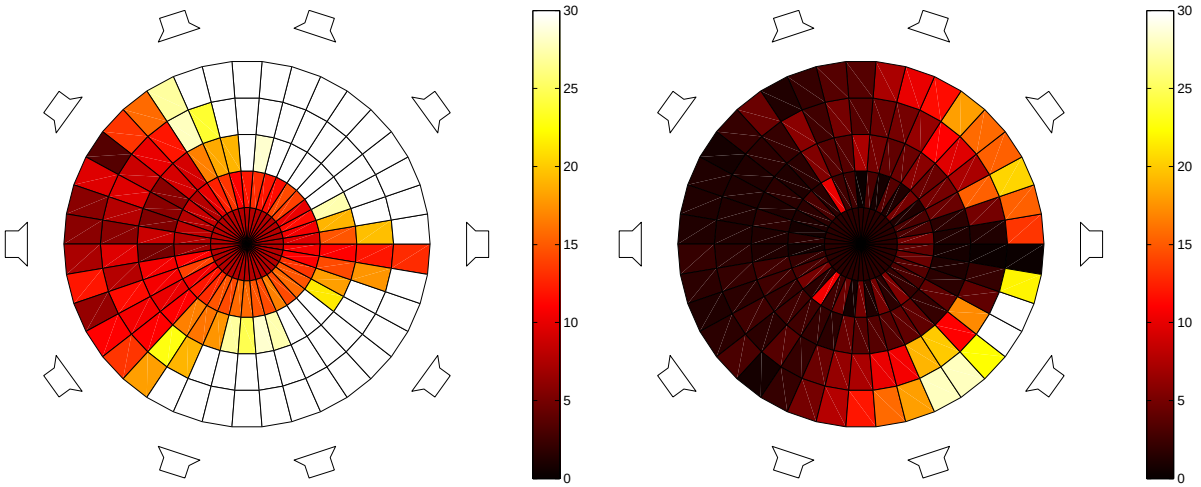


(d) WFS ($N_L = 40$, M = 4)



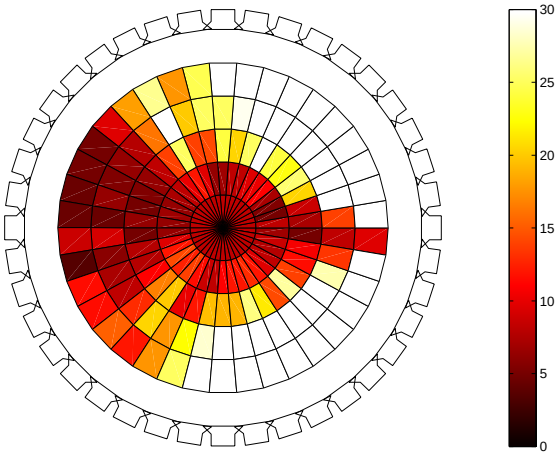
(e) HOA ($N_L = 40$, M = 19)

FIG. 2.36 – ILD (dB) évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 1.25$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.

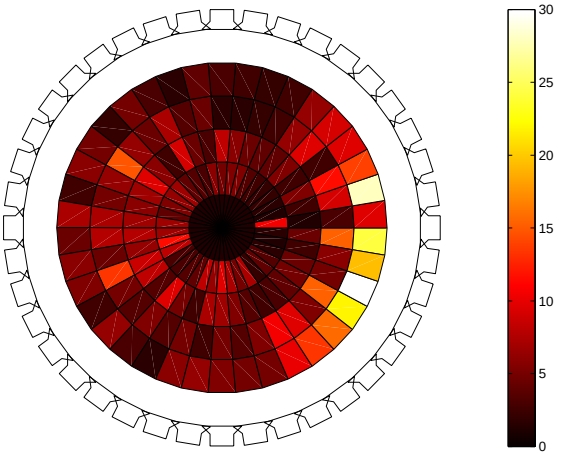


(a) WFS ($N_L = 10$, M = 4)

(b) HOA ($N_L = 10$, M = 4)

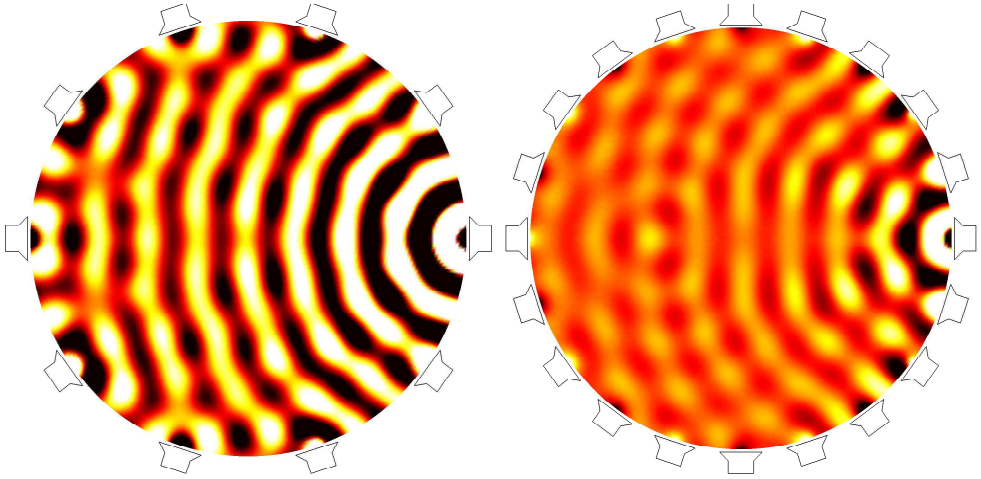


(c) WFS ($N_L = 40$, M = 19)



(d) HOA ($N_L = 40$, M = 19)

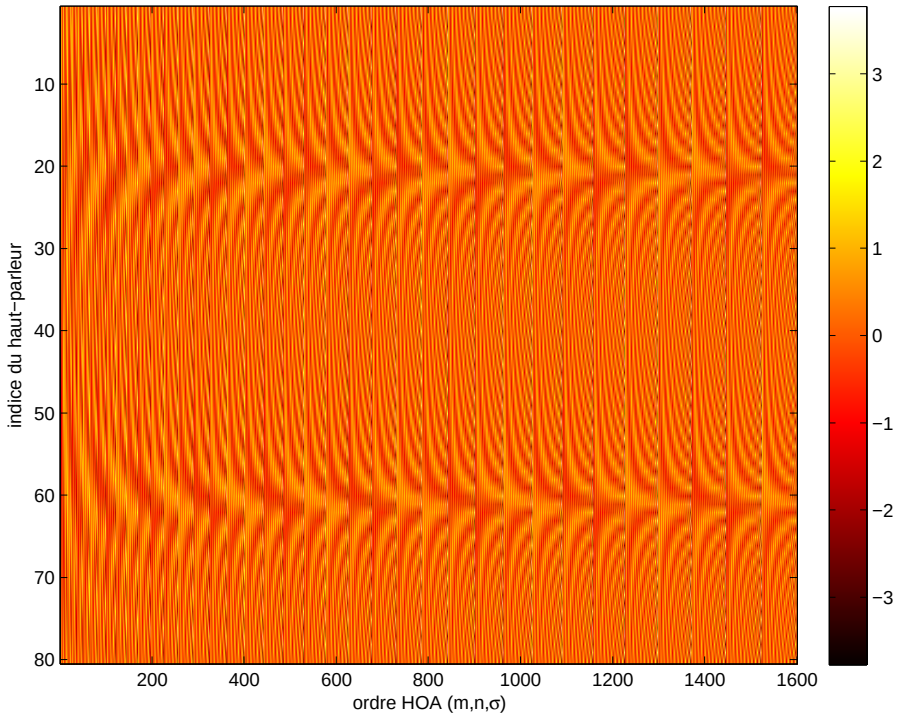
FIG. 2.37 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par les systèmes WFS (OS) et HOA (OP) : Onde sphérique d'azimut $\phi = 0°$ située à une distance $r_S = 1.25$ m. Pour chaque position, la tête de l'auditeur pointe dans la direction $\vec{v}(1, 0, 0)$.



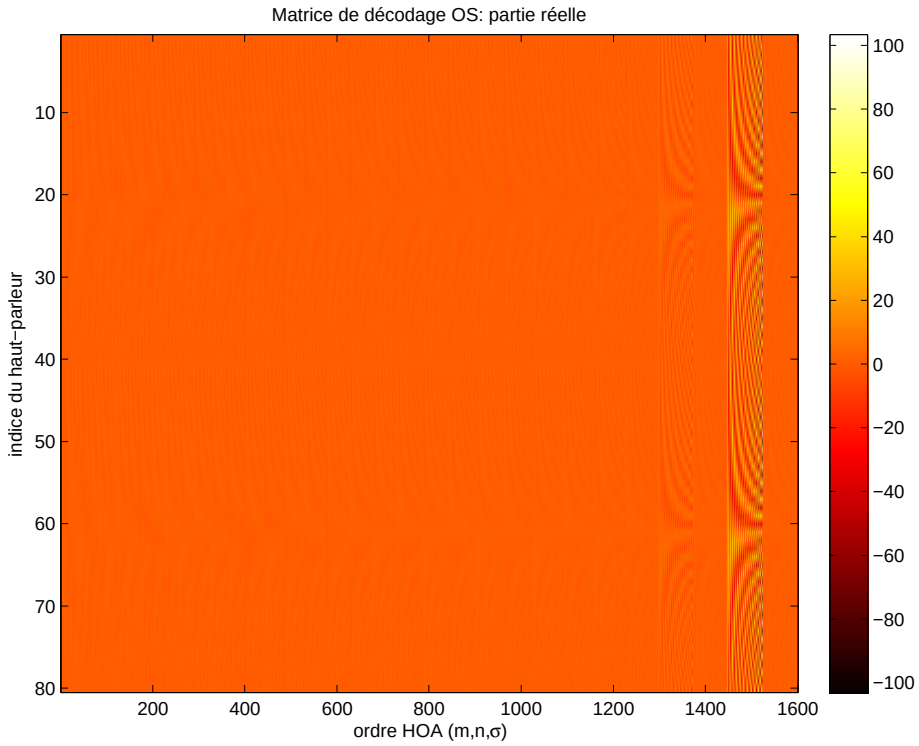
(a) $N_L = 10$, M = 4 (b) $N_L = 20$, M = 9

(c) $N_L = 40$, M = 19 (d) $N_L = 80$, M = 39

FIG. 2.38 – Illustration des ondes synthétisées par le système HOA : Synthèse HOA par ondes sphériques dite synthèse HOA OS (onde plane d'azimut $\phi = 0°$, fréquence : f = 1 kHz).

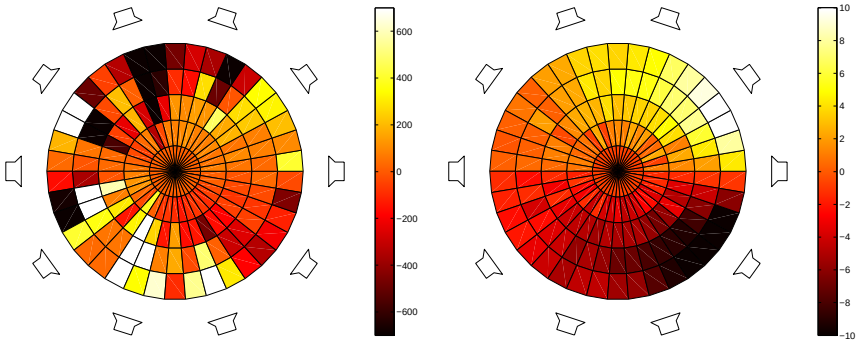


(a) HOA OP



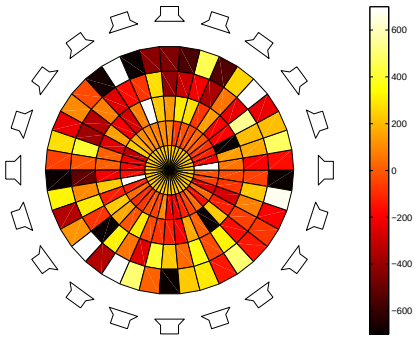
(b) HOA OS

FIG. 2.39 – Matrice de décodage dans le cas d'une synthèse HOA OP et HOA OS (N_L = 80, M = 39) : Mise en évidence du filtrage passe-haut des composantes HOA dans le cas d'une synthèse HOA OS (onde plane d'azimut $\phi = 0°$, fréquence : f = 1 kHz).

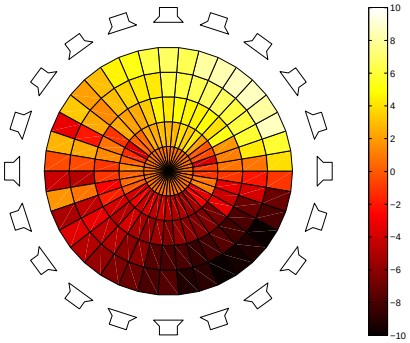


(a) ITD ($N_L = 10$, M = 4)

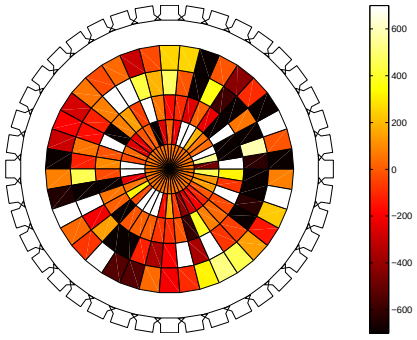
(b) ILD ($N_L = 10$, M = 4)



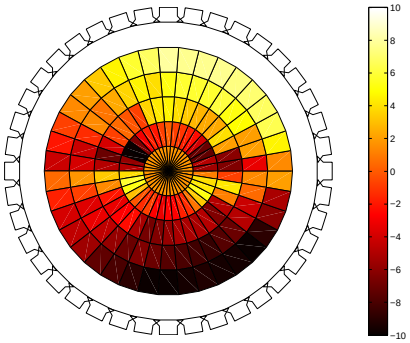
(c) ITD ($N_L = 20$, M = 9)



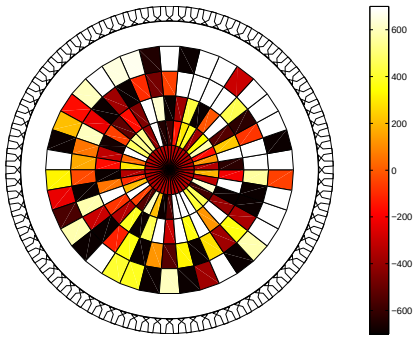
(d) ILD ($N_L = 20$, M = 9)



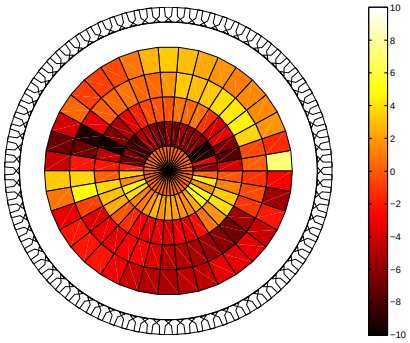
(e) ITD ($N_L = 40$, M = 19)



(f) ILD ($N_L = 40$, M = 19)

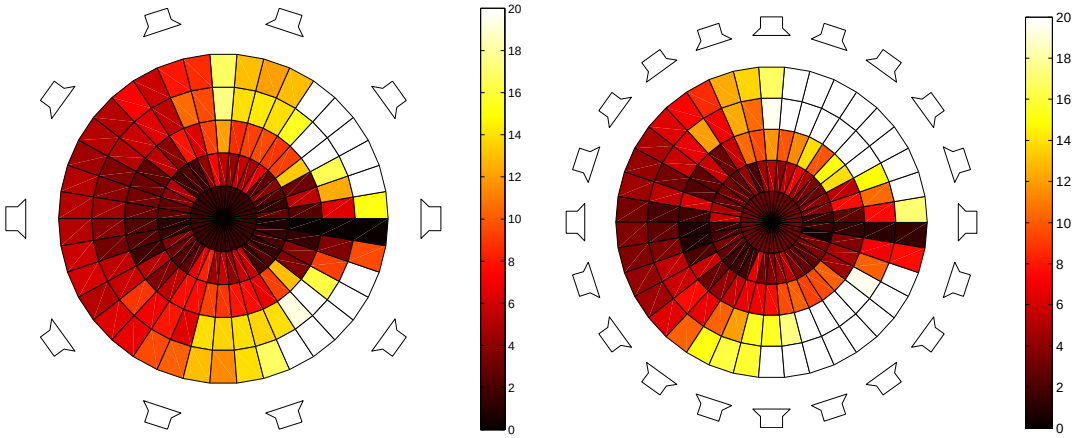


(g) ITD ($N_L = 80$, M = 39)



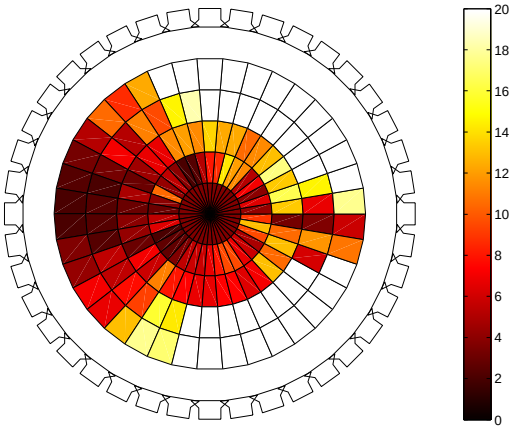
(h) ILD ($N_L = 80$, M = 39)

FIG. 2.40 – ITD (μs) et ILD (dB) évaluées sur la zone d'écoute pour les ondes synthétisées par le système HOA OS (onde plane d'azimut $\phi = 0°$, tête de l'auditeur pointée dans la direction $\vec{v}(1, 0, 0)$).

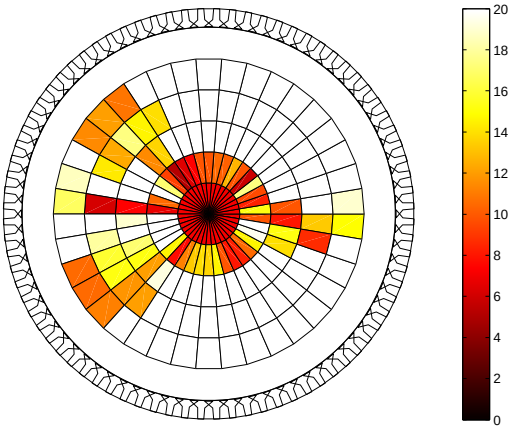


(a) ISSD ($N_L = 10$, M = 4)

(b) ISSD ($N_L = 20$, M = 9)

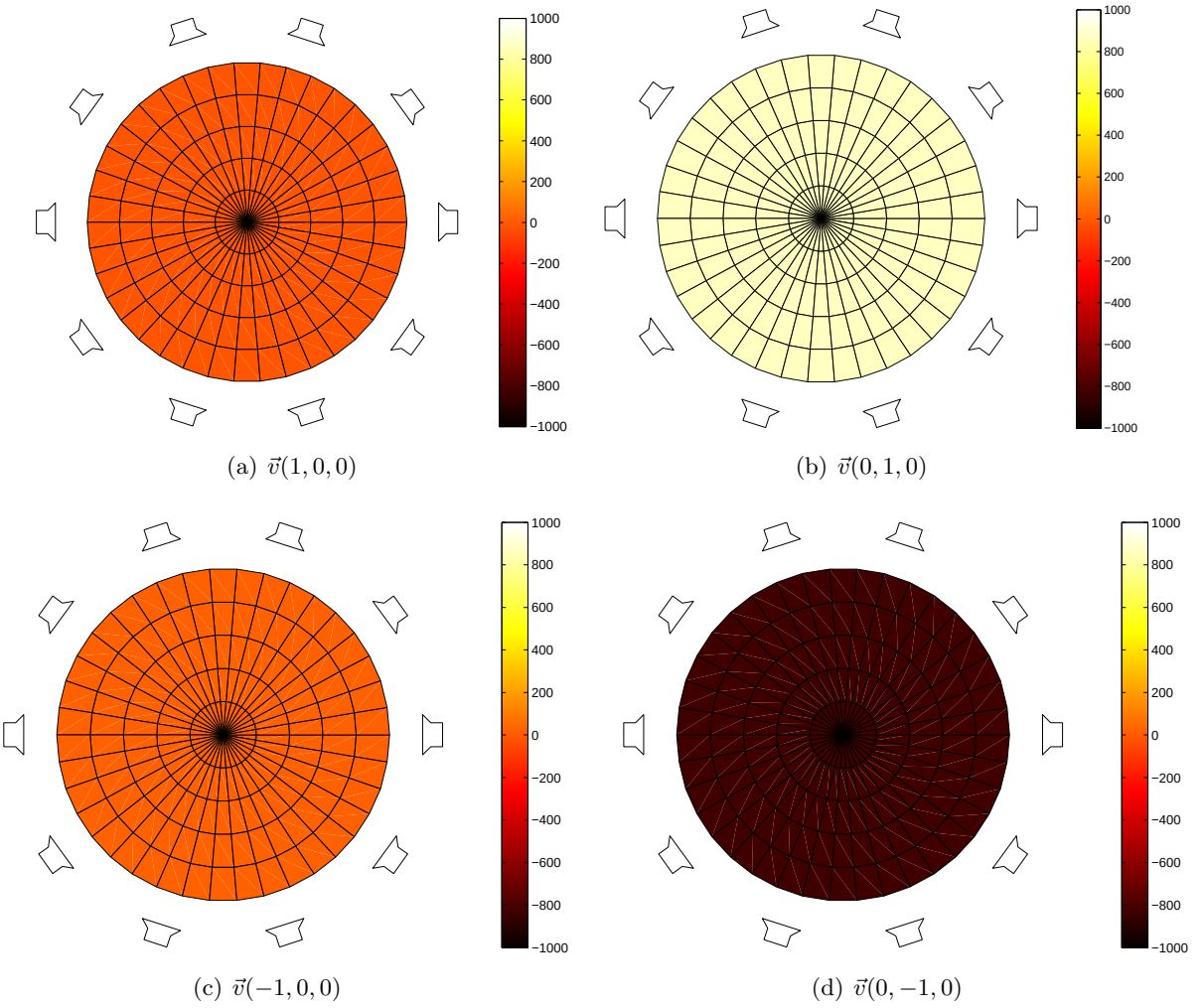


(c) ISSD ($N_L = 40$, M = 19)



(d) ISSD ($N_L = 80$, M = 39)

FIG. 2.41 – ISSD évaluée sur la zone d'écoute pour les ondes synthétisées par le système HOA OS (onde plane d'azimut $\phi = 0°$, tête de l'auditeur pointée dans la direction $\vec{v}(1, 0, 0)$).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

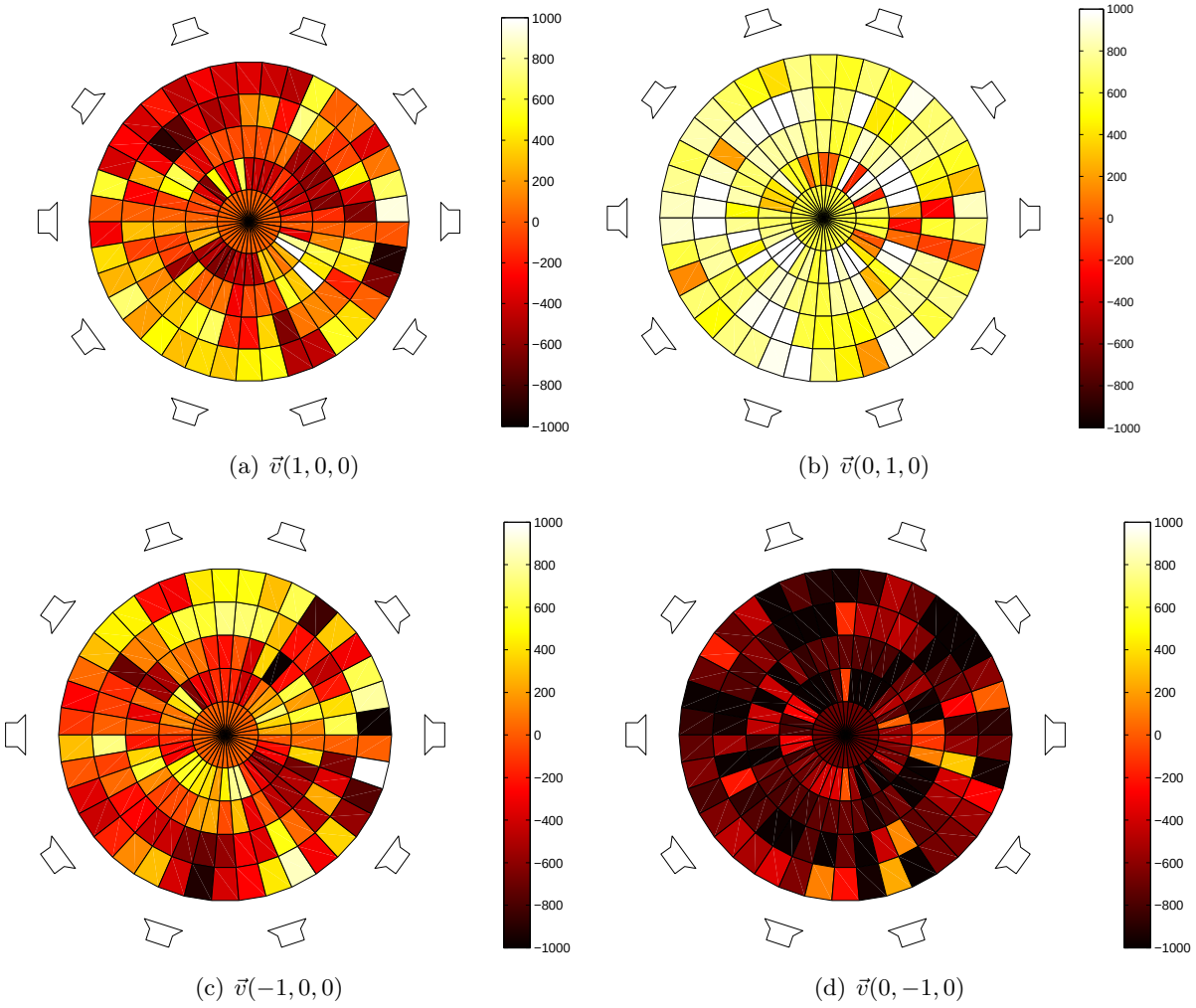
(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

FIG. 2.42 – ITD (μs) évaluée sur la zone d'écoute pour l'onde plane cible en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$).

note qu'elles dépendent fortement de l'orientation de la tête de l'auditeur. Pour les deux systèmes, elles sont les plus faibles lorsque l'onde se propage perpendiculairement à l'axe interaural, mais augmentent dans des proportions dramatiques pour les autres orientations, ce qui constitue un réel handicap pour la qualité du rendu. Ce qu'on observe sur les simulations signifie en effet qu'au moindre mouvement de l'auditeur le timbre des sources est modifié, ce qui, outre l'inconfort auditif, tend à dégrader, voire anéantir, l'illusion de la source virtuelle (affectant notamment les attributs perceptifs de présence et de naturel). La synthèse HOA avec 40 haut-parleurs semble réduire les détimbrages par rapport à $N_L = 10$, en comparaison de WFS.

2.4 Conclusions

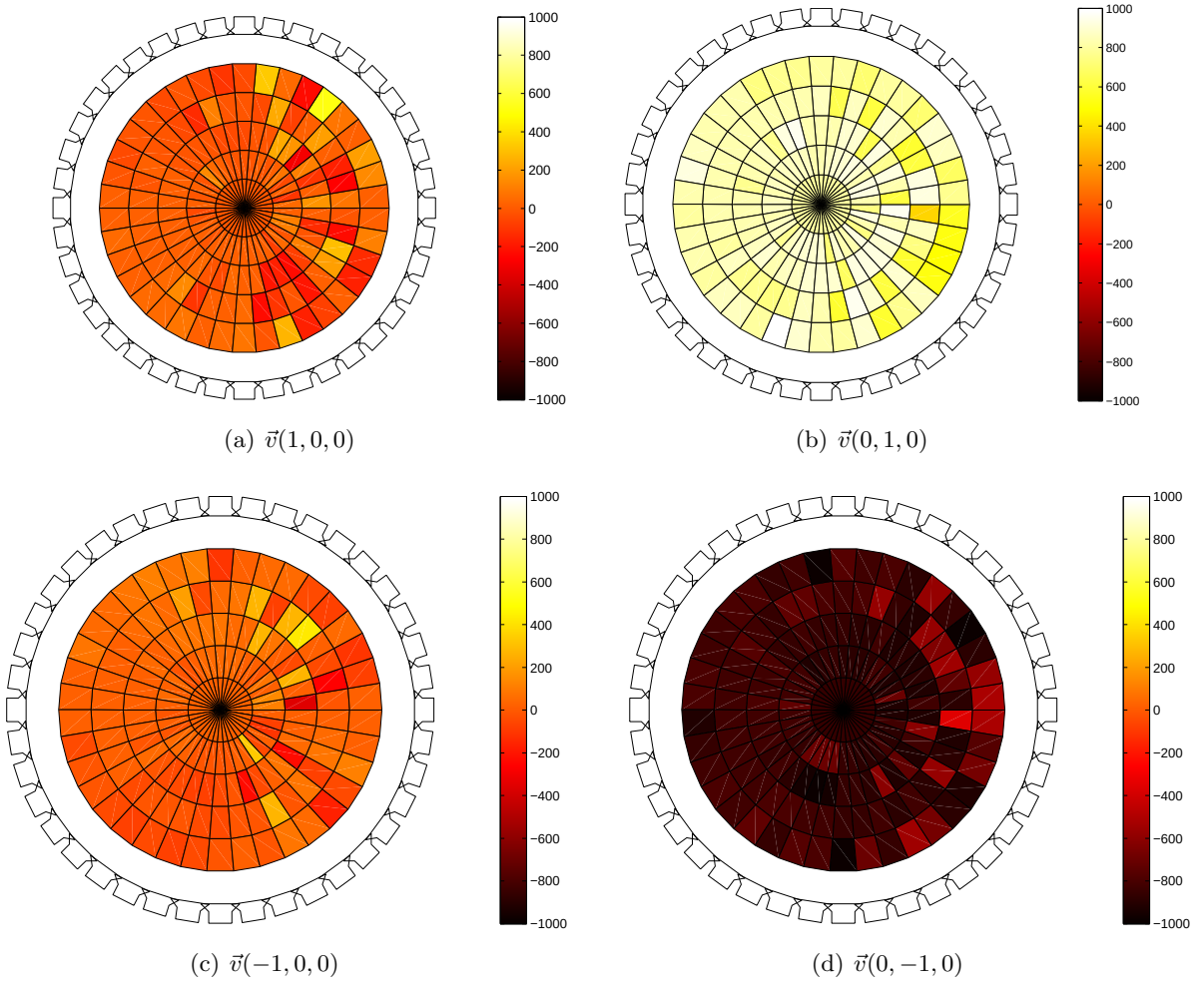
L'étude qui vient d'être réalisée laisse encore de nombreuses questions ouvertes, il est clairement nécessaire de consolider et de compléter les premiers résultats qu'elle dégage. Cependant son premier résultat est de valider la pertinence des outils et de la méthodologie utilisés pour évaluer les technologies WFS et HOA. Comme premiers éléments de réponse, on retiendra que :



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

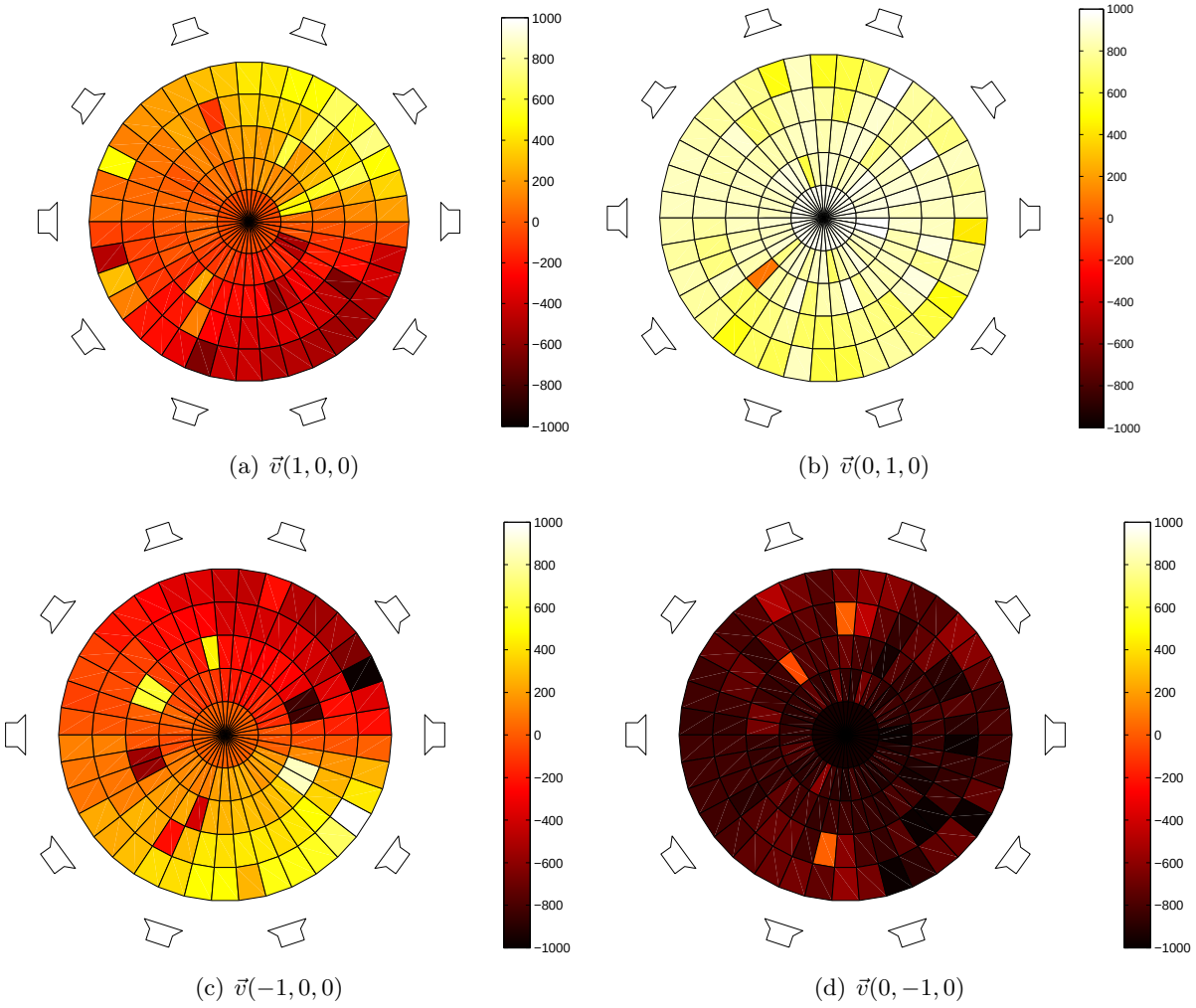
FIG. 2.43 – ITD (μs) évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

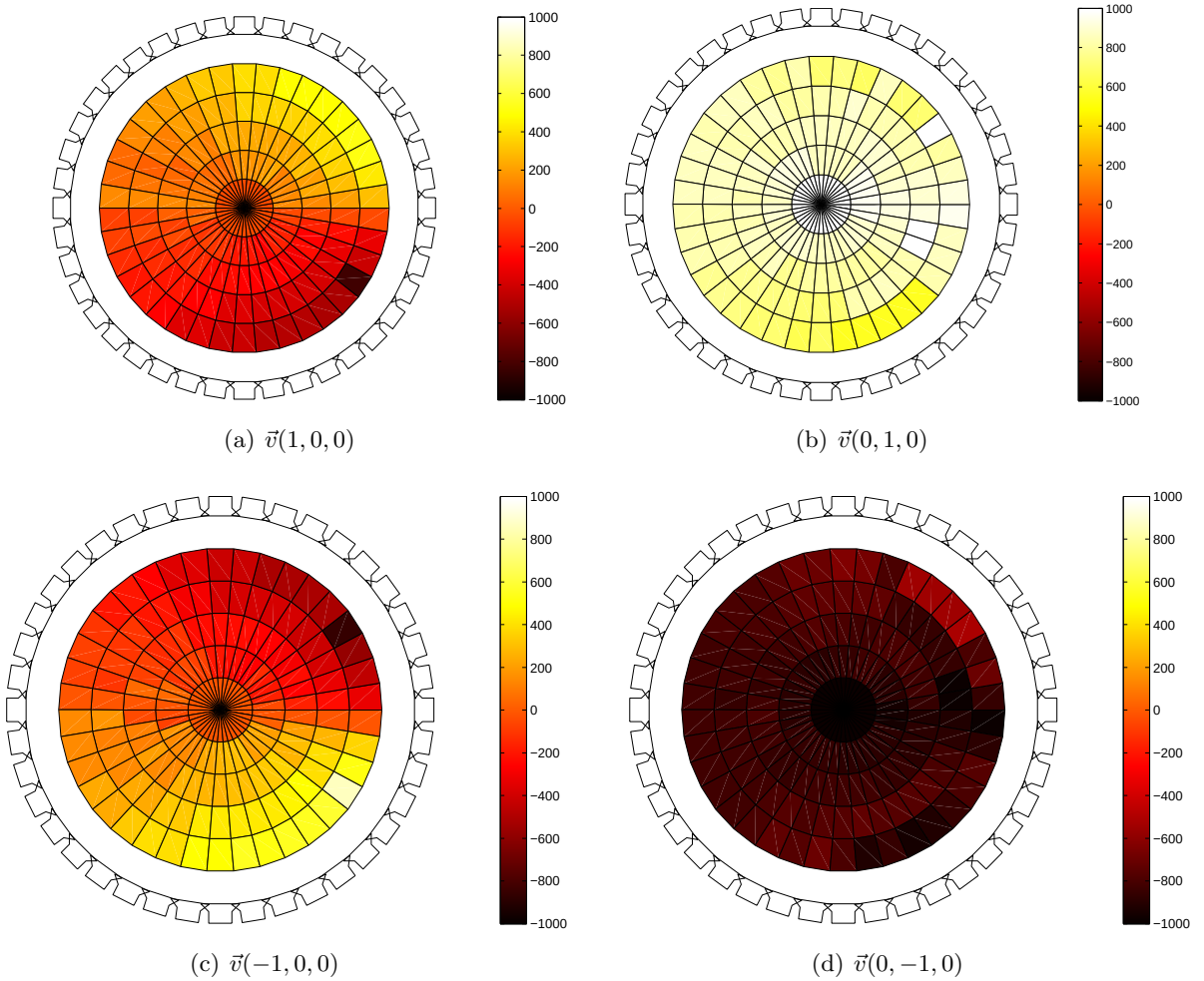
FIG. 2.44 – ITD (μs) évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

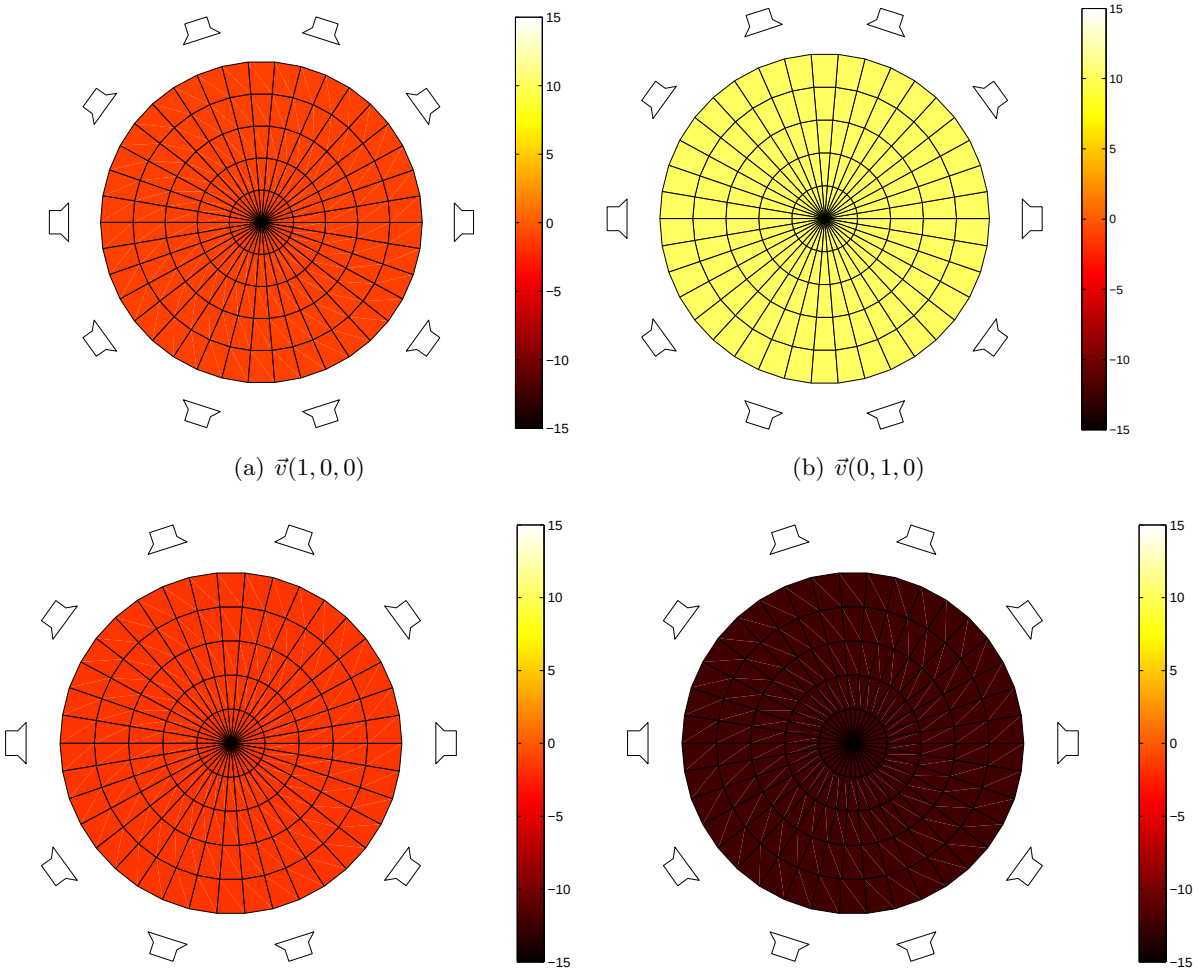
FIG. 2.45 – ITD (μs) évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$, M = 4).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

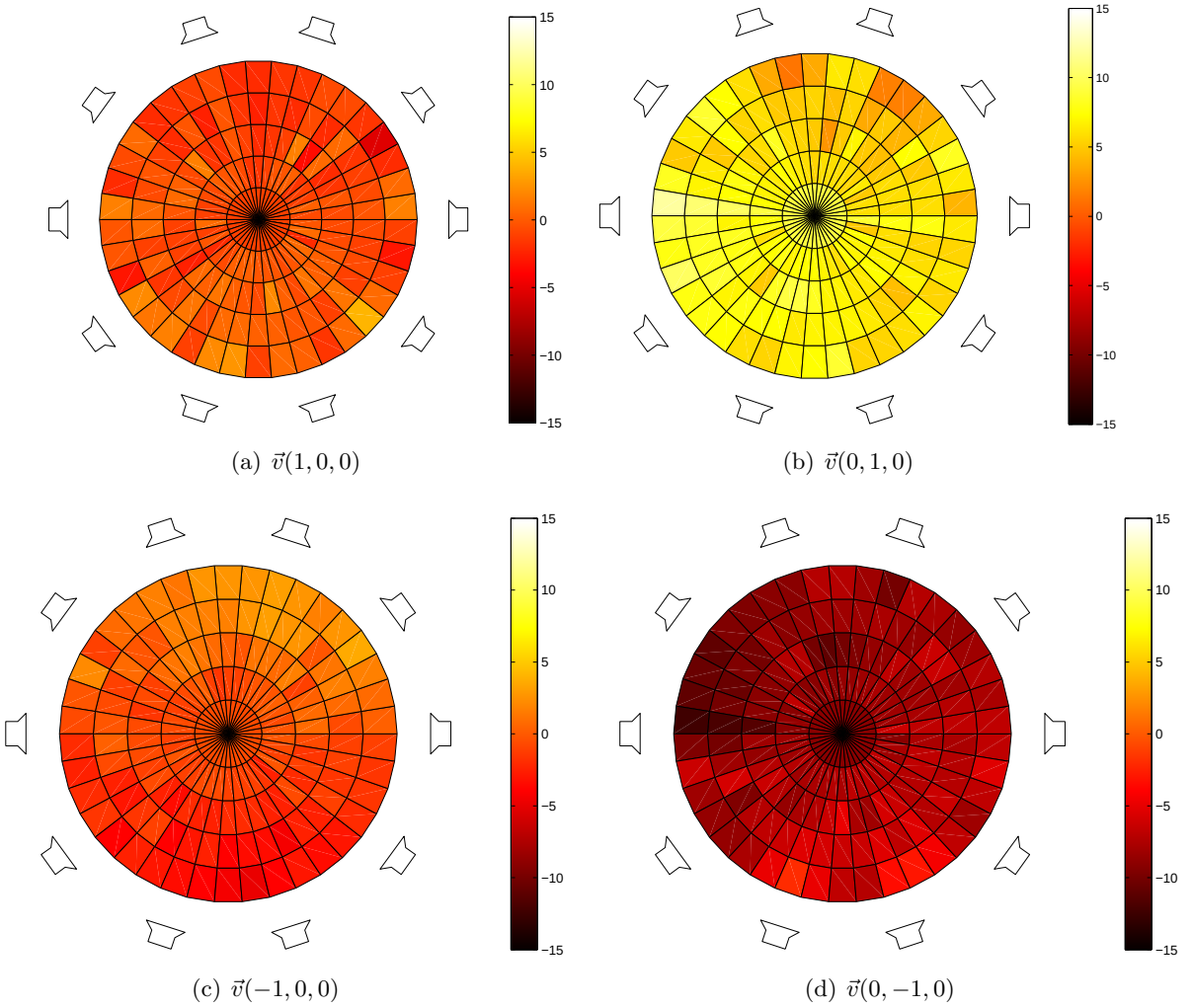
FIG. 2.46 – ITD (μs) évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$, M = 19).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

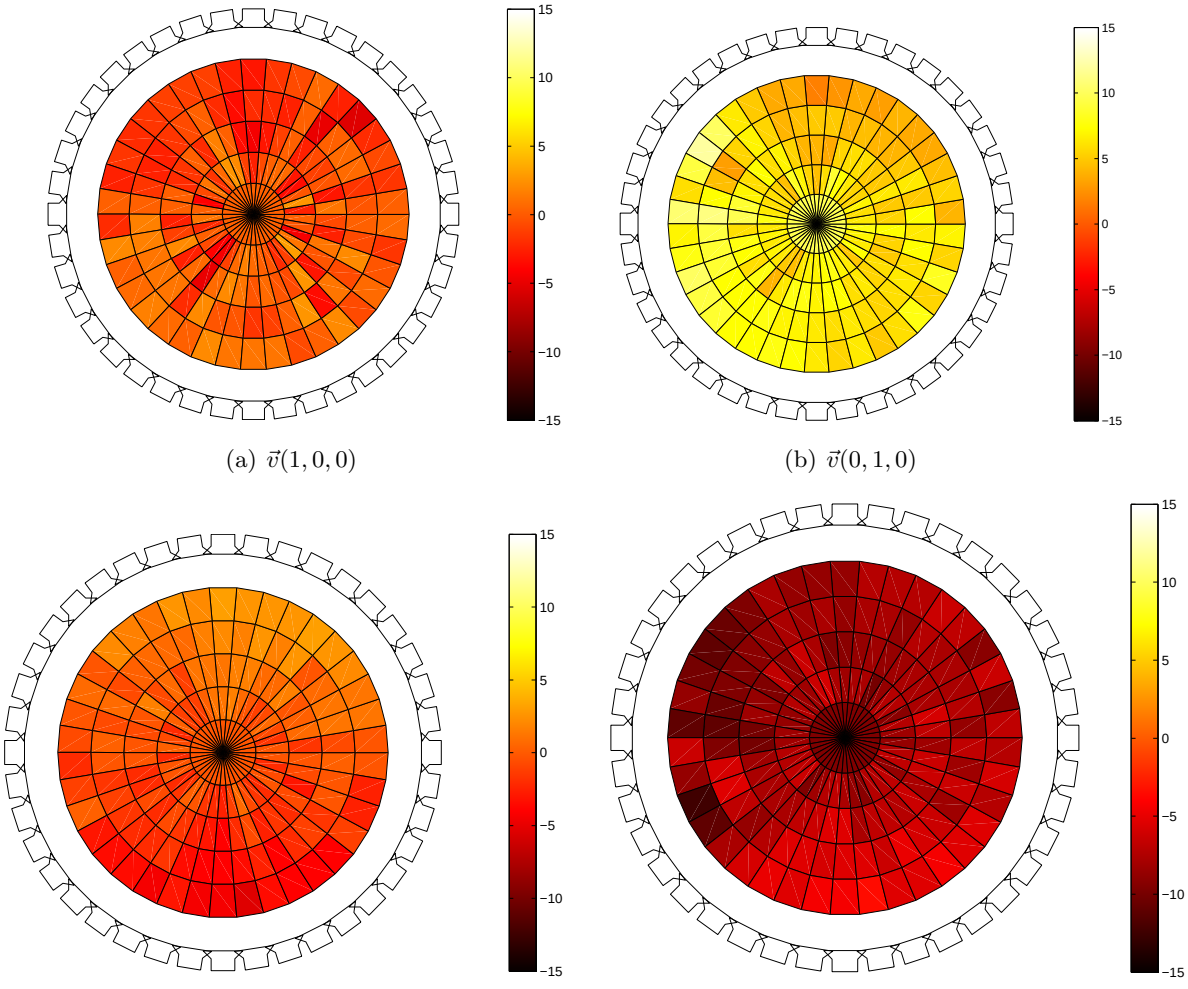
FIG. 2.47 – ILD (dB) évaluée sur la zone d'écoute pour l'onde plane cible en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

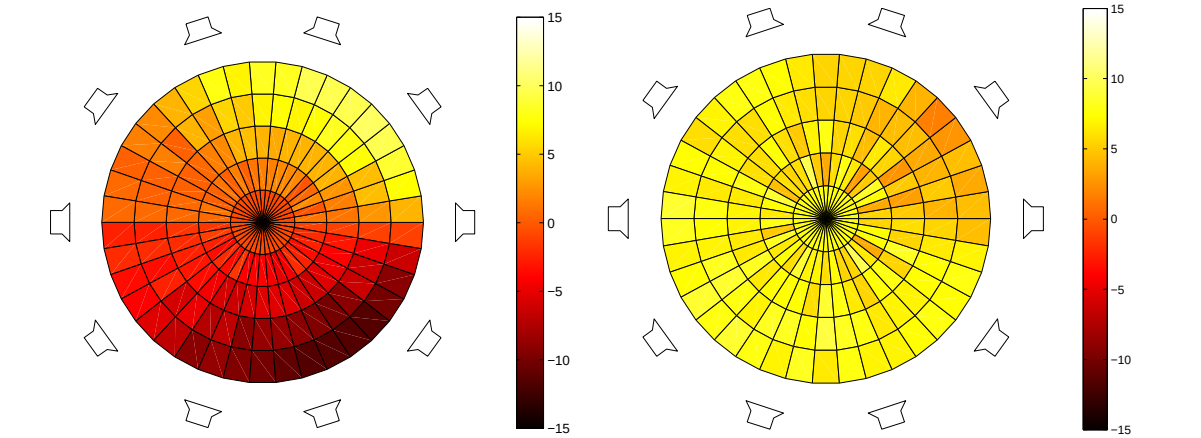
FIG. 2.48 – ILD (dB) évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$).



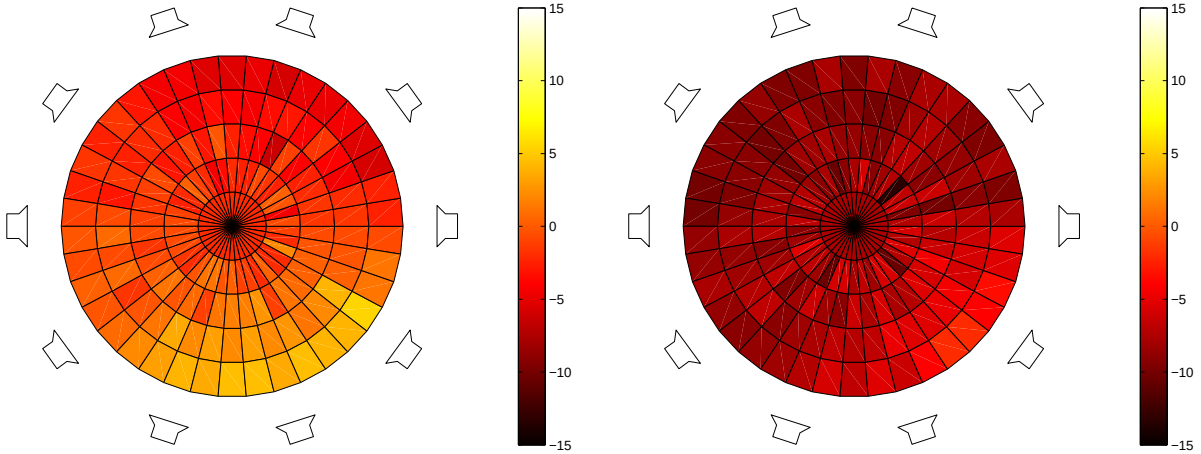
(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

FIG. 2.49 – ILD (dB) évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$).

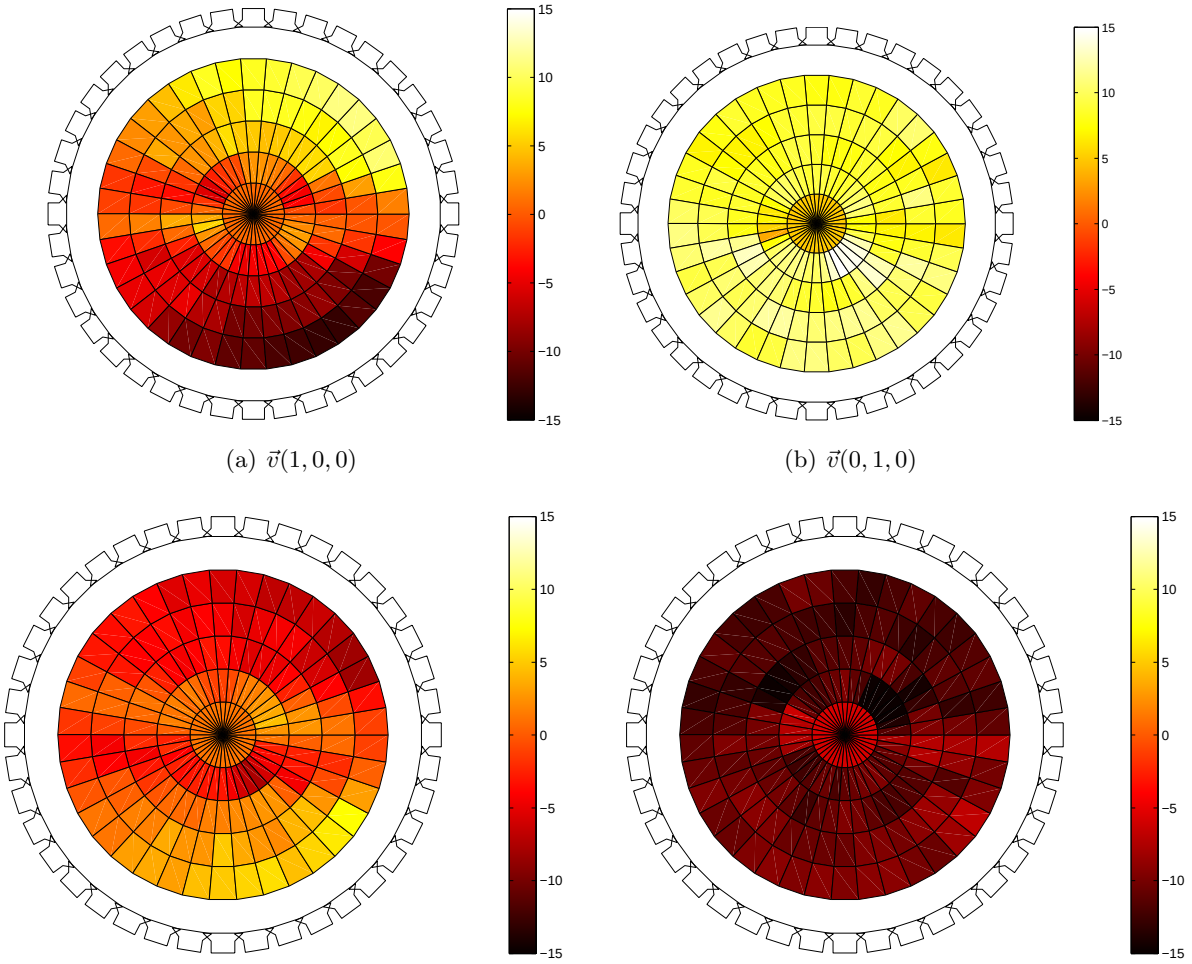


(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$



(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

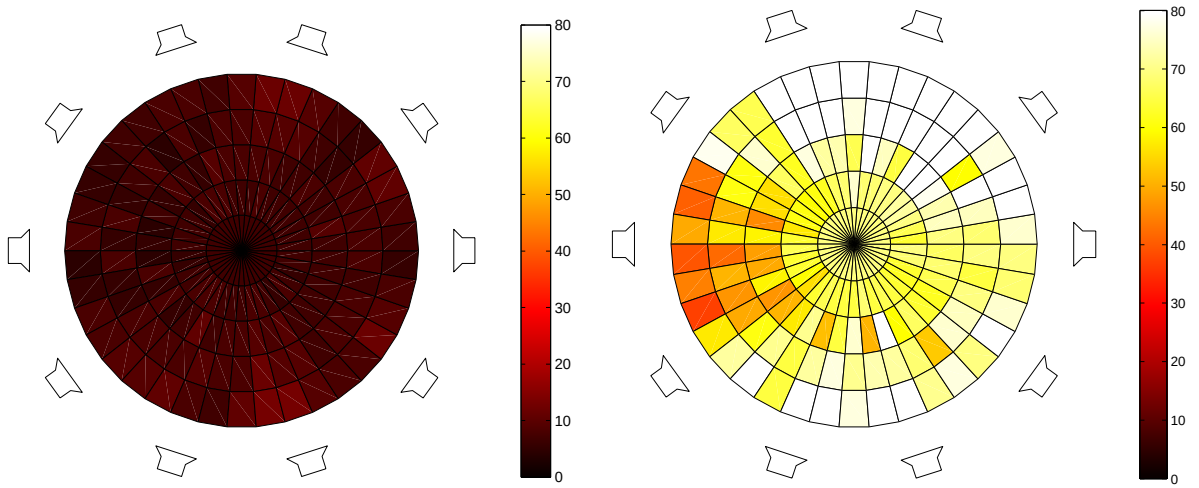
FIG. 2.50 – ILD (dB) évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$, M = 4).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

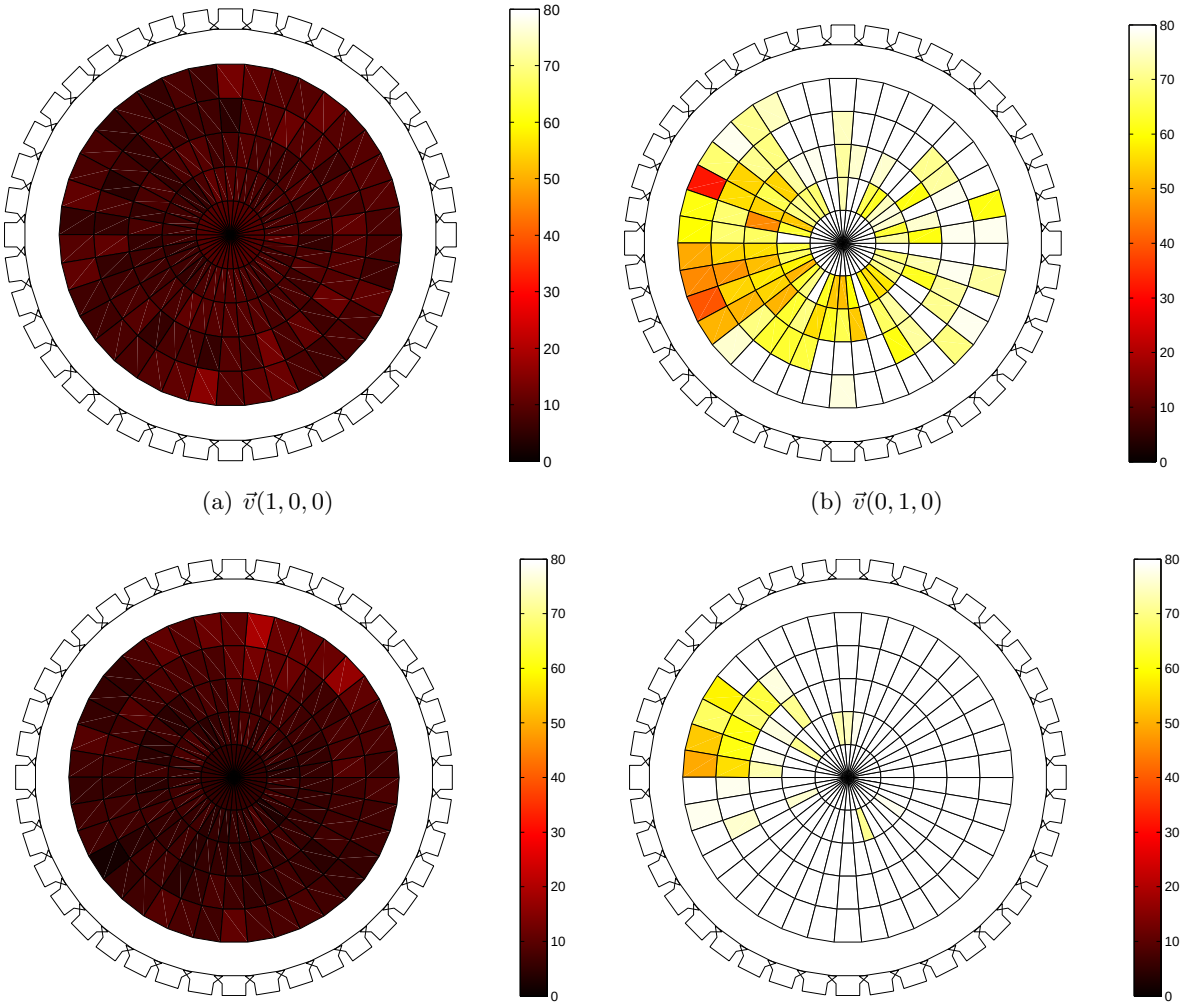
FIG. 2.51 – ILD (dB) évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$, M = 19).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

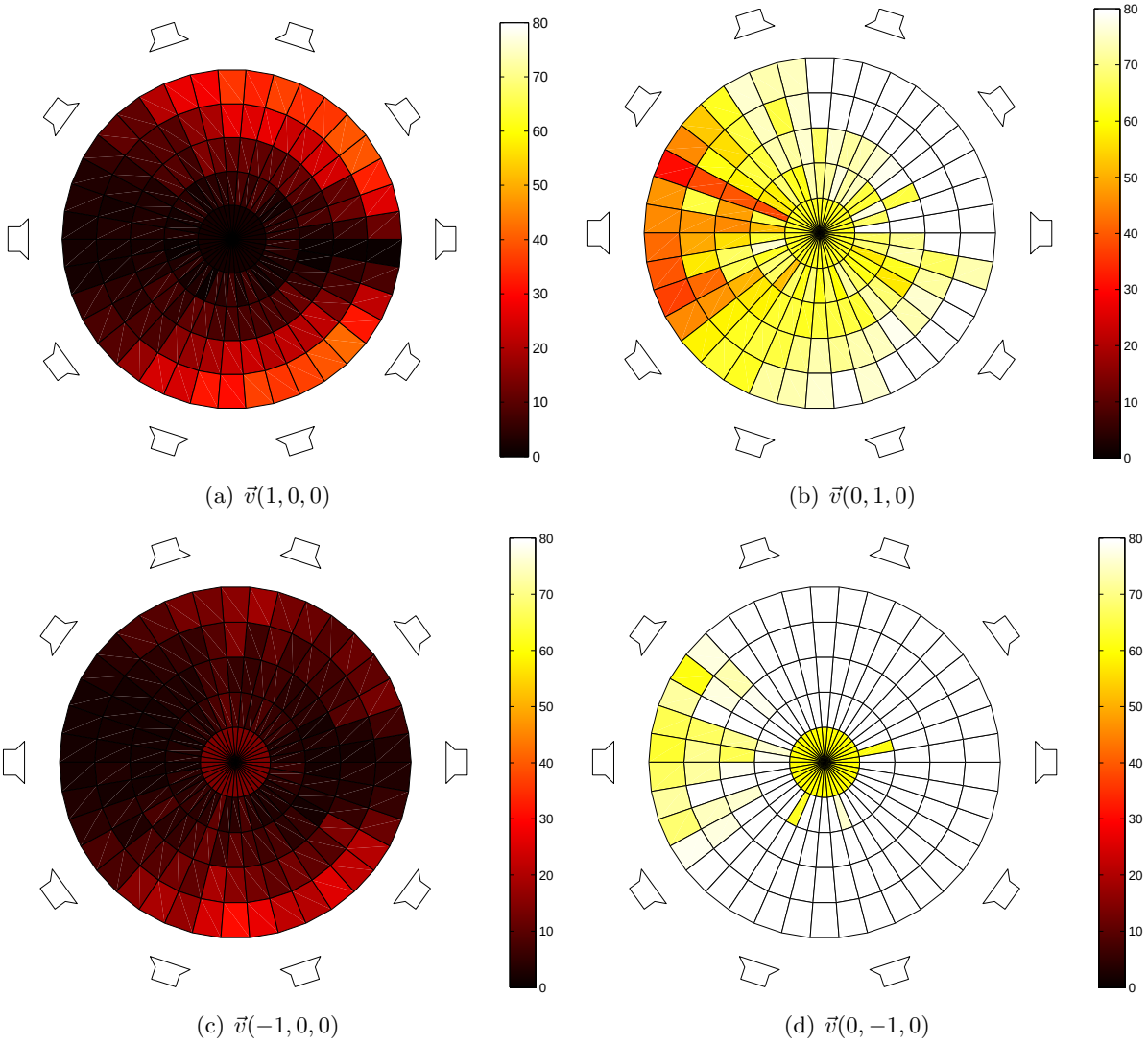
FIG. 2.52 – ISSD évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

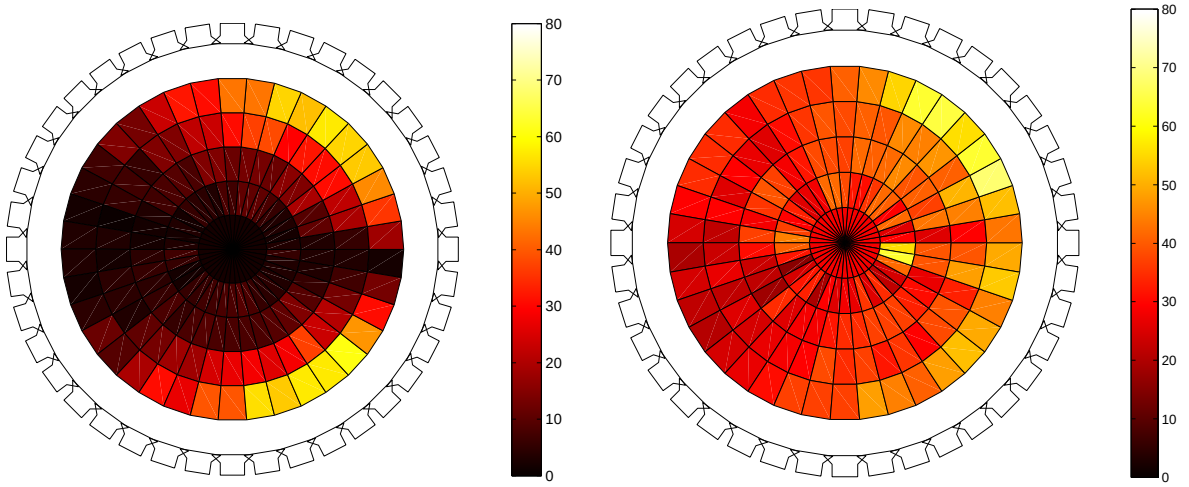
FIG. 2.53 – ISSD évaluée sur la zone d'écoute pour l'onde synthétisée par WFS en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$).



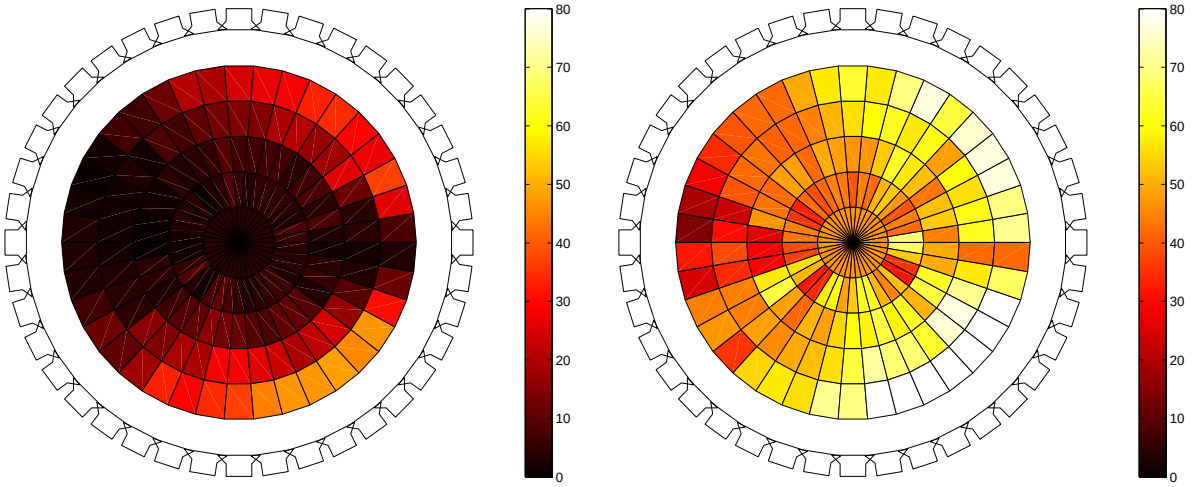
(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$

(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

FIG. 2.54 – ISSD évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 10$, M = 4).



(a) $\vec{v}(1,0,0)$ (b) $\vec{v}(0,1,0)$



(c) $\vec{v}(-1,0,0)$ (d) $\vec{v}(0,-1,0)$

FIG. 2.55 – ISSD évaluée sur la zone d'écoute pour l'onde synthétisée par HOA OP en fonction de l'orientation $\vec{v}$ de la tête de l'auditeur (onde plane d'azimut $\phi = 0°$, $N_L = 40$, M = 19).

- Les technologies WFS et HOA sont très sensibles à la nature des sources sonores à synthétiser (position, onde plane, onde sphérique...), et par extension au contenu de la scène sonore. C'est là un problème majeur. La stabilisation des performances de spatialisation quelle que soit la scène sonore à reproduire apparaît comme le premier défi à relever.
- L'ITD et l'ILD semblent mieux restituées par WFS, ce qui constitue un atout pour cette dernière (notamment en ce qui concerne l'ITD).
- En revanche la synthèse WFS souffre des distorsions spectrales (probablement en raison du repliement spectral) qui atteignent souvent un niveau bien supérieur à ceux obtenus pour HOA. C'est là aussi un problème majeur à résoudre d'autant que le détimbrage est sujet à évoluer en fonction de l'orientation de la tête de l'auditeur, ce qui contribue à le rendre d'autant plus audible et gênant.
- Pour les deux technologies, les indices dynamiques de localisation sont correctement reproduits, ce qui est très positif.
- Les performances de la synthèse HOA tendent à se dégrader avec un nombre trop élevé de haut-parleurs, tandis que la synthèse WFS s'améliore en raison de la diminution du repliement spatial, ce qui suggère que HOA est une solution optimale quand on dispose d'un faible nombre de haut-parleurs. Dès qu'on mise sur un nombre élevé de haut-parleurs, WFS apparaît en revanche préférable.

Il reste à réexaminer les résultats observés sur les indicateurs psycho-acoustiques (ITD, ILD, ISSD) à la lumière des propriétés physiques des ondes synthétiques.

Par ailleurs, parmi les pistes d'amélioration, le premier point à traiter est la mise en œuvre de décodages HOA autres que le décodage basique qui a été l'option de cette étude en raison de sa simplicité. Il serait intéressant d'évaluer l'apport de décodages optimisés sur les indices de localisation.

Chapitre 3

Synthèse binaurale

3.1 Concepts généraux et questions fondamentales

3.1.1 Encodage binaural

Alors que les technologies WFS et HOA utilisent un encodage spatial issu d'une représentation de la scène sonore dans l'*espace physique*, les technologies binaurales se fondent sur un encodage spatial défini sur la base d'une représentation des ondes acoustiques dans l'*espace perceptif*. Il s'agit même de la méthode de spatialisation sonore la plus proche de la perception : fondamentalement les technologies binaurales ne font ni plus ni moins qu'imiter les mécanismes de localisation auditive utilisés en situation d'écoute naturelle. Au quotidien, pour percevoir une scène sonore en 3 dimensions, les deux signaux captés au niveau de chaque tympan de l'auditeur suffisent en effet à décrire l'information spatiale du point de vue du système auditif. Les technologies binaurales sont basées sur cette idée : la scène sonore est ainsi représentée par seulement deux signaux (ou deux canaux) qui correspondent aux signaux perçus au niveau des tympans, ce qui constitue certainement la représentation la plus efficace en termes de compression. L'information spatiale est encodée au travers des indices de localisation : différences interaurales de temps et d'intensité et indices spectraux. Les HRTF (Head Related Transfer Function) définissent les fonctions de transfert qui décrivent la propagation acoustique entre la source sonore et les oreilles de l'auditeur (Fig. 3.1). Ces HRTF rassemblent sous une forme compacte l'ensemble des indices mis à disposition du système auditif pour localiser les sons. Ainsi l'encodage spatial binaural repose uniquement sur les HRTF. Par suite, cet encodage comprend :

- des différences de temps et d'intensité entre les deux canaux (cf. Fig. 3.2 & 3.3) : ces différences sont susceptibles de dépendre de la fréquence,
- un filtrage fréquenciel du spectre de la source sonore pour chaque canal.

Tous ces paramètres d'encodage dépendent de la position de la source. Le système auditif est capable de les interpréter pour localiser les sons. En situation d'écoute naturelle, cet encodage est réalisé par l'interaction de l'onde acoustique avec le corps de l'auditeur, principalement à travers les phénomènes de réflexion et diffraction avec le pavillon de l'oreille, la tête et le haut du torse de l'auditeur [Shaw & Teranishi, 1968] [Algazi et al., 2001a]. Ces phénomènes dépendent fortement de la morphologie de l'auditeur : l'encodage est la traduction acoustique de l'empreinte morphologique de l'auditeur avec toutes ses spécificités individuelles (taille de la tête, forme et taille du pavillon, ...). Il en résulte que l'encodage binaural est individuel, c'est à dire que l'encodage des informations spatiales ne vaut que pour un individu, ce qui est une sévère limitation. Le caractère individuel de l'encodage binaural est une première spécificité des technologies binaurales. La seconde spécificité est la nature spectrale de cet encodage : si les différences de temps et d'intensité sont des paramètres

d'encodage communs à d'autres technologies de spatialisation sonore (telles que la stéréophonie ou le multicanal), l'encodage par filtrage fréquenciel est l'apanage des technologies binaurales. Cet encodage spectral est cependant un obstacle intrinsèque à l'exigence de transparence de reproduction d'une scène sonore, puisqu'il implique d'altérer le spectre original des sources sonores. Néanmoins, en situation d'écoute naturelle, nous subissons ces distorsions spectrales sans ressentir de gêne. Ce paradoxe reste difficile à expliquer : il faut sans doute en chercher la raison dans des processus cognitifs permettant de compenser les modifications spectrales, en procédant par exemple par recoupement soit en se basant sur les différents spectres perçus au cours de déplacements de la source ou de mouvements de l'auditeur, soit en exploitant le jeu des réflexions sur les parois de la salle.

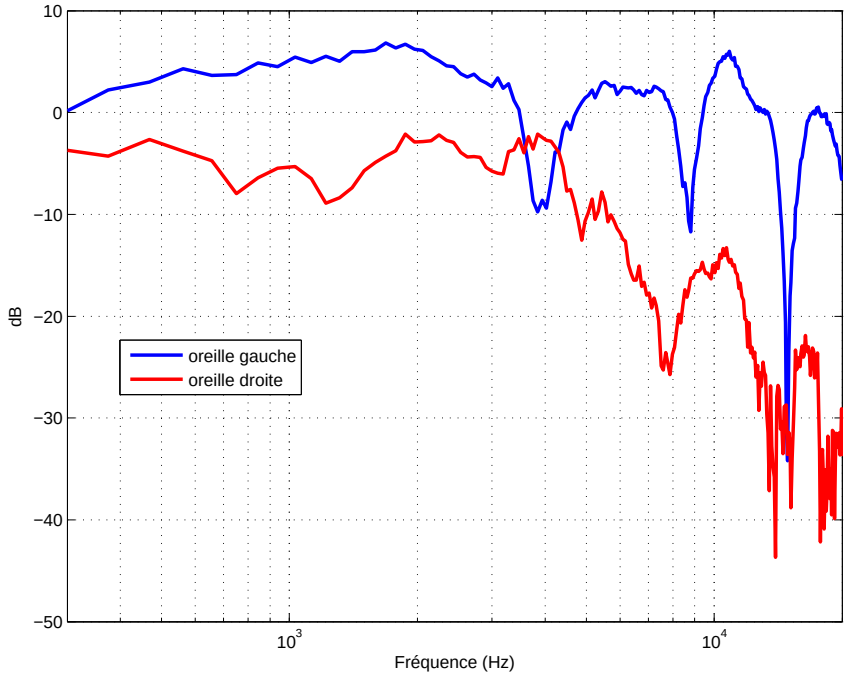
En pratique, les technologies binaurales se déclinent sous deux formes :

- encodage *naturel* : Les signaux binauraux sont acquis au moyen d'un enregistrement en plaçant une paire de microphones à l'entrée des conduits auditifs d'un individu ou d'un mannequin (têtes artificielles, cf. Fig. 3.4 & 3.5). Cette déclinaison trouve son application dans la captation de scènes sonores pour le partage d'ambiance ou le concept de carte postale sonore. Le principal inconvénient est l'impossibilité de modifier la scène sonore a posteriori.
- encodage *artificiel* : Les signaux binauraux sont obtenus par **synthèse binaurale** en convoluant un signal monophonique représentant le signal émis par la source sonore par une paire de filtres modélisant les HRTF associées aux oreilles gauche et droite en relation avec une position de source donnée (Fig. 3.6). Potentiellement les HRTF peuvent prendre en compte l'effet de salle lié à l'environnement acoustique des sources sonores. Contrairement à un enregistrement, la synthèse binaurale offre toute liberté dans le positionnement et le contrôle des sources sonores. Elle permet aussi de coupler le rendu binaural à un dispositif de suivi de mouvements de tête de l'auditeur (*head-tracking*) afin de les compenser et de conserver un positionnement stable des sources quels que soient l'orientation et les mouvements de tête de l'auditeur. On parle alors de *synthèse binaurale dynamique*.

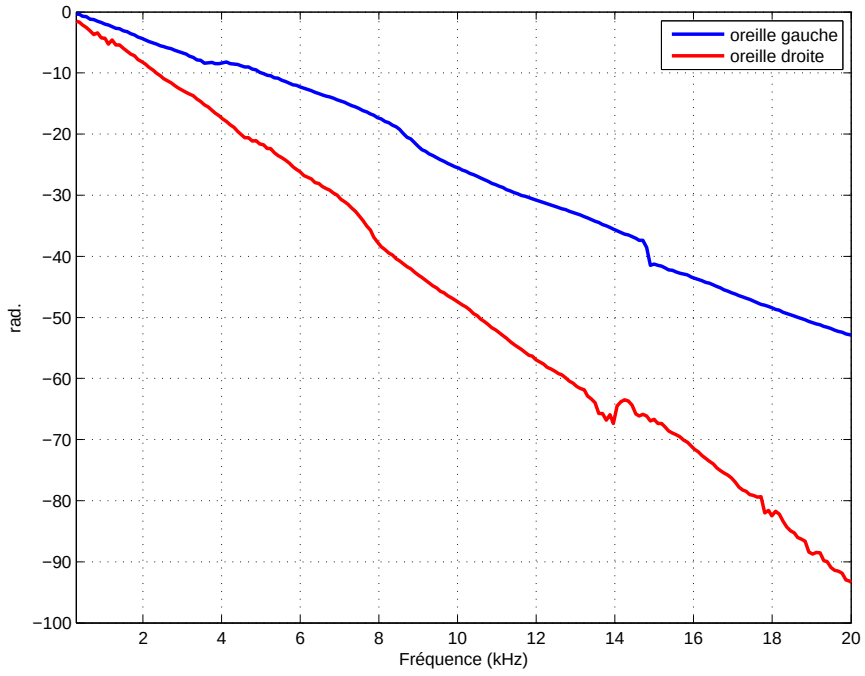
3.1.2 Décodage binaural

Décodage sur casque

Pour le décodage, le mode privilégié est l'écoute des signaux binauraux avec un casque qui permet de restituer les signaux à l'endroit où ils ont été captés. Le seul défaut qu'on peut reprocher au rendu binaural sur casque est l'absence des vibrations mécaniques perçues par le corps et qui participent à notre perception en situation d'écoute naturelle. Dans l'étape de décodage, le premier requis est de veiller à corriger la réponse du casque. Cette opération constitue la **calibration du casque** et consiste à compenser la fonction de transfert entre le casque et l'entrée des conduits auditifs de l'auditeur (fonction de transfert désignée sous le nom de *Head-Phone Transfert Function* ou HPTF). La calibration est cependant délicate à mettre en œuvre, car, outre qu'elle nécessite de mesurer la HPTF du casque d'écoute, il faut avoir conscience que la HPTF dépend aussi du **positionnement** du casque sur les oreilles et de l'**individu**. Autant la calibration adaptée à l'individu semble possible, la dépendance au positionnement du casque est un problème majeur qui a été mis en avant par Kulkarni & Colburn [Kulkarni & Colburn, 2000], d'autant qu'une calibration *moyenne* ne semble pas satisfaisante. La Figure 3.7 illustre les HPTF obtenues pour 10 positionnements successifs du même casque (Sennheiser HD600) pour deux individus. En fonction du positionnement, on observe principalement des modifications de l'amplitude des pics et des creux des HPTF (amplification ou atténuation) à partir de 7 kHz, plus rarement des décalages de ces pics et creux. Il apparaît que les variations individuelles sont beaucoup plus marquées que les variations d'un positionnement à l'autre pour un même individu. McAnally & Martin soulignent en outre que ces dernières présentent une variance bien moindre que les colorations *utiles* des HRTF, c'est à dire



(a) Module



(b) Phase

FIG. 3.1 – Exemple de HRTF gauche et droite mesurées sur un individu (base de HRTF *Jean-Marie Pernaux* - base privée d'Orange Labs -, sujet RN, direction (ϕ, θ)=(-141°,-11°) en coordonnées polaires verticales)

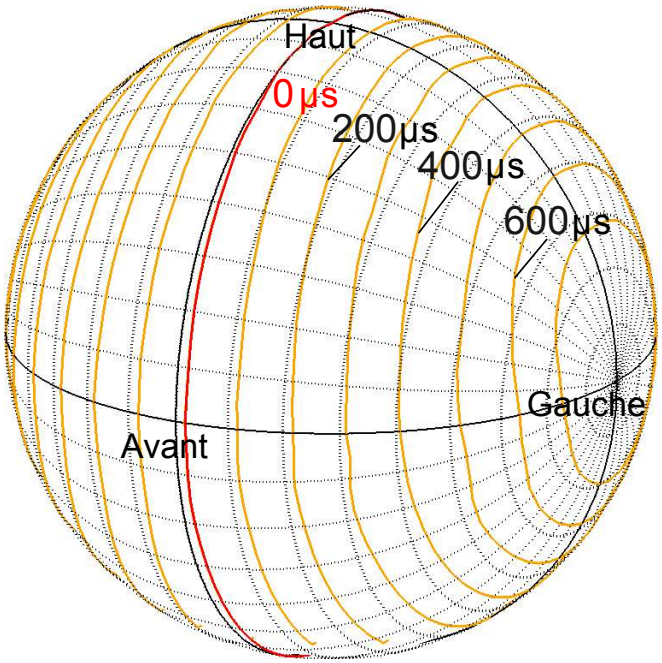


FIG. 3.2 – Evolution sur la sphère des différences interaurales de temps (ITD pour *Interaural Time Difference*) extraites des HRTF [Guillon, 2009] : les lignes relient les directions correspondant à une même valeur et représentent donc des lignes iso-ITD (estimation de l'ITD par régression linéaire de la phase aux basses fréquences, base *Jean-Marie Pernaux*, sujet ME).

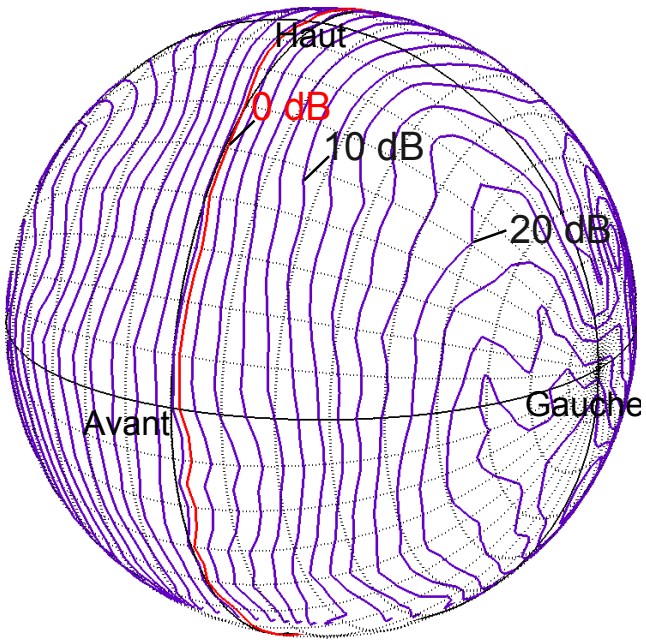


FIG. 3.3 – Evolution sur la sphère des différences interaurales d'intensité (ILD pour *Interaural Level Difference*)extraites des HRTF [Guillon, 2009] : les lignes relient les directions correspondant à une même valeur et représentent donc des lignes iso-ILD (estimation de l'ILD par l'équation 3.9 dans laquelle $[f_1 - f_2] = [1.5 - 10kHz]$, base *Jean-Marie Pernaux*, sujet ME).

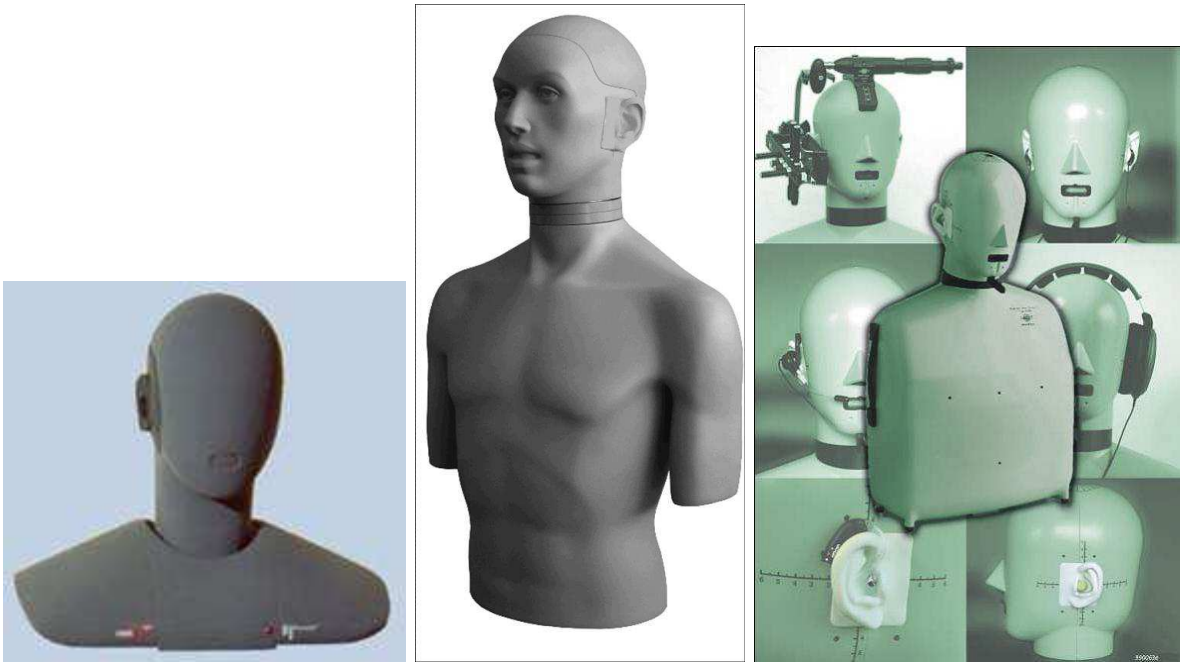


FIG. 3.4 – Exemples de têtes artificielles, de gauche à droite : Head Acoustics HMSIII, KEMAR, Brüel & Kjær HATS

FIG. 3.5 – Enregistrements binauraux in situ : travaux de l'association Omnihead [Rueff & Blum, 2003].

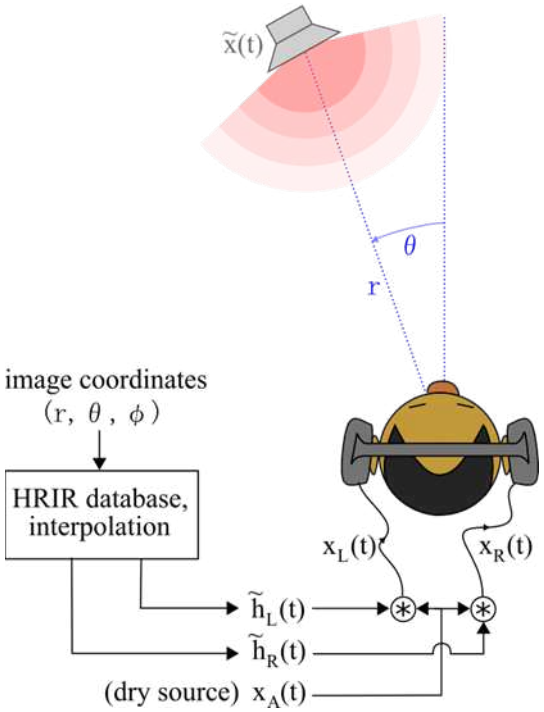


FIG. 3.6 – Principe de la synthèse binaurale : Un signal monophonique et anéchoïque $x_A(t)$ est convolué par la paire de fonctions de transfert $\tilde{h}_L$ et $\tilde{h}_L$ associée à la direction où on souhaite créer la source virtuelle. Les HRTF utilisées sont soit directement les HRTF mesurées, soit le résultat d'une interpolation si la direction désirée n'est pas disponible dans la base de données.

les colorations spectrales portant l'information de localisation des sons [McAnally & Martin, 2002]. Ces observations amènent à la conclusion suivante : à défaut de corriger les variations liées au positionnement, il faut impérativement appliquer une **calibration individuelle**, c'est à dire adaptée à chaque individu [Møller, 1992]. Des études montrent en effet qu'une calibration non individuelle engendrent potentiellement des dégradations équivalentes à l'utilisation de HRTF non individuelles [Pralong & Carlile, 1996]. Certains auteurs avancent même que l'individualisation de la calibration du casque serait plus importante que celle des HRTF, du moins en termes d'externalisation des sources virtuelles [Kim & Choi, 2005].

Le **choix du casque** doit aussi être considéré avec soin pour garantir un rendu binaural de qualité. Möller [Møller, 1992] recommande les casques de type ouvert, plus précisément de type FEC (*Free-air Coupling Equivalent*), qui se caractérisent par le fait qu'ils offrent les conditions de rayonnement en champ libre en termes d'impédance vue par le tympan, comme si l'auditeur ne portait pas de casque. Möller a proposé un critère pour évaluer la pertinence d'un casque pour la restitution binaurale : le PDR (Pressure Division Ratio) qui compare les conditions d'impédance vue par le tympan en présence du casque aux conditions de rayonnement en champ libre. Le casque idéal pour un rendu binaural doit satisfaire un PDR égal à 1 et répond alors au label FEC. La majorité des études sur la mise en œuvre de casques pour le binaural ne considère pas le cas particulier des écouteurs intra-auriculaires. C'est une lacune qui mériterait d'être comblée en raison de la généralisation de l'usage des écouteurs dans le contexte des téléphones mobiles ou des lecteurs MP3. Se posent alors les questions de leur calibration et de leur labellisation FEC.

Synthèse binaurale dynamique

Si les signaux binauraux sont délivrés aux oreilles de l'auditeur sans autre précaution que la correction des transducteurs, on se place dans le mode de restitution binaurale *statique*, au sens où les éventuels mouvements de tête de l'auditeur ne sont pas compensés par le dispositif de restitution. Ce mode constitue encore aujourd'hui le cas le plus courant. Dans ces conditions, lorsque l'auditeur tourne la tête, l'ensemble de la scène sonore pivote, ce qui tend à dégrader le réalisme de l'illusion sonore et à réduire l'externalisation des sources virtuelles. L'alternative consiste à annuler les mouvements de tête de l'auditeur afin de conserver des sources virtuelles fixes quelle que soient la position et l'orientation de l'auditeur : il s'agit du mode de restitution binaurale *dynamique*. Ce mode n'est possible que dans le cas d'un encodage artificiel par synthèse binaurale, car il nécessite de modifier en temps réel la position des sources virtuelles en fonction des mouvements de l'auditeur, ce qui ne peut être réalisé a posteriori sur un enregistrement binaural[1]. On parle donc communément de *synthèse binaurale dynamique*. Ce mode dynamique implique de coupler le rendu binaural à un système de suivi de mouvements de tête (*head-tracking*) chargé d'informer en temps réel le moteur de synthèse binaurale sur la position et l'orientation de la tête de l'auditeur afin d'actualiser en conséquence les filtres binauraux. L'objectif est de conserver des sources sonores virtuelles fixes dans un référentiel absolu indépendant de l'auditeur. La synthèse binaurale dynamique suppose de connaître potentiellement les filtres binauraux pour n'importe quelle direction de l'espace, pour laquelle les HRTF ont été mesurées ou non. Pour les directions pour lesquelles les HRTF ne sont pas connues, les filtres binauraux sont obtenus par **interpolation** des filtres des directions voisines.

Il existe différentes technologies disponibles pour réaliser les systèmes de *head-tracking* [Faure, 2004] :

– **acoustiques** : Le système se compose d'un émetteur (fixe) d'ultrasons et d'un récepteur

[1]Dans le cas d'un encodage naturel avec une tête artificielle, lorsque la prise et la restitution sonore sont effectuées simultanément, il est possible d'asservir le mannequin de prise de son aux mouvements de l'auditeur [Mackensen, 2004], ce qui permet un mode dynamique.

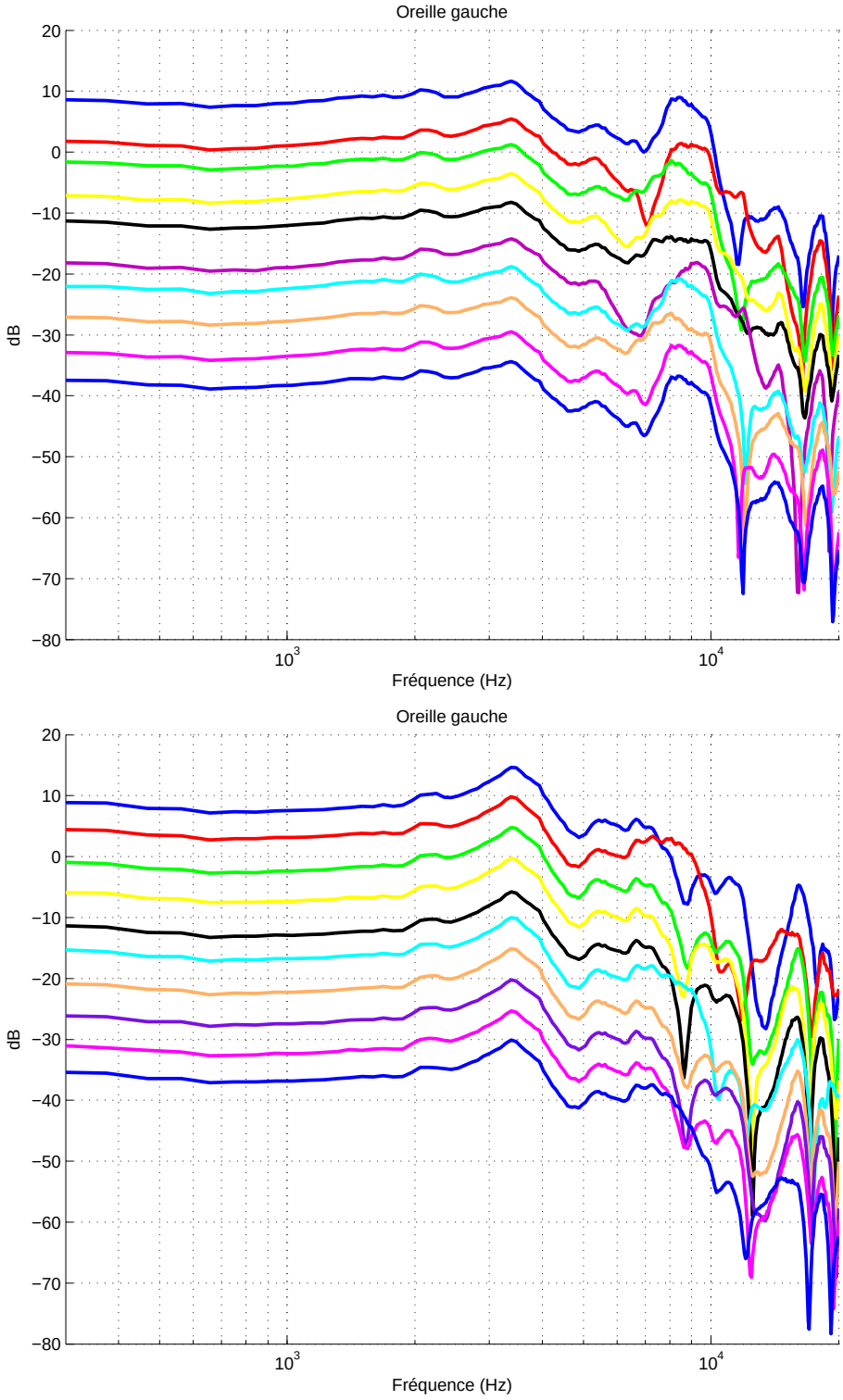


FIG. 3.7 – HPTF mesurées pour 10 positionnements du casque (Sennheiser HD600) et pour deux individus (base *Jean-Marie Pernaux* : sujets ME en haut et VM en bas) : les différentes fonctions de tranfert ont été décalées de 5 dB pour une meilleure lisibilité.

associé qui est solidaire de la tête. La portée est limitée en raison de l'absorption des ultrasons par l'air.
- **inertiels** : Un capteur inertiel combine un gyroscope (mesure de l'orientation de la tête par simple intégration) et un accéléromètre (mesure de la position par double intégration). Ces systèmes souffrent d'une faible précision à vitesse lente, ainsi que de problèmes de stabilité.
- **optiques** : Le procédé repose sur l'utilisation de caméra(s) et d'algorithmes d'analyse d'images.
- **magnétiques** : Le système se base sur un émetteur (fixe) et un récepteur (attaché à l'auditeur) de champ magnétique[2]. Cette technologie donne la meilleure précision, mais souffre d'une portée limitée et d'une forte sensibilité aux perturbations électromagnétiques.

Outre le dispositif de *head-tracking*, la mise en œuvre de la synthèse binaurale dynamique soulève deux problèmes spécifiques :
- La **latence** du système (c'est à dire le temps qui s'écoule entre l'instant où l'auditeur effectue un mouvement et l'instant où les filtres binauraux sont effectivement mis à jour), doit être très faible afin d'offrir à l'auditeur le rendu le plus naturel et le plus transparent possible. Pour les sons longs, une latence de 250 ms suffirait [Wenzel, 1999], tandis que pour les sons courts le seuil s'abaisse à 75 ms [Brungart et al., 2004].
- La mise à jour des filtres requiert de **commuter** rapidement entre deux jeux de coefficients, ce qui peut entraîner des discontinuités dans les signaux et, par suite, des artefacts audibles [Larcher, 2001]. Une solution simple pour réaliser une commutation inaudible est un fondu-enchaîné entre les signaux avant et après actualisation des filtres.

Le principal intérêt de la synthèse binaurale dynamique réside dans l'ajout des indices dynamiques de localisation. L'apport du mode dynamique par rapport au mode statique a été évalué par plusieurs études qui s'accordent sur la diminution du taux des confusions avant/arrière [Wenzel, 1995] [Wenzel, 1999] [Begault et al., 2001] [Faure, 2005]. Une étude menée à Orange Labs [Faure, 2005] montre qu'en outre le mode dynamique tend à réduire l'erreur de localisation en azimut, principalement pour des sujets non experts de la spatialisation sonore, ainsi que la dispersion des jugements de localisation dans la zone frontale. En revanche le mode dynamique ne permet pas d'améliorer le rendu des sources frontales au sens où la difficulté à percevoir ces sources précisément dans l'espace frontal devant l'auditeur persiste. Ces travaux mettent en évidence la complémentarité entre le mode dynamique et l'utilisation de filtres binauraux individuels, c'est à dire adaptés à l'encodage morphologique de l'auditeur [Faure, 2005]. Si, en mode statique, les filtres individuels permettent de réduire les confusions avant/arrière par rapport à des filtres non individuels, l'ajout du mode dynamique avec des filtres individuels diminue tout aussi significativement les confusions. L'apport du mode dynamique est cependant prépondérant : le bénéfice de l'ajout seul du mode dynamique est supérieur à celui de l'utilisation seule de filtres individuels[3]. Par ailleurs il semble qu'un head-tracking limité à un degré de liberté (rotation en azimut) suffise. Cette étude comporte d'une part un test de localisation et des tests d'écoute basés sur d'autres méthodes d'évaluation :
- Un test d'évaluation *indirecte*, dans lequel on demande au sujet de décrire la scène sonore qu'il a perçue : la fiabilité des informations qu'il rapporte est considérée comme un indicateur de la qualité de la spatialisation perçue,
- Un test d'évaluation *directe*, dans lequel le sujet doit juger la scène sonore sur une grille de 4 attributs spatiaux (précision spatiale, externalisation, enveloppement, réalisme) complétés par un jugement de préférence.

[2] A défaut le champ magnétique terrestre peut être utilisé.

[3] Toutefois il convient de rester prudent dans les conclusions de cette étude, étant donné que les conditions filtre individuels et non individuels correspondent à deux groupes distincts de sujets. Pour étayer les résultats, il faudrait que le même groupe de sujet soit soumis aux 4 conditions : filtres non individuels + mode statique, filtres non individuels + mode dynamique, filtres individuels + mode statique, filtres individuels + mode dynamique.

Il ressort que le mode dynamique apporte une amélioration significative à la fois en termes de précision spatiale, d'externalisation, d'enveloppement, de réalisme et de la préférence globale.

Rendu sur haut-parleurs

L'alternative au casque est l'écoute sur un système de deux haut-parleurs. Cependant, si l'on alimente directement les haut-parleurs par les signaux binauraux, on est confronté au problème des trajets croisés : le signal binaural gauche (respectivement droit) qui est destiné uniquement à l'oreille gauche (respectivement droite) est perçu non seulement par l'oreille gauche (respectivement droite), mais aussi par l'oreille droite (respectivement gauche) modulo le contournement de la tête (cf. Fig. 3.8). Cette **diaphonie** entre les deux oreilles vient complètement détruire l'illusion de la scène sonore virtuelle. Pour un rendu équivalent au casque, il convient donc de l'éliminer. Une solution intuitive consiste à placer un écran acoustique devant l'auditeur et perpendiculairement à l'axe interaural, afin de supprimer l'onde de contournement. La solution générale est basée sur un prétraitement des signaux binauraux en amont de la diffusion par les haut-parleurs (cf. Fig. 3.8) : le signal parasite résultant du trajet croisé est injecté en opposition de phase au signal binaural original, de façon à annuler l'onde de contournement lors de la diffusion. Il s'agit du procédé d'**annulation des trajets croisés** (en anglais *crosstalk canceller* [Gardner, 1997]). La solution théorique s'exprime comme suit. Soient B_L et B_R les signaux binauraux originaux, X_L et X_R les signaux alimentant les haut-parleurs gauche et droit respectivement, Y_L et Y_R les signaux perçus en entrée des oreilles gauche et droite respectivement de l'auditeur. Tous ces signaux sont exprimés dans le domaine fréquenciel en fonction de la fréquence f. L'objectif est d'adapter les signaux des haut-parleurs X_L et X_R afin que les signaux perçus s'identifient aux signaux binauraux B_L et B_R comme en situation d'écoute au casque :

$$\begin{cases} Y_L(f) \equiv B_L(f) \\ Y_R(f) \equiv B_R(f) \end{cases} \tag{3.1}$$

Or, les signaux Y_L et Y_R résultent de la propagation acoustique (trajets direct et croisé) entre chaque haut-parleur et chaque oreille :

$$\begin{cases} Y_L(f) = X_L(f) \times H_{1L}(f) + X_R(f) \times H_{2L}(f) \equiv B_L(f) \\ Y_R(f) = X_L(f) \times H_{1R}(f) + X_R(f) \times H_{1L}(f) \equiv B_R(f) \end{cases} \tag{3.2}$$

Dans cette expression, les fonctions de transfert H_{1L} et H_{2R} désignent les trajets directs entre les haut-parleurs gauche et droit et les oreilles gauche et droite respectivement, tandis que les fonctions de transfert H_{1R} et H_{2L} définissent les trajets croisés entre les haut-parleurs gauche et droit et les oreilles droite et gauche respectivement. En résolvant le système 3.2, on montre comment modifier les signaux X_L et X_R pour reproduire les signaux B_L et B_R au niveau des oreilles de l'auditeur :

$$\begin{cases} X_L(f) = \frac{B_L(f) \times H_{1L}(f) - B_R(f) \times H_{2L}(f)}{H_{1L}(f) \times H_{2R}(f) - H_{1R}(f) \times H_{2L}(f)} \\ X_R(f) = \frac{B_R(f) \times H_{2R}(f) - B_L(f) \times H_{1R}(f)}{H_{1L}(f) \times H_{2R}(f) - H_{1R}(f) \times H_{2L}(f)} \end{cases} \tag{3.3}$$

L'effet de l'annulation des trajets croisés est illustré sur la Figure 3.9. L'effet est double : il permet d'une part de corriger la réponse des haut-parleurs et de compenser la propagation directe entre chaque haut-parleur et l'oreille ipsilatérale, de telle sorte qu'une impulsion parfaite est restituée au niveau de l'oreille ipsilatérale. D'autre part les trajets croisés sont annulés : ainsi la contribution de chaque haut-parleur sur l'oreille controlatérale est nulle. Le système *transaural* est un exemple de réalisation d'annulation des trajets croisés [Atal & Schroeder, 1966] [Cooper & Bauck, 1989]. Le procédé peut aussi s'appliquer à un dispositif de quatre haut-parleurs [Guastavino et al., 2007].

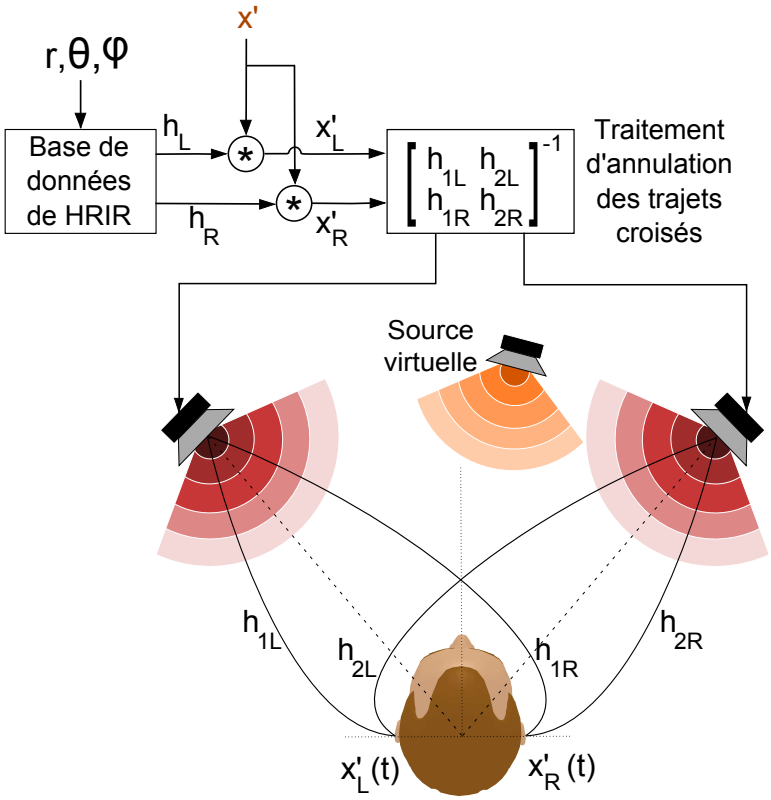


FIG. 3.8 – Rendu binaural sur un dispositif de deux haut-parleurs : illustration des trajets directs (H_{1L} et H_{2R}) et des trajets croisés (H_{1R} et H_{2L}).

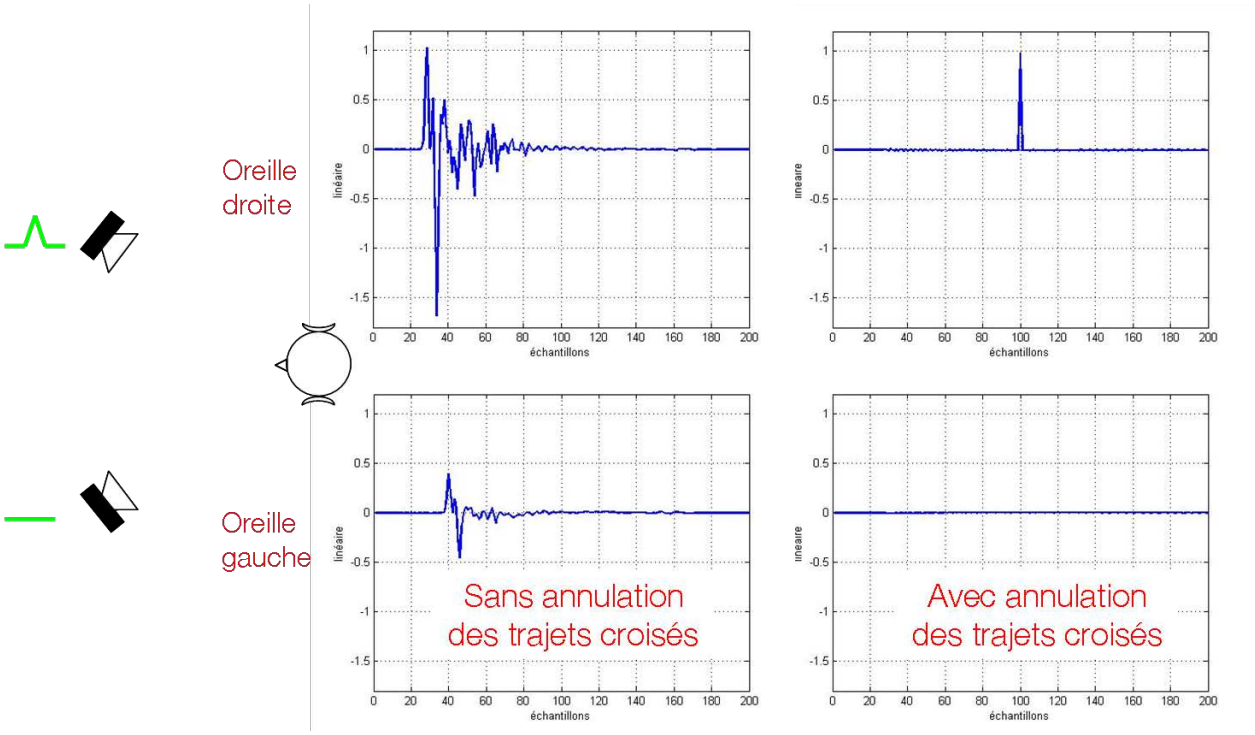


FIG. 3.9 – Mise en oeuvre de l'annulation des trajets croisés : Le haut-parleur droit émet une impulsion, tandis que la haut-parleur gauche reste muet. En l'absence de l'annulation des trajets croisés, l'oreille droite perçoit l'impulsion en propagation directe et l'oreille gauche l'impulsion après contournement de la tête. Avec l'annulation des trajets croisés, la fonction de transfert entre le haut-parleur droit et l'oreille droite est compensée de façon à restituer une impulsion parfaite au niveau de l'oreille droite. Pour l'oreille gauche, le signal émis par le haut-parleur droit est annulé.

Une autre déclinaison de système de rendu binaural sur haut-parleurs est le stéréo dipôle (*stereo dipole*) proposé par Kirkeby [Kirkeby et al., 1997]. Dans cette solution, les haut-parleurs ne sont plus disposés selon la configuration stéréophonique, mais parallèlement à l'axe interaural avec un très faible écart angulaire de l'ordre de 5 à 10° (cf. Fig. 3.10). L'avantage de cette configuration est de minimiser les interactions entre les deux haut-parleurs, ce qui simplifie d'autant le travail d'annulation des trajets croisés. Les auteurs montrent qu'ainsi le traitement de compensation des trajets croisés est plus robuste, la zone d'écoute est sensiblement élargie, ce qui autorise les mouvements de tête de l'auditeur. La configuration du stéréo dipôle se prête particulièrement bien à l'intégration du rendu binaural sur les petits terminaux terminaux individuels, tels que les ordinateurs portables (qui, typiquement, sont équipés de deux haut-parleurs faiblement espacés et disposés parallèlement à l'axe interaural) ou les téléphones portables.

3.1.3 HRTF

Le support de l'encodage spatial binaural

Les HRTF (ou leur équivalent temporel HRIR pour Head Related Impulse Response) constituent le concept fondamental des technologies binaurales, au sens où elles matérialisent le support de l'encodage spatial intrinsèque à ces technologies. La position d'une source sonore est encodée par la fonction de transfert associée à sa direction et qui traduit l'ensemble des phénomènes

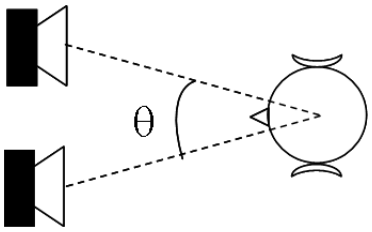


FIG. 3.10 – Stéréo Dipôle : Les deux haut-parleurs sont disposés parallèllement à l'axe interaural et forment un angle θ de l'ordre de 5° à 10°.

de propagation des ondes acoustiques entre la source et l'entrée des conduits auditifs. Ces phénomènes [Guillon, 2007] comprennent à la fois la propagation en champ libre, la diffraction par la tête de l'auditeur [Duda & Martens, 1998] [Algazi et al., 2001a], les réflexions sur les épaules et le haut du torse [Algazi et al., 2002a] [Algazi et al., 2002b], et surtout le jeu des résonances liées au pavillon (réflexions et diffraction par la conque) [Batteau, 1967] [Shaw & Teranishi, 1968] [Hebrank & Wright, 1974]. Les HRTF sont totalement déterminées par la morphologie de l'individu. Le rôle particulier de certains éléments morphologiques a été mis en évidence [Algazi et al., 2002a]. Pour cette question, la modélisation numérique de type BEM ouvre des perspectives d'investigation systématique et approfondie du lien entre HRTF et morphologie [Katz, 1998] [Iwaya & Suzuki, 2008] [Fels & Vorländer, 2009], même si au final les interactions sont complexes et multiples, ce qui rend souvent difficile, voire illusoire[4] de chercher à établir une relation simple et prédictible entre les propriétés des HRTF et un élement particulier de la morphologie. Une question qui est examinée en particulier est d'identifier les éléments morphologiques qui exercent un rôle prédominant dans la formation des HRTF [Fels & Vorländer, 2009]. D'un point de vue qualitatif, le rôle-clef du pavillon est établi par de nombreuses études [Guillon, 2009]. D'un point de vue plus quantitatif, outre la distance interaurale dont l'importance est évidente, une récente étude [Fels & Vorländer, 2009] met en avant les paramètres anthropométriques suivants : d'une part, la distance entre l'oreille et l'épaule, la largeur de la tête et l'excursion arrière de la tête par rapport à l'oreille, pour décrire la morphologie globale de la tête et du haut du torse, et d'autre part, l'orientation du pavillon, la largeur et la profondeur de la conque, pour décrire les détails de la morphologie du pavillon.

Fonctions de transfert et fonctions de directivité

Pour un individu, les HRTF se composent d'un ensemble de données constituées de N fonctions de transfert exprimées pour M bins fréquenciels[5]. N désigne ici le nombre de directions pour lesquelles les HRTF ont été acquises. Pour décrire au mieux la sphère 3D entourant l'auditeur, les N directions doivent représenter un échantillonnage homogène de la sphère. En général chaque direction est décrite par un couple d'angles (ϕ_i : angle d'azimut, θ_i : angle d'élévation, i variant de 1 à N), dans un système de coordonnées sphériques soit polaire-interaural, soit polaire-vertical [Guillon, 2007]. Pour un individu, les HRTF comportent donc deux principales dépendances :

- dépendances **fréquencielles** : évolution des données en fonction de la fréquence,
- dépendances **spatiales** : évolution des données en fonction des angles ϕ_i et θ_i.

[4] C'est un peu la même difficulté qu'on rencontre en acoustique des salles. Des salles aux géométries et caractéristiques relativement différentes présentent parfois des qualités très proches, tandis qu'à l'inverse des salles qui "semblent" a priori peu différentes donnent des résultats percetifs très éloignés.

[5] On considère ici des données numériques, c'est à dire complètement discrétisées à la fois dans le domaine des coordonnées d'espace et dans le domaine des fréquences.

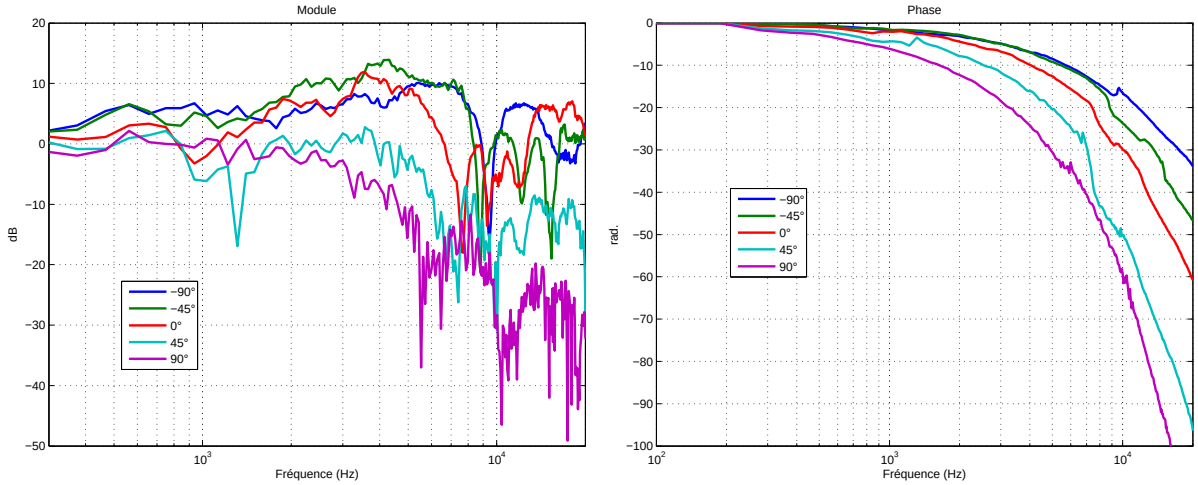


FIG. 3.11 – Illustration des variations fréquencielles des HRTF : Module (à gauche) et phase (à droite) des fonctions de transfert pour 5 directions dans le plan horizontal (ϕ = -90°, -45°, 0°, 45°, 90°). HRTF mesurées pour l'oreille gauche du sujet RN de la base *Jean-Marie Pernaux*. On observe les pics et les creux caractéristiques des résonances du pavillon pour les directions ipsilatérales (ϕ = -90°, -45°, 0°) et l'effet de la diffraction de la tête avec l'atténuation des hautes fréquences pour les directions controlatérales (ϕ = 45°, 90°). Sur la réponse en phase, on remarque que la pente augmente avec l'angle d'azimut, ce qui traduit l'augmentation du temps de propagation et donc du temps d'arrivée de l'impulsion à l'oreille controlatérale.

Par suite, il existe deux façons de considérer les HRTF : soit comme une collection de N fonctions de tranfert (fonctions dépendant de la fréquence, cf. Fig. 3.11 & 3.12), soit comme une collection de M fonctions de directivité (fonctions dépendant des coordonnées d'espace, cf. Fig. 3.13). Les fonctions de directivité sont aussi désignées dans la littérature comme des *Spectral Frequency Response Surfaces* (SFRS) [Cheng & Wakefield, 1999] [Cheng & Wakefield, 2000]. Nous y reviendrons par la suite dans l'analyse des propriétés des HRTF en relation avec les indices de localisation.

Individualité des HRTF

Les données d'HRTF dépendent non seulement de la fréquence et de la direction, mais aussi (et surtout) de l'**individu**. C'est sans doute là leur dépendance fondamentale et la plus critique. Les HRTF sont en effet déterminées pour l'essentiel par l'interaction entre les ondes acoustiques et le corps de l'auditeur. Elles vont donc varier avec sa morphologie, ce qui leur donne leur caractère individuel. La Figure 3.14 illustre les variations inter-individuelles des HRTF mesurées pour 8 individus dans une même direction. On observe des différences marquées d'un individu à l'autre : la fréquence et l'amplitude des pics et des creux sont décalés, leur nombre varie aussi selon l'individu. Ces différences traduisent la spécificité individuelle de l'encodage binaural. En situation d'écoute naturelle, nous construisons par apprentissage le décodeur associé à l'encodeur déterminé par notre morphologie. Ce décodeur présente une certaine **flexibilité**, afin de s'adapter aux évolutions de notre morphologie, ce qui explique que notre capacité à localiser des sons ne soit pas complètement perdue à chaque fois que, par exemple, nous changeons de vêtements ou de coiffure ! Cette flexibilité repose en grande partie sur la plasticité du cerveau en général et du système auditif en particulier. Néanmoins elle reste limitée : lorsque nous sommes confrontés aux signaux générés par l'encodeur d'un autre individu, notre localisation des sons est fortement perturbée [Hofman et al., 1998]. Les

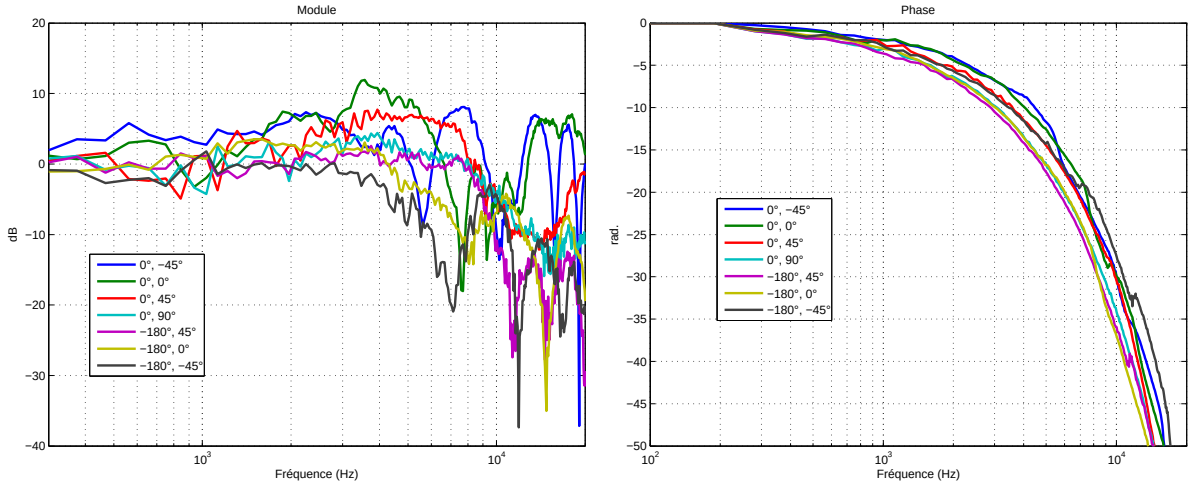


FIG. 3.12 – Illustration des variations fréquencielles des HRTF : Module (à gauche) et phase (à droite) des fonctions de transfert pour 7 directions dans le plan médian, (ϕ,θ) = (0°,-45°), (0°,0°), (0°,45°), (0°,90°), (180°,45°), (180°,0°), (180°,-45°). HRTF mesurées pour l'oreille gauche du sujet RN de la base *Jean-Marie Pernaux*. On observe une variabilité des modules spectraux nettement moindre que dans le plan horizontal, du moins en termes de dynamique. Les directions arrières (180°,45°), (180°,0°), (180°,-45°) subissent l'influence de la diffraction par le pavillon avec une atténuation des hautes fréquences. Les principales évolutions portent sur la fréquence et l'amplitude des pics et des creux. Les courbes en phase varient très peu étant donné que le chemin de propagation ne varie quasiment pas.

principaux artefacts alors observés sont :

- une augmentation des confusions avant/arrière,
- une dégradation de l'externalisation des sources virtuelles se traduisant par une perception intracrânienne des sources,
- une distorsion de la localisation en élévation,
- une perte de la frontalisation correspondant à une difficulté à localiser correctement les sources frontales qui sont localisées en général au-dessus de la tête,

Ces phénomènes surviennent aussi bien dans le cas d'enregistrements binauraux qu'en synthèse binaurale. Avec l'individualisation de l'encodage spatial binaural, c'est toute la crédibilité de la scène virtuelle binaurale qui est donc en jeu. Il a été montré qu'il est possible de créer par synthèse binaurale des sources virtuelles non discriminables de sources réelles [Kulkarni & Colburn, 1998], mais l'individualisation de l'encodage est considéré comme une condition sine qua non pour atteindre ce résultat. Dans cette spécification individuelle de l'encodage binaural, les **indices spectraux** jouent un rôle majeur : le défaut de leur individualisation est la principale origine des artefacts décrits précédemment. Les indices interauraux et la localisation en latéralisation qui y est associée s'avèrent en effet plus robustes à la non-individualisation [Wenzel et al., 1993].

Acquérir des HRTF individuelles

Pour acquérir des HRTF individuelles, une première solution est la mesure acoustique [Pernaux, 2003] [Busson, 2006]. Le système de mesure de HRTF développé au TNO de Soesterberg est illustré sur la Figure 3.15 [Bronkhorst, 1995]. Ce système a été utilisé pour la campagne de mesures de HRTF pour constituer la base *Jean-Marie Pernaux* [Pernaux, 2003]. Même si la me-

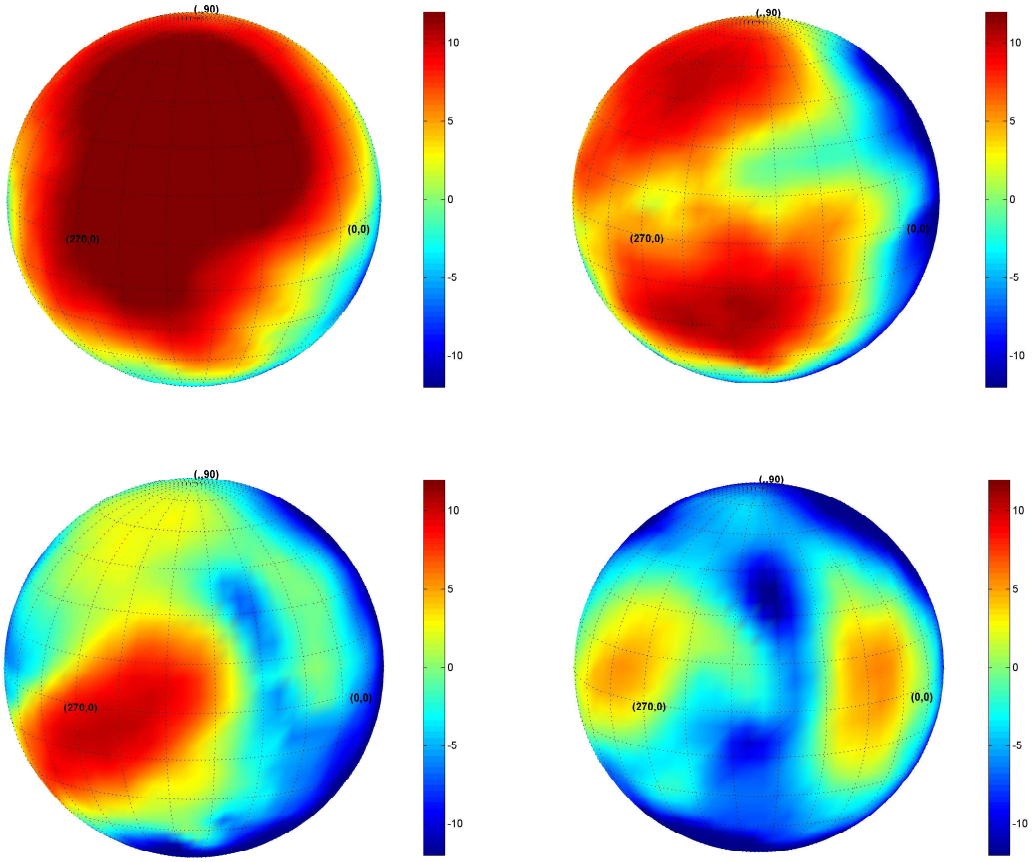


FIG. 3.13 – Illustration des variations spatiales des HRTF : Fonctions de directivité pour 4 fréquences (4875 Hz, 7500 Hz, 9375 Hz et 11812.5 Hz, de haut en bas et de droite à gauche). HRTF mesurées pour l'oreille gauche du sujet ME de la base *Jean-Marie Pernaux*. Pour une fréquence donnée, la fonction de directivité se caractérise par un ou plusieurs maximums qui sont potentiellement exploités par le système auditif pour localiser les sons.

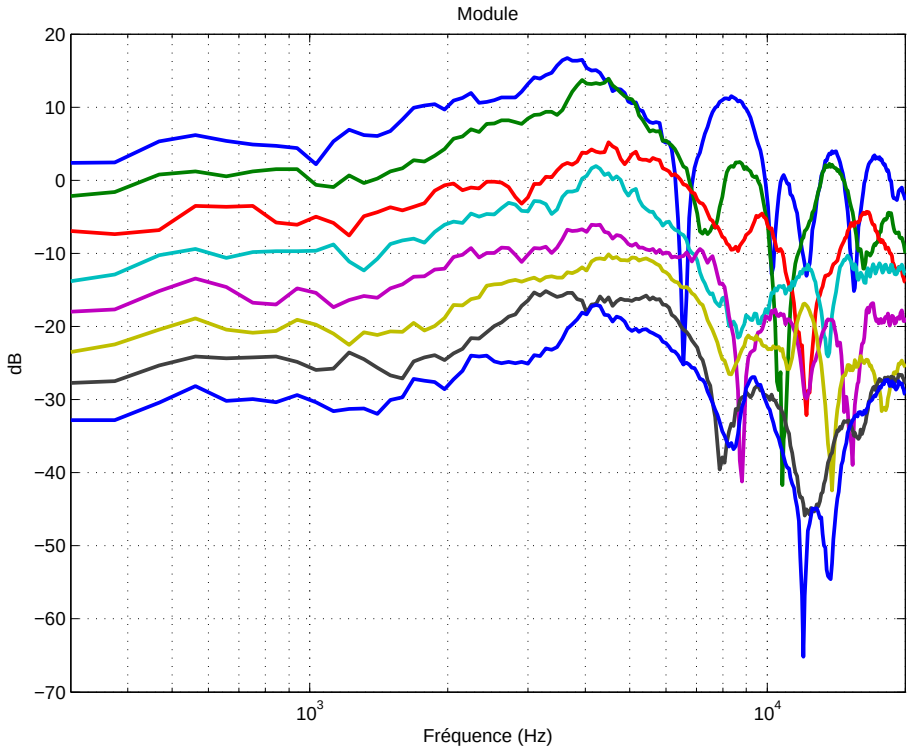
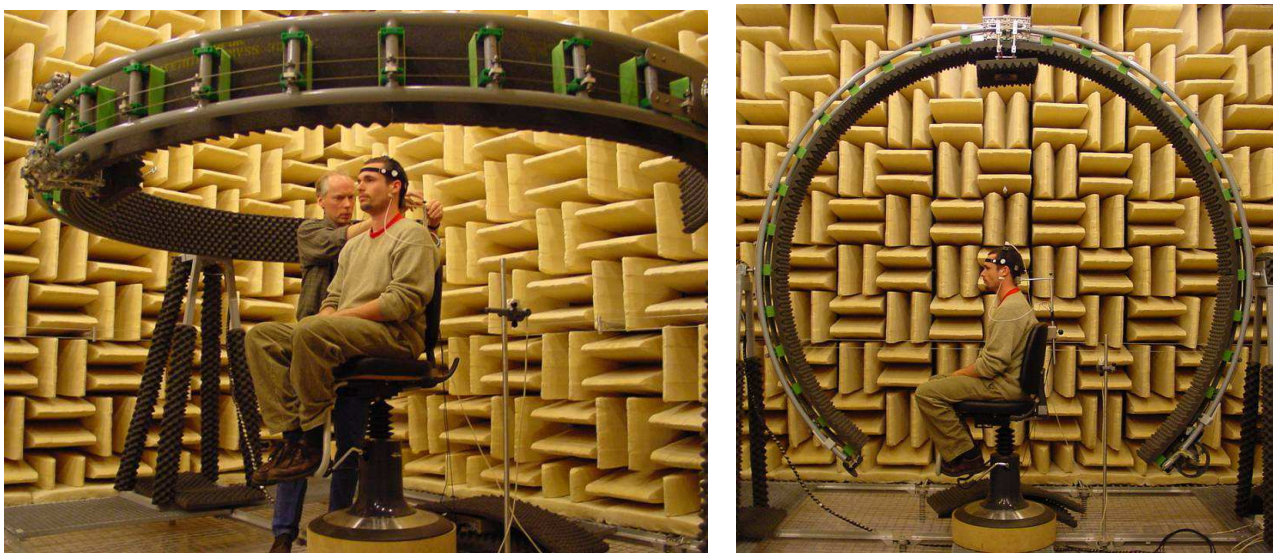


FIG. 3.14 – Illustration des variations inter-individuelles des HRTF (base *Jean-Marie Pernaux*) : Module des fonctions de transfert pour 8 individus dans la même direction (ϕ,θ) = (-45°,0°). Les HRTF ont été décalées de 5 dB pour une meilleure lisibilité des courbes. On observe comment d'un individu à l'autre, le nombre, l'amplitude et la fréquence des pics et des creux diffèrent.



(a) Vue d'ensemble du système : installation du sujet et structure mécanique pour le positionnement du haut-parleur

(b) Système pour le suivi et le contrôle des mouvements du sujet

(c) Microphone de mesure : positionnement du microphone dans un moulage du conduit auditif pour une mesure de type "conduit bloqué"

FIG. 3.15 – Système de mesure de HRTF du TNO utilisé pour l'acquistion de la base *Jean-Marie Pernaux*.

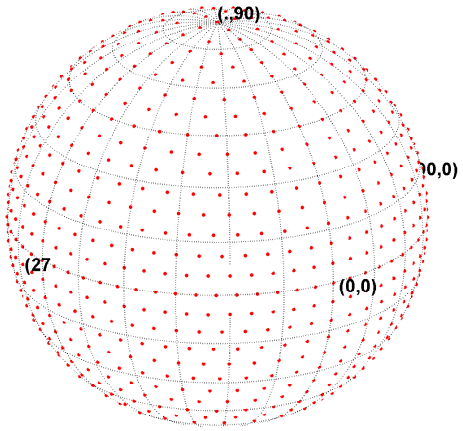


FIG. 3.16 – Visualisation des 965 directions mesurées pour la base *Jean-Marie Pernaux*.

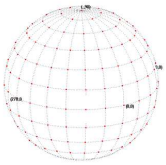
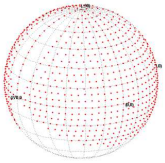
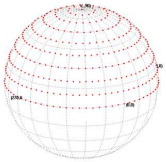
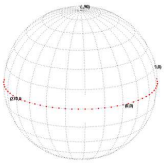
sure de HRTF reste la méthode recommandée pour obtenir des HRTF individuelles, cette méthode souffre de nombreuses contraintes qui en limitent l'utilisation :

- Le dispositif est relativement lourd à mettre en œuvre (cf. Fig. 3.15) et nécessite notamment une chambre anéchoïque.
- la séance de mesure est une réelle "épreuve" pour le sujet qui doit rester immobile le temps que soit collecté l'ensemble des HRTF associées à toutes les directions à mesurer pour couvrir la sphère 3D, ce qui peut représenter une durée totale de plus de deux heures pour un millier de directions [Pernaux, 2003]. Plus récemment des solutions ont été proposées pour réduire la durée de mesure : appliquer le principe de réciprocité [Zotkin et al., 2004], jouer sur les signaux de mesure de façon à émettre simultanément des bandes fréquencielles distinctes qui seront séparées à l'analyse par des techniques d'analyse temps-fréquence [Majdak et al., 2007].

Il existe aujourd'hui un certain nombre de bases de données publiques de HRTF mesurées. Elles sont répertoriées dans le tableau 3.1. La base *Jean-Marie Pernaux* est constituée de 8 sujets décrits par 965 directions (cf. Fig. 3.16) [Pernaux, 2003].

L'alternative à la mesure est la modélisation par éléments finis utilisant une méthode BEM (*Boundary Element Method*) à partir d'un maillage de la morphologie de l'auditeur. Cette solution a été validée pour la première fois par Katz qui a montré qu'il était possible de calculer des HRTF par modélisation BEM pour des fréquences inférieures à 5 kHz [Katz, 1998]. La méthode a été définitivement validée par Kahana qui a obtenu une excellente concordance entre HRTF calculées et HRTF mesurées jusqu'à la fréquence de 15 kHz [Kahana, 2000]. Malgré sa fiabilité, la modélisation des HRTF par éléments finis n'est pas forcément plus facile à mettre en œuvre que la mesure de HRTF, car elle pose les problèmes suivants [Busson, 2006] :

- La modélisation BEM représente un coût de calcul très important qui requiert des calculateurs très performants dès lors qu'on veut une modélisation précise dans les hautes fréquences. En effet la résolution du maillage de la morphologie augmente avec la fréquence désirée. Or, c'est surtout à partir de 5 kHz qu'interviennent les indices spectraux et qu'ils contribuent à l'individualité de l'encodage binaural [Guillon, 2007]. Même si, avec les progrès informatiques, la fréquence limite est chaque jour repoussée, un calcul de HRTF sur toute la bande audible reste encore très délicat.
- L'acquisition des maillages de la morphologie de l'auditeur nécessite un matériel spécifique (scan 3D ou Imagerie à Résonance Magnétique) dont l'utilisation est soumis à des conditions et des compétences restrictives. De plus l'obtention d'une résolution satisfaisante pour les

Propriétaire	Nombre de sujets	Nombre de directions	Lien
IRCAM (France)	51 sujets	187 directions 	`http://recherche.ircam.fr/equipes/salles/listen/`
CIPIC (UC Davis, USA)	45 sujets	1250 directions 	`http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm` [Algazi et al., 2001d]
E. Grassi, University of Maryland (USA)	7 sujets	1093 directions 	`http://www.isr.umd.edu/Labs/NSL/`
Pr. Suzuki, Tohoku University (Japon)	3 sujets	454 directions 	`http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/index.html`
Pr. Itakura, Nagoya University (Japon)	96 sujets	72 directions (plan horizontal) 	`http://www.itakura.nuee.nagoya-u.ac.jp/HRTF/`

TAB. 3.1 – Bases de données publiques de HRTF

calculs dans les hautes fréquences reste une gageure.
Ainsi il apparaît que les méthodes de calcul de HRTF par éléments finis sont des méthodes performantes, mais qui restent limitées aux basses fréquences.

Des HRTF aux filtres binauraux : mise en œuvre de la synthèse binaurale, modèle du filtre à phase minimale associé à un retard pur

La principale utilisation des HRTF est la synthèse binaurale (cf. Fig. 3.6), qu'elle soit statique ou dynamique. La synthèse binaurale consiste à créer une source sonore virtuelle en convoluant le signal source par la paire de HRTF associée à la position à simuler. Pour leur implémentation, les HRTF sont modélisées par des filtres binauraux. Le modèle le plus usuel [Kistler & Wightman, 1992] [Kulkarni et al., 1995] se compose :

- d'un **filtre à phase minimale** qui reproduit le module spectral de la HRTF,
- et d'un **retard pur** qui représente l'information temporelle contenue dans la HRTF.

Soit $H(f)$, la HRTF qu'on souhaite modéliser, le filtre à phase minimale $H_{phasemin}(f)$ associé s'obtient comme suit :

$$\begin{cases} |H_{phasemin}(f)| &= |H(f)| \\ \angle H_{phasemin}(f) &= \Im[TH(-\log(|H(f)|))] \end{cases} \tag{3.4}$$

où TH désigne la Transformée de Hilbert. Le filtre à phase minimale est ainsi uniquement déterminé par le module spectral de la HRTF. Quant au calcul du retard pur, il est obtenu à partir du retard estimé de la HRTF ou de la HRIR. Par commodité on préfère souvent implémenter un seul retard pour une paire de filtres binauraux associés à une direction de l'espace. Ce retard correspond alors à la différence[6] entre les retards des HRTF gauche et droite et il est affecté au filtre du côté controlatéral, tandis que le filtre du côté ipsilatéral est associé à un retard nul.

Il convient cependant de prendre le temps de s'arrêter sur la validité du modèle du filtre à phase minimale et retard pur. Concernant les informations contenues dans le module spectral des HRTF, force est de constater qu'elles sont parfaitement et intégralement restituées par le filtre à phase minimale. La seule approximation introduite par le modèle porte sur la phase spectrale des HRTF, mais elle est de taille. On sait en effet que le phase des HRTF présente des variations fréquencielles et que le retard de phase associé dépend fortement de la fréquence, notamment dans les basses fréquences, ce qui rend la phase d'une HRTF difficile, du moins d'un point de vue purement physique, à modéliser par un retard pur. Sur la base d'une tête modélisée par une sphère et une onde plane incidente, Kuhn [Kuhn, 1977] a étudié la différence de phase (IPD pour *Interaural Phase Difference*) entre les ondes diffractées évaluées au niveau des oreilles gauche et droite. L'ITD se déduit de l'IPD par la relation :

$$ITD = \frac{IPD}{2\pi f} \tag{3.5}$$

L'analyse théorique de l'onde diffractée par la sphère met en évidence une évolution fréquencielle marquée de l'ITD caractérisée par :

- Aux basses fréquences, dans l'hypothèse où $(ka)^2 \ll 1$ (k : nombre d'onde, a : rayon de la sphère), l'ITD est donnée par :

$$ITD_{bf} = \frac{3a}{c}\sin\phi \tag{3.6}$$

On note que cette limite "basses fréquences" ne dépend pas de la fréquence.

[6] On parle alors un peu abusivement d'ITD au lieu de retard, mais pour éviter tout confusion il semble préférable de limiter le terme d'ITD à l'indice de latéralisation.

– Aux hautes fréquences, Kuhn montre que l'onde diffractée revient à une onde atténuée se propageant à une vitesse proche de c. L'ITD s'obtient alors comme la différence de trajet de l'onde en prenant en compte le contournement de la sphère. Il vaut :

$$ITD_{hf} = \frac{a}{c}(\sin\phi + \phi) \tag{3.7}$$

ce qui correspond à la formule de Woodworth [Woodworth & Schlosberg, 1954] qui est ainsi identifiée comme une modélisation "hautes fréquences" de l'ITD. Comme pour le modèle *basses fréquences*, cette valeur ne dépend pas de la fréquence. Pour des angles proches du plan médian, comme :

$$\phi \approx \sin\phi \ ,$$

il vient :

$$ITD_{hf} \approx \frac{2a}{c}\sin\phi = \frac{2}{3}ITD_{bf} \tag{3.8}$$

ce qui révèle un effet d'amplification "basses fréquences" de l'ITD qui théoriquement est augmenté de 150% par rapport à la valeur limite "hautes fréquences".

– Aux fréquences intermédiaires, l'ITD décroît progressivement à partir de ITD_{bf} pour atteindre ITD_{hf}.

Cette évolution théorique est confirmée par la mesure [Kuhn, 1977]. Cette dépendance fréquencielle traduit la dispersion de l'onde acoustique par les phénomènes de diffraction de l'onde acoustique autour de la tête [Constan & Hartmann, 2003]. L'observation de la phase de HRTF mesurées met aussi en évidence ce comportement dispersif. La Figure 3.17 reproduit l'ITD estimée en fonction de la fréquence à partir des retards de phase[7] calculés sur la phase de HRTF issues de la base *Jean-Marie Pernaux*. On vérifie que l'ITD augmente dans les basses fréquences. Des valeurs limites ITD_{bf} et ITD_{hf} peuvent être calculées en moyennant les valeurs de l'ITD obtenue aux basses fréquences [0-500 Hz] et aux hautes fréquences [3-7 kHz], ce qui permet d'évaluer le ratio $\frac{ITD_{bf}}{ITD_{hf}}$ qui, sur ces données, varie entre 1.17 et 1.84, avec de nombreuses valeurs proches de la valeur théorique de 1.5 (cf. Tab. 3.2).

L'observation des signaux physiques établit donc que la phase des HRTF présente des fortes variations fréquencielles, ce qui n'est pas compatible avec une modélisation par un retard pur. Cependant, il n'est pas évident que toute cette information fréquencielle puisse être exploitée par le système auditif. On sait notamment que l'information de phase n'est pas exploitable dans les hautes fréquences, d'une part à cause de l'ambiguité de l'interprétation l'IPD en termes d'ITD au dessus de 1.5 kHz, et d'autre part en raison de l'incapacité du système nerveux central à encoder les différences de phase en dehors des basses fréquences. Il reste néanmoins qu'au vu du comportement de la phase des HRTF même en se restreignant aux basses fréquences, il est impossible de substituer à la phase des HRTF un retard pur sans valider cette modélisation au préalable. Une première validation a été réalisée par un test de localisation comparant les performances de localisation entre les HRTF originales et leur modélisation par un filtre à phase minimale et un retard pur [Kistler & Wightman, 1992]. Il est montré que la modélisation n'affecte pas les performances de localisation en termes d'erreur, de précision et de taux de confusion avant/arrière. Une étude plus fondamentale a été proposée dans [Kulkarni et al., 1999] où la sensibilité du système auditif à l'évolution fréquencielle de la phase est examinée. Il est conclu que remplacer la phase naturelle des HRTF par un retard pur est transparent pour le système auditif qui ne semble donc pas exploiter l'information des variations fréquencielles fines de la phase. Ce résultat est d'autant plus édifiant qu'il est obtenu avec un test de discrimination très sèvère où les auteurs se placent dans les

[7] Si $\psi(f)$ désigne la phase spectrale de la HRTF, le retard de phase τ_{phase} s'obtient comme : $\tau_{phase}(f) = \frac{\psi(f)}{2\pi f}$.

Angle d'azimut (°)	16.875	28.125	45	61.875	73.125	90
Sujet ME	1.75	1.73	1.61	1.54	1.35	1.17
Sujet JMP	1.61	1.67	1.54	1.44	1.33	1.18
Sujet VM	1.70	1.68	1.61	1.43	1.37	1.26
Sujet JD	1.75	1.80	1.67	1.50	1.38	1.29
Sujet RN	1.75	1.66	1.44	1.45	1.39	1.30
Sujet MA	1.84	1.73	1.63	1.50	1.49	1.32
Sujet PA	1.70	1.67	1.51	1.03	1.35	1.19
Sujet NC	1.36	1.61	1.59	1.43	1.45	1.21

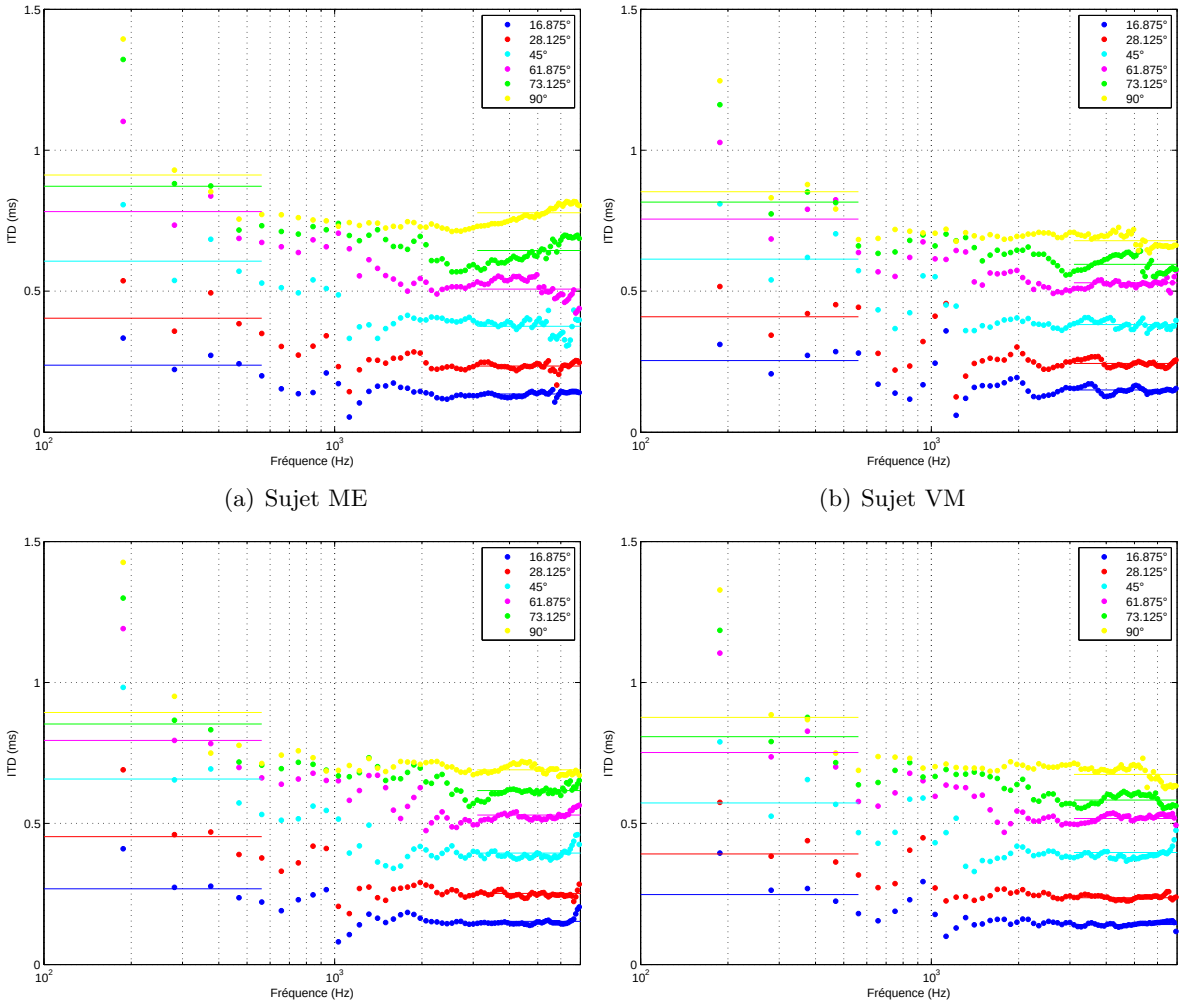
TAB. 3.2 – Rapport entre les valeurs limites ITD_{bf} et ITD_{hf} de l'ITD pour les 8 sujets de la base *Jean-Marie Pernaux*.

conditions les plus critiques. Il est confirmé par une étude plus récente [Constan & Hartmann, 2003] qui montre que le système auditif ne discrimine pas des signaux de bruits présentant une ITD constante de signaux caractérisés par une ITD dépendante de la fréquence.

Maintenant que le modèle d'implémentation est validé dans son principe, il reste une question ouverte : quelle est la valeur du retard pur à appliquer pour garantir la meilleure équivalence entre les HRTF et les filtres binauraux ? C'est le problème de l'estimation du retard des HRTF mesurées. Le problème est double : il s'agit d'une part d'identifier une méthode pour estimer une valeur de retard à partir des fonctions de transfert, et d'autre part de choisir la fréquence à laquelle estimer le retard, étant donnée la dépendance fréquencielle de ce dernier. Concernant cette seconde question, [Kulkarni et al., 1999] émettent une recommandation : il faut s'assurer que les filtres binauraux conservent l'ITD moyenne aux basses fréquences, plus exactement sur la bande [0-2 kHz]. Il est en effet établi par plusieurs études que l'ITD est un indice de latéralisation exploité principalement aux basses fréquences[8] [Wightman & Kistler, 1992] [Macpherson & Middlebrooks, 2002], ce qui justifie de porter l'effort de modélisation sur cette gamme de fréquences. Pour autant faut-il considérer que la valeur limite ITD_{bf} est la valeur de retard pur à associer au filtre à phase minimale ? Wightman et Kistler avancent que l'amplification de près de 150% de l'ITD aux basses fréquences n'est pas pertinente d'un point de vue perceptif [Wightman & Kistler, 1997]. Ils se fondent sur les résultats d'un test de localisation où ils comparent les performances entre les HRTF originales présentant une ITD naturelle (c'est à dire variant naturellement en fonction de la fréquence) et leur implémentation par un filtre à phase minimale et un retard pur estimé par le maximum de la fonction d'intercorrélation [Kistler & Wightman, 1992]. Cependant cette comparaison, du fait de la méthodologie utilisée (offrant une discrimination nettement moins sévère que celle adoptée par Kulkarni *et al*), n'est pas capable de mettre en évidence des différences fines. Néanmoins une étude de Constan et Hartmann [Constan & Hartmann, 2003] confirme que la valeur ITD_{bf} ne possèderait pas d'utilité perceptive. A l'instar de Kulkarni, les auteurs préconisent une valeur moyenne[9] de l'ITD. La question de la **fréquence d'estimation du retard pur** reste donc partiellement posée.

[8] Plusieurs raisons concourent à la prédominance dans les basses fréquences de l'ITD. D'abord les différences de phase ne sont plus exploitables pour les fréquences supérieures à 1.5 kHz en raison de l'ambiguïté de la phase [Mills, 1972]. Ensuite la capacité du système auditif à encoder les différences de phase semblent limitées aux basses fréquences [Palmer & Russell, 1986] [Zwislocki & Feldman, 1956]. Dans les hautes fréquences, même si le système auditif pourrait utiliser les retards d'enveloppe, il apparaît que l'ILD devient l'indice de latéralisation prédominant au détriment de l'ITD [Wightman & Kistler, 1992].

[9] Moyenne pondérée par une fonction proposée par [Raatgever, 1980] correspondant à une gaussienne asymétrique.



(a) Sujet ME

(b) Sujet VM

(c) Sujet JD

(d) Sujet RN

FIG. 3.17 – Evolution de l'ITD en fonction de la fréquence (plan horizontal : azimut variant de 16° à 90°) : L'ITD est estimée à partir des retards de phase calculés sur les phases des HRTF. Les lignes horizontales indiquent les limites basses fréquences ITD_{bf} (estimée comme l'ITD moyenne sur la bande [0-500 Hz]) et hautes fréquences ITD_{hf} (estimée comme l'ITD moyenne sur la bande [3-7 kHz]). Illustration pour 4 sujets de la base *Jean-Marie Pernaux* (sujets ME, VM, JD et RN).

Examinons à présent les différentes méthodes disponibles aujourd'hui pour estimer le retard d'une HRTF[10]. Classiquement on dénombre 4 principales familles [Busson, 2006] :

- **Estimation de la pente de la phase de la composante à excès de phase** : Cette méthode suppose que la phase soit une fonction linéaire de la fréquence, ce qui n'est jamais le cas du moins sur une gamme étendue de fréquences. On effectue donc une régression linéaire en se limitant à une bande de fréquences : par exemple [1 - 5 kHz] [Jot et al., 1995] ou [500 Hz - 2 kHz] [Huopaniemi & Smith, 1999]. Minnaar *et al* montrent cependant que la phase n'est plus linéaire au delà de 1.5 kHz [Minnaar et al., 2000]. Par suite il est pertinent de limiter l'utilisation de cette méthode aux fréquences inférieures à 1.5 kHz.
- **Détection d'un seuil de montée de la HRIR** : On détermine à quel instant la HRIR atteint un seuil donné qui est défini par rapport à la valeur de son maximum. Par exemple ce seuil peut être 10% [Minnaar et al., 2000] ou 20% [Algazi et al., 2001b] du maximum. Cet instant détermine le temps d'arrivée de l'onde acoustique à l'entrée du canal auditif. Pour une meilleure précision, la HRIR peut être suréchantillonnée d'un facteur 8 [Algazi et al., 2001b] ou 10 [Minnaar et al., 2000]. De plus, comme souligné dans [Kulkarni et al., 1999], il convient d'évaluer en parallèle le retard de la composante à phase minimale afin de le retrancher de celui estimé sur la HRIR globale pour obtenir le retard destiné à être associé à la composante à phase minimale. Ainsi ce dernier ne prend bien en compte que le retard correspondant à la composante à excès de phase.
- **Calcul de la fonction d'intercorrélation des HRIR gauche et droite** : Le retard relatif entre les HRIR gauche et droite s'obtient comme le maximum de la fonction d'intercorrélation des enveloppes[11] des HRIR [Kistler & Wightman, 1992]. Il s'agit d'une méthode assez communément utilisée, peut-être parce qu'elle a été proposée dans les premières mises en œuvre du modèle de filtre à phase minimale et retard pur [Kistler & Wightman, 1992]. Minnaar *et al* recommandent d'appliquer un suréchantillonnage d'un facteur 10 [Minnaar et al., 2000]. Comme précédemment, le maximum d'intercorrélation des composantes à phase minimale doit être aussi estimé et soustrait du retard estimé sur les HRIR pour une estimation correcte du retard à associer au filtre à phase minimale [Kulkarni et al., 1999]. Une nouvelle déclinaison de cette méthode a été récemment proposée dans [Nam et al., 2008]. Elle consiste à déterminer le maximum de la fonction d'intercorrélation de la HRIR et de sa composante à phase minimale associée. En d'autres termes, on évalue le retard de la composante à excès de phase séparément pour chaque oreille. L'intérêt de cette approche repose sur l'idée que la HRIR est plus "proche", en termes de similarité du point de vue signal, de sa composante à phase minimale que de la HRIR de l'autre oreille, et ce d'autant plus que la source sonore s'éloigne du plan médian. Cette ressemblance laisse augurer que l'intercorrélation entre les deux fonctions présente un maximum mieux défini, ce qui améliore l'estimation du retard. Les auteurs montrent que cette méthode revient à la méthode d'estimation de la pente de la phase dans lequel on aurait introduit une pondération fréquencielle par le module de la HRTF, ce qui permet de porter l'effort de modélisation sur les bandes fréquencielles riches en énergie. Une première évaluation de cette nouvelle méthode indique un apport significatif de l'estimation du retard par rapport à la méthode simple du maximum de la fonction d'intercorrélation des HRIR gauche et droite, principalement pour les positions fortement latéralisées où, en raison de la diffraction par la tête, la HRIR controlatérale est très atténuée, ce qui

[10] Rappelons que sont décrites dans ce qui suit les méthodes destinées à estimer le retard pur à associer au filtre à phase minimale. Il ne s'agit donc pas à proprement parlé du retard de la HRTF, mais de celui de la *composante à excès de phase* correspondant à la composante résiduelle de la HRTF une fois que la composante à phase minimale est extraite.

[11] Enveloppe au sens défini par Hiranaka et Yamasaki dans [Hiranaka & Yamasaki, 1983].

rend délicate la détection du maximum d'intercorrélation.

– **Estimation du retard de groupe de la composante à excès de phase** [Minnaar et al., 2000] : Plogsties *et al* ont proposé une nouvelle méthode consistant à estimer le retard comme le retard de groupe[12] de la composante à excès de phase évalué à la fréquence de 0 Hz [Plogsties et al., 2000]. Comme dans [Nam et al., 2008], ce nouvel estimateur est motivé par l'échec du modèle de filtre à phase minimale et retard pur pour les positions fortement latéralisées où les HRTF modélisées deviennent discriminables des HRTF originales. Les auteurs montrent qu'il n'est pas discriminé si le retard est estimé par le retard de groupe à O Hz. Cependant, cette solution soulève des difficultés en pratique, du fait que la composante continue des HRTF mesurées est rarement exploitable [Nam et al., 2008]. Une alternative séduisante est avancée dans [Nam et al., 2008] : le retard de groupe est évalué pour l'ensemble des fréquences de la bande [500-2000 Hz], puis pondéré fréquenciellement par le module de la HRTF, avant d'être moyenné sur les fréquences.

Au final on se rend compte qu'on dispose d'une large variété d'estimateurs du retard pur, si l'on combine le choix de la méthode et de la fréquence d'estimation. Le *meilleur* estimateur, au sens de l'estimateur qui donne la valeur de retard offrant la perception des sources virtuelles la plus naturelle et la plus proche de la perception de sources réelles, n'est malheureusement pas clairement identifié. C'est une question que j'ai abordée avec Sylvain Busson dans le cadre de ses travaux de thèse [Busson, 2006]. Cette étude est présentée en Section 3.3.

Des HRTF aux indices de localisation

Ces discussions relatives à l'estimation du retard des HRTF et à l'ITD conduisent à la question du lien entre les HRTF et les indices de localisation. S'il est évident que les HRTF contiennent les indices de localisation, l'extraction des indices de localisation dans l'information fournie par les HRTF est déjà moins claire. Commençons par le plus simple. L'ITD contenue dans les HRTF se calcule comme la différence des retards estimés sur les HRTF gauche et droite (cf. Fig. 3.17). On note que cette ITD dépend de la fréquence. Quand on parle d'ITD, il convient donc de spécifier la fréquence associée, ou s'il s'agit de l'ITD basses fréquences ou hautes fréquences, ou encore s'il s'agit d'une ITD moyenne sur une bande de fréquences à préciser. Quant à l'ILD, elle s'exprime comme la différence des spectres d'énergie associés aux HRTF gauche et droite. L'ILD ainsi obtenue dépend de la fréquence. Larcher a proposé une ILD indépendante de la fréquence en effectuant une intégration sur une bande de fréquences $[f_1 - f_2]$ [Larcher, 2001] :

$$ILD = 10 \log_{10} \left[\frac{\int_{f_1}^{f_2} |H_L(f)|^2 df}{\int_{f_1}^{f_2} |H_R(f)|^2 df} \right] \tag{3.9}$$

En général, l'intégration est réalisée sur la bande de fréquence [1kHz-5kHz] sur laquelle l'ILD joue un rôle perceptif prépondérant.

A présent vient le cas des Indices Spectraux (IS). La première étape est d'abord de les identifier. Qu'est-ce qu'un **indice spectral** de localisation ? Un IS se définit comme une caractéristique spectrale présente dans le module des HRTF et qui est détectée, analysée et utilisée par le système auditif pour localiser les sons (localisation en élévation). L'existence des IS repose sur l'hypothèse d'un *encodage fréquenciel* de l'élévation des sons, c'est à dire que la fréquence véhicule l'information de l'élévation. Il s'agit d'une sorte de représentation qui serait le miroir de la *tonotopie cochléaire* dans laquelle l'information de fréquence est encodée par la localisation du maximum de

[12] Le retard de groupe $\tau_{groupe}(f)$ à la fréquence f se définit comme la dérivée de la phase des fonctions de transfert par rapport à la fréquence : $\tau_{groupe}(f) = \frac{d\psi(f)}{2\pi df}$.

résonance sur la membrane basilaire [Leipp, 1997]. Ici c'est la fréquence qui, à son tour, représente l'information de localisation. Le jeu des réflexions, diffractions et résonances engendrées par la morphologie de l'auditeur contribue à élaborer une cartographie (en anglais *mapping*) fréquencielle[13] de l'espace sonore. Cependant, pour que cet encodage fréquenciel soit efficace et pertinent, il faut qu'il soit le moins ambigu possible, c'est à dire que deux directions différentes soient bien représentées par deux caractéristiques fréquencielles distinctes de façon univoque. Le module des HRTF contient une grande richesse d'informations : parmi ces informations, quelle(s) caractéristique(s) en particulier constitue(nt) des IS ? En termes de support fréquenciel, tout d'abord, les IS sont principalement situés dans la bande [4 - 16 kHz] [Hebrank & Wright, 1974] [King & Oldfield, 1997] [Langendijk & Bronkhorst, 2002]. Il existe des IS aux fréquences inférieures à 4 kHz (IS engendrés par la tête et le torse), mais leur rôle est jugé secondaire [Han, 1994] [Brown & Duda, 1998] [Langendijk & Bronkhorst, 2002] [Morimoto et al., 2003]. Dans l'identification des IS, deux théories s'affrontent :

- **Spectre global** des HRTF : C'est le spectre dans son intégralité et sa continuité fréquencielle qui est exploité par le système auditif pour localiser les sons [Middlebrooks, 1992] [Opstal & Esch, 2003]. Le jugement de localisation se fonderait sur une comparaison (de type corrélation spectrale par exemple) entre le spectre perçu de la source et celui des HRTF stockées en mémoire.
- **Caractéristiques locales** du spectre des HRTF : Ce sont les accidents du spectre, tels que les **creux** ou les **pics**, qui seraient utilisés par le système auditif [Bloom, 1977] [Blauert, 1970].

Tant que les mécanismes par lesquels le système auditif interprête les IS ne sont pas mieux connus, il est difficile en l'état actuel des connaissances de privilégier l'une ou l'autre des théories. Des expériences menées en laboratoire prouvent que ces deux catégories d'IS sont susceptibles d'être interprêtées en termes d'élévation perçue. Il est d'ailleurs probable qu'en situation naturelle d'écoute, le spectre global et les caractéristiques locales sont utilisés de façon conjointe et complémentaire. Le rôle des caractéristiques locales est beaucoup mis en avant par la littérature. Concernant les creux, l'examen des HRTF indique que certains creux spectraux présentent la propriété intéressante que la fréquence du creux augmente de façon monotone en fonction de l'élévation [Guillon, 2007], ce qui en fait un candidat valide au titre d'IS. Et en effet, des expériences d'illusion auditive montrent qu'en manipulant la fréquence de creux spectraux introduits dans des stimuli de bruit, on est capable de modifier et de contrôler la perception de l'élévation de la source sonore [Bloom, 1977] [Watkins, 1978] [Hebrank & Wright, 1974] [Iida & Itoh, 2006]. Dans les tests de localisation, l'élévation perçue est ainsi fortement corrélée à la fréquence du creux, indépendamment de la position physique de la source. Cependant il semble que les pics spectraux soient mieux détectés que les creux [Moore et al., 1989], ce qui en ferait des IS plus fiables. Comme pour les creux, des illusions auditives, comme par exemple la fameuse expérience des bandes directives de Blauert [Blauert, 1970] témoignent que la perception de l'élévation peut être contrôlée par la fréquence des pics [Hebrank & Wright, 1974] [Butler & Helwig, 1983]. Toutefois, quand on parle de pics, il convient de distinguer les **maxima fréquenciels** (*overt peaks*) et les **maxima spatiaux** (*covert peaks*) [Butler, 1987] [Butler et al., 1990]. Les premiers correspondent au maximum de la fonction de transfert dans une direction donnée (cf. Fig. 3.11 ou 3.12), tandis que les seconds se définissent comme le maximum de la fonction de directivité à une fréquence donnée (cf. Fig. 3.13). Butler *et al.* [Butler et al., 1990] montrent que les jugements de localisation suivent fidèlement les maxima spatiaux, ce qui suggère que ces derniers sont plus pertinents du point de la localisation Les maxima

[13] Cependant des études neurophysiologiques récentes suggèrent qu'il existerait au niveau central une cartographie *spatiale* de l'espace, au sens où chaque neurone répond spécifiquement à une direction privilégiée de l'espace [Sterbing et al., 2003]. De plus, ces neurones sont regroupés par affinité spatiale selon une organisation topographique. Le jugement de localisation se baserait ainsi sur la distribution spatiale de l'activité neuronale [Campbell et al., 2006].

spatiaux ou *covert peaks* se visualisent sous la forme de CPA (Covert Peak Area) qui regroupent les directions de l'espace pour lesquelles la fonction de directivité est égale à l'amplitude du maximum à 1 dB près. L'évolution des CPA en fonction de la fréquence suit une trajectoire caractéristique (cf. Fig. 3.18) [Guillon, 2007] : migration d'avant en arrière (avec ou sans montée) de 4 à 7-8 kHz, saut discontinu vers le bas entre 7 et 9 kHz, migration le plus souvent vers l'arrière et vers le haut jusqu'à 12 kHz, saut discontinu vers l'avant autour de 12-13 kHz et migration vers le haut et/ou l'arrière jusqu'à 16 kHz. Les discontinuités (*breakpoint* en anglais) semblent liées au caractère dipolaire des résonances du pavillon [Shaw & Teranishi, 1968]. Même si l'on observe des différences d'un individu à l'autre (cf. Fig. 3.18), il est remarquable que, quels que soient le sujet et son individualité, la trajectoire suive une loi sensiblement identique [Cheng & Wakefield, 1999]. Ce trait commun transverse à l'individualité de la localisation auditive constitue un premier élément de réponse encourageant pour l'individualisation de la synthèse binaurale.

Les IS soulèvent une dernière question : dans l'analyse des IS par le système auditif, les IS issus des deux oreilles sont-ils traités indépendamment (traitement *monaural*) ou conjointement (traitement *binaural*) ? Dans l'hypothèse d'un traitement binaural, il reste aussi à déterminer si ce sont les IS gauche et droit qui sont exploités de façon conjointe, mais sur la base de leur valeurs absolues, ou si c'est leur différence qui compte (notion d'ISD pour *Interaural Spectral Difference*) ? Le rôle de l'ISD semble à écarter : outre que les différences interaurales d'IS ne présentent pas les propriétés qui feraient d'elles de bons candidats comme indice de localisation, car elles restent faibles et ne varient pas de façon monotone avec l'élévation [Middlebrooks et al., 1989] [Musicant et al., 1990] [Carlile & Pralong, 1994], des tests de localisation montrent que les ISD seules ne suffisent à préserver des performances de localisation en élévation [Wightman & Kistler, 1997] [Macpherson, 1996] [Macpherson, 1998] [Rakerd, 1999]. Ce sont donc bien sur les spectres gauche et droit considérés en absolu que se fonde le jugement de localisation. Il s'agit ainsi d'indices monauraux. Cependant les deux oreilles ne contribuent pas de façon équivalente. L'oreille ipsilatérale joue un rôle prédominant, du moins lorsque la source s'écarte de plus de 30° du plan médian [Morimoto, 2001] [Hofman & Van Opstal, 2003]. Néanmoins la contribution de l'oreille controlatérale ne peut être négligée pour les positions situées autour du plan médian. Elle pourrait même suppléer à l'oreille ipsilatérale lorsque cette dernière ne fournit pas d'indices exploitables [Musicant & Butler, 1984] [Humanski & Butler, 1988]. Le traitement des IS est ainsi à la fois monaural et binaural selon la position de la source sonore.

3.1.4 Axe de recherche : Quels filtres binauraux pour un espace auditif virtuel de qualité ?

A partir de maintenant nous nous focaliserons sur le cas de la *synthèse binaurale*. L'objectif est de créer un espace auditif virtuel convaincant qui replace l'auditeur dans les conditions d'une écoute naturelle, avec notamment une perception spatialisée des sources sonores, car c'est cet aspect que nous avons choisi d'étudier. La question fondamentale qu'il convient alors de se poser est la suivante : comment spécifier et obtenir les filtres binauraux qui garantissent ce résultat (ou, plus raisonnablement, permettent de s'en rapprocher au mieux) ? L'état de l'art des recherches sur la synthèse binaurale qui viet d'être présenté, permet d'isoler les points qui ne font (quasiment) plus débat des questions encore ouvertes. La première condition est d'effectuer avec le plus grand soin toutes les étapes de traitement associées à l'encodage et au décodage. Dans l'ensemble, on sait comment procéder : modélisation par un filtre à phase minimale associé à un retard pur des filtres binauraux à partir des HRTF, égalisation individuelle du casque etc... En revanche il demeure un problème fondamental qui n'est pas encore résolu : *comment préserver, dans le cadre d'une synthèse binaurale, les caractéristiques individuelles de l'encodage binaural* ? En d'autres

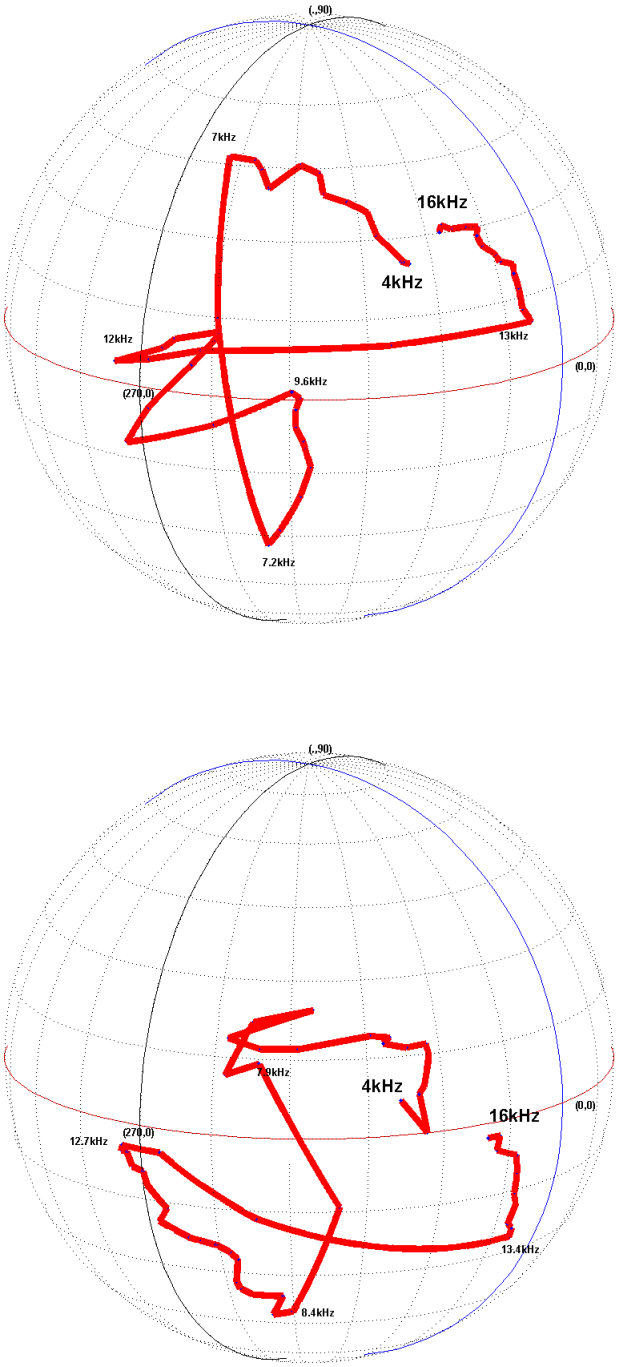


FIG. 3.18 – Visualisation des maxima spatiaux : trajectoire des CPA pour les sujets ME (en haut) et RN (en bas) de la base *Jean-Marie Pernaux*.

termes, comment proposer à n'importe quel individu une synthèse binaurale individuelle, c'est à dire reproduisant les spécificités de l'empreinte de sa morphologie dans son encodage spatial ? C'est à cette question que se sont principalement attachés mes travaux de recherche, notamment au travers de l'encadrement des travaux de thèses de Jean-Marie Pernaux, Sylvain Busson et Pierre Guillon [Pernaux, 2003] [Busson, 2006] [Guillon, 2009]. Comme nous avons choisi de placer ce problème dans un contexte d'applications destinées au grand public, une contrainte supplémentaire est à intégrer : nous recherchons des solutions d'individualisation simples à mettre en œuvre, ce qui exclut les méthodes classiques de mesure ou de modélisation par éléments finis, du moins dans leur intégralité.

La question connexe à la modélisation de HRTF individuelles est l'évaluation de la qualité de cette modélisation, évaluation qui peut être abordée d'un double point de vue :

- **objectif** : Les performances de modélisation peuvent être évaluées par une mesure de (dis)similarité (par exemple un critère d'erreur) entre les HRTF individuelles mesurées au préalable et les HRTF modélisées. On fait ici l'hypothèse que les HRTF mesurées sur l'individu constituent la référence à atteindre, ce qui suppose que le processus de mesure est parfait. Cependant il faut bien avoir conscience que cette mesure objective de (dis)similarité n'apporte qu'une information partielle tant qu'elle n'est pas étalonnée sur la base de la perception. Pour être correctement interprétable, il faut savoir pour une valeur donnée de similarité observée entre 2 HRTF si elle est discriminable ou non par le système auditif. Cet ancrage perceptif des mesures de (dis)similarité peut être réalisé soit par des tests perceptifs permettant de relier les valeurs objectives aux jugements subjectifs des sujets, soit en interprétant les valeurs objectives en termes des seuils de discrimination ou JND (*Just Noticeable Difference*).
- **subjectif** : L'alternative est de donner à "écouter" les HRTF modélisées à l'auditeur pour le laisser juger de leur adéquation à son espace perceptif. La méthodologie classique d'évaluation subjective repose sur des tests de localisation auditive qui apportent une information certes utile, mais qui mériterait d'être précisée et complétée par de nouvelles approches.

Un dernier point mérite une attention particulière : l'**externalisation** des sources virtuelles restituées par les techniques binaurales. Parmi les défauts reprochés aux technologies binaurales, si le non respect des timbres (détimbrage) lié à la nature de l'encodage binaural basé justement des indices spectraux est perçu comme l'aspect plus gênant, la perception intra-crânienne des sons est aussi un point souvent relevé. Ces défauts constituent les principales raisons freinant l'acceptation des technologies binaurales dans des contextes en dehors des activités de recherche (ingénieurs du son, grand public). Le défaut d'externalisation est un phénomène bien spécifique caractérisé en ce qu'il concerne principalement la perception des sources frontales et qu'il se définit comme une distorsion de la localisation des sources virtuelles qui sont localisées non pas devant l'auditeur, mais d'une part très proches ou à l'intérieur de la tête (*in-head localization*), en termes de distance, et d'autre part au dessus ou derrière la tête, en termes de direction. Ce phénomène ne doit pas être confondu avec un défaut de perception des indices de distance [Moore, 2009]. Il existe de nombreuses causes possibles provoquant l'apparition de ce phénomène [Moore, 2009] :

- une incohérence de l'ensemble des indices de localisation [Hartmann & Wittenberg, 1996], telle qu'une contradiction entre ITD et ILD ou une ILD nulle,
- la similarité entre les signaux des oreilles gauche et droite de l'auditeur [Toole, 1970] [Blauert, 1983] [Brookes & Treble, 2005],
- l'utilisation de HRTF non individuelles,
- l'absence d'indices de localisation dynamique dans le cas du mode binaural statique,
- l'influence des indices cognitifs, liés notamment à l'absence d'indices visuels associés à la perception de la source sonore, au port du casque d'écoute qui est susceptible de biaiser la perception de l'auditeur et de restreindre la scène virtuelle à l'espace compris entre les deux

écouteurs, ainsi qu'au degré de familiarité de l'auditeur avec des signaux binauraux et une écoute spatialisée.

A l'inverse, on observe que le défaut d'externalisation peut être corrigé par :

- l'ajout d'un effet de salle [Kendall, 1995],
- le mode binaural dynamique [Ianaga et al., 1995] [Begault et al., 2001] [Faure, 2005].

Cependant le problème n'est pas aussi simple qu'il pourrait le paraître. On a vu que les facteurs déclenchant ou corrigeant le défaut d'externalisation sont multiples. De plus ils ne sont pas systématiques et leur effet reste pour la plupart très controversé. Les phénomènes dépendent aussi largement des individus. S'ajoute la difficulté de juger et quantifier l'externalisation des sources virtuelles. Ces questions ont fait l'objet d'un travail de collaboration entre Orange Labs et l'Université de York et a constitué le travail de thèse d'A. H. Moore [Moore, 2009]. A l'origine, le projet visait l'identification des **indices d'externalisation** permettant de personnaliser des HRTF génériques à un auditeur donné en corrigeant le défaut d'externalisation des sources frontales. L'étude a montré que la recherche d'indices spécifiques et clairement identifiés était vaine. En revanche, la question fondamentale qui a émergé et sur laquelle se sont focalisés les travaux a porté sur l'évaluation du degré d'externalisation perçu par l'auditeur : pour corriger un phénomène, encore faut-il correctement le mesurer ! Un protocole basé sur un test de discrimination offrant une comparaison directe entre une source réelle et une source virtuelle est apparu comme la meilleure solution, plutôt que de demander au sujet de quantifier directement l'externalisation. Comme le défaut d'externalisation se rencontre aussi sur des sources réelles, l'objectif ici n'est plus de créer des sources virtuelles parfaitement externalisées dans l'absolu, mais aussi bien externalisées que des sources réelles. Les HRTF sont décrites sous la forme d'une représentation paramétrique (par exemple une représentation de type ACP) et leur adaptation consiste à ajuster les paramètres pour obtenir des sources virtuelles non discriminables de sources réelles. Ces travaux ne sont pas détaillés dans la suite du mémoire.

Les sections qui suivent décrivent mes travaux sur l'ensemble des ces problématiques. Ces travaux s'organisent autour de 2 thèmes principaux regroupant 7 études :

- retard pur associé au filtre à phase minimale pour l'implémentation des filtres binauraux en synthèse binaurale :
  1. mesure du seuil de discrimination du retard [Busson, 2006],
  2. évaluation perceptive des estimateurs de retard proposés dans la littérature [Busson, 2006],
  3. proposition, mise en œuvre et validation d'un modèle d'ITD individualisée basée sur une tête sphérique avec individualisation du rayon de la sphère et du positionnement des oreilles [Busson, 2006],
- modélisation des IS individuels :
  4. modélisation de HRTF par éléments finis utilisant des géométries simplifiées de la morphologie tête-cou-torse [Pernaux, 2003],
  5. modélisation de HRTF par des réseaux de neurones [Busson, 2006],
  6. adaptation morphologique de HRTF non-individuelles par une transformation combinant une homothétie sur l'axe fréquenciel et une rotation du système de coordonnées [Guillon, 2009],
  7. reconstruction individuelle de HRTF à partir d'un faible nombre de directions mesurées, en utilisant apprentissage et reconnaissance de formes pour exploiter l'information d'une base de données comportant un large nombre d'individus et de directions [Guillon, 2009].

3.2 Mesure du seuil de discrimination de l'ITD en contexte de synthèse binaurale

3.2.1 Motivations

Le retard à implémenter en complément du filtre à phase minimale est obtenu soit par **estimation** dans les HRTF mesurées, soit par **modélisation**. Ces questions seront traitées dans les deux sections qui suivent. Cependant, quelle que soit la méthode de calcul du retard, elle comporte un biais inévitable. La question préalable, pour déterminer l'impact de cette erreur, est donc de connaître le seuil de discrimination du système auditif en termes d'ITD, dans le but de savoir si l'erreur d'estimation ou de modélisation est perceptible ou non.

La JND (*Just Noticeable Difference*) de l'ITD définit la plus petite différence d'ITD qui soit perçue par le système auditif. Plusieurs études l'ont mesurée [Klumpp & Eady, 1956] [Domnitz, 1973] [Hershkowitz & Durlach, 1969] [Hafter & Maio, 1975] [Domnitz & Colburn, 1977]. Cependant, dans toutes ces études, l'ITD est considérée dans un contexte de pure latéralisation, c'est à dire un contexte d'écoute dichotique relativement *artificiel* (au sens *écologique* du terme) où les deux oreilles de l'auditeur sont excitées par des signaux qui ne diffèrent que par une différence de temps (ITD) [Klumpp & Eady, 1956] [Hershkowitz & Durlach, 1969] [Hafter & Maio, 1975], éventuellement combinée à une différence d'intensité (ILD) indépendante de la fréquence [Domnitz, 1973] [Domnitz & Colburn, 1977]. Les différentes études s'accordent sur la valeur minimale de la JND qui vaut de l'ordre de 10 μs pour une ITD nulle lorsque le stimulus est un bruit large bande. Mais le seuil de discrimination de l'ITD dépend de plusieurs facteurs :

- la valeur de l'ITD : le JND augmente pour atteindre respectivement 29 et 50 μs quand l'ITD passe à 430 et 790 μs [Klumpp & Eady, 1956],
- la valeur de l'ILD associée : les effets d'interaction avec l'ILD restent faibles tant que ITD et ILD se renforcent en termes de latéralisation, mais la JND de l'ITD augmente s'ils s'opposent [Domnitz, 1973],
- le type de stimulus (nature du signal et bande passante) : par exemple, pour un son pur de 90 Hz, on observe un seuil de 75 μs pour une ITD de 0 μs [Klumpp & Eady, 1956].

Dans le cadre de nos travaux, les stimuli appliqués aux oreilles de l'auditeur se rapprochent d'une situation d'écoute *naturelle*. Par rapport aux études décrites précédemment, ces stimuli comportent non seulement une ITD, mais aussi un filtrage fréquenciel différent pour chaque oreille, correspondant à la paire de filtres à phase minimale censés reproduire les IS. La question posée est alors la suivante : la présence de ce filtrage vient-elle interagir avec la résolution de l'ITD, c'est à dire modifier la valeur de la JND ? Cette question est d'autant plus légitime que l'on sait que la JND de l'ITD dépend du type de stimulus sonore, dont notamment son spectre [Klumpp & Eady, 1956]. Cette possible interaction mérite d'être étudiée par une mesure de la JND de l'ITD en contexte de synthèse binaurale. C'est l'objectif de l'expérience décrite à présent.

3.2.2 Dispositif expérimental

On veut donc mesurer la plus petite différence d'ITD discriminable par le système auditif plongé dans un espace auditif virtuel créé par synthèse binaurale. Les stimuli sonores[14] sont obtenus en convoluant un bruit blanc de type gaussien par une paire de filtres binauraux correspondant à la synthèse d'une source virtuelle dans une direction donnée de l'espace. Chaque filtre binaural est constitué d'un filtre à phase minimale et d'un retard pur. Le retard pur associé à l'ITD est appliqué

[14]Le lecteur est invité à se reporter au document de thèse de Sylvain Bussson [Busson, 2006] pour plus de détails sur l'expérience et son protocole.

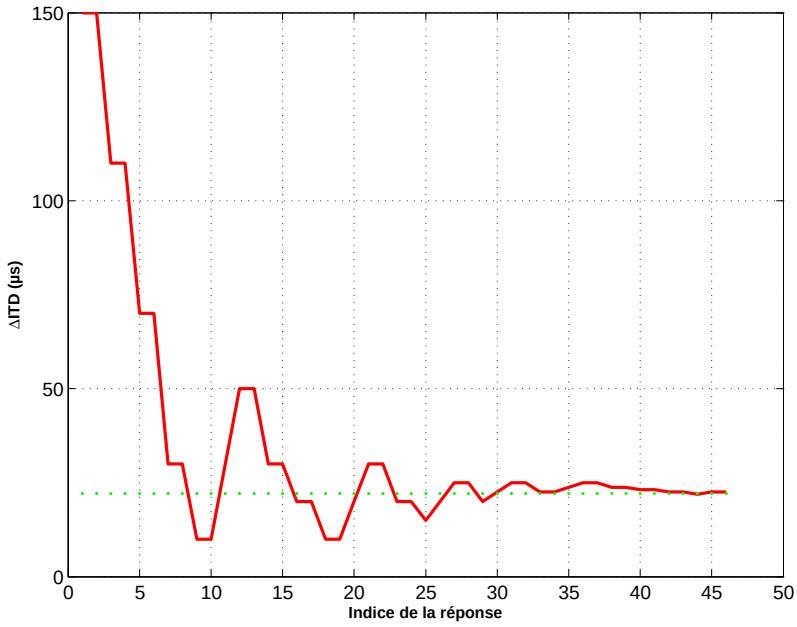


FIG. 3.19 – Mesure de la JND de l'ITD par une méthode adaptative de type *two-down, one-up* [Levitt, 1971] : Oscillations de la différence ΔITD (maximum de 6 oscillations) et estimation du JND (droite horizontale en pointillé vert).

dans le domaine fréquenciel afin d'offrir un contrôle aussi fin que possible des valeurs de retard. Le protocole expérimental utilise une méthode adaptative [Levitt, 1971] avec une procédure de type 2I-2AFC [Marvit et al., 2003]. On fait entendre au sujet deux stimuli qui ne diffèrent que par la valeur de l'ITD :

- Pour un des stimuli, l'ITD correspond à l'ITD estimée à partir des HRTF par une méthode de détection de seuil [Algazi et al., 2001b]. Cette ITD définit ce que nous appellerons dans la suite l'ITD de référence ITD_{ref}.
- Pour l'autre stimulus, l'ITD appliquée (ITD_{var}) correspond à la valeur de l'ITD de référence modifiée d'une valeur additionnelle ΔITD, soit : $ITD_{var} = ITD_{ref} + \Delta ITD$. Cette variation peut être une augmentation ou une diminution de l'ITD de référence.

La tâche du sujet consiste à identifier le stimulus localisé le plus à gauche, comme dans l'expérience de Klumpp & Eady [Klumpp & Eady, 1956]. Cette tâche n'est donc pas une simple discrimination, mais fait intervenir un jugement de localisation. Le principe de la méthode adaptative est le suivant : la valeur de la différence ΔITD entre les stimuli est initialisée à $\Delta ITD = 150\mu s$. Tant que le sujet identifie correctement le stimulus le plus à gauche, cette différence est diminuée. En revanche, à chaque réponse erronée, la valeur de ΔITD est augmentée. La différence d'ITD suit ainsi une succession d'oscillations (cf. Fig. 3.19) autour de la valeur correspondant au seuil de discrimination, c'est à dire la JND de l'ITD. Le pas de variation de ΔITD s'affine au cours des oscillations pour décroître de 40 μs (première oscillation) à 1 μs (dernière oscillation). La JND est estimée comme la moyenne des valeurs médianes des 4 dernières remontées (*mid-run estimate*) [Levitt, 1971] (cf. Fig. 3.19), les premières oscillations n'étant pas prises en compte car elles sont fortement sensibles aux erreurs initiales, en raison notamment du pas élevé de variation (40 et 20 μs). Cette estimation correspond à un pourcentage de 70.7% de réponses correctes de la fonction psychométrique selon [Levitt, 1971].

Indice d'élévation	Azimut 1	Azimut 2	Azimut 3
1	(0°, -51°)	(23°, -65°)	(58°, -62°)
2	(0°, 0°)	(22.5°, 0°)	(56°, 0°)
3	(0°, 45°)	(23°, 50°)	(58°, 33°)
4	(0°, 90°)	(22.5°, 90°)	(56°, 90°)
5	(0°, 135°)	(23°, 130°)	(58°, 147°)
6	(0°, 180°)	(22.5°, 180°)	(56°, 180°)
7	(0°, 231°)	(23°, 245°)	(58°, 242°)

TAB. 3.3 – Positions des sources virtuelles de mesure de la JND de l'ITD (coordonnées polaires interaurales) : couple des azimut et élévation pour les 3 plans d'azimut constant.

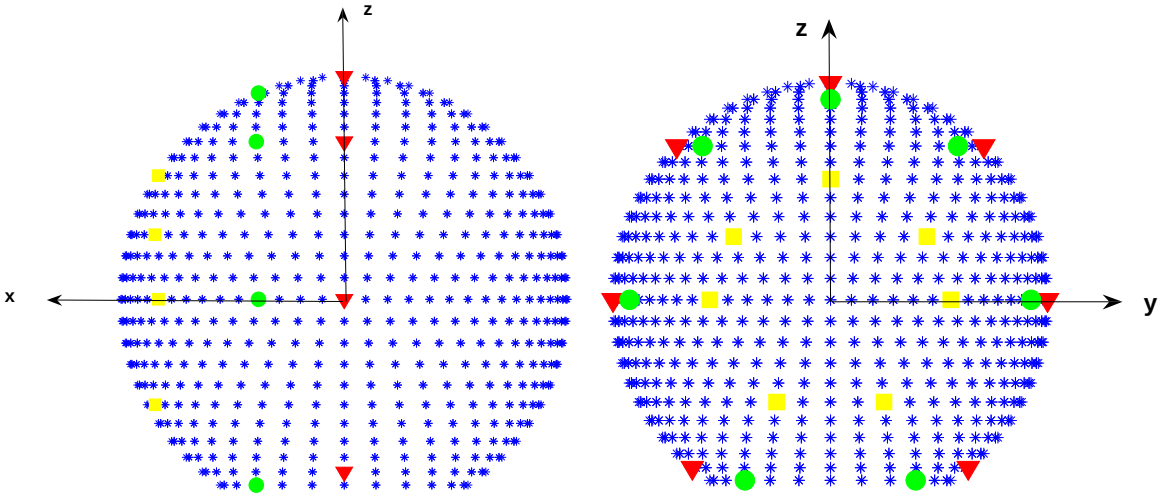


FIG. 3.20 – Sélection des positions des sources virtuelles de mesure de la JND de l'ITD parmi les directions mesurées (*) pour la base *Jean-Marie Pernaux* : vue de face et vue de côté. Les points associés aux plans d'azimut $\phi = 0$, 22.5° et 56,25° sont respectivement représentés par les triangles rouges, les cercles verts et les carrés jaunes.

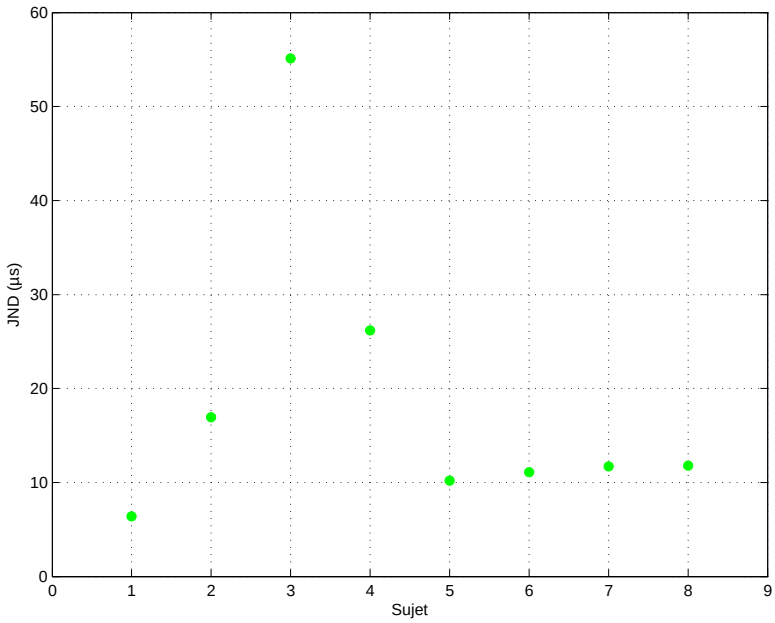


FIG. 3.21 – Expérience de contrôle : Mesure de la JND de l'ITD en fonction du sujet pour un ITD_{ref} nul sans filtre à phase minimale.

Huit sujets ont participé à l'expérience : 4 sujets avec leurs HRTF individuelles et 4 sujets avec des HRTF non individuelles, ces dernières étant les HRTF d'un bon localisateur (sujet ME de la base *Jean-Marie Pernaux*) identifié dans un précédent test de localisation [Pernaux, 2003]. Toutes les HRTF sont issues de la base *Jean-Marie Pernaux*. Le seul paramètre expérimental est la position de la source virtuelle synthétisée : la JND de l'ITD est mesurée pour un total de 21 positions qui se répartissent sur 3 plans d'azimut constant correspondant aux azimuts[15] $\phi = 0°$ (plan médian), 22.5° et 56.25°. Tous ces plans sont situés dans l'hémisphère droit[16]. Pour chaque plan d'azimut, 7 élévations sont considérées et couvrent l'étendue du plan vertical. Les coordonnées exactes des positions sont indiquées dans le tableau 3.3 (cf. Fig. 3.20).

3.2.3 Expérience de contrôle

Afin de valider le protocole expérimental, une expérience de contrôle est réalisée en préliminaire de l'expérience principale. Dans cette expérience de contrôle, la JND de l'ITD est mesurée dans un contexte identique aux études antérieures, au sens où aucun filtre à phase minimale n'est appliqué. Les stimuli sonores sont générés en ne contrôlant que le retard interaural. L'objectif est de valider le protocole de l'expérience en vérifiant que les seuils obtenus dans cette condition sont conformes aux valeurs reportées dans la littérature. Pour cette validation, le seuil n'est mesuré que pour une seule valeur d'ITD_{ref} : $ITD_{ref} = 0\mu s$. Les JND mesurées pour les 8 sujets sont reproduits sur la Figure 3.21. Cinq sujets présentent un seuil proche de 10 μs, ce qui est en parfait accord avec les résultats de la littérature [Klumpp & Eady, 1956]. Cependant, pour les trois autres sujets, on observe un

[15] En coordonnées polaires interaurales.

[16] Certaines études suggèrent une possible asymétrie des performances de localisation auditive entre les deux hémisphères [Sonoda et al., 2001] [Savel et al., 2006]. En toute rigueur, les valeurs de JND mesurées dans l'hémisphère droit ne peuvent donc être extrapolées à l'hémisphère gauche. Des positions dans l'hémisphère gauche n'ont pas été mesurées afin que la durée de l'expérience reste raisonnable, sachant qu'en se limitant à 21 positions l'expérience dure déjà de l'ordre de 5 heures.

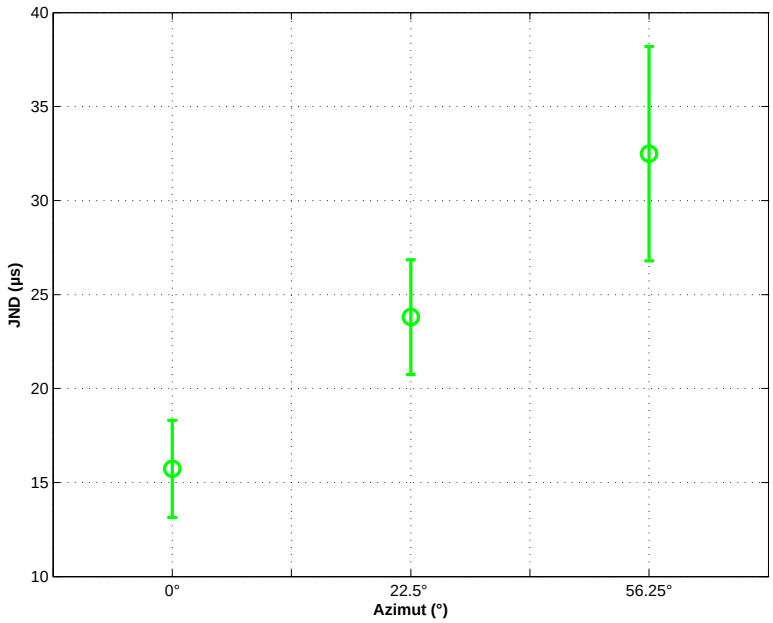


FIG. 3.22 – Mesure de la JND de l'ITD en fonction de l'azimut de la source virtuelle : moyenne sur les sujets et les élévations.

seuil sensiblement plus élevé : 17, 26 et 55 μs, ce qui dénote une forte variabilité inter-individuelle qui a déjà été signalée dans la littérature [Domnitz, 1973].

3.2.4 Expérience principale

Passons à présent à l'expérience principale. La valeur moyenne des JND de l'ITD collectées pour l'ensemble des positions vaut 20.21 μs avec un intervalle de confiance à 95% de $\pm$ 5.66 μs. On observe que la JND augmente avec l'azimut de la source virtuelle (cf. Fig. 3.22) : le seuil moyen vaut 15.72 μs dans le plan d'azimut 0° et s'élève à 32.5 μs dans le plan d'azimut 56.25°. Ce résultat est en accord avec les observations de Klumpp & Eady [Klumpp & Eady, 1956] qui montrent effectivement une augmentation de la JND avec l'ITD de référence. En revanche la JND ne semble pas dépendre de l'élévation (cf. Fig. 3.23), du moins pas de manière significative. Il en ressort que la discrimination de l'ITD ne semble pas affectée par les filtres à phase minimale et les modifications spectrales qu'ils engendrent. Comme dans l'expérience de contrôle, on note une forte variabilité inter-individuelle (cf. Fig. 3.24). On isole deux groupes de sujets : les sujets 1, 5, 6, 7 et 8 se caractérisent par une JND faible et une variance modérée, tandis que les sujets 2, 3 et 4 présentent des seuils notablement plus élevés avec une variance importante. On retrouve bien les tendances observées lors de l'expérience de contrôle (cf. Fig. 3.21). L'ensemble des jugements a été analysé selon une procédure d'ANOVA (*ANalyse Of VAriance*). Le calcul de l'ANOVA s'est basé sur :

- deux facteurs expérimentaux : l'azimut et l'élévation de la source virtuelle,
- un facteur aléatoire : le sujet.

L'ANOVA confirme que seuls l'azimut et le sujet ont un effet significatif (cf. Tab. 3.4). Pour la suite de l'étude, nous considérerons les valeurs de JND reportées dans [Klumpp & Eady, 1956], ce qui signifie que seule la dépendance en azimut sera prise en compte. La dépendance individuelle n'est pas reproduite, ce qui serait d'ailleurs impossible étant donné que les bases de données de HRTF utilisées ne disposent pas de données de JND.

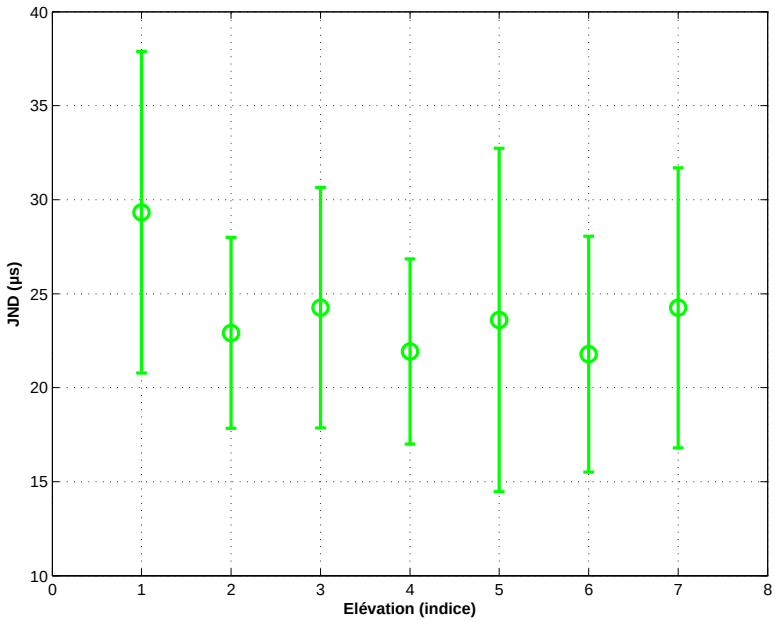


FIG. 3.23 – Mesure de la JND de l'ITD en fonction de l'élévation de la source virtuelle : moyenne sur les sujets et les azimuts.

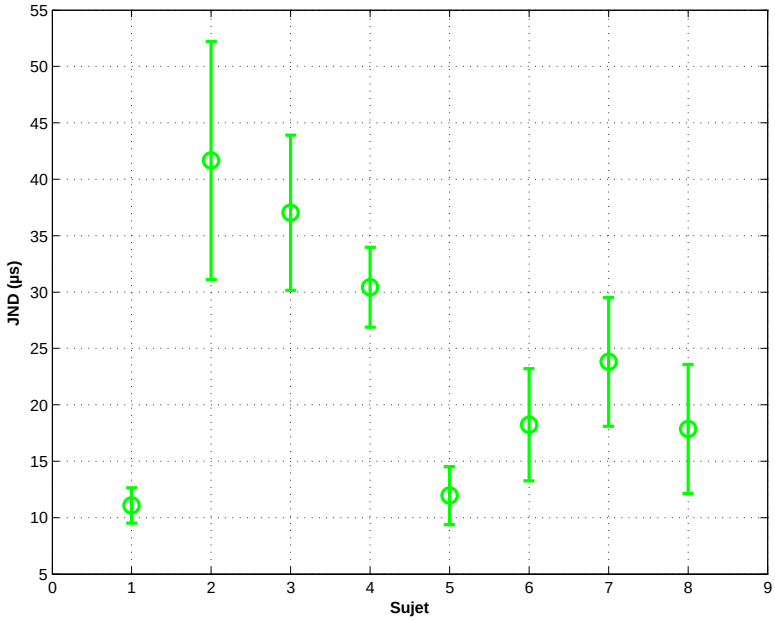


FIG. 3.24 – Mesure de la JND de l'ITD en fonction du sujet : moyenne sur toutes les positions.

Facteur expérimental	Somme des carrés	Degrés de liberté	Moyenne des carrés	F	p
Sujet	19028.1	7	2718.3	9.62	0.0002
Azimut	7882.5	2	3941.24	14.02	0.0005
Elévation	936.1	6	156.02	1.73	0.1387
Sujet x Azimut	3934.3	14	281.02	3.17	0.0005
Sujet x Elévation	3796.9	42	90.4	1.02	0.4602
Azimut x Elevation	2089.3	12	174.11	1.96	0.038
Erreur	7452.3	84	88.72		

TAB. 3.4 – Mesure de la JND de l'ITD : Résultats de l'ANOVA.

3.3 Evaluation perceptive des estimateurs de retard des HRTF

Nous avons vu qu'il existe différentes méthodes pour estimer le retard contenu dans les HRTF en vue de leur implémentation en synthèse binaurale. La question qu'on veut traiter ici est d'évaluer ces méthodes d'estimation et de déterminer quelle est la meilleure. Il faut d'abord préciser ce que signifie "meilleur estimateur". La finalité des estimateurs de retard est de fournir la valeur de retard qui, associée au filtre à phase minimale, donnera une perception, principalement en termes de spatialisation, équivalente à une écoute naturelle. Partant de là, le "meilleur estimateur" pour notre étude se définit comme l'estimateur fournissant la valeur de retard garantissant cette équivalence perceptive, ou si l'équivalence ne peut être obtenue, s'en rapprochant au mieux. La première étape consiste donc à quantifier cette cible perceptive. Dans ce but, une expérience a été menée afin de déterminer la valeur de retard telle que le filtre binaural résultant de l'association de ce retard au filtre à phase minimale soit perceptivement équivalent à la HRTF brute implémentée directement. En d'autres termes, cette expérience fournit une estimation *"perceptive"* du retard, à laquelle nous opposerons les estimations *"mathématiques"* (c'est à dire basées sur une analyse des HRIR ou des HRTF) que nous désirons évaluer. La seconde étape vise à comparer les différentes valeurs données par les estimateurs mathématiques à la valeur obtenue par l'estimateur perceptif. L'estimateur jugé le meilleur est alors celui qui fournit la valeur la plus proche de l'estimation perceptive. Cette section commence par rappeler les motivations de l'expérience, en illustrant la divergence des méthodes mathématiques dans leur estimation du retard. Le protocole expérimental est ensuite brièvement décrit, avant de présenter les résultats, avec notamment l'évaluation comparée des différents estimateurs.

3.3.1 Divergence des estimateurs mathématiques du retard

Il s'agit ici de bien poser le problème. Avant de mener l'expérience, il convient de quantifier l'ampleur du phénomène : quel est l'ordre de grandeur des écarts d'estimation entre les différentes méthodes ? Ces écarts sont d'abord à mettre en relation avec les seuils perceptifs de discrimination de l'ITD définissant la tolérance du système auditif aux potentielles altérations de l'ITD. Ces écarts seront aussi comparés aux variations observées d'un individu à l'autre (variations inter-individuelles de l'ITD).

Six méthodes d'estimation du retard, représentatives de l'éventail des estimateurs proposés à ce jour, sont évaluées[17] :

[17]Ne figure pas dans les méthodes sélectionnées pour cette évaluation, l'estimateur proposé par Minnaar *et al* [Minnaar et al., 2000] basé sur le retard de groupe évalué à 0 Hz. Il s'avère en effet que cette méthode est de mise en oeuvre limitée, étant donné que la fonction de transfert est rarement exploitable pour la composante continue,

Méthode 1 **Estimation du retard de phase moyen dans les basses fréquences [Kulkarni et al., 1999]** : Cette méthode n'est pas identifiée à proprement parler comme un estimateur de retard, mais elle repose sur la recommandation émise par Kulkarni *et al* [Kulkarni et al., 1999] (cf. page 141). Elle consiste à calculer le retard de phase de la composante à excès de phase et à le moyenner sur la plage de fréquences [0 - 2 kHz] :

$$\tau_{L,R} = \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} \tau_{phase(L,R)}(f) df \text{ où : } f_1 = 0 \text{ Hz}, f_2 = 2 \text{ kHz}. \tag{3.10}$$

Méthode 2 **Estimation de la pente de la phase de la composante à excès de phase** : Les paramètres d'estimation sont les bornes fréquencielles de la plage de fréquences sur laquelle est estimée la pente de la phase. Pour notre étude, ces bornes ont été fixées à [500 - 2000 Hz], conformément aux résultats de la littérature [Minnaar et al., 2000].

Méthode 3 **Détection du seuil de montée de la HRIR** : Le principal paramètre d'estimation est la valeur du seuil. Dans notre cas, un seuil de 20% du maximum de la HRIR est choisi. Aucun suréchantillonnage des HRIR n'est effectué. De plus, il est apparu que la détection de seuil échouait pour les HRIR controlatérales en raison d'un rapport signal à bruit insuffisant. Ce défaut a été corrigé en appliquant une fenêtre de Hann de longueur $N_W = 31$ échantillons et centrée sur la maximum de la HRIR, le reste de la réponse impulsionnelle étant mis à zéro, afin d'aider le travail de détection de seuil et de le focaliser sur la partie "utile" de la réponse impulsionnelle en éliminant les segments ne contenant que du bruit. Le retard pris en compte par les composantes à phase minimale est également estimé par détection de seuil pour être retranché au retard obtenu pour les HRIR, ce qui donne le retard des composantes à excès de phase.

Méthode 4 **Identification du maximum de la fonction d'intercorrélation des HRIR gauche et droite** : La différence de retard entre les oreilles gauche et droite s'obtient simplement comme le maximum de la fonction d'intercorrélation des enveloppes des HRIR gauche et droite. Les HRIR sont suréchantillonnées au préalable, d'un facteur de 8 dans notre cas. C'est le seul paramètre d'estimation à ajuster de cette méthode. Comme précédemment, les retard des composantes à phase minimale sont aussi estimés en détectant le maximum d'intercorrélation afin de les soustraire à celui des HRIR et de se ramener au retard des composantes à excès de phase.

Méthode 5 **Identification du maximum de la fonction d'intercorrélation de la HRIR et de sa composante à phase minimale** : Cet estimateur est basé sur la nouvelle méthode proposé par Nam [Nam et al., 2008] consistant à estimer le retard de la HRIR (ou plus exactement de sa composante à excès de phase) en calculant son intercorrélation avec sa composante à phase minimale (cf. page 143). Les HRIR sont suréchantillonnées d'un facteur 8.

Méthode 6 **Estimation du retard de groupe moyen dans les basses fréquences** : Dans leur étude, Nam *et al* montrent que leur nouvelle méthode revient à estimer la pente de la phase en appliquant une pondération fréquencielle par le module spectral de la HRTF associée [Nam et al., 2008]. En résulte la proposition d'une méthode alternative consistant à estimer le retard de groupe de la composante à excès de phase et à le moyenner sur la plage de fréquences [500 - 2000 Hz]. Au préalable, le retard est pondéré fréquenciellement par le

comme l'a remarqué Nam *et al* [Nam et al., 2008].

module spectral de la HRTF associée.

$$\tau_{L,R} = \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} \tau_{groupe(L,R)}(f) w(f) df \text{ où : } f_1 = 500 \text{ Hz}, f_2 = 2 \text{ kHz}$$
$$\text{avec : } w(f) = \frac{A^2(f)}{\int_{f_1}^{f_2} A^2(f) df} \tag{3.11}$$

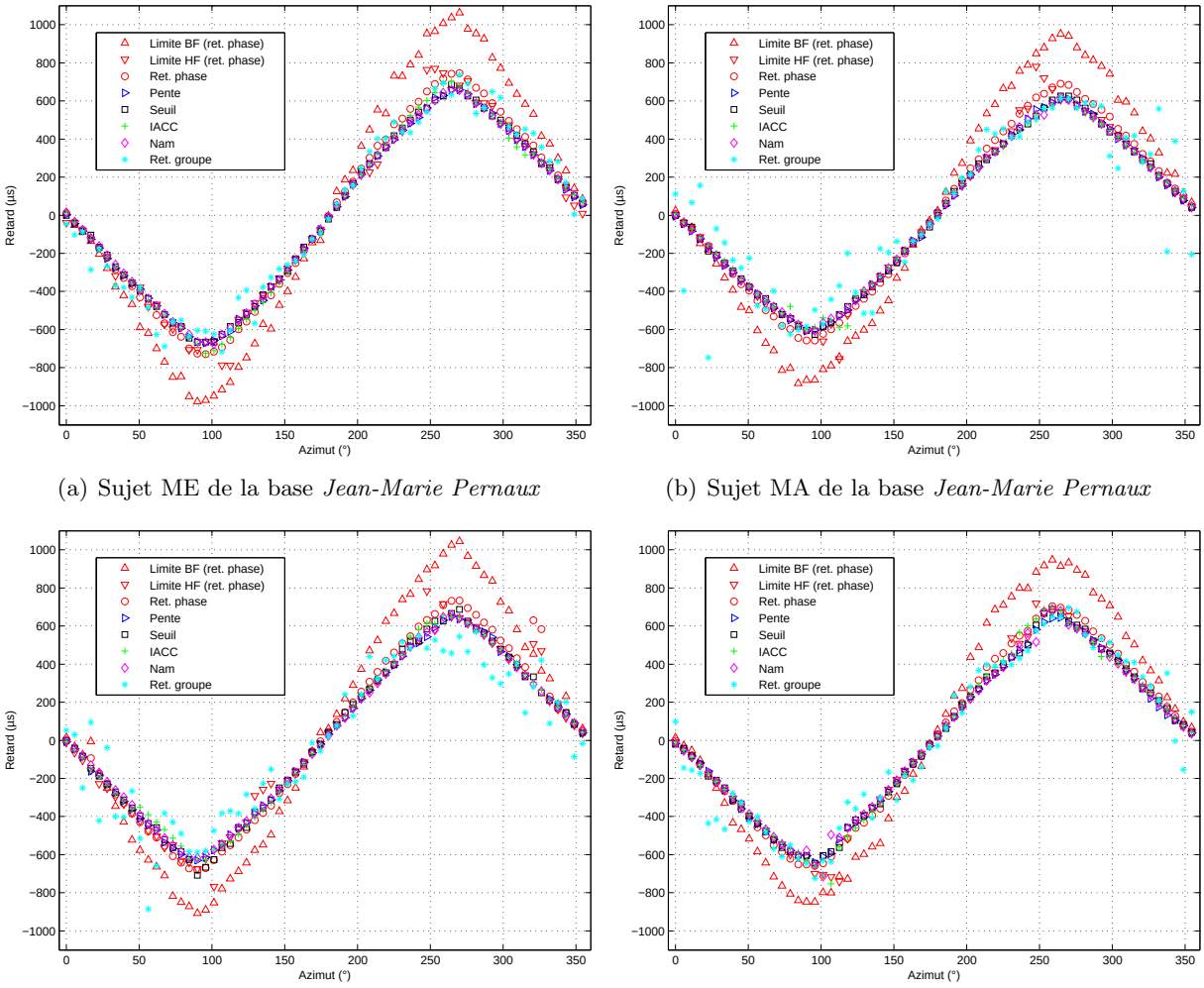
Pour évaluer ces différents estimateurs, 5 base de données de HRTF ont été considérées : il s'agit des bases d'Orange Labs, de l'IRCAM, du CIPIC, de l'Université du Maryland (E. Grassi) (cf. Tab. 3.1) et de F.L. Wightman [Wightman & Kistler, 1989a] [Wightman & Kistler, 1989b]. Nous avons souhaité prendre en compte différentes bases de données, afin de tester la robustesse des estimateurs selon la méthodologie de mesure des HRTF. Au total, 112[18] individus ont été considérés, ce qui a permis de collecter une quantité importante de statistiques. Pour chaque individu, les retards ont été estimés dans le plan horizontal, en couvrant tous les azimuts compris entre 0 et 360° selon les données disponibles dans chaque base. Les retards estimés pour 4 sujets de la base *Jean-Marie Pernaux* sont illustrés sur la Figure 3.25. Dans l'ensemble, on observe une relative bonne convergence des estimateurs. Se démarquent la méthode du retard de phase qui a tendance à surestimer le retard et celle du retard de groupe qui présente une forte divergence. Les résultats pour les autres bases de données sont représentés sur les figures 3.26, 3.27 et 3.28. Le comportement des estimateurs est très similaire à celui obtenu sur la base *Jean-Marie Pernaux*. La méthode du retard de groupe dénote par son instabilité et s'écarte parfois assez fortement des autres estimateurs. Pour les sujets de la base d'E. Grassi, la méthode de retard de phase tend à sousestimer le retard contrairement à ce qui a été observé pour la base *Jean-Marie Pernaux*.

Pour la base *Jean-Marie Pernaux* sont également reproduites les valeurs limites basses et hautes fréquences (cf. Fig. 3.25) estimées à partir du retard de phase et correspondant aux ancres mises en évidence par les travaux de Kuhn [Kuhn, 1977]. On vérifie l'allongement du retard aux basses fréquences. Il en ressort aussi que tous les estimateurs donnent des valeurs proches de la limite hautes fréquences, et peuvent donc être considérés comme des estimateurs *hautes fréquences* au sens des résultats de Kuhn. Il s'avère ainsi que l'augmentation de l'ITD aux très basses fréquences mise en évidence par Kuhn n'est probablement jamais prise en compte en synthèse binaurale, du moins dans les implémentations de type filtre à phase minimale et retard pur. Cette conclusion avait déjà été tirée par Wightman et Kistler [Wightman & Kistler, 1997].

Il est à noter que les paramètres d'estimation sont identiques pour toutes les bases de données, c'est à dire qu'il n'a pas été nécessaire de les adapter en fonction des données. La relative bonne concordance des estimateurs est dans ces conditions assez remarquable. La principale difficulté rencontrée concerne l'estimation du retard de phase. Le retard de phase s'obtient en divisant la phase des HRTF par la fréquence, mais cette opération requiert que la phase soit déroulée. Classiquement le déroulement de la phase consiste, à partir d'une phase comprise entre $-\pi$ et π, à supprimer les sauts de phase en considérant que les différences de phase supérieures à un seuil donné[19] correspondent à un saut et sont réduites en ajoutant $\pm 2\pi$. Cependant, dans le cas de données mesurées, le bruit de mesure est susceptible d'introduire des accidents de phase qui peuvent être considérés à tort comme des sauts. Inversement, si le temps d'arrivée est long, la phase oscille tellement rapidement en fonction de la fréquence que les sauts de phase n'atteignent pas le seuil suffisant pour être détecté. Le problème s'est posé pour la base de l'IRCAM (cf. Fig. 3.29). La Figure 3.29a

[18] Chaque base a été utilisée en intégralité, à l'exception de la base de l'IRCAM dans laquelle les sujets référencés 09, 34, 44 et 55 ont été écartés car l'observation des données a révélé des artefacts de mesure, ne permettant pas une exploitation fiable.

[19] Par exemple pour la routine Matlab *unwrap*, le seuil par défaut vaut π.



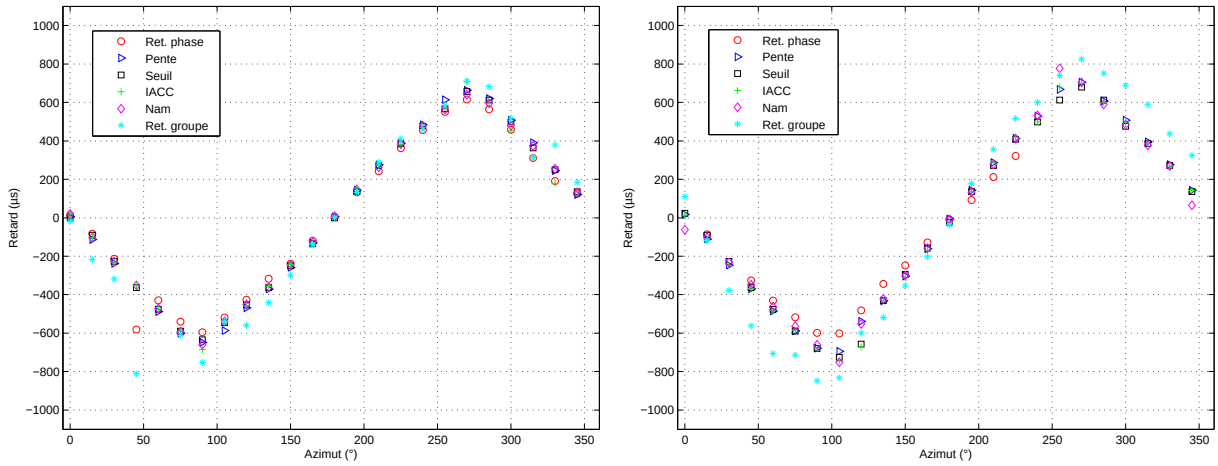
(a) Sujet ME de la base *Jean-Marie Pernaux*

(b) Sujet MA de la base *Jean-Marie Pernaux*

(c) Sujet NC de la base *Jean-Marie Pernaux*

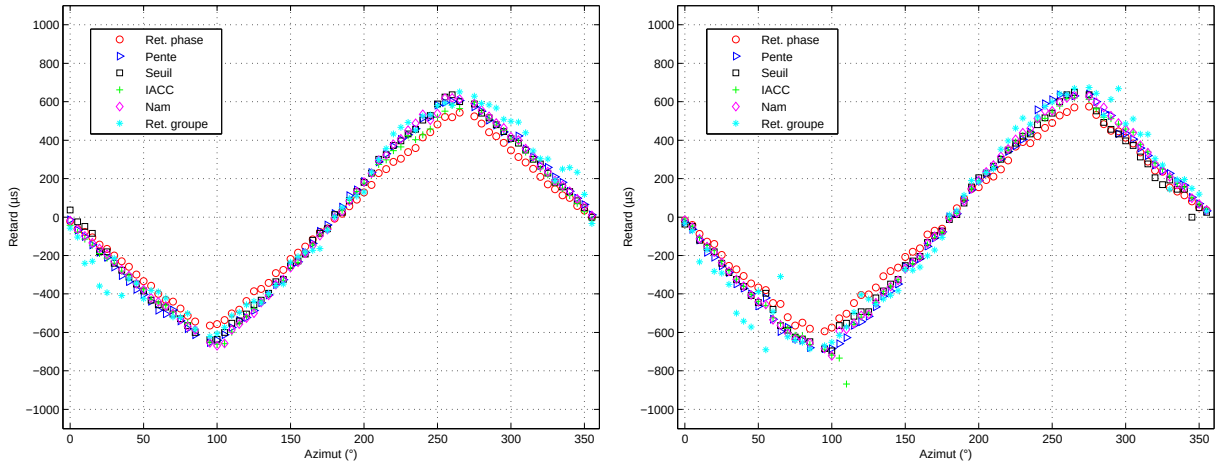
(d) Sujet VM de la base *Jean-Marie Pernaux*

FIG. 3.25 – Estimation des retards pour 4 sujets de la base *Jean-Marie Pernaux* par les 6 méthodes sélectionnées (”Ret. phase” : estimation du retard de phase moyen dans les basses fréquences, ”Pente” : estimation de la pente de la phase de la composante à excès de phase, ”Seuil” : détection du seuil de montée de la HRIR, ”IACC” : identification du maximum de la fonction d’intercorrélation des HRIR gauche et droite, ”Nam” : identification du maximum de la fonction d’intercorrélation de la HRIR et de sa composante à phase minimale, ”Ret. groupe” : estimation du retard de groupe moyen dans les basses fréquences). Sont aussi présentées les valeurs limites du retard de phase moyen pour les basses (”Limite BF (ret. phase)”) et les hautes (”Limite HF (ret. phase)”) fréquences de la composante à excès de phase (cf. page 140). Dans tous les cas, la valeur de retard représente en fait la différence entre les retards gauche et droit des composantes à excès de phase.



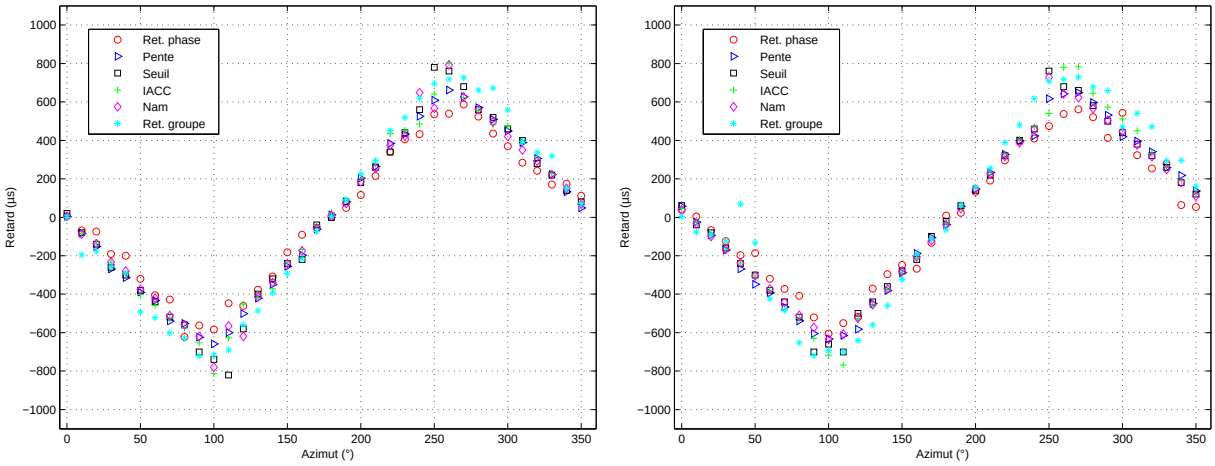
(a) Sujet 07 de la base de l'IRCAM (b) Sujet 55 de la base de l'IRCAM

FIG. 3.26 – Estimation des retards pour 2 sujets de la base de l'IRCAM par les 6 méthodes sélectionnées (se reporter à la Figure 3.25 pour plus de détails).



(a) Sujet EG de la base d'E. Grassi (b) Sujet TH de la base d'E. Grassi

FIG. 3.27 – Estimation des retards pour 2 sujets de la base d'E. Grassi par les 6 méthodes sélectionnées (se reporter à la Figure 3.25 pour plus de détails).



(a) Sujet SJX de la base de F.L. Wightman

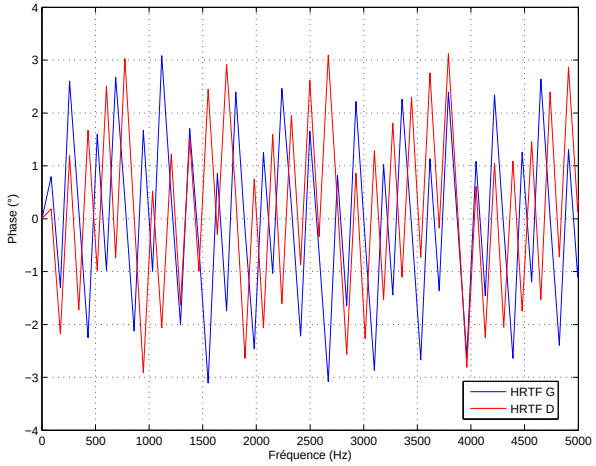
(b) Sujet SOU de la base de F.L. Wightman

FIG. 3.28 – Estimation des retards pour 2 sujets de F.L. Kistler par les 6 méthodes sélectionnées (se reporter à la Figure 3.25 pour plus de détails).

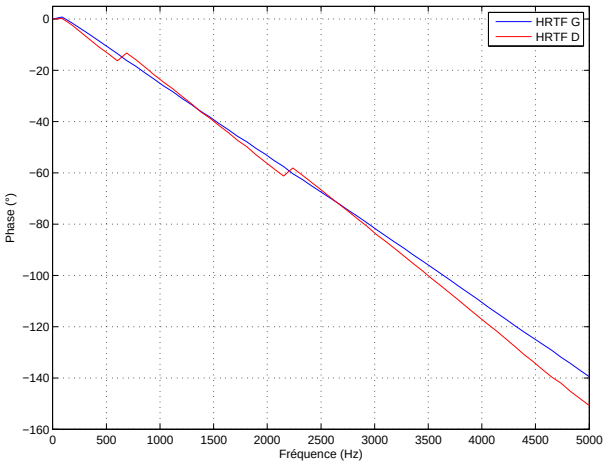
illustre la phase non déroulée des HRTF gauche et droite. On observe un enchaînement très rapide de sauts dont l'amplitude est souvent inférieure à π. Il en résulte que la phase déroulée de la HRTF droite (cf. Fig. 3.29b) présente encore des sauts, ce qui rend impossible une estimation fiable du retard entre les HRTF gauche et droite à partir de l'information de phase. Pour résoudre ce problème, nous avons appliqué une technique de *zero padding* sur les HRIR avant de calculer la HRTF, afin d'améliorer la résolution fréquencielle et ainsi augmenter l'amplitude des sauts de phase pour favoriser leur détection. L'effet du *zero padding* est illustré sur la Figure 3.29c. On vérifie que l'amplitude des sauts et bien amplifiée. Il en résulte sur la Figure 3.29d que le déroulement de la phase est beaucoup plus satisfaisant. Il reste qu'il est impossible de déterminer avec certitude si une rupture de phase est véritablement un saut de phase ou une "évolution naturelle" de la phase. La phase n'est connue qu'à 2π près, et cette contrainte vient amoindrir la fiabilité de l'information portée par la phase déroulée dont l'évolution fréquencielle ne peut jamais être connue exactement. Ce constat fragilise tous les estimateurs basés sur la phase, principalement celui utilisant le retard de phase, mais aussi celui consistant à estimer la pente de la phase. L'expérience montre que des 2 estimateurs, l'estimation du retard de phase est la plus problématique, notamment dans les basses fréquences, en raison de la division par la fréquence qui tend à amplifier les erreurs sur la phase. Il convient aussi de noter que la phase considérée pour l'estimation du retard est celle de la composante à excès de phase. Or, on observe que la composante à phase minimale peut avoir une phase relativement accidentée. Lorsqu'elle est soustraite à la phase originelle de la HRTF, elle introduit ainsi des irrégularités qui viennent en quelque sorte troubler l'évolution fréquencielle de l'excès de phase. Un exemple est donné sur la Figure 3.30. Sur la phase originelle de la HRTF brute, la HRTF droite est bien en retard sur la HRTF gauche jusque dans les basses fréquences. En revanche, sur les composantes à excès de phase, pour les premiers bins fréquenciels, la phase de la HRTF droite devient supérieure à celle de la HRTF gauche, ce qui vient fausser toute estimation du retard dans les très basses fréquences.

Pour quantifier l'amplitude des variations d'un estimateur à l'autre, on définit un critère de *divergence* Div_{est} qui correspond à l'écart maximum observé entre les 6 estimateurs pour une estimation donnée :

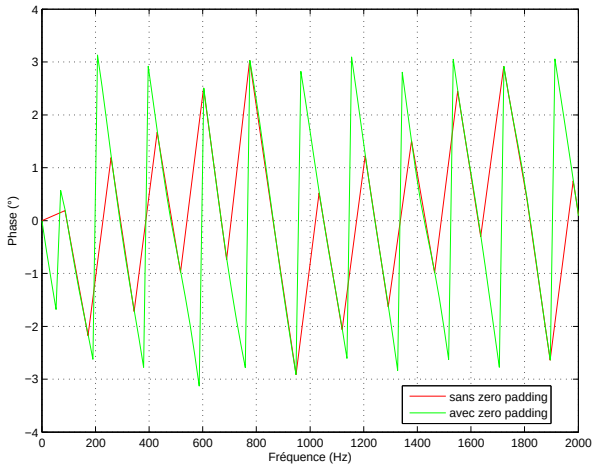
$$Div_{est}(ind, \phi) = \max_{|estimateur}[\tau(ind, \phi)] - \min_{|estimateur}[\tau(ind, \phi)] \tag{3.12}$$



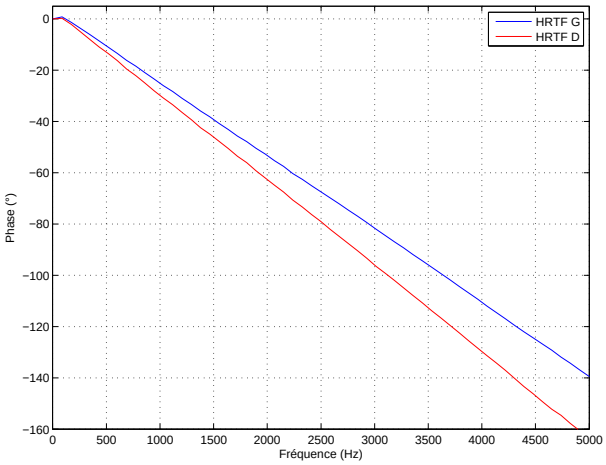
(a) Phase de la HRTF originale non déroulée



(b) Phase de la HRTF originale déroulée

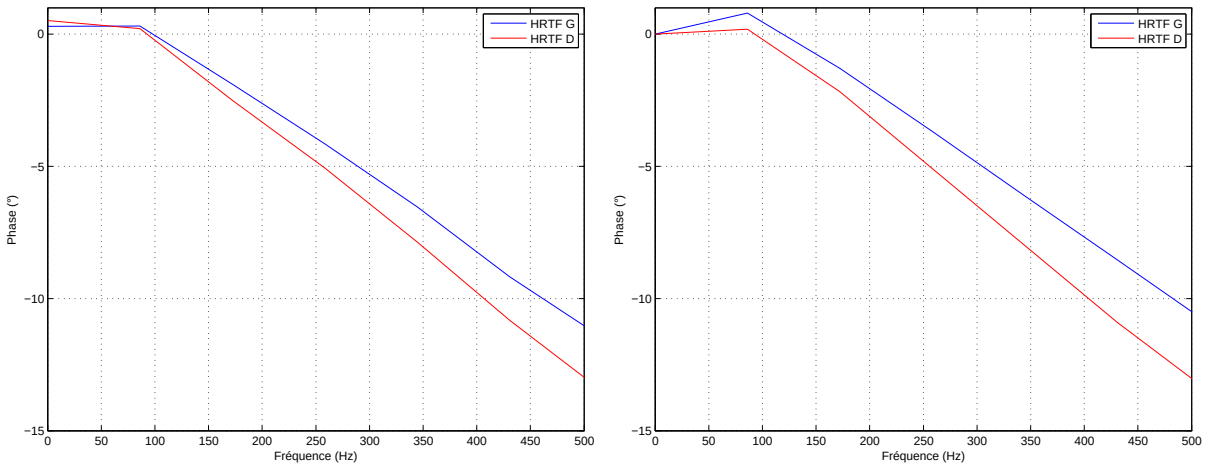


(c) Phase de la HRTF non déroulée, avec et sans correction par *zero padding* (HRTF droite)



(d) Phase de la HRTF originale déroulée après *zero padding*

FIG. 3.29 – Estimation du retard de phase pour un sujet (sujet 02) de la base de l'IRCAM (HRTF gauche et droite pour la direction $[\phi = 90°, \theta = 0°]$).



(a) Phase de la HRTF brute

(b) Phase de la composante à excès de phase

FIG. 3.30 – Evolution fréquencielle de la phase : comparaison des phases de la HRTF brute et de la composante à excès de phase (sujet 02 de la base de l'IRCAM, HRTF gauche et droite pour la direction $[\phi = 90°, \theta = 0°]$).

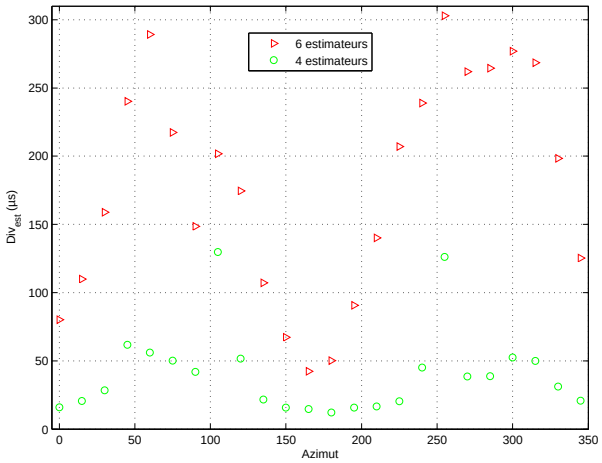


FIG. 3.31 – Divergence des estimateurs de retard : en considérant les 6 estimateurs ou en éliminant les estimateurs 1 (Estimation du retard de phase moyen dans les basses fréquences) et 6 (Estimation du retard de groupe moyen dans les basses fréquences).

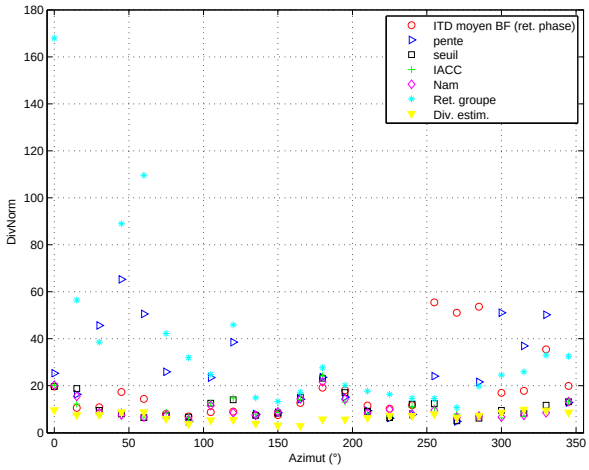


FIG. 3.32 – Divergence liée à l'estimateur et divergence résultant des variations interindividuelles pour chaque estimateur (divergence normalisée par la JND, moyenne sur les individus).

où l'indice *ind* repère le sujet considéré, ϕ désigne l'angle d'azimut, et τ le retard estimé. Les valeurs maximale et minimale sont calculées sur l'ensemble des 6 estimations. Ce critère est évalué pour les 112 individus couvrant les 5 bases de données. Pour chaque individu, 24 directions sont considérées dans le plan horizontal. Selon les positions mesurées dans la base de données, les directions sélectionnées se rapprochent au mieux des 24 azimuts correspondant à un échantillonnage régulier entre 0 et 345° avec un pas de 15°. Le critère de divergence moyenné sur l'ensemble des individus est reproduit en fonction de l'azimut sur la Figure 3.31. On observe que la divergence est minimale dans le plan médian et croît pour les positions latérales. Elle varie entre 42 ($\phi = 180°$) et 303 μs ($\phi = 255°$), ce qui est assez considérable. Sur les figures 3.25, 3.26, 3.27 et 3.28, on a relevé que les estimateurs basés sur le retard de phase (méthode 1) et le retard de groupe (méthode 6) présentaient une divergence plus prononcée. La divergence des estimateurs a été recalculée en éliminant ces 2 estimateurs, ce qui donne la deuxième courbe de la Figure 3.31. La réduction de la divergence est très sensible puisqu'on observe alors des valeurs comprises entre 12 ($\phi = 180°$) et 130 μs ($\phi = 105°$), ce qui confirme le caractère marginal des estimateurs 1 et 6.

Cependant les valeurs de divergence ne signifient rien si on ne les interprète pas en termes de perception, afin de déterminer si les différences observées entre les estimateurs sont discriminables par le système auditif. On se propose donc de normaliser la divergence par la JND (Just Noticeable Difference) de l'ITD :

$$DivNorm_{est}(ind, \phi) = \frac{Div_{est}(ind, \phi)}{JND[\overline{\tau}(\phi)]} \tag{3.13}$$

où :

$$\overline{\tau}(\phi) = \frac{1}{N_{sujet} N_{estimateur}} \sum_{ind=1}^{N_{sujet}} \sum_{j=1}^{N_{estimateur}} \tau_j(ind, \phi) \ .$$

Les valeurs de JND sont obtenues par interpolation linéaire à partir des résultats de l'étude de Klumpp & Eady [Klumpp & Eady, 1956], en considérant pour chaque azimut la JND associée au retard moyen $\overline{\tau}$ pour l'ensemble des estimateurs et individus. La divergence normalisée par la JND (moyenne sur les individus) est représentée en fonction de l'azimut sur la Figure 3.32. Une divergence normalisée inférieure à 1 correspond à des écarts d'ITD inférieurs au seuil de discrimination, donc non perceptibles par le système auditif. A l'opposé, toutes les valeurs supérieures à 1 correspondent à des écarts d'ITD qui sont détectables par le système auditif. Une fois normalisée par

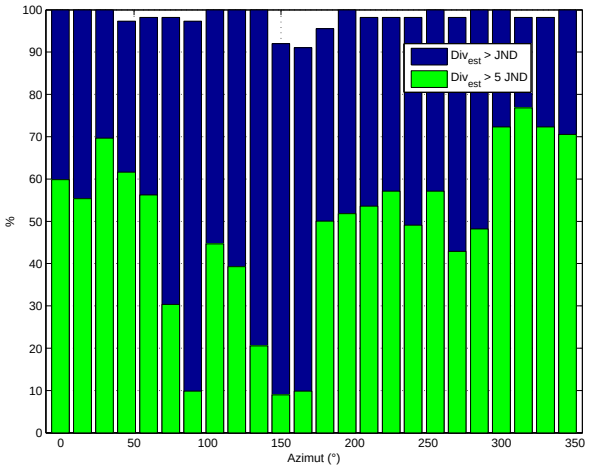


FIG. 3.33 – Pourcentage de valeurs de la divergence des estimateurs dépassant une et cinq fois la JND en fonction de l'azimut.

la JND, la divergence tend à être plus homogène quel que soit l'azimut. Cependant au minimum elle vaut 2.8 JND, soit près de 3 fois le seuil d'audibilité : les différences entre les estimateurs sont donc parfaitement perceptibles par le système auditif. Pour chaque azimut, l'ensemble des valeurs collectées sur les individus permet d'estimer une fonction de répartition empirique $f_S(D)$ qui donne la probabilité P que la divergence D reste inférieure à un seuil S donné :

$$f_S(D) = P(D \leq S) \tag{3.14}$$

Dans notre cas, le seuil qui nous intéresse principalement est défini par la valeur de 1, correspondant au seuil de discrimination (JND) pour la divergence normalisée. La probabilité P s'interprête en termes de pourcentage $P\%$ de valeurs dépassant le seuil S par la relation suivante :

$$P\% = 100(1 - P) \tag{3.15}$$

Ce pourcentage pour un seuil de 1 (c'est à dire : pourcentage de valeurs dépassant la JND) et de 5 (c'est à dire : pourcentage de valeurs dépassant 5 fois la JND) est reproduit sur la Figure 3.33. On constate que le seuil de discrimination est atteint dans près de 100% des cas quel que soit l'azimut. De plus, pour la plupart des azimuts, près de 50% des valeurs sont supérieures à 5 fois la JND. Les écarts d'estimation sont donc clairement audibles.

Enfin il est aussi intéressant de comparer la divergence entre les estimateurs aux variations interindividuelles. Un critère de divergence entre les individus est ainsi défini :

$$Div_{ind}(\phi) = \max_{|ind}[\tau(ind, \phi)] - \min_{|ind}[\tau(ind, \phi)] \tag{3.16}$$

qu'il convient aussi de normaliser par la JND :

$$DivNorm_{ind}(ind, \phi) = \frac{Div_{ind}(ind, \phi)}{JND[\overline{\tau}(\phi)]} \tag{3.17}$$

La divergence interindividuelle normalisée est illustrée pour les 6 estimateurs sur la Figure 3.32 et peut ainsi être comparée à la divergence entre les estimateurs. Il ressort que les variations interindividuelles s'avèrent majoritairement supérieures au flou des estimateurs[20]. En conséquence

[20]De plus, on observe au passage que les écarts inter-individuels dépendent sensiblement de l'estimateur, ce qui suggère un critère potentiel pour évaluer la qualité des estimateurs de retard.

il semblerait préférable d'utiliser un "mauvais" estimateur que d'avoir au recours aux HRTF d'un autre individu. Néanmoins les incertitudes d'estimation sont discriminables. Il reste maintenant à positionner les différents estimateurs par rapport à la cible perceptive offrant un rendu non discriminable de la HRTF originale.

3.3.2 Protocole expérimental

Le protocole expérimental est brièvement rappelé ici. Pour de plus amples détails, le lecteur est invité à se reporter au document de thèse de Sylvain Busson [Busson, 2006]. L'objectif de l'expérience est de déterminer la valeur d'ITD qui corresponde à une équivalence perceptive entre la HRTF originale et sa modélisation en un filtre à phase minimale associé à un retard pur. Dans ce but, l'expérience réalisée est la suivante : le sujet écoute une séquence de deux stimuli, le premier stimulus constitue la cible, c'est à dire le son synthétisé avec la HRTF originale, tandis que le second stimulus est obtenu avec un filtre à phase minimale associé à un retard pur. La tâche confiée au sujet consiste à ajuster le retard du second stimulus jusqu'à ce qu'il perçoive les deux stimuli comme identiques. Le protocole experimental est basé sur une méthode adaptative [Levitt, 1971] avec une procédure de type 2I-2AFC (*two-interval - two-alternative forced choice*) [Marvit et al., 2003]. Les HRTF considérées sont les HRTF individuelles des sujets. Le seul paramètre expérimental est la position de la source virtuelle : 12 positions situées dans le plan horizontal sont testées, soit 12 angles d'azimut balayant la plage de -180° à +135°. L'étude intègre les HRTF issues de deux bases de données : celles de l'IRCAM et d'Orange Labs. Ving sujets ont participé à l'expérience : 14 sujets de la base de l'IRCAM et 6 sujets de la base *Jean-Marie Pernaux*. Chaque condition a été évaluée cinq fois.

L'expérience s'est décomposée en deux parties : une expérience de contrôle (10 sujets) et l'expérience principale (10 sujets). Dans l'expérience de contrôle, le protocole est identique à celle de l'expérience principale, à la seule différence que le stimulus cible est lui aussi synthétisé avec un filtre à phase minimal associé à un retard pur. L'objectif de cette expérience est de valider le protocole de test : en théorie le sujet doit converger vers la valeur de retard appliqué au stimulus cible.

3.3.3 Résultats : Expérience de contrôle

Le premier résultat de l'expérience de contrôle concerne la **faisabilité de la tâche** qu'on demande au sujet : a-t-il réussi à converger vers une valeur de retard quelle que soit la position de la source virtuelle ? On sait en effet que la modélisation en filtre à phase minimale associé à un retard pur tend à n'être plus valide pour les positions fortement latéralisées, notamment pour la HRTF controlatérale [Avendano et al., 1999] [Kulkarni et al., 1999]. La synthèse binaurale par des filtres basés sur cette modélisation risque donc d'introduire des artefacts de perception susceptibles de perturber le sujet. Mais surtout, des études ont montré que le positionnement relatif de sources sonores est une tâche qui peut s'avérer difficile pour des positions latérales, même en champ libre [Mills, 1958] [Braasch & Hartung, 2002]. Le succès de la tâche est donc contrôlé en vérifiant que la valeur de retard reportée par le sujet est proche de la valeur cible. Si la valeur donnée par le sujet est de 50% supérieure ou inférieure au retard cible, on considère que la tâche a échoué. Le nombre d'échecs est comptabilisé, ce qui donne une mesure de la faisabilité de la tâche. Pour la majorité des positions testées, on n'observe aucun échec. Seules les positions latérales ($\phi = \pm$ 90°, -75°, 105°) présentent un taux d'échec non nul, mais qui reste inférieur ou égal à 6% des jugements. Ce résultat est une première validation du protocole expérimental. Le retard estimé par les sujets est illustré sur la Figure 3.34.

Il reste à vérifier que la valeur de retard trouvée par le sujet concorde bien avec le retard cible. La Figure 3.35 représente la réponse du sujet en fonction de la valeur cible. Pour l'ensemble des

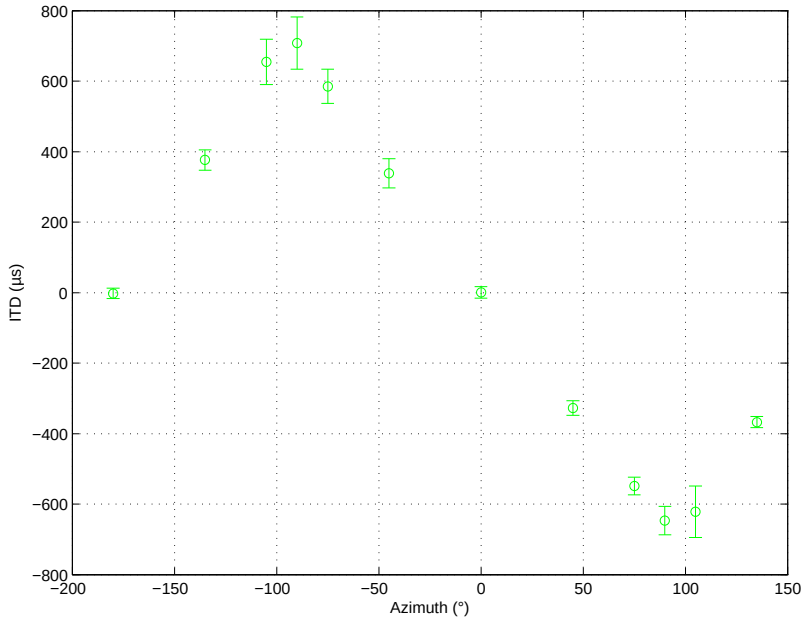


FIG. 3.34 – Expérience de contrôle : Retard estimé par les sujets en fonction de l'azimut (valeur moyenne sur les 10 sujets et intervalle de confiance à 95% associé).

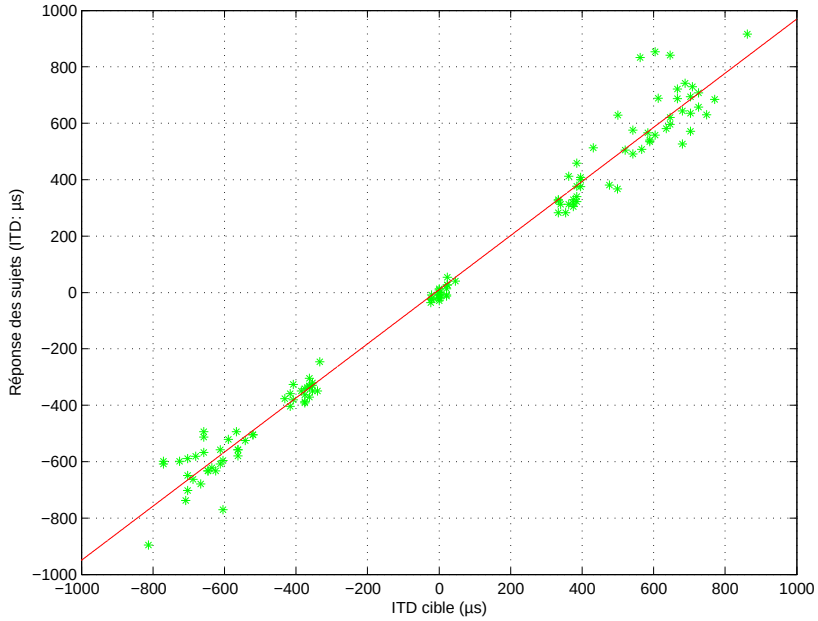


FIG. 3.35 – Expérience de contrôle : Retard estimé par les sujets en fonction du retard cible pour l'ensemble des 10 sujets et des 12 azimuts. La droite représente la régression linéaire calculée sur l'ensemble des données.

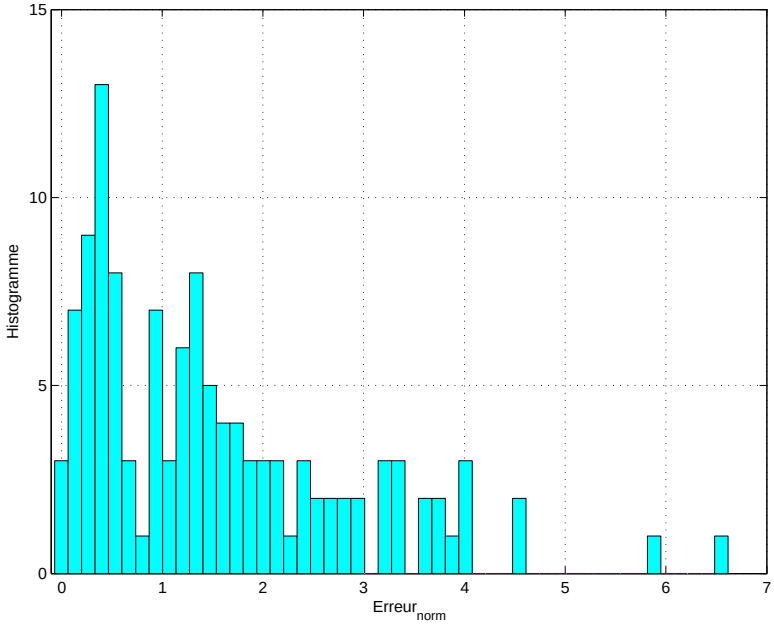


FIG. 3.36 – Expérience de contrôle : Histogramme de l'erreur normalisée $Erreur_{norm}$ pour l'ensemble des 120 jugements (12 azimuts x 10 sujets).

jugements (total de 120 jugements pour 12 azimuts et 10 sujets), le retard estimé est proche du retard cible, ce qui se traduit sur la Figure 3.35 par des points situés au voisinage de la diagonale. Une régression linéaire calculée sur les données donne un coefficient de corrélation de 0.96. L'erreur d'estimation est définie comme la valeur absolue de la différence entre le retard estimé par le sujet (soit le retard perceptif $\tau_{percept}$) et le retard cible (τ_{cible}) :

$$Erreur(ind, \phi) = |\tau_{percept}(ind, \phi) - \tau_{cible}(ind, \phi)| \tag{3.18}$$

Afin de quantifier cette erreur en termes de perception, comme précédemment on la normalise par la JND de l'ITD [Klumpp & Eady, 1956] associée à la position considérée :

$$Erreur_{norm}(ind, \phi) = \frac{Erreur(ind, \phi)}{JND[\tau_{percept}(ind, \phi)]} \tag{3.19}$$

Ainsi une valeur de $Erreur_{norm}$ inférieure à 1 peut être considérée comme non perceptible. Un second intérêt de la normalisation par la JND est que les valeurs d'erreur obtenues pour des azimuts différents sont désormais directement comparables, et ce quelle que soit la direction considérée, car ramenées à l'unité de la JND. Il est alors possible de calculer des statistiques sur l'ensemble des données collectées. L'histogramme de l'erreur normalisée pour l'ensemble des 120 jugements est illustré sur la Figure 3.36. La majorité des erreurs observées correspond à des valeurs inférieures à 2, avec une proportion importante inférieure à 1, soit en dessous du seuil de perception. L'erreur normalisée vaut en moyenne $\overline{Erreur}_{norm} = 1.55$ (intervalle de confiance à 95% : $\pm$ 0.24) sur les 120 jugements. On en conclut que le retard estimé par les sujets s'avère proche du retard cible. L'écart est souvent de l'ordre du JND, même si, pour les positions fortement latéralisées (cf. Fig. 3.34 & 3.35), l'erreur d'estimation excède le seuil perceptif de discrimination.

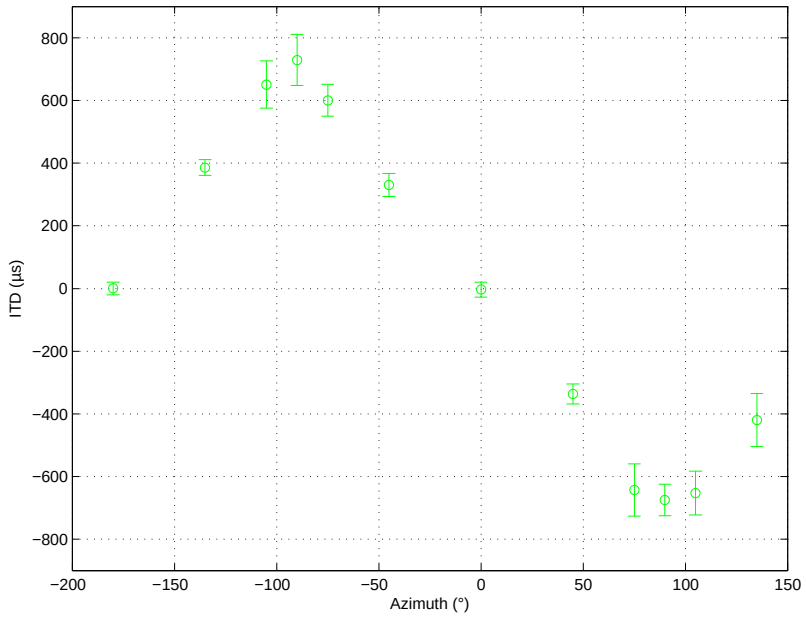


FIG. 3.37 – Expérience principale : Retard perceptif en fonction de l'azimut (moyenne sur les 11 sujets).

3.3.4 Résultats : Expérience principale

L'expérience de contrôle a permis de valider le protocole expérimental en montrant qu'il conduisait à une estimation fiable du retard imposé comme cible. A présent, dans l'expérimentation principale, le retard estimé représente véritablement le *retard perceptif* offrant l'équivalence en termes de perception entre le filtre binaural (modèle de filtre à phase minimale associé à un retard pur) et la HRTF originale. Le retard perceptif obtenu en moyenne pour les 10 sujets de l'expérience principale est illustré sur la Figure 3.37. Le retard perceptif ainsi estimé par les sujets est dans la suite considéré comme le retard de référence auquel seront comparés les retards obtenus par les estimateurs mathématiques. La Figure 3.38 représente le retard perceptif en fonction des retards fournis par les différents estimateurs mathématiques. A l'instar de la Figure 3.35, une parfaite adéquation entre les estimations perceptives et mathématiques correspondrait à des points situés sur la diagonale. On est d'abord frappé par le comportement marginal de l'estimateur basé sur le retard de phase. Dans l'ensemble, ses estimations restent proches de la diagonale, à l'exception de quelques points qui se trouvent situés dans les quadrants inférieur droit et supérieur gauche et qui traduisent par la même des inversions du retard. Ces erreurs d'estimation dramatiques, même si elles sont peu nombreuses, viennent dévier la régression linéaire. Elles sont imputables au problème de fiabilité de l'information du retard de phase mentionné précédemment (cf. page 158). Au vu de ces résultats, sauf mention contraire, la suite de l'étude se focalise sur les 5 autres estimateurs. En ce qui concerne ces derniers, les données sont très proches de la diagonale et les différents estimateurs présentent des comportements similaires. Pour chaque estimateur, une régression linéaire est calculée. Les coefficients de la régression sont donnés dans le Tableau 3.5. Globalement, les estimations des 5 méthodes offrent une bonne corrélation avec le retard perceptif. La corrélation la plus élevée est obtenue par la méthode de linéarisation de la phase et l'estimateur proposé par Nam (C= 0.985). Les estimateurs basés sur la détection de seuil et le maximum de la fonction d'intercorrélation en sont néanmoins très proches (respectivement C= 0.984 et 0.983). Il est raisonnable de considérer que ces 4 estimateurs présentent des performances équivalentes. L'estimateur basé sur le retard

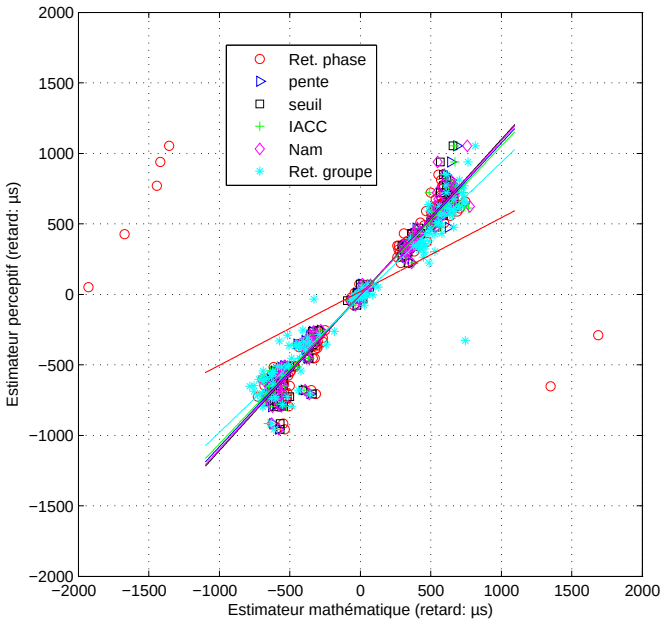


FIG. 3.38 – Expérience principale : Retard perceptif en fonction des retards obtenus par les différents estimateurs mathématiques pour l'ensemble des 10 sujets et des 12 azimuts. Les droites représentent les régressions linéaires calculées sur l'ensemble des données pour chaque estimateur.

de groupe obtient une corrélation légèrement plus faible (C= 0.961). L'observation de la droite de régression linéaire indique que les 4 premiers estimateurs conduisent à une sous-estimation du retard (pente de la droite de régression > 1), tandis que l'estimateur basé sur le retard de groupe correspond à une surestimation (pente de la droite de régression < 1). Il est important de noter ici que l'information apportée par la régression entre le retard perceptif et les estimations mathématiques offre la possibilité de corriger a posteriori les valeurs estimées afin de se ramener à la valeur perceptive. Il suffit d'appliquer les coefficients de la régression.

Le seul critère de régression linéaire ne suffit pas à donner une évaluation complète des estimateurs mathématiques. Il convient de les passer au crible de critères complémentaires. On souhaite notamment évaluer l'amplitude des erreurs commises par chaque estimateur. Dans ce but, l'*erreur normalisée* définie à la Section précédente (cf. Equ. 3.19) est calculée. L'erreur normalisée moyenne pour les 10 sujets de l'expérience est reproduite sur la Figure 3.39. On constate d'abord que quels

Estimateur	C	P (%)	$\overline{Erreur}_{norm}$
Ret. phase	0.561	85	6.13 ± 3.29
Pente	0.985	58	1.74 ± 0.28
Seuil	0.984	54	1.82 ± 0.30
IACC	0.983	64	1.87 ± 0.28
Nam	0.985	55	1.75 ± 0.28
Ret. groupe	0.961	79	3.40 ± 0.82

TAB. 3.5 – Evaluation comparée des estimateurs mathématiques selon trois critères : coefficient (C) de régression linéaire, pourcentage (P) de valeurs estimées dans l'intervalle [ITD_{percep} - JND, ITD_{percep} + JND], erreur normalisée moyennée sur les 120 observations et intervalle de confiance à 95% associé.

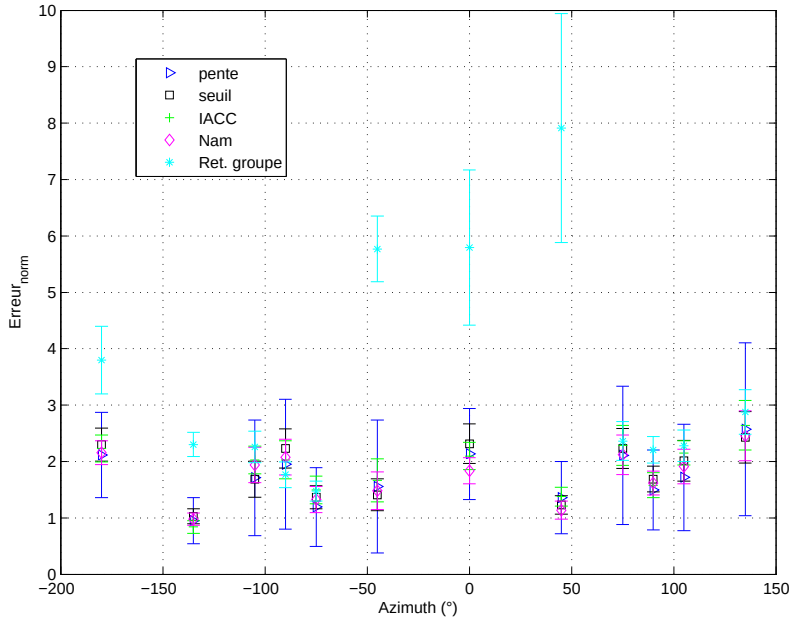


FIG. 3.39 – Expérience principale : Erreur normalisée $Erreur_{norm}$ (moyenne sur les 10 sujets et intervalle de confiance à 95% associé) en fonction de l'azimut pour les 5 estimateurs mathématiques.

que soient l'estimateur et l'azimut, l'erreur normalisée est supérieure à 1, c'est à dire que l'erreur d'estimation dépasse dans tous les cas le seuil de discrimination. Elle est donc perceptible. Cependant, si l'on met à part l'estimateur basé sur le retard de groupe, l'erreur reste comprise entre 1 et 3, ce qui s'avère assez faible. On observe aussi que l'erreur est relativement homogène en fonction de l'azimut. En dehors du fait que l'estimateur basé sur l'estimation de la pente de la phase se démarque par une forte variance, les estimateurs 2 à 5 présentent des résultats très proches. En revanche, l'estimateur basé sur le retard de groupe offre un comportement marginal avec une erreur $Erreur_{norm}$ supérieure à 5 autour du plan médian, ce qui est cohérent avec les observations de la Figure 3.38 et sa corrélation plus faible. L'erreur normalisée moyenne (cf. Tab. 3.5) reflète ces observations. Pour les estimateurs 2 à 5, l'erreur normalisée moyenne est légèrement inférieure à 2. Elle est minimale pour l'estimateur basé sur la pente de la phase, mais l'estimateur proposé par Nam en est très proche. Viennent ensuite l'estimateur basé sur la détection de seuil, puis celui basé sur le maximum de la fonction d'intercorrélation. On se rend compte que l'estimateur proposé par Nam et présenté comme une amélioration du calcul du maximum de la fonction d'intercorrélation [Nam et al., 2008] atteint effectivement cet objectif puisqu'il réduit l'erreur d'estimation de 1.87 à 1.75, ce qui l'amène presqu'à la valeur du meilleur estimateur (1.74). Il faut cependant noter qu'au vu des intervalles de confiance, les différences entre les estimateurs 2 à 5 sont faiblement significatives. Néanmoins, on retiendra que, quel que soit l'estimateur, l'erreur d'estimation est perceptible par le système auditif.

Enfin il est intéressant d'examiner la distribution de l'erreur normalisée selon l'estimateur. Les histogrammes des valeurs obtenues sont reproduits sur la Figure 3.40. On observe deux types de profil :

– La distribution de l'erreur présente son maximum pour une erreur de l'ordre du JND, puis elle décroît relativement rapidement pour les valeurs croissantes d'erreur. Les estimateurs 2, 3 et 5 appartiennent à cette catégorie. Ce résultat est cohérent avec leur erreur moyenne qui est faible (cf. Tab. 3.5).

– L'erreur se distribue de façon plus homogène sur l'intervalle [0 - 5 JND], ce qui se traduit par une erreur moyenne plus élevée (cf. Tab. 3.5). C'est le cas des estimateurs 4 et 6.

Pour terminer, considérons un dernier critère consistant à comptabiliser le nombre de fois où l'ITD estimé sort de la plage [[ITD_{percep} - JND, ITD_{percep} + JND], ce qui définit une erreur d'estimation perceptible au sens de la JND de l'ITD. Ce nombre peut être traduit en pourcentage d'observations réalisées. Le Tableau 3.5 donne les pourcentages obtenus par chaque estimateur. Au sens de ce critère, le meilleur estimateur est celui basé sur la détection de seuil (54%), mais il distance de très peu l'estimateur proposé par Nam (55%). Viennent ensuite dans l'ordre : l'estimateur basé sur la pente de la phase (58%), l'estimateur basé sur le maximum de la fonction d'intercorrélation (64%), l'estimateur basé sur le retard de groupe (79%) et l'estimateur basé sur le retard de phase (85%). Ces deux derniers se démarquent par leurs piètres performances.

Pour conclure, 3 estimateurs se distinguent par la qualité de leur estimation sur l'ensemble des critères considérés : l'estimateur basé sur la pente de la phase, celui basé sur la détection de seuil et celui proposé par Nam. Cependant aucun de ces estimateurs n'est parfaitement satisfaisant dans la mesure où la valeur estimée ne garantit pas, dans la plupart des cas, l'équivalence perceptive avec la HRTF originale, ce qui reste problématique. Néanmoins, les calculs de régression entre la cible perceptive offrant cette équivalence et les estimateurs mathématiques suggèrent une piste intéressante pour corriger ces derniers et obtenir des valeurs de retard offrant un rendu plus conforme aux HRTF originales.

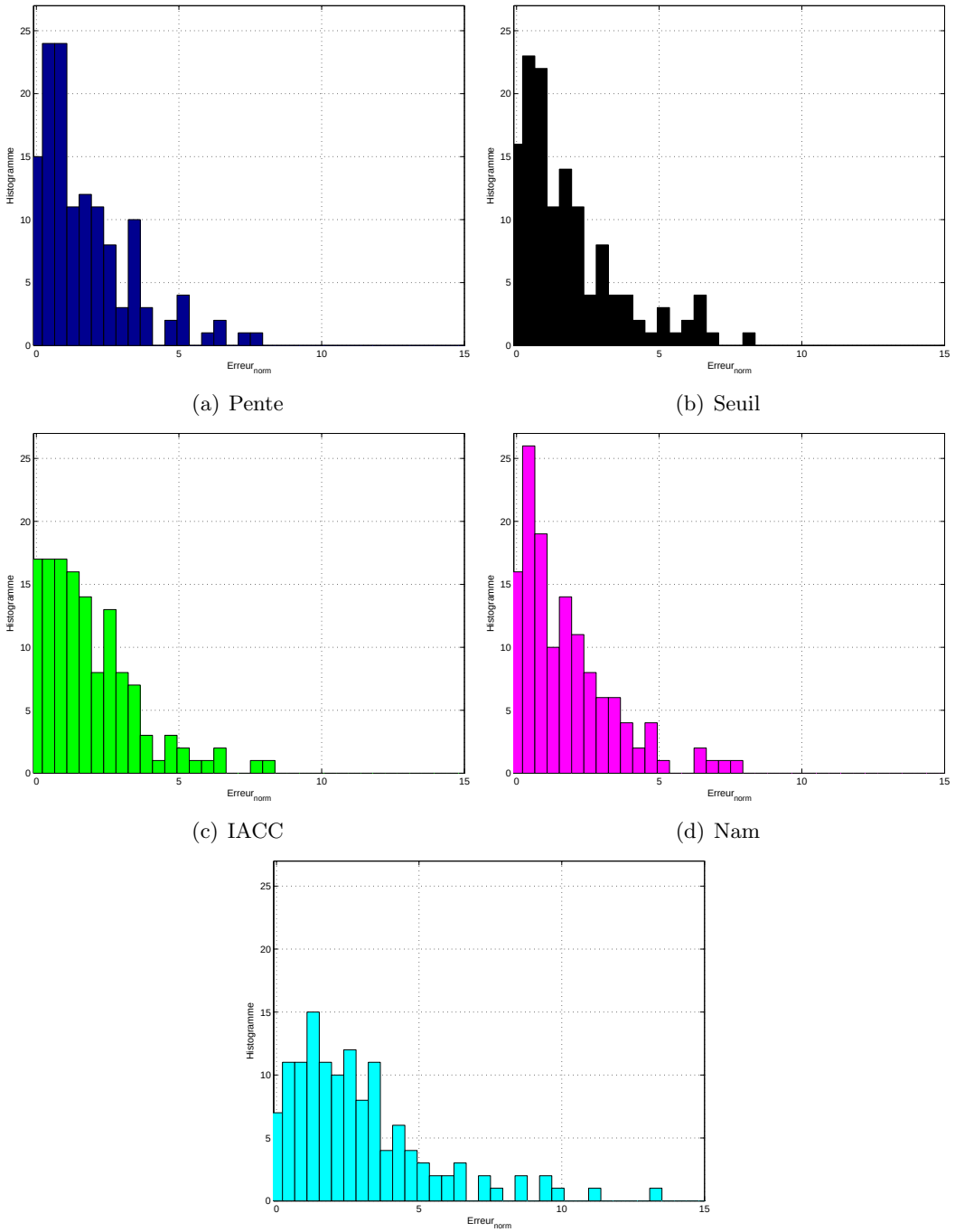
3.4 Modélisation de l'ITD

Dans la section précédente, on s'est intéressé aux estimateurs mathématiques destinés à extraire le retard des HRTF mesurées sur un individu. A présent, nous nous plaçons dans la situation où les HRTF mesurées de l'individu ne sont pas disponibles. Il convient alors de se doter d'un outil de modélisation des filtres binauraux pour faire alternative aux mesures acoustiques. Pour résoudre ce problème, on se propose de séparer le problème en deux et de considérer d'une part la modélisation du retard et d'autre part celle des IS. Cette section est consacrée à la modélisation du retard. Après un bref rappel des modèles existants, les lacunes de ces modèles, notamment pour rendre compte des variations fines de l'ITD en fonction d'élévation, vont être mises en évidence, sur la base de l'ITD estimée pour plusieurs bases de données. L'audibilité de ces variations en élévation est alors évaluée. Afin de les modéliser, un nouveau modèle est ensuite proposé, introduisant un nouveau degré de liberté dans la modélisation concernant le positionnement des oreilles de l'auditeur pour une reproduction plus précise de l'évolution de l'ITD en élévation. Il reste enfin à valider ce modèle en comparaison des modèles existants. Une validation objective est menée. L'erreur de modélisation est interprétée en termes de la JND de l'ITD, comme précédemment, afin de déterminer si elle est perceptible.

3.4.1 Etat de l'art des modèles de l'ITD

Par modèle d'ITD, on entend ici des formules mathématiques permettant de calculer l'ITD à partir d'une géométrie plus ou moins simplifiée de la morphologie de l'auditeur. Plusieurs formules ont été proposées dans la littérature [Blauert, 1983], mais nous ne retiendrons que les deux modèles qui présentent le plus d'intérêt pour notre étude : le modèle de Woodworth [Woodworth & Schlosberg, 1954] et le modèle SHM (*Spherical Head Model*) proposé dans [Algazi et al., 2001b].

Le modèle de Woodworth suppose que la tête de l'auditeur est une sphère de rayon R sur laquelle les oreilles sont disposées de façon diamétralement opposées. Si l'onde incidente est une



(a) Pente

(b) Seuil

(c) IACC

(d) Nam

(e) Ret. groupe

FIG. 3.40 – Expérience principale : Distribution de l'erreur normalisée $Erreur_{norm}$ pour les estimateurs 2 à 6.

onde plane se propageant selon la direction[21] (ϕ, θ), l'ITD est calculée sur la base d'une différence de marche pour la propagation directe, complétée d'un trajet de contournement pour la propagation par diffraction dans l'ombre de la tête [Blauert, 1983] :

$$ITD_{Woodworth}(\phi) = \frac{R}{c}(\sin\phi + \phi) \tag{3.20}$$

Compte tenu de la symétrie sphérique du modèle utilisé, cette formule ne présente qu'une dépendance en azimut ϕ. En d'autres termes, l'ITD résultante reste constante sur un plan d'azimut constant et ne présente pas de variations en fonction de l'élévation θ de la source. De plus, dans l'équation 3.20, on considère un rayon de tête correspondant à une moyenne anthropométrique R = 8.75 cm. Le modèle SHM propose d'individualiser le rayon de la sphère à partir de paramètres morphologiques décrivant les principales dimensions de la tête de l'auditeur :

$$ITD_{SHM}(\phi, ind) = \frac{R_{ind}}{c}(\sin\phi + \phi) \tag{3.21}$$

Le rayon individualisé R_{ind} s'obtient comme une combinaison linéaire de 3 paramètres anthropométriques x_1, x_2, x_3 décrivant respectivement la demi-largeur, la demi-hauteur et la demi-profondeur de la tête de l'individu [Algazi et al., 2001b] :

$$R_{ind} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \tag{3.22}$$

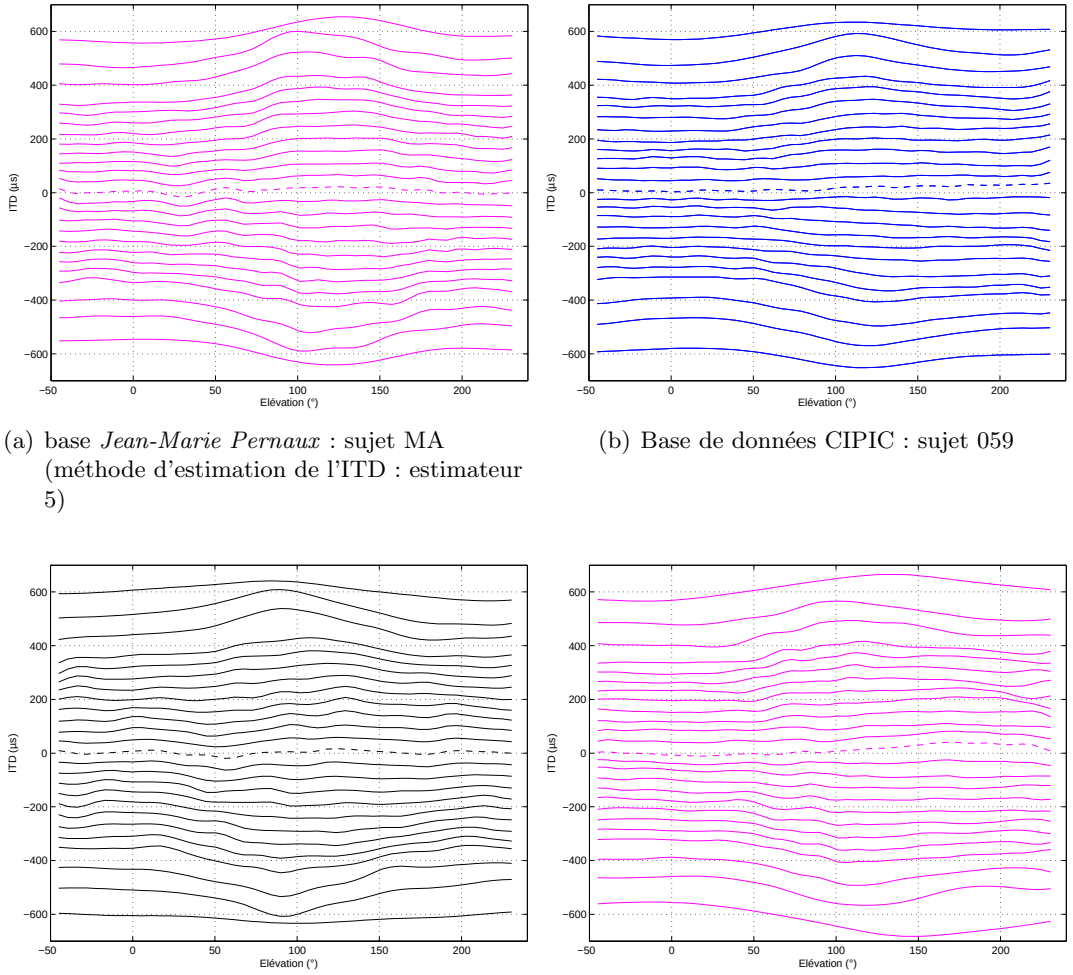
avec $w_0 = 0.032$, $w_1 = 0.0051$, $w_2 = 0.00019$, $w_3 = 0.0018$. Les valeurs des coefficients w_i ont été obtenus par minimisation de l'erreur quadratique entre l'ITD modélisée et l'ITD estimée sur les HRTF d'individus issus de la base de données du CIPIC (cf. Tab. 3.1) [Algazi et al., 2001b]. Le seul apport du modèle SHM par rapport au modèle de Woodworth est la dépendance individuelle avec l'adaptation du rayon de la sphère à l'individu.

3.4.2 Observation de l'ITD sur la sphère 3D

Aucun des modèles précédents ne permet de reproduire d'éventuelles variations de l'ITD en fonction de l'élévation. Or, l'observation de l'ITD estimée sur des HRTF mesurées montre que l'ITD n'est pas constante en fonction de l'élévation. La Figure 3.41 illustre l'ITD obtenue pour 4 sujets appartenant à différentes bases de données. L'ITD est représentée en fonction de l'élévation par plan d'azimut constant. On remarque que l'ITD présente un maximum autour de l'élévation 90°, en général pour des élévations supérieures et comprises entre 90° et 120°. L'amplitude de cette bosse augmente avec l'azimut jusqu'à ±65°. Cette évolution caractéristique s'observe quelle que soit la base de données, l'individu ou l'estimateur. Le fait que ce maximum d'ITD se situe autour de l'élévation 90° suggère une asymétrie morphologique de type haut/bas.

Même si les courbes des figures précédentes sont parfaitement représentatives des tendances générales, on constate quelques comportements divergents qui ont plusieurs origines. On note tout d'abord sur certains sujets une asymétrie entre les azimuts gauches et droits : par exemple, pour un sujet de la base de l'IRCAM illustré sur la Figure 3.42a, le maximum de l'ITD se situe pour les azimuts gauches (ITD > 0) à l'élévation 50°, tandis que pour les azimuts droits il apparaît autour de 120°. De même, pour un sujet de la base du CIPIC, on voit sur la Figure 3.43a que le maximum de l'ITD se décale de 90° pour les azimuts gauches à 120° pour les azimuts droits. L'ITD est

[21] Coordonnées polaires interaurales. Sauf mention contraire, dans toute cette section (Section 3.4), les angles d'azimut et d'élévation sont spécifiés en coordonnées polaires interaurales ou horizontales. Ce système de coordonnées est en effet parfaitement adapté à une étude par plan d'azimut constant pour l'observation des variations de l'ITD en élévation. Le système de coordonnées polaires interaurales est le système adopté pour la base de données de HRTF du CIPIC.



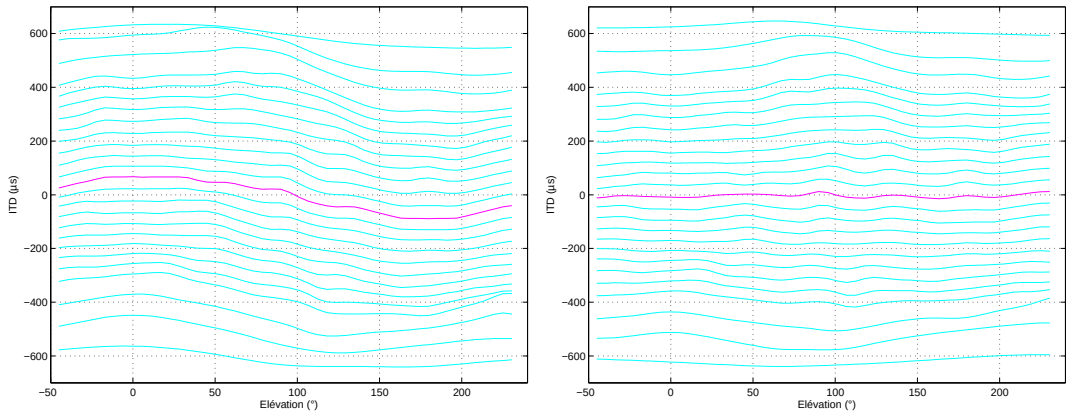
(a) base *Jean-Marie Pernaux* : sujet MA (méthode d'estimation de l'ITD : estimateur 5)

(b) Base de données CIPIC : sujet 059

(c) Base de données IRCAM : sujet 12 (méthode d'estimation de l'ITD : estimateur 3)

(d) Base de données Wightman : sujet AFW (méthode d'estimation de l'ITD : estimateur 5)

FIG. 3.41 – Evolution de l'ITD en fonction de l'élévation pour différents plans d'azimut constant de -80° à 80° (coordonnées polaires interaurales). Chaque courbe décrit l'ITD associée à un azimut donné. L'ITD est estimée sur des HRTF mesurées pour différents individus par l'une des méthodes décrites en Section 3.3.1, à l'exception de la base du CIPIC pour laquelle l'ITD représentée correspond aux valeurs d'ITD fournies dans la base de données et qui ont été obtenues par une méthode d'estimation de type 3 décrite dans [Algazi et al., 2001b].



(a) ITD non corrigée (b) ITD après correction

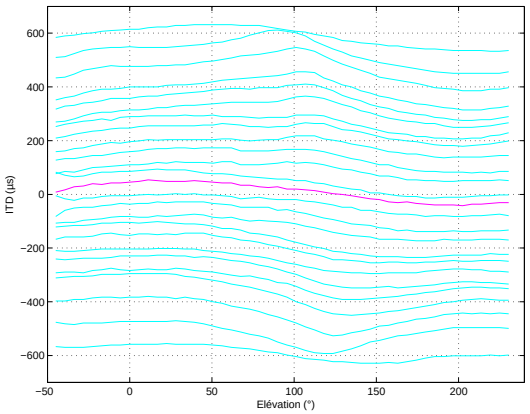
FIG. 3.42 – Illustration des asymétries entre les azimuts gauches et droits, avant et après correction par une procédure de réalignement basé sur une rotation des coordonnées d'espace (base de données IRCAM : sujet 06, estimateur 3). La courbe en magenta correspond à l'azimut 0°, les courbes en cyan aux autres azimuts.

déterminée par la morphologie de l'auditeur. Dans le problème physique, aucun élement ne peut justifier cette asymétrie gauche/droite. La visualisation des courbes iso-ITD sur la sphère 3D (cf. Fig. 3.43c) suggère une explication plausible : l'ensemble des courbes semble avoir subi un basculement gauche/droite comme si le sujet avait incliné la tête par rapport au plan médian. Afin de vérifier la validité de cette hypothèse, une opération de réalignement est testée : elle consiste à déterminer la rotation de la sphère qui maximise la symétrie de l'ITD [Guillon, 2009]. Les résultats obtenus après application de cette rotation sont illustrés sur les figures 3.42b, 3.43b et 3.43d. Le réalignement des données est très convaincant : les variations de l'ITD suivent à présent des évolutions parfaitement symétriques pour les azimuts gauches et droits. Pour cette raison, la correction de réalignement par rotation a été systématiquement appliquée sur toutes les données présentées dans cette section[22] (Section 3.4). Une autre anomalie observée chez certains individus est une ITD non nulle dans le plan médian. Tant que ces variations restent inférieures à $\pm 10\mu s$ (ordre de grandeur de la JND de l'ITD) pour une ITD nulle, elles ne sont pas problématiques. Mais dans certains cas comme sur la Figure 3.44, on relève un écart important qui là encore n'a pas de justification physique, si ce n'est une éventuelle translation du sujet par rapport à la référence du repère. Cependant il est souvent difficile d'identifier un décalage constant et systématique quel que soit l'azimut. Pour cette raison, ce défaut, qui reste heureusement marginal, n'a pas été corrigé.

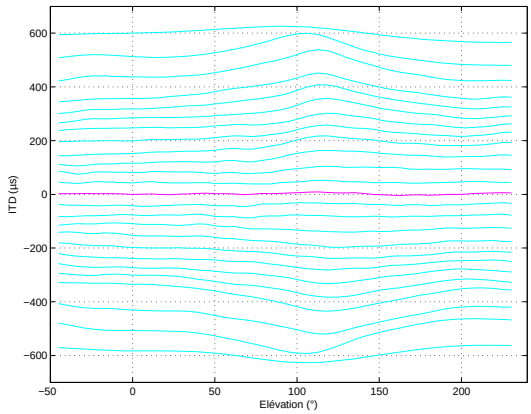
Un dernier aspect à prendre en compte est l'impact potentiel du choix de l'estimateur de l'ITD. Pour cette étude, 5 bases de données ont été analysées (*Jean-Marie Pernaux*, IRCAM, CIPIC, Université du Maryland, Wightman), représentant un total de 116 individus. Pour chacune de ces

[22] En particulier l'ITD illustrée sur la Figure 3.41. Pour récapituler, les données d'ITD considérées dans toute la présente section ont subi les prétraitements suivants [Guillon, 2009] :

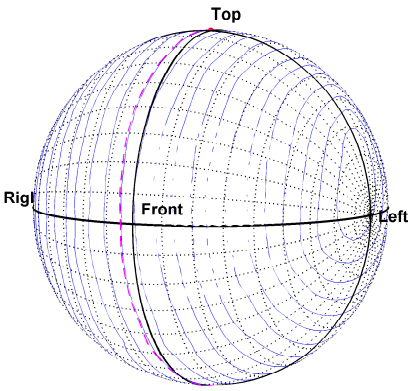
- interpolation par les fonctions *Spline* de type STPS (*Spherical Thin Plane Spline*) afin de compléter les données qui n'ont pas été mesurées sur la calotte sphérique inférieure, données qui sont nécessaires à une décomposition sur les harmoniques sphériques,
- rotation du système de coordonnées dans le domaine des harmoniques sphériques, pour corriger les asymétries gauche/droite,
- rééchantillonnage des données sur une grille spatiale commune correspondant aux 1250 directions de mesure de la base du CIPIC.



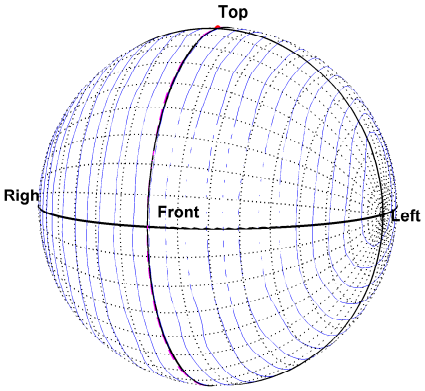
(a) ITD non corrigée



(b) ITD après correction



(c) ITD non corrigée : courbes iso-ITD visualisées sur la sphère 3D



(d) ITD après correction : courbes iso-ITD visualisées sur la sphère 3D

FIG. 3.43 – Illustration des asymétries entre les azimuts gauches et droits, avant et après correction par une procédure de réalignement basé sur une rotation des coordonnées d'espace (base de données CIPIC : sujet 015, estimateur défini dans [Algazi et al., 2001b]). La courbe en magenta correspond à l'azimut 0°, les courbes en cyan aux autres azimuts.

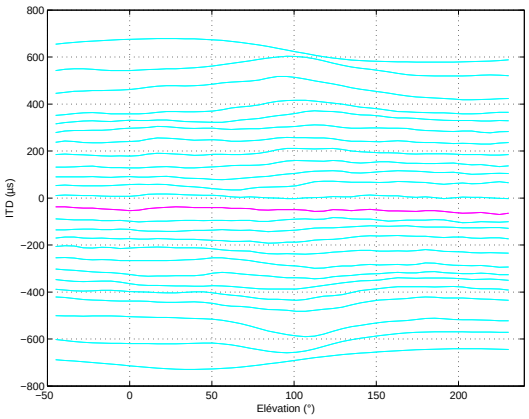
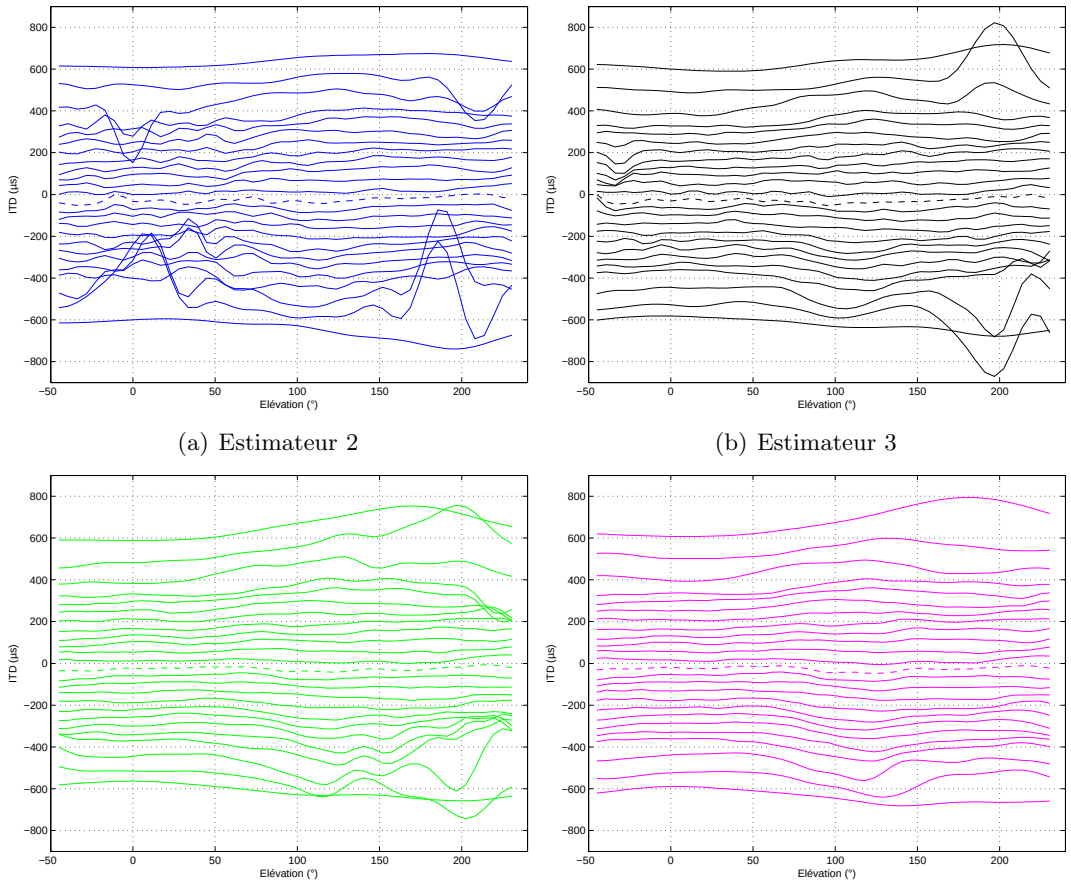


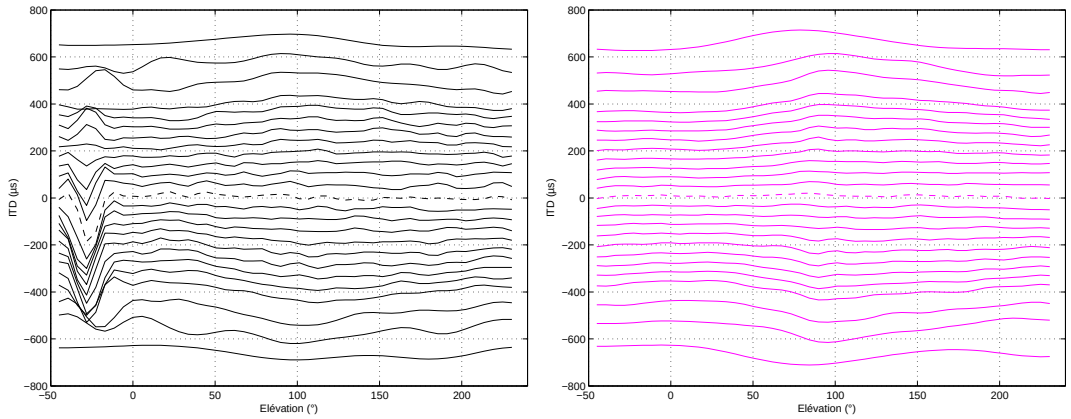
FIG. 3.44 – Illustration d'une ITD non nulle dans le plan médian (base de données CIPIC : sujet 009, estimateur [Algazi et al., 2001b]). La courbe en magenta correspond à l'azimut 0°, les courbes en cyan aux autres azimuts.



(a) Estimateur 2 (b) Estimateur 3

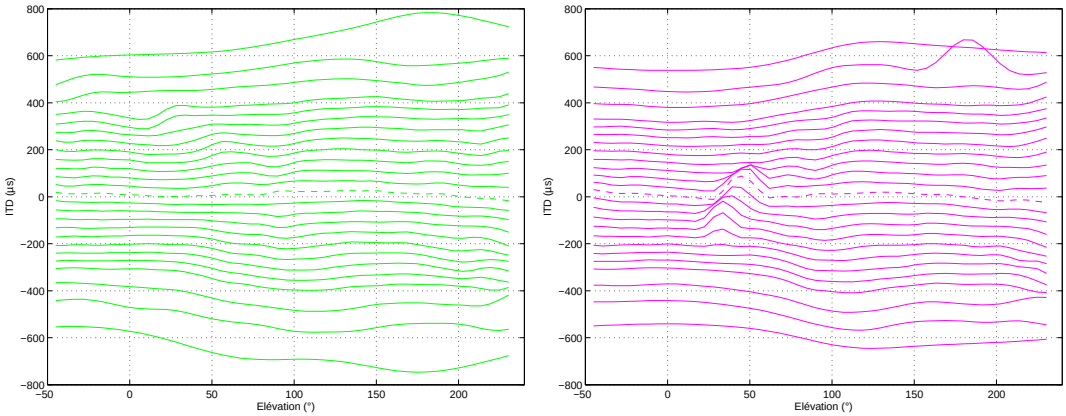
(c) Estimateur 4 (d) Estimateur 5

FIG. 3.45 – Erreurs d'estimation de l'ITD : résultats des 4 estimateurs pour le même sujet (base de données du CIPIC, sujet 065).



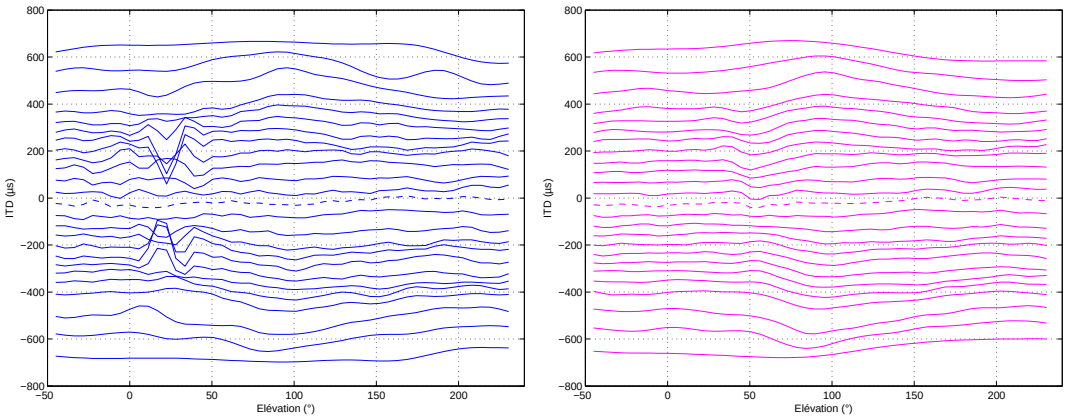
(a) Estimateur 3 (b) Estimateur 5

FIG. 3.46 – Erreurs d'estimation de l'ITD : résultats de 2 estimateurs pour le même sujet (base de données *Jean-Marie Pernaux*, sujet JMP).



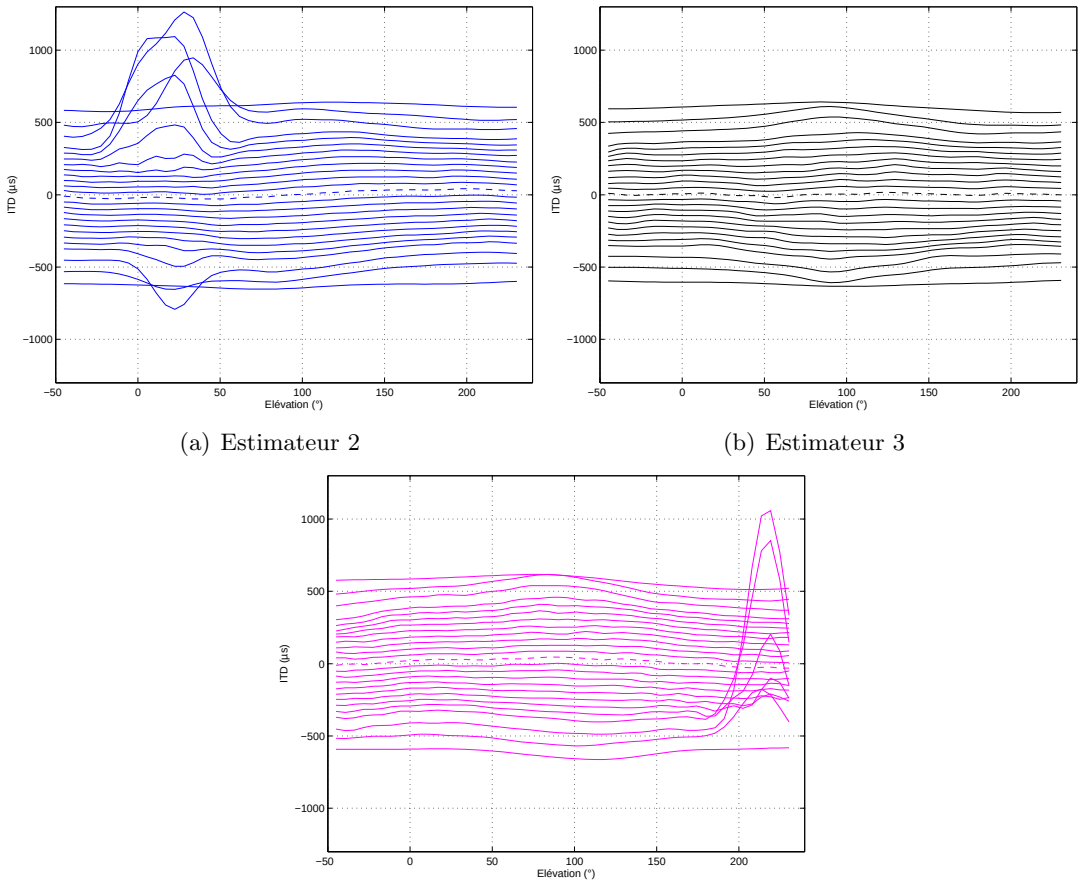
(a) Estimateur 4 (b) Estimateur 5

FIG. 3.47 – Erreurs d'estimation de l'ITD : résultats de 2 estimateurs pour le même sujet (base de données de Wightman, sujet SOU).



(a) Estimateur 2 (b) Estimateur 5

FIG. 3.48 – Erreurs d'estimation de l'ITD : résultats de 2 estimateurs pour le même sujet (base de données de l'Université du Maryland, sujet NM).



(a) Estimateur 2

(b) Estimateur 3

(c) Estimateur 5

FIG. 3.49 – Erreurs d'estimation de l'ITD : résultats de 3 estimateurs pour le même sujet (base de données de l'IRCAM, sujet 12).

bases, l'ITD a été estimée par 4 méthodes différentes (estimateurs 2 à 5 décrits dans la Section 3.3.1) pour une sélection de 1250[23] directions réparties uniformément sur la sphère 3D. L'ensemble des données ainsi collectées illustrent en quoi l'estimation de l'ITD reste une opération délicate. La fiabilité de l'estimation n'est jamais garantie quel que soit l'estimateur. Les figures 3.45 à 3.49 illustrent des exemples d'échec d'estimation. Dans certains cas, tous les estimateurs sont mis en difficulté (cf. Fig. 3.45), mais souvent au moins un des estimateurs parvient à extraire une ITD ne présentant pas d'anomalies évidentes (cf. Fig. 3.46b, 3.47a, 3.48b et 3.49b). En général, pour un individu donné, c'est un estimateur en particulier qui rencontre des difficultés (cf. Fig. 3.46a, 3.47b et 3.48a). On retiendra un estimateur préférentiel pour chaque base de données : estimateur 5 pour les bases d'Orange Labs et de Wightman, estimateurs 1 ou 2 pour la base de l'IRCAM. Pour la base du CIPIC, l'estimateur élaboré par Algazi *et al.* [Algazi et al., 2001b] s'avère le plus robuste et semble ainsi avoir été adapté spécifiquement aux données. Néanmoins, même si l'estimation de l'ITD semble fiable, on observe chez certains sujets une forte variabilité entre les estimateurs, notamment sur les positions fortement latéralisées, correspondant aux plans d'azimut $|\phi| \geq 65°$, comme le montre la Figure 3.50. C'est justement pour ces azimuts qu'on observe les plus fortes variations de l'ITD en fonction de l'élévation. Il importe donc d'adopter la plus grande prudence dans les conclusions tirées à partir des données obtenus pour un estimateur en particulier.

3.4.3 Quantification des variations de l'ITD en fonction de l'élévation

Avant de chercher à modéliser les variations de l'ITD en fonction de l'élévation, il importe de quantifier leur amplitude et d'évaluer si elles sont potentiellement perceptibles par le système auditif. Il faut d'abord se doter d'un critère permettant de refléter l'ordre de grandeur des variations de l'ITD en élévation. On se propose de considérer l'évolution de l'ITD en fonction de l'élévation sur un plan d'azimut constant. Un descripteur des variations en fonction de l'élévation pourrait être la variance associée à la valeur moyenne de l'ITD observée sur ce plan. Cependant on constate que la distribution des valeurs d'ITD n'est pas gaussienne, mais de type uniforme, c'est à dire que chaque valeur d'ITD est atteinte de façon égale. Un calcul de variance n'est donc pas adapté. On préfère calculer l'écart entre les valeurs minimale et maximale atteinte par l'ITD, ce qui décrit bien l'amplitude des variations de l'ITD :

$$MinMax_{elev}(ind, \phi) = \max_{|\theta \in [0-2\pi]}[ITD(ind, \phi, \theta)] - \min_{|\theta \in [0-2\pi]}[ITD(ind\phi, \theta)] \tag{3.23}$$

Comme précédemment, cet écart peut être interprêté en termes de JND de l'ITD. Afin d'obtenir une valeur directement exprimée en nombre de JND, on normalise le critère $MinMax_{elev}$ par la JND :

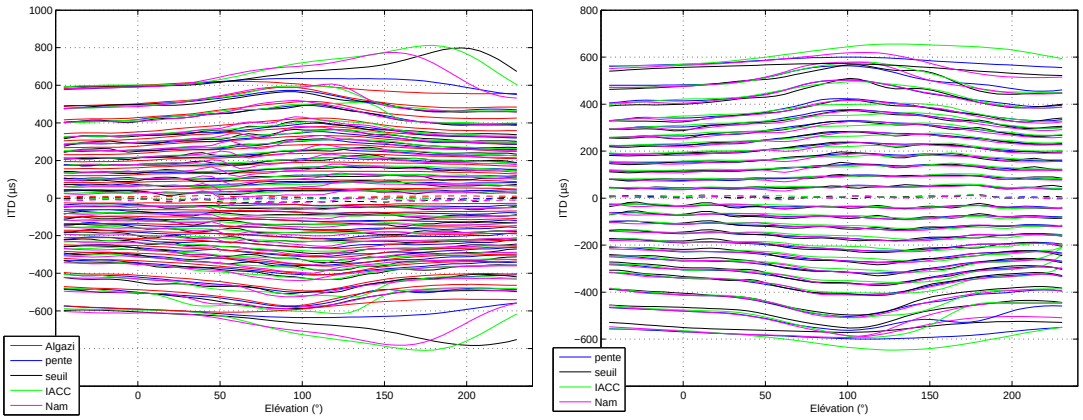
$$MinMaxNorm_{elev}(ind, \phi) = \frac{MinMax_{elev}(ind, \phi)}{JND[\overline{ITD}(ind, \phi)]} \tag{3.24}$$

où $\overline{ITD}(ind, \phi)$ définit l'ITD moyenne (moyenne sur les N élévations mesurées θ_i) calculée sur le plan d'azimut constant ϕ pour l'individu ind :

$$\overline{ITD}(ind, \phi) = \frac{1}{N} \sum_{i=1}^{N} ITD(ind, \phi, \theta_i) \ .$$

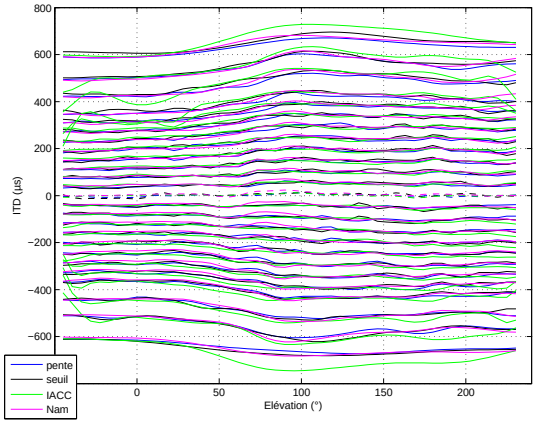
Tant que ce critère $MinMaxNorm_{elev}$ garde des valeurs inférieures ou égales à 1, on admettra que les variations de l'ITD en fonction de l'élévation ne sont pas discriminables par le système auditif.

[23] Il s'agit des 1250 directions de mesure de la base du CIPIC permettant une observation de l'ITD sur 25 plans d'azimut constant de -80° à +80° [Algazi et al., 2001d].



(a) Base de données du CIPIC : sujet 165

(b) Base de données de l'IRCAM : sujet 39



(c) Base de données *Jean-Marie Pernaux* : sujet ME

FIG. 3.50 – Ecarts entre les estimateurs de l'ITD : illustration pour 3 individus.

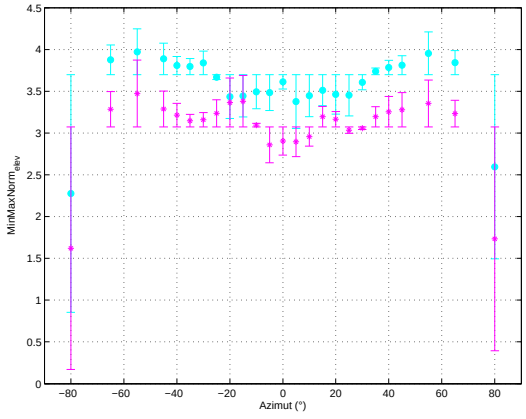
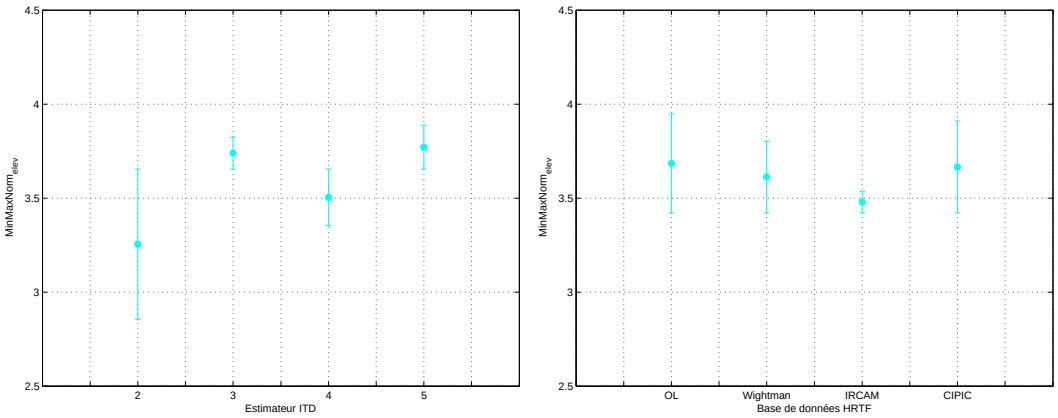


FIG. 3.51 – Quantification des variations de l'ITD en fonction de l'élévation : le critère $MinMaxNorm_{elev}$ est représenté en fonction de l'azimut (moyenne sur les individus et les estimateurs avec l'intervalle de confiance à 95%). La courbe cyan correspond aux résultats de la première étude incluant 4 base de données, un total de 34 sujets et 4 méthodes d'estimation de l'ITD. La courbe magenta correspond aux résultats de la seconde étude qui ne concerne que la base du CIPIC (37 sujets, estimateur défini dans [Algazi et al., 2001b]).



(a) (b)

FIG. 3.52 – Quantification des variations de l'ITD en fonction de l'élévation : influence de l'estimateur de l'ITD et de la base de données de HRTF (moyenne sur les individus & intervalle de confiance à 95%, données de la première étude incluant 4 base de données, un total de 34 sujets et 4 méthodes d'estimation de l'ITD).

Deux études ont été menées, elles ne diffèrent pas par la méthodologie mais par les données considérées. La première étude considère une ensemble de 34 individus sélectionnés[24] à partir de 4 bases de données (*Jean-Marie Pernaux*, IRCAM, Wightman, CIPIC). Pour chaque individu, le critère $MinMaxNorm_{elev}$ est évalué pour les 25 plans d'azimut constant compris entre -80° et +80°. La valeur moyenne de $MinMaxNorm_{elev}$ sur l'ensemble des données ainsi collectées vaut 3.57 (avec un intervalle de confiance à 95% : IC95 = $\pm$0.0387), ce qui correspond à plus de 3 fois le seuil de discrimination. Les variations de l'ITD en fonction de l'élévation sont donc perceptibles par le système auditif. Si l'on dénombre le nombre d'occurrences où la valeur du critère dépasse la JND (c'est à dire $MinMaxNorm_{elev}(ind, \phi) > 1$), il s'avère que la quasi totalité des valeurs obtenues (99.92%) est supérieure au seuil de discrimination. La Figure 3.51 représente le critère $MinMaxNorm_{elev}$ en fonction de l'azimut. L'amplitude des variations en élévation est, en termes de JND, relativement constante en fonction de l'azimut, ce qui s'explique par le fait que, même si l'amplitude des variations croît avec l'azimut, cette augmentation est compensée d'un point de vue perceptif par celle de la JND. L'amplitude des variations présente un léger maximum pour les azimuts [-65°, -45°] et [45°, 65°]. On note que l'intervalle de confiance pour les plans d'azimut $\pm$80 deg est trés élevé, ce qui semble dû à la forte divergence d'estimation de l'ITD (cf. Fig. 3.50). Le critère moyen obtenu pour chaque base de donnée est reproduit sur la Figure 3.52a. Compte tenu des intervalles de confiance, on ne relève pas de différence significative entre les bases de données. La valeur moyenne du critère a aussi été calculée par estimateur et est illustrée sur la Figure 3.52b. On observe un effet faiblement significatif de l'estimateur : les variations de l'ITD en fonction de l'élévation sont maximales pour les estimateurs 3 et 5 et minimales pour l'estimateur 2, pour lequel on note cependant un intervalle de confiance important.

[24] Les sujets non sélectionnés ont été écartés en raison d'anomalies d'estimation de l'ITD pour au moins un des estimateurs, les sujets retenus devant présenter une ITD fiable sur toute la sphère 3D pour les 4 estimateurs, ce qui est très sélectif. La liste des sujets sélectionnés est la suivante :

- base *Jean-Marie Pernaux* : ME, JD, RN,
- base de Wightman : AFW,
- base de l'IRCAM : 02, 08, 15, 22, 28, 29, 30, 32, 33, 39, 41, 43, 45, 52, 54, 57, 58, 59,
- base du CIPIC : 011, 015, 017, 019, 028, 040, 059, 124, 126, 135, 155, 165.

La seconde étude se focalise sur la base du CIPIC (37 sujets sélectionnés[25]) en considérant les valeurs d'ITD fournies dans la base de données (estimateur de type 3 décrit dans [Algazi et al., 2001b]). Cette étude corrobore parfaitement les résultats précédents. On obtient une valeur moyenne du critère $MinMaxNorm_{elev}$ de 3.05 (pour un intervalle de confiance à 95% : IC95 = $\pm$0.0660). Sur l'ensemble des valeurs, 99.68% sont supérieures à la JND. L'évolution du critère moyen en fonction de l'azimut est donnée sur la Figure 3.51. Les variations de l'ITD en élévation sont donc potentiellement audibles.

Il faut cependant souligner que le critère $MinMaxNorm_{elev}$ offre une vision majorée de l'amplitude des variations puisqu'il représente l'amplitude maximale. Il convient de compléter cette évaluation par un autre critère. Le modèle SHM de l'ITD (cf. Section 3.4.1) reproduit une ITD individualisée mais constante en élévation. Une autre façon d'évaluer l'importance des variations de l'ITD en élévation consiste à évaluer l'erreur de modélisation introduite par ce modèle lorsqu'on remplace l'ITD évoluant naturellement sur la sphère 3D (c'est à dire l'ITD estimée sur les HRTF) par une ITD indépendante de l'élévation (ITD issue du modèle SHM) :

$$Erreur_{SHM}(ind, \phi, \theta) = |ITD(ind, \phi, \theta) - ITD_{SHM}(ind, \phi, \theta)| \tag{3.25}$$

On définit l'erreur normalisée par la JND associée :

$$ErreurNorm_{SHM}(ind, \phi, \theta) = \frac{Erreur_{SHM}(ind, \phi, \theta)}{JND[\overline{ITD}(ind, \phi, \theta)]} \tag{3.26}$$

Cette erreur de modélisation est calculée sur une sélection de 17 individus[26] appartenant tous à la base de CIPIC afin de disposer des données anthropométriques permettant de mettre en oeuvre le modèle SHM. L'erreur moyenne obtenue pour l'ensemble des 17 individus et des 1250 directions vaut 1.07 (pour un intervalle de confiance à 95% : IC95 = $\pm$0.011). L'erreur de modélisation est donc de l'ordre de la JND et se situe au niveau "juste" audible. Sur la totalité des données, on compte 44.73% d'erreurs supérieures à la JND, comme en témoigne l'histogramme de l'erreur représenté sur la Figure 3.53. La Figure 3.54 précise l'évolution de l'erreur en fonction de l'azimut. L'erreur est inaudible pour les azimuts $\phi \leq -65°$ et $\phi \geq 55°$, mais elle passe au dessus du seuil de discrimination sur la plage [-55°, 45°]. Les résultats obtenus du point de vue de l'erreur $ErreurNorm_{SHM}$ conduisent à nuancer l'impact potentiel des variations de l'ITD en élévation par rapport au critère $MinMaxNorm_{elev}$. Le risque d'artefacts audibles s'avère critique pour un peu moins de la moitié des cas contre quasiment 100% des cas avec le premier critère. Néanmoins ce risque demeure élevé et est présent pour une large portion de la sphère 3D puisqu'il concerne les plans d'azimut compris entre -55° et 45°, ce qui justifie la recherche d'un nouveau modèle d'ITD permettant de prendre en compte ses variations en élévation.

3.4.4 Proposition d'un modèle d'ITD basé sur une tête sphérique avec individualisation du rayon de la sphère et du positionnement des oreilles

Pour reproduire les variations de l'ITD en élévation, on se propose d'introduire un degré de liberté supplémentaire dans le modèle SHM, concernant le positionnement des oreilles sur la sphère. Cette piste est en effet jugée prometteuse dans [Duda et al., 1999] [Algazi et al., 2001a] pour un

[25] Comme précédemment, le critère de sélection est une ITD fiable sur toute la sphère 3D. La liste des sujets sélectionnés est la suivante : 003, 008, 009, 010, 011, 012, 015, 017, 019, 021, 027, 028, 033, 040, 050, 051, 058, 059, 060, 061, 065, 119, 124, 126, 127, 133, 134, 135, 137, 147, 148, 155, 156, 158, 162, 163, 165.

[26] Les critères de sélection ont été les suivants : données anthropométriques disponibles et ITD estimée ne comportant pas d'évolution spatiale aberrante. La liste des sujets sélectionnés est la suivante : 003, 010, 021, 027, 028, 040, 050, 058, 059, 060, 126, 127, 131, 133, 147, 155, 165.

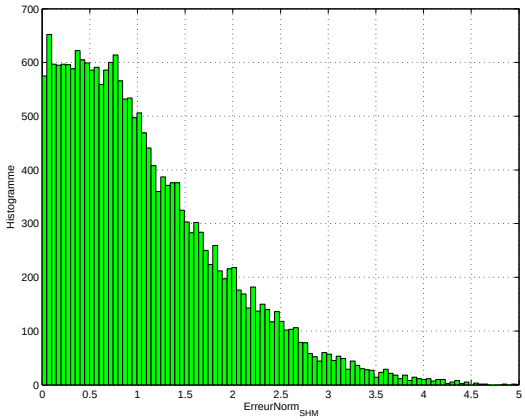


FIG. 3.53 – Histogramme de l'erreur de modélisation $ErreurNorm_{SHM}$ du modèle SHM (17 sujets, 1250 directions, base de données du CIPIC, estimateur décrit dans [Algazi et al., 2001b]).

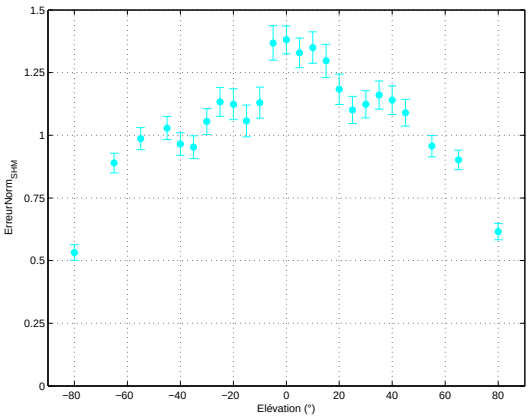
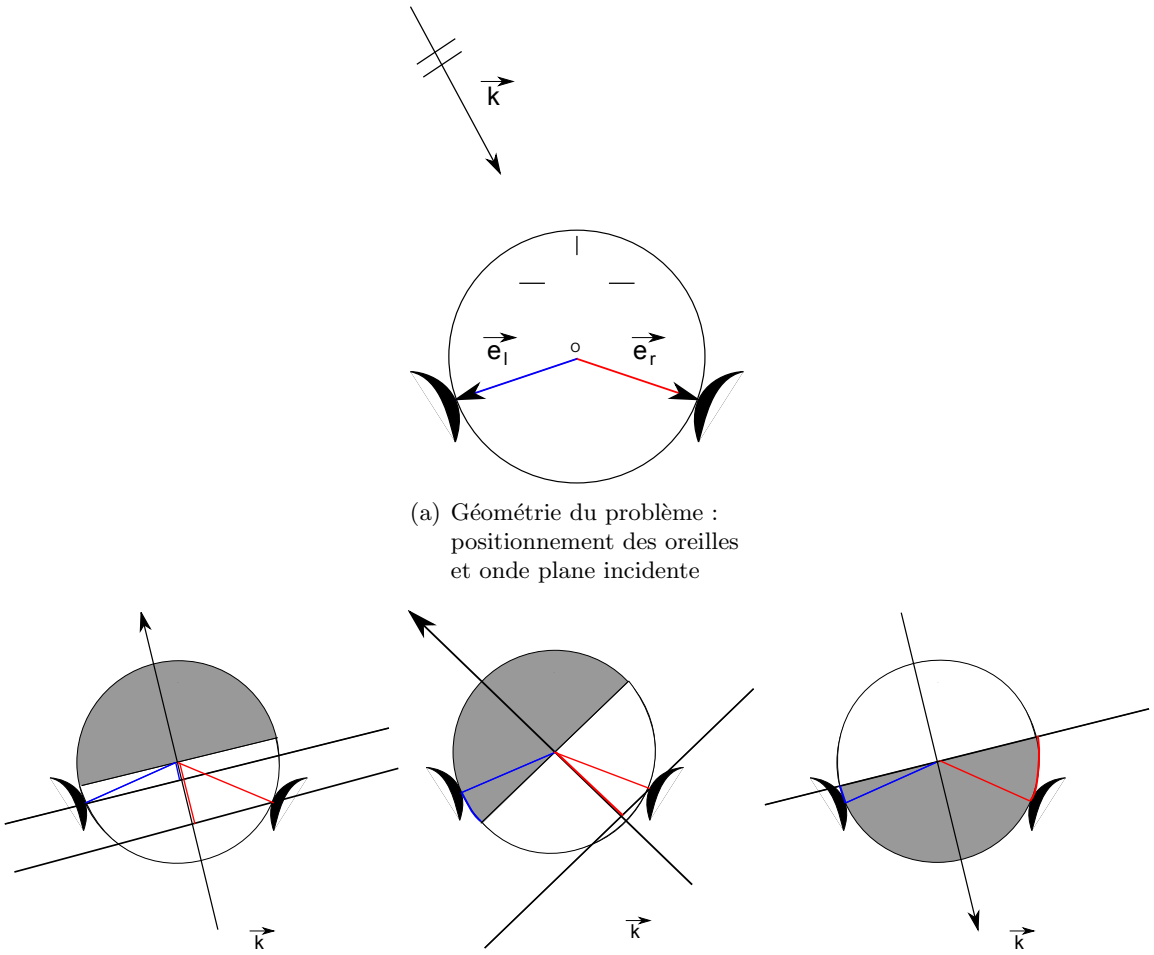


FIG. 3.54 – Erreur de modélisation $ErreurNorm_{SHM}$ du modèle SHM en fonction de l'azimut ϕ (17 sujets, base de données du CIPIC, estimateur décrit dans [Algazi et al., 2001b]). Moyenne sur les individus et l'angle d'élévation.



(a) Géométrie du problème : positionnement des oreilles et onde plane incidente

(b) Cas 1 : les deux oreilles sont éclairées par l'onde acoustique incidente

(c) Cas 2 : une oreille est éclairée et l'autre dans l'ombre

(d) Cas 3 : les deux oreilles sont situées dans l'ombre

FIG. 3.55 – Modèle SHMWOE : illustration géométrique des 3 cas considérés.

modèle de tête sphérique ou ellipsoïdal. Les auteurs montrent qu'en décalant de quelques centimètres les oreilles par rapport à un positionnement centré, c'est à dire diamétralement opposé, une dépendance de l'ITD en élévation est obtenue, ce qui est confirmé par [Pernaux, 2003]. Dans [Algazi et al., 2001a] [Pernaux, 2003], l'ITD est extraite de HRIR calculées à partir d'une résolution analytique de la diffraction d'une onde acoustique par une sphère. Il manque une formulation analytique de l'ITD. Duda *et al* posent le problème de façon plus rigoureuse, mais comme il considère un modèle ellipsoïdal de tête, aucune solution analytique ne peut être exprimée [Duda et al., 1999]. Notre objectif est d'établir une formule analytique permettant de calculer directement l'ITD, à la manière des modèles de Woodworth (Equ. 3.20) ou SHM (Equ. 3.21). Le point de départ est donc le modèle de tête sphérique mis au point par Woodworth [Blauert, 1983]. Dans notre modèle, on considère que les oreilles peuvent être situées n'importe où sur la sphère. Leurs positions sont repérées par les vecteurs $\vec{e_l}$ et $\vec{e_r}$ respectivement pour l'oreille gauche et droite (cf. Fig. 3.55.a). La contrainte de la symétrie par rapport au plan médian est conservée, c'est à dire qu'une fois que la position d'une des oreilles est fixée, la position de l'autre oreille s'en déduit par symétrie par rapport au plan médian. En d'autres termes le décalage des oreilles est contrôlé par un seul paramètre : le vecteur[27] $\vec{e_o}$ correspondant indifféremment soit à $\vec{e_l}$, soit $\vec{e_r}$, l'autre vecteur s'en déduisant par symétrie. L'onde acoustique incidente est une onde plane se propageant dans la direction (ϕ, θ) repérée par son vecteur d'onde $\vec{k}$ (cf. Fig. 3.55). Conformément au modèle de Woodworth [Blauert, 1983], cette onde est diffractée par la sphère qui est alors décomposée en deux moitiés (cf. Fig. 3.55) : la demi-sphère *éclairée* pour laquelle s'applique une propagation directe et la demi-sphère qui reste dans l'ombre et pour laquelle l'onde se propage par *contournement*, c'est à dire en suivant la surface de la sphère. Selon les positions des oreilles et l'incidence de l'onde, trois cas doivent être distingués :

- Cas 1 : les deux oreilles sont situées sur la demi-sphère éclairée, cf. Fig. 3.55.b,
- Cas 2 : une oreille est située dans la zone éclairée, et l'autre dans l'ombre, cf. Fig. 3.55.c,
- Cas 3 : les deux oreilles sont placées dans l'ombre, cf. Fig. 3.55.d.

L'ITD correspondant à ces 3 cas s'exprime :

$$ITD(\phi,\theta) = \begin{cases} \frac{R}{c}\left[\vec{e_l}\cdot\frac{\vec{k}}{k} - \vec{e_r}\cdot\frac{\vec{k}}{k}\right] & \text{Cas 1} \\ \frac{R}{c}\left[\vec{e_r}\cdot\frac{\vec{k}}{k} - \frac{\pi}{2} + \arccos(\vec{e_l}\cdot\frac{\vec{k}}{k})\right] & \text{Cas 2 : oreille gauche éclairée} \\ \frac{R}{c}\left[\frac{\pi}{2} - \arccos(\vec{e_r}\cdot\frac{\vec{k}}{k}) - \vec{e_l}\cdot\frac{\vec{k}}{k}\right] & \text{Cas 2 : oreille droite éclairée} \\ \frac{R}{c}\left[\arccos(\vec{e_r}\cdot\frac{\vec{k}}{k}) - \arccos(\vec{e_l}\cdot\frac{\vec{k}}{k})\right] & \text{Cas 3} \end{cases} \tag{3.27}$$

où :

$$k = |\vec{k}|.$$

Ces équations définissent le nouveau modèle d'ITD que nous avons proposé [Busson, 2006] et que nous désignerons par la suite sous l'acronyme SHM-WOE pour *Spherical Head Model - With Offset Ears* en référence à sa première description dans [Algazi et al., 2001a]. Les 2 paramètres de ce modèle (R, $\vec{e_o}$) dépendent de l'individu.

Les figures 3.56 et 3.57 illustrent l'influence des paramètres R et $\vec{e_o}$ du modèle sur l'ITD modélisée. Pour une meilleure compréhension, le décalage des oreilles est exprimé en termes de différentiel d'azimut $\Delta\phi$ et d'élévation $\Delta\theta$ par rapport à la position de référence correspondant aux oreilles diamétralement opposées ($\Delta\phi = 0°$, $\Delta\theta = 0°$). Ainsi un décalage de $\Delta\phi = +15°$ définit un déplacement des oreilles vers l'avant (à l'opposé $\Delta\phi = -15°$ correspond à un déplacement vers l'arrière), tandis que $\Delta\theta = +15°$ définit un déplacement des oreilles vers le haut (à l'opposé $\Delta\theta = -15°$ correspond

[27] Le vecteur $\vec{e_o}$ en lui-même est décrit (en termes de coordonnées sphériques) par 2 paramètres : un angle d'azimut et un angle d'élévation, le rayon étant imposé par le rayon R de la sphère.

à un déplacement vers le bas). Sur la Figure 3.56a, on vérifie qu'en l'absence de décalage, le modèle donne une ITD indépendant de l'élévation. Dès qu'on introduit un décalage des oreilles, apparaît une bosse dont la localisation en élévation coïncide avec les angles de décalage :

- décalage vers l'avant ($\Delta\phi = 15°$, $\Delta\theta = 0°$) : bosse à l'élévation 180° (cf. Fig. 3.56b).
- décalage vers le haut ($\Delta\phi = 0°$, $\Delta\theta = 15°$) : bosse à l'élévation 270° (cf. Fig. 3.56c).
- décalage vers l'arrière ($\Delta\phi = -15°$, $\Delta\theta = 0°$) : bosse à l'élévation 0° (cf. Fig. 3.56d).
- décalage vers le bas ($\Delta\phi = 0°$, $\Delta\theta = -15°$) : bosse à l'élévation 90° (cf. Fig. 3.56e).

Si l'on compare ces courbes à celles obtenues à partir des HRTF mesurées (cf. Fig. 3.41), il ressort qu'un décalage vers le bas est le plus plausible. Sur la Figure 3.57, on vérifie que l'ITD augmente avec le rayon de la sphère modélisant la tête.

3.4.5 Mise en œuvre et validation objective du modèle SHM-WOE

Calcul des paramètres optimaux du modèle pour chaque azimut

Dans une première étape d'évaluation des capacités du modèle à reproduire les variations spatiales d'une ITD naturelle, les paramètres de décalage *optimaux* $(R, \Delta\phi, \Delta\theta)_{opt}$, c'est à dire correspondant à l'erreur minimale entre l'ITD estimée sur les mesures et l'ITD modélisée, sont calculés pour chaque plan d'azimut pris isolément. En d'autres termes, on obtient un jeu de paramètres de décalage $(R, \Delta\phi, \Delta\theta)_{opt}(\phi_i)$ pour chaque azimut ϕ_i. Cette mise en œuvre du modèle n'est pas réaliste au sens où d'un point de vue physique et morphologique, un jeu unique de paramètres doit contrôler le modèle quel que soit l'azimut. Cependant, dans la Section 3.4.2, on a constaté que les variations de l'ITD en élévation pouvaient différer sensiblement d'un azimut à l'autre. De plus il est apparu que les variations de l'ITD en élévation, notamment la bosse de l'ITD autour de $\theta = 90°$, sont plus lisibles sur certains plans d'azimut (en général $\phi = \pm 65°$) que pour d'autres. On veut donc ici observer comment les paramètres de décalage évoluent avec l'azimut et s'il convient de choisir un azimut en particulier pour l'optimisation des paramètres du modèle. Quoi qu'il en soit, compte tenu des observations faites en Section 3.4.2, une optimisation *globale* des paramètres du modèle, cherchant le jeu de paramètres qui minimise l'erreur de modélisation conjointement pour tous les azimuts, est écartée. Le principal apport du modèle est la bosse de l'ITD au voisinage de $\theta = 90°$. Or cette bosse n'est souvent bien définie que pour les azimuts $\phi = \pm 65°$ ou leur voisinage immédiat. L'information issue des azimuts plus proches du plan médian apparaît davantage noyée dans le bruit de mesure et d'estimation. Dans ces conditions, considérer globalement l'ensemble des azimuts risque de polluer l'information utile issue des azimuts les plus lisibles par l'information bruitée des autres azimuts. Pour ces raisons nous avons opté pour une stratégie de calcul des paramètres du modèle considérant séparément chaque plan d'azimut.

Pour un individu donné, les paramètres optimaux de décalage $(R, \Delta\phi, \Delta\theta)_{opt}(\phi_i)$ sont calculés pour chaque azimut par minimisation de l'erreur quadratique moyenne :

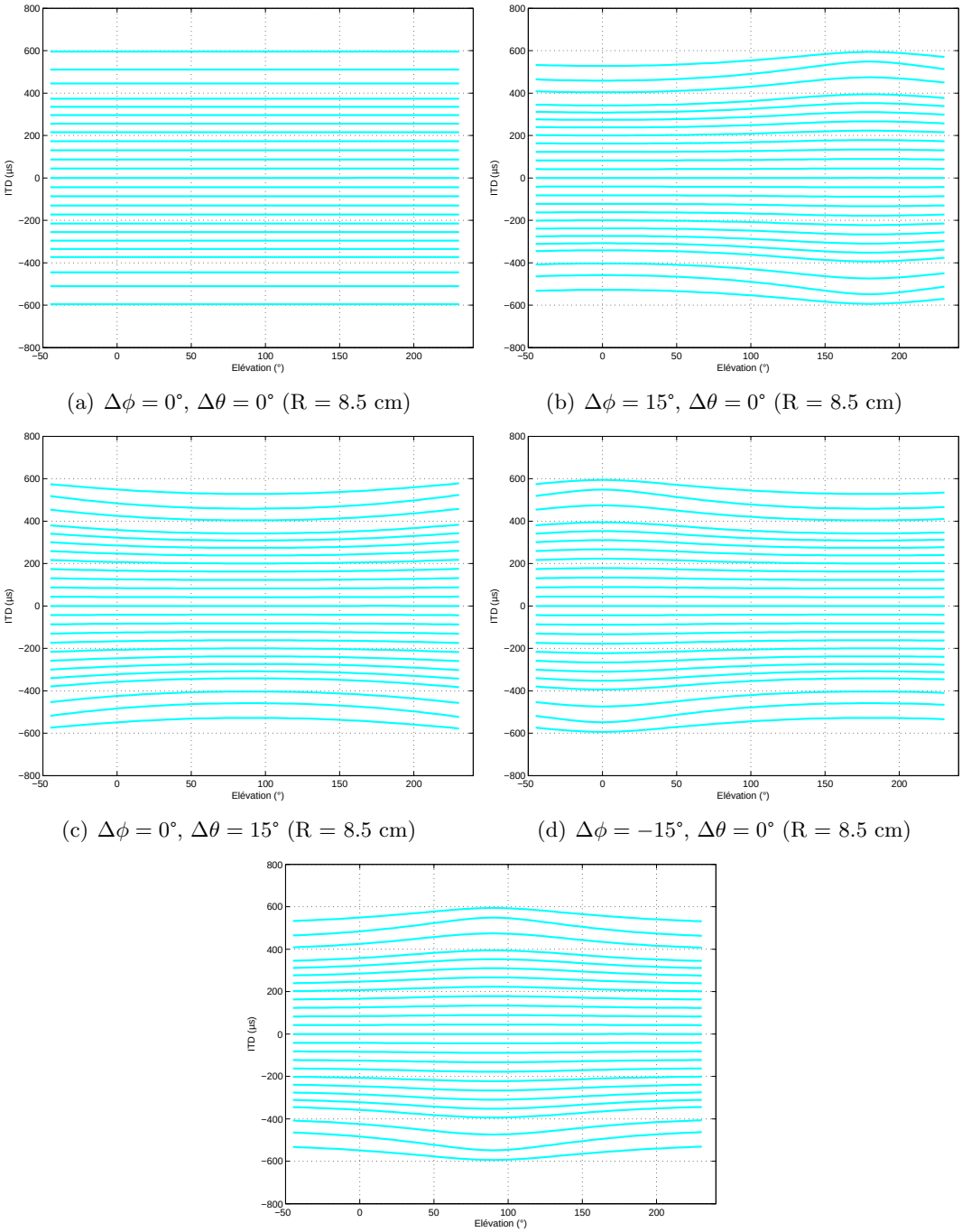
$$E_q(ind, \phi_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N} |ITD(ind, \phi_i, \theta_j) - ITD_{SHM}(ind, \phi_i, \theta_j)|^2} \tag{3.28}$$

où N désigne le nombre d'élévations (N = 50). L'étude porte sur un total de 36 individus extraits de 3 bases[28] de données. Pour évaluer les performances de modélisation, on se dote du critère d'erreur

[28] Il s'agit des bases suivantes :

- base *Jean-Marie Pernaux*(estimateur d'ITD 5) : sujets ME, JMP, JD, RN, MA, NC,
- base de Wightman (estimateur 5 d'ITD) : sujets AFW et SOW,
- base de l'IRCAM (estimateur 3 d'ITD) : sujets 02, 03, 04, 05, 06, 07, 12, 13, 15, 16, 17, 21, 22, 23, 26, 30, 33, 40, 41, 43, 44, 45, 47, 51, 52, 54, 57, 58.

.



(a) $\Delta\phi = 0°$, $\Delta\theta = 0°$ (R = 8.5 cm)

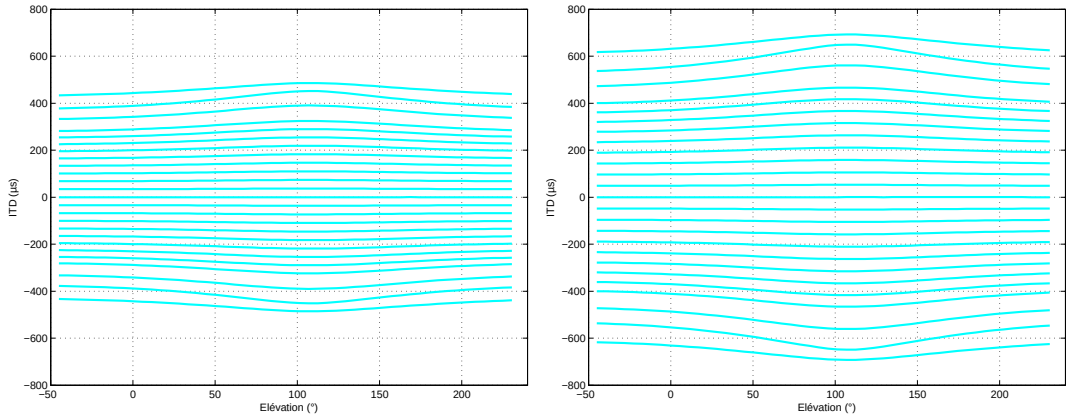
(b) $\Delta\phi = 15°$, $\Delta\theta = 0°$ (R = 8.5 cm)

(c) $\Delta\phi = 0°$, $\Delta\theta = 15°$ (R = 8.5 cm)

(d) $\Delta\phi = -15°$, $\Delta\theta = 0°$ (R = 8.5 cm)

(e) $\Delta\phi = 0°$, $\Delta\theta = -15°$ (R = 8.5 cm)

FIG. 3.56 – Impact du décalage des oreilles sur l'ITD obtenue par le modèle SHM-WOE : ITD représentée en fonction de l'angle d'élévation pour les 25 plans d'azimut compris entre -80° et +80°.



(a) R = 7 cm ($\Delta\phi = 5°$, $\Delta\theta = -15°$) (b) R = 10 cm ($\Delta\phi = 5°$, $\Delta\theta = -15°$)

FIG. 3.57 – Impact du rayon sur l'ITD obtenue par le modèle SHM-WOE : ITD représentée en fonction de l'angle d'élévation pour les les 25 plans d'azimut compris entre -80° et +80°.

Modèle	$ErreurNorm_{SHM-WOE}$	$P\%_{>1}$
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{opt}(\phi_i)$	0.35 [± 0.0031]	4.84
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[20^o-65^o]}$	0.64 [± 0.0049]	21.44
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[20^o-55^o]}$	0.65 [± 0.0051]	22.08
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[55^o-65^o]}$	0.76 [± 0.0054]	29.23
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{55^o}$	0.64 [± 0.0048]	21.65
$(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{65^o}$	0.93 [± 0.068]	37.62

TAB. 3.6 – Performances de modélisation du modèle SHM-WOE : erreur moyenne de modélisation $ErreurNorm_{SHM-WOE}$ (moyenne sur les 36 individus et les 1250 directions) et pourcentage $P\%_{>1}$ de valeurs d'erreurs supérieures à 1 (seuil de discrimination). La configuration du modèle spécifiée par les paramètres $(R, \Delta\phi, \Delta\theta)_{opt}(\phi_i)$ correspond à la mise en œuvre des paramètres optimaux dépendant de l'azimut. Dans toutes les autres configurations, un jeu unique de paramètres est appliqué pour l'ensemble des azimuts.

suivant :

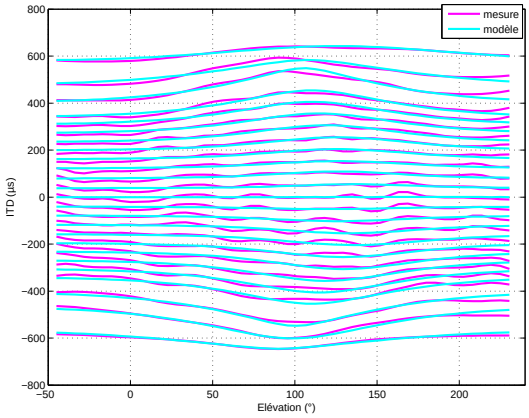
$$Erreur_{SHM-WOE}(ind,\phi,\theta) = |ITD(ind,\phi,\theta) - ITD_{SHM-WOE}(ind,\phi,\theta)| \tag{3.29}$$

ainsi que l'erreur associée après normalisation par la JND :

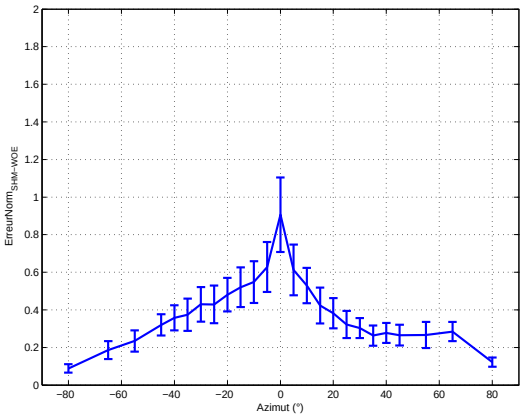
$$ErreurNorm_{SHM-WOE}(ind,\phi,\theta) = \frac{Erreur_{SHM-WOE}(ind,\phi,\theta)}{JND[\overline{ITD}(ind,\phi,\theta)]} \tag{3.30}$$

En complément est calculé le pourcentage de cas où l'erreur $ErreurNorm_{SHM-WOE}$ dépasse la valeur de 1. Pour l'ensemble des données disponibles (36 individus, 1250 directions), l'erreur moyenne $ErreurNorm_{SHM-WOE}$ vaut 0.35 (cf. Tab. 3.6), ce qui est très en dessous du seuil de discrimination. On dénombre un peu moins de 5% d'erreurs supérieures à 1 (cf. Tab. 3.6). Le modèle SHM-WOE offre donc d'excellentes perfomances de modélisation : l'erreur peut être considérée comme inaudible, ce qui signifie que, d'un point de vue perceptif, le modèle est transparent. Les figures 3.58a-b, 3.59a-b, 3.60a-b et 3.61a-b illustrent l'ITD modélisée en comparaison de l'ITD estimée sur les mesures, ainsi que l'erreur de modélisation pour 4 individus. On vérifie combien le modèle SHM-WOE réussit à reproduire les variations individualisées de l'ITD en élévation. L'erreur de modélisation est représentée en fonction de l'azimut. L'erreur est maximale dans le plan médian, notamment en raison du seuil de discrimination de l'ITD qui est très faible et qui vient pénaliser la moindre erreur, aussi faible soit-elle. Ainsi la présence d'oscillations dans le plan d'azimut $\phi = 0\,\text{deg}$, qui semblent dues au biais de l'estimateur d'ITD, vient grever l'erreur de modélisation. Pour les autres azimuts, l'erreur décroît rapidement dès qu'on s'écarte du plan médian. Dans l'ensemble l'erreur reste majoritairement inférieure au seuil de 1. Il convient de nuancer ces résultats en rappelant qu'ils sont les meilleurs qu'on puisse espérer obtenir, puisque les paramètres de décalage sont ajustés séparément pour chaque plan d'azimut. Dans une utilisation réelle du modèle, où un jeu unique de paramètres sera appliqué par individu, il faut s'attendre à une dégradation des performances de modélisation.

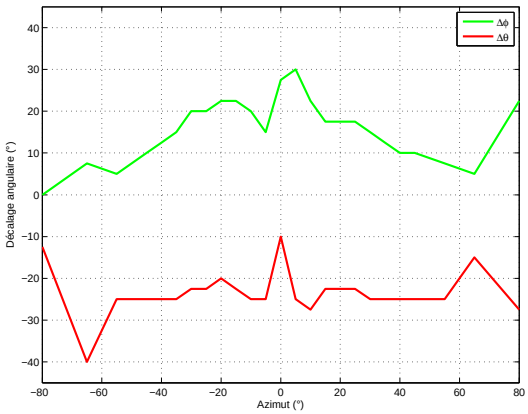
Les figures 3.58.c-e, 3.59.c-e, 3.60.c-e et 3.61.c-e représentent les paramètres optimaux de décalage $\Delta\phi$ et $\Delta\theta$, ainsi que le rayon optimal de la sphère en fonction de l'azimut. On vérifie d'abord que les paramètres sont bien symétriques entre la gauche et la droite. De légères asymétries peuvent être relevées (cf. Fig. 3.60c-e) mais on vérifie que l'ITD estimée présente déjà une asymétrie (cf. Fig. 3.60a) et le modèle ne fait que la reproduire. On observe que le décalage $\Delta\theta$ présente une plage d'azimuts (de l'ordre de $|\phi| \in [20^o, ..., 65^o]$) où il prend des valeurs quasiment constantes quel que soit l'azimut. On a en effet remarqué précédemment que les variations de l'ITD en élévation sont en général les plus lisibles sur cette plage. Cette plage de valeurs uniformes est en soi une preuve de la fiabilité du modèle. Le décalage observé correspond à un glissement vers le bas de 15 à 25°, soit 3 à 4 cm, ce qui est cohérent avec les observations morphologiques. En revanche, au voisinage du plan médian et dans une moindre mesure pour les azimuts les plus latéralisés (principalement $\phi = 80^o$), les paramètres de décalage présentent des valeurs incohérentes, voire aberrantes, alliées à une forte instabilité (cf. Fig. 3.60c-e). Contrairement au décalage haut/bas, le décalage $\Delta\phi$ ne présente pas de valeurs stables en fonction de l'azimut (cf. Fig. 3.59c-d). Pour le sujet 03 de la base de l'IRCAM, on obtient même un décalage de 5 cm vers l'avant dans la plan médian contre un décalage de 4 cm vers l'arrière à l'azimut $\phi = \pm 80^o$. Le paramètre $\Delta\phi$ semble plus difficile à déterminer à partir de l'ITD estimée que $\Delta\theta$. D'ailleurs l'allure générale de l'évolution de l'ITD en fonction de l'élévation indique que la bosse autour de $\theta = 90^o$ est principalement due à un décalage vers le bas (cf. Fig. 3.41 & 3.56e). Il est possible que, du fait qu'il soit faible et par suite très sensible au bruit de mesure et d'estimation, le décalage $\Delta\phi$ soit délicat à identifier. De manière générale, l'optimisation des paramètres du modèle reste sensible aux erreurs d'estimation de l'ITD. Pour le sujet 05 de la base de l'IRCAM (cf. Fig. 3.61a,c-e), on observe ainsi comment les oscillations présentes sur l'ITD estimée viennent entâcher d'instabilités les paramètres de décalage et le rayon.



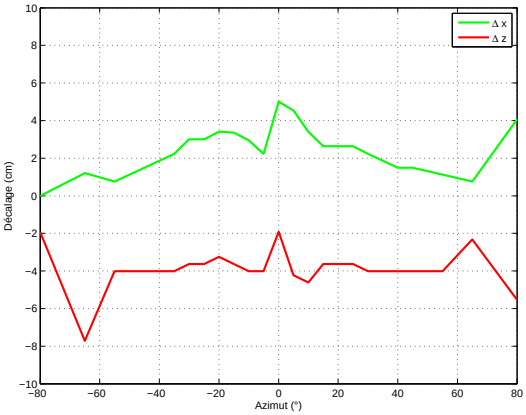
(a) ITD estimée et ITD modélisée



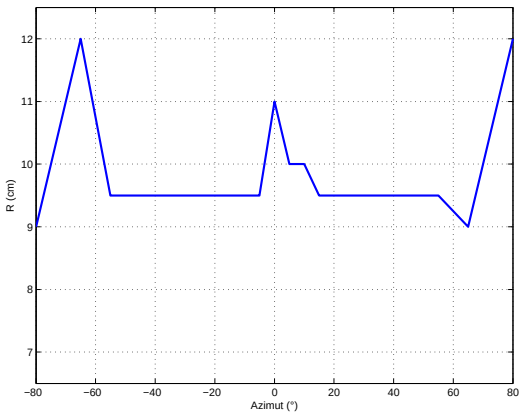
(b) Erreur de modélisation en fonction de l'azimut



(c) Paramètres optimaux du modèle en fonction de l'azimut : $\Delta\phi$ et $\Delta\theta$

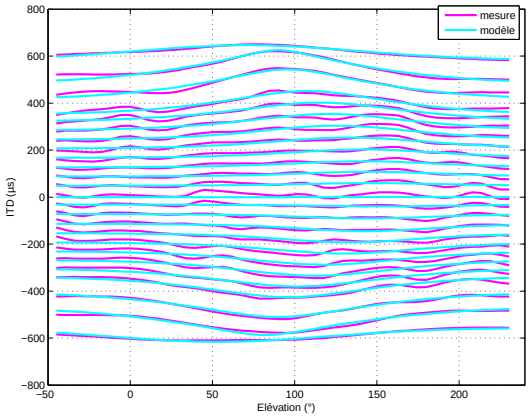


(d) Paramètres optimaux du modèle en fonction de l'azimut : traduction des décalages angulaires en termes de coordonnées cartésiennes

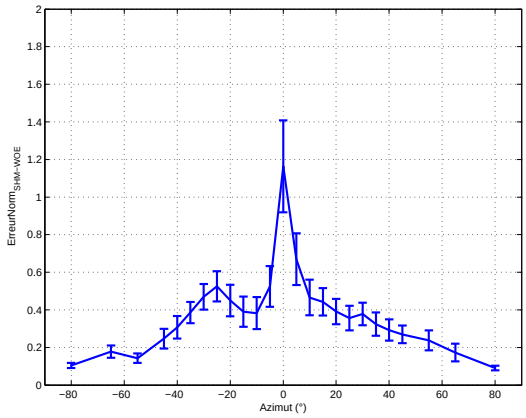


(e) Paramètres optimaux du modèle en fonction de l'azimut : rayon R de la sphère

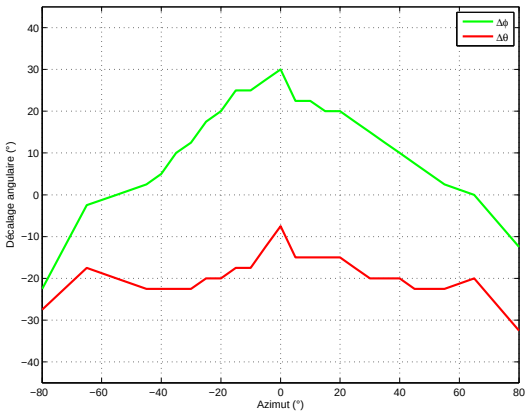
FIG. 3.58 – Modélisation de l'ITD du sujet 17 de la base de l'IRCAM : erreur de modélisation (moyenne sur les 50 élévations et intervalle de confiance à 95% associé) et paramètres optimaux du modèle en fonction de l'azimut.



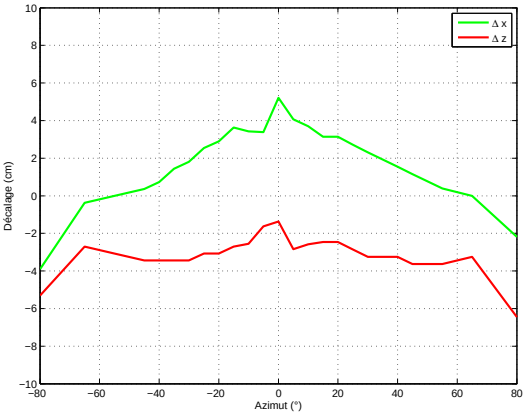
(a) ITD estimée et ITD modélisée



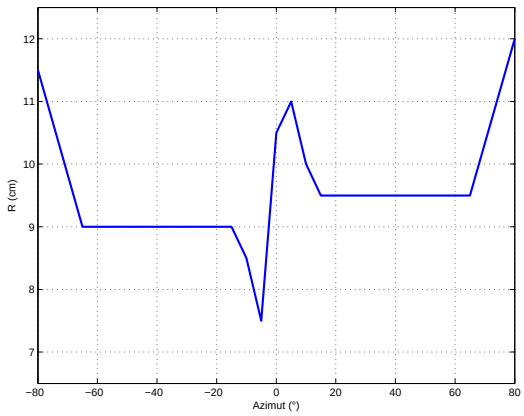
(b) Erreur de modélisation en fonction de l'azimut



(c) Paramètres optimaux du modèle en fonction de l'azimut : $\Delta\phi$ et $\Delta\theta$

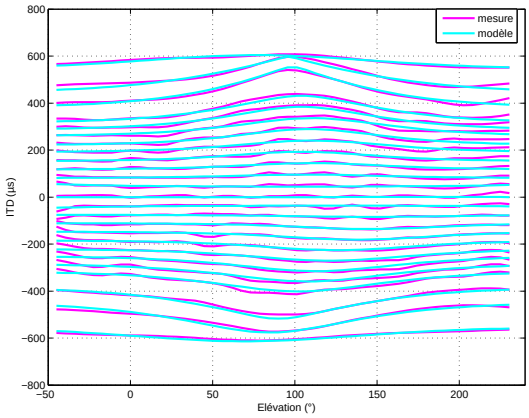


(d) Paramètres optimaux du modèle en fonction de l'azimut : traduction des décalages angulaires en termes de coordonnées cartésiennes

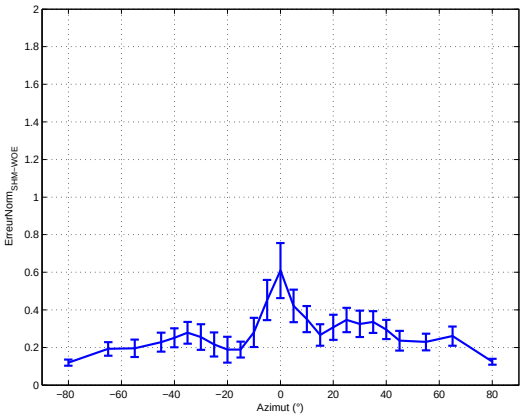


(e) Paramètres optimaux du modèle en fonction de l'azimut : rayon R de la sphère

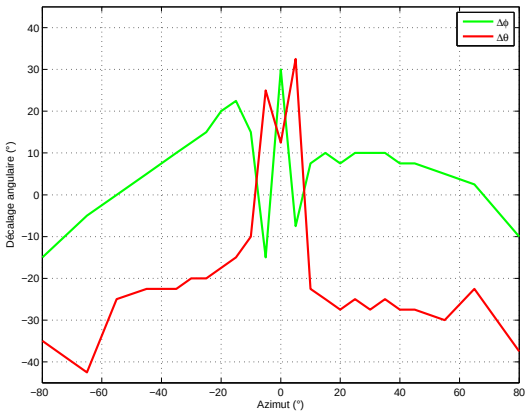
FIG. 3.59 – Modélisation de l'ITD du sujet 21 de la base de l'IRCAM : erreur de modélisation (moyenne sur les 50 élévations et intervalle de confiance à 95% associé) et paramètres optimaux du modèle en fonction de l'azimut.



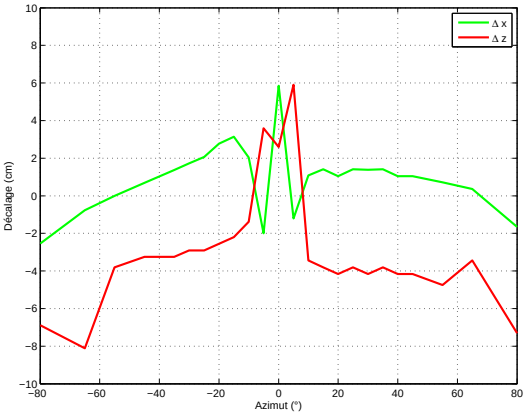
(a) ITD estimée et ITD modélisée



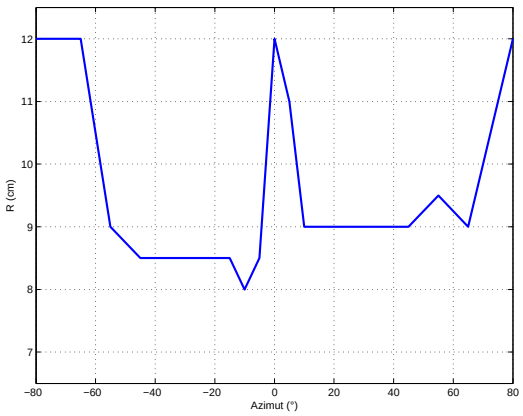
(b) Erreur de modélisation en fonction de l'azimut



(c) Paramètres optimaux du modèle en fonction de l'azimut : $\Delta\phi$ et $\Delta\theta$

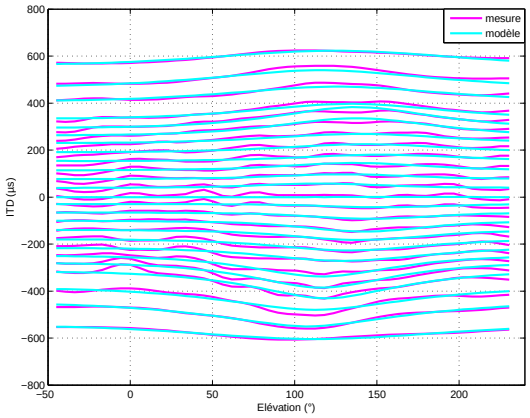


(d) Paramètres optimaux du modèle en fonction de l'azimut : traduction des décalages angulaires en termes de coordonnées cartésiennes

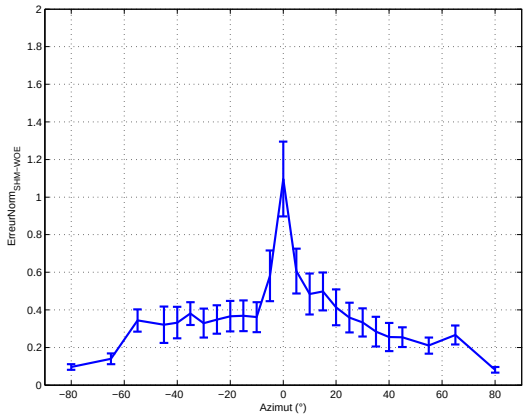


(e) Paramètres optimaux du modèle en fonction de l'azimut : rayon R de la sphère

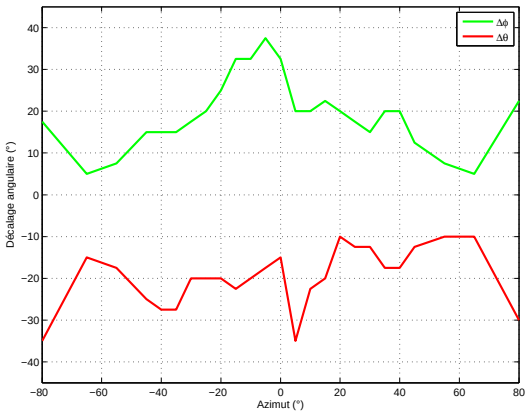
FIG. 3.60 – Modélisation de l'ITD du sujet 03 de la base de l'IRCAM : erreur de modélisation (moyenne sur les 50 élévations et intervalle de confiance à 95% associé) et paramètres optimaux du modèle en fonction de l'azimut.



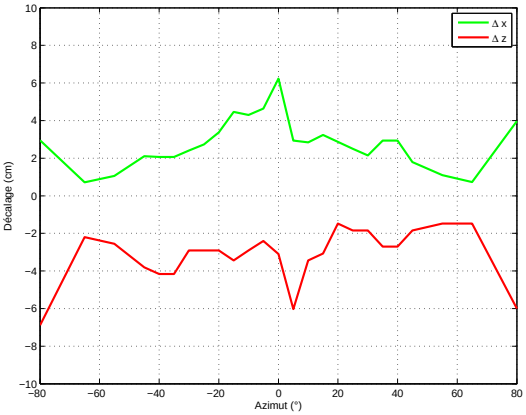
(a) ITD estimée et ITD modélisée



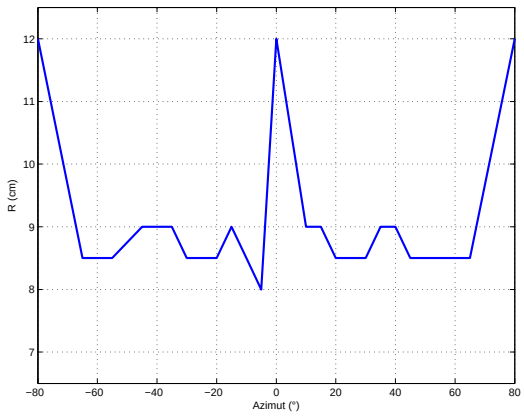
(b) Erreur de modélisation en fonction de l'azimut



(c) Paramètres optimaux du modèle en fonction de l'azimut : $\Delta\phi$ et $\Delta\theta$



(d) Paramètres optimaux du modèle en fonction de l'azimut : traduction des décalages angulaires en termes de coordonnées cartésiennes



(e) Paramètres optimaux du modèle en fonction de l'azimut : rayon R de la sphère

FIG. 3.61 – Modélisation de l'ITD du sujet 05 de la base de l'IRCAM : erreur de modélisation (moyenne sur les 50 élévations et intervalle de confiance à 95% associé) et paramètres optimaux du modèle en fonction de l'azimut.

Extraction d'un jeu unique de paramètres pour un individu donné

Le modèle SHM-WOE est contrôlé par 3 paramètres indépendants : rayon R de la sphère, décalages angulaires $\Delta\phi$ et $\Delta\theta$ des oreilles. Dans ce qui précède, les valeurs optimales de ces paramètres ont été calculées pour chaque azimut pris isolément. A présent nous allons examiner comment, à partir de ces valeurs dépendant de l'azimut, identifier un jeu unique de paramètres permettant pour un individu donné de modéliser au mieux l'ITD. Compte tenu des fortes instabilités des paramètres en dehors de la plage $|\phi| \in [20^o, ..., 65^o]$, on ne retient que les valeurs situées dans cette plage. Il existe plusieurs solutions pour extraire une valeur unique, correspondant à différentes sélections d'azimuts pour calculer une valeur moyenne des paramètres. Ainsi, par exemple, pour le rayon R de la tête, on peut extraire les valeurs suivantes :

$$\overline{R}_{[20^o-65^o]} = \frac{1}{N}\sum_{i=1}^{N} R(\phi_i, \phi_i \in [-65^o, -55^o, ..., -20^o, 20^o, ..., 55^o, 65^o] \quad (3.31)$$

$$\overline{R}_{[20^o-55^o]} = \frac{1}{N}\sum_{i=1}^{N} R(\phi_i, \phi_i \in [-55^o, ..., -20^o, 20^o, ..., 55^o] \quad (3.32)$$

$$\overline{R}_{[55^o-65^o]} = \frac{1}{N}\sum_{i=1}^{N} R(\phi_i, \phi_i \in [-65^o, -55^o, 55^o, 65^o] \quad (3.33)$$

$$\overline{R}_{55^o} = \frac{1}{N}\sum_{i=1}^{N} R(\phi_i, \phi_i \in [-55^o, 55^o] \quad (3.34)$$

$$\overline{R}_{65^o} = \frac{1}{N}\sum_{i=1}^{N} R(\phi_i, \phi_i \in [-65^o, 65^o] \quad (3.35)$$

$$(3.36)$$

La même loi étant appliquée aux 2 autres paramètres $\Delta\phi$ et $\Delta\theta$. Au final, on a évalué 5 propositions de jeu de paramètres (cf. Tab. 3.6). L'erreur la plus faible est atteinte par le jeu $(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[20^o-65^o]}$. Elle vaut 0.64, soit un peu moins du double de l'erreur optimale. Elle demeure largement au dessous du seuil de discrimination et le modèle peut donc être considéré comme transparent. On compte un peu plus de 20% de valeurs supérieures à 1. On note qu'au lieu de faire la moyenne sur l'intervalle $[20\,\text{deg}, ..., 65\,\text{deg}]$, on peut se contenter de calculer la moyenne des paramètres à +55° et -55°, qui semble une aussi bonne estimation des paramètres du modèle, les performances de modélisation étant sensiblement équivalentes. L'erreur de modélisation est représentée en fonction de l'azimut sur la Figure 3.62. Elle reste proche de l'erreur optimale, sauf pour les azimuts les plus latéralisés ($|\phi| \geq 55\,\text{deg}$). Une hypothèse pour expliquer ce comportement marginal est la présence du pavillon dont le rôle devient plus sensible au fur et à mesure où l'incidence de l'onde acoustique se rapproche de l'axe interaural. Or le modèle SHM-WOE ne permet pas de rendre compte des perturbations du trajet de propagation engendrées par le pavillon, ce qui explique la dégradation des performances de modélisation.

Lien entre les paramètres du modèle et la morphologie

Il reste à prédire les paramètres du modèle à partir des données morphologiques de l'auditeur : ainsi par une simple observation de la morphologie d'un individu, il sera possible de calculer son ITD individualisée pour toute la sphère 3D. Comme seule la base du CIPIC dispose de données anthropométriques, nous nous focalisons sur cette base pour cette étude en reprenant la sélection des

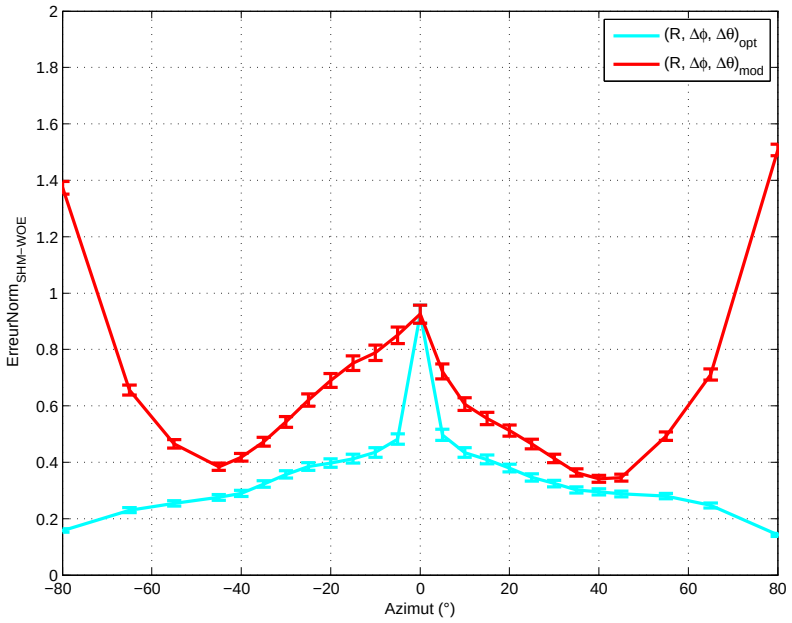


FIG. 3.62 – Erreur moyenne de modélisation du modèle SHM-WOE en fonction de l'azimut : erreur obtenue avec les paramètres optimisés séparément pour chaque azimut $(R, \Delta\phi, \Delta\theta)_{opt}$ et avec un jeu unique (c'est à dire indépendant de l'azimut) de paramètres $(R, \Delta\phi, \Delta\theta)_{mod}$ correspondant à la moyenne des paramètres optimaux calculée sur la plage d'azimuts [-65°,...,-20°,20°,...,65°]. Moyenne sur les élévations et les individus (36 individus extraits des bases d'Orange Labs, de Wightman et de l'IRCAM).

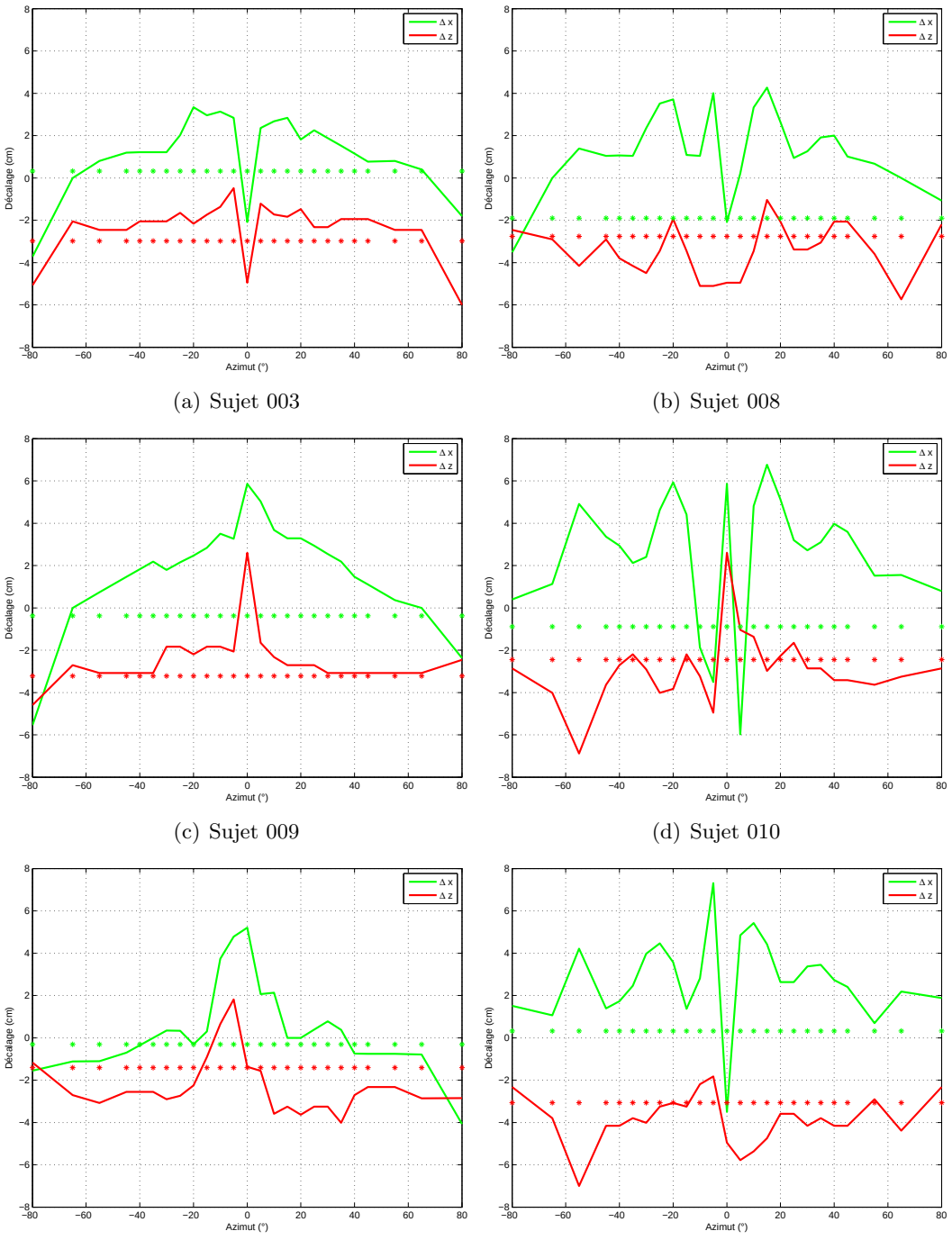
17 sujets qui a été utilisée auparavant pour l'évaluation du modèle SHM (cf. p.184). Les paramètres $\Delta\phi$ et $\Delta\theta$ du modèle sont les angles de décalage des positions des oreilles sur la sphère. Le bon sens physique voudrait qu'ils soient proches des décalages des oreilles observés sur la morphologie des sujets. Les angles peuvent être traduits, au moyen d'un passage des coordonnées sphériques en coordonnées cartésiennes, en termes de translation Δx et Δz selon les axes avant/arrière (axe $o\vec{x}$) et haut/bas (axe $o\vec{z}$). Les paramètres Δx et Δz sont alors directement comparables aux paramètres anthropométriques x_5 et x_4 fournis dans la base du CIPIC. Les figures 3.63, 3.64 et 3.65 reproduisent les paramètres de décalage optimaux du modèle SHM-WOE en fonction de l'azimut, en comparaison des paramètres anthropométriques x_4 et x_5, pour l'ensemble des 17 sujets. On observe une relative concordance pour le décalage haut/bas (c'est à dire entre Δz et x_4), du moins sur la plage $|\phi| \in [20\,\text{deg}, ..., 65\,\text{deg}]$ où Δz est le plus constant. D'un façon générale, Δz tend à osciller autour de la valeur de x_4. En revanche, pour le décalage avant/arrière, on constate, outre une variabilité assez forte du paramètre Δx en fonction de l'azimut, que les valeurs de Δx et x_5 coïncident rarement. On remarque que la valeur de x_5 est très souvent proche de zéro et que sur l'ensemble des sujets le décalage avant/arrière observé sur les données anthropométriques est très faible, voire nul. Ce résultat corrobore ce qui a été observé sur le modèle SHM-WOE (cf. Fig. 3.56e) : l'évolution de l'ITD en fonction de l'élévation suggère un décalage principalement de type haut/bas associé à un décalage avant/arrière très faible ou nul. Le paramètre Δx étant faible[29], il est difficile à estimer car sensible aux erreurs de mesure.

L'erreur de modélisation $ErreurNorm_{SHM-WOE}$ est évaluée pour les 2 jeux de paramètres suivants :

- un jeu de paramètres déduit directement de la morphologie $(R, \Delta x, \Delta z)_{morph}$: $R_{morph} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ (cf. Equ. 3.22
  $\Delta x_{morph} = -x_5$
  $\Delta z_{morph} = -x_4$
  où les paramètres $x_i, i \in [1, 2, 3, 4, 5]$ sont les données anthropométriques fournies dans la base du CIPIC,
- le jeu de paramètres $(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[20^o-65^o]}$.

L'erreur moyenne globale est donnée dans le tableau 3.7, tandis que la Figure 3.66 reproduit l'erreur en fonction de l'azimut. Les performances du modèle SHM sont rappelées à titre de comparaison (cf. Section 3.4.3). L'utilisation directe des paramètres morphologiques $(R, \Delta x, \Delta z)_{morph}$ est décevante : l'erreur moyenne globale vaut 1.02 contre 1.07 pour le modèle SHM. En revanche l'application des paramètres optimaux moyens $(\overline{R}, \overline{\Delta\phi}, \overline{\Delta\theta})_{[20^o-65^o]}$ apporte une amélioration sensible avec une erreur globale réduite à 0.84 et un peu moins d'un tiers de valeurs supérieures à 1. Cependant le succès de la modélisation est dans l'ensemble moins satisfaisant sur ce panel d'individus que dans l'étude précédente (cf. Tab. 3.6), alors que les 2 études obtiennent des résultats comparables avec les paramètres optimaux ajustés séparément pour chaque azimut. Il est difficile de conclure sur le paramétrage du modèle à partir des données anthropométriques. Sur la Figure 3.66, on se rend compte que, bien que le modèle SHM-WOE démontre sa capacité à reproduire finement une ITD individualisée (utilisation des paramètres $(R, \Delta\phi, \Delta\theta)_{opt}$), et ce bien en deça du seuil d'audibilité, ses performances sont très sensibles à l'ajustement de ses paramètres d'entrée $(R, \Delta\phi, \Delta\theta)$. Clairement les informations directement issues de la morphologie $(R, \Delta\phi, \Delta\theta)_{morph}$ ne conviennent pas. Mais ici on est peut-être confronté à la réelle difficulté d'identifier et de mesurer le décalage des oreilles de l'auditeur sur sa morphologie. En effet, le décalage est évalué par rapport à un point de référence

[29] Le décalage avant/arrière apparaît si faible qu'on peut se demander s'il ne serait pas pertinent de le fixer à zéro, ce qui revient à réduire à 2 le nombre de paramètres du modèle : R et $\Delta\theta$. Tout l'effort de modélisation pourrait alors se porter sur l'optimisation du décalage haut/bas. Cette idée a été évaluée, mais n'a pas donné d'amélioration concluante de la modélisation de l'ITD.

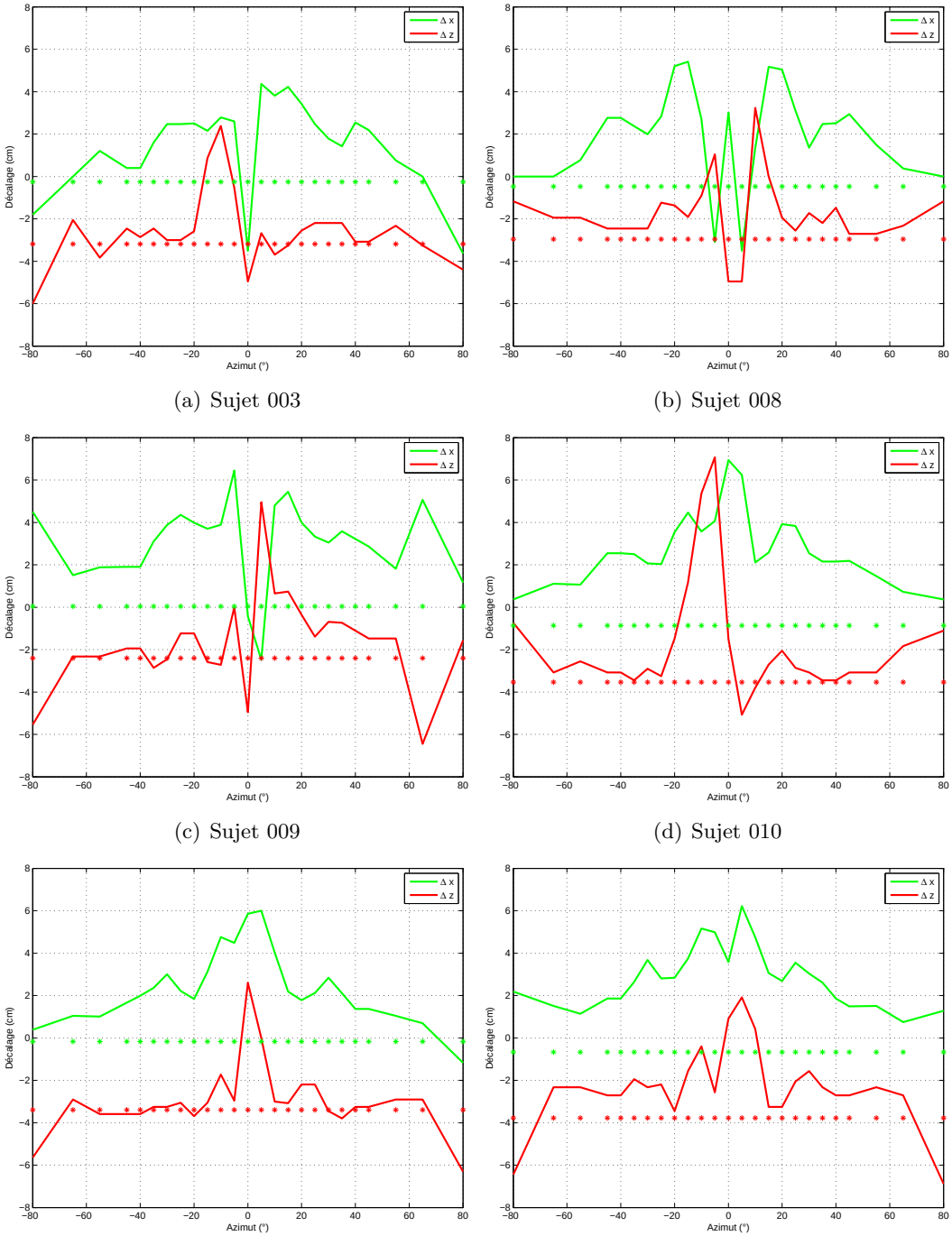


(a) Sujet 003 (b) Sujet 008

(c) Sujet 009 (d) Sujet 010

(e) Sujet 011 (f) Sujet 012

FIG. 3.63 – Comparaison entre les paramètres de décalage optimaux du modèle SHM-WOE (courbes en trait continu) et les paramètres issus de la morphologie de l'individu (*). Base de données du CIPIC.

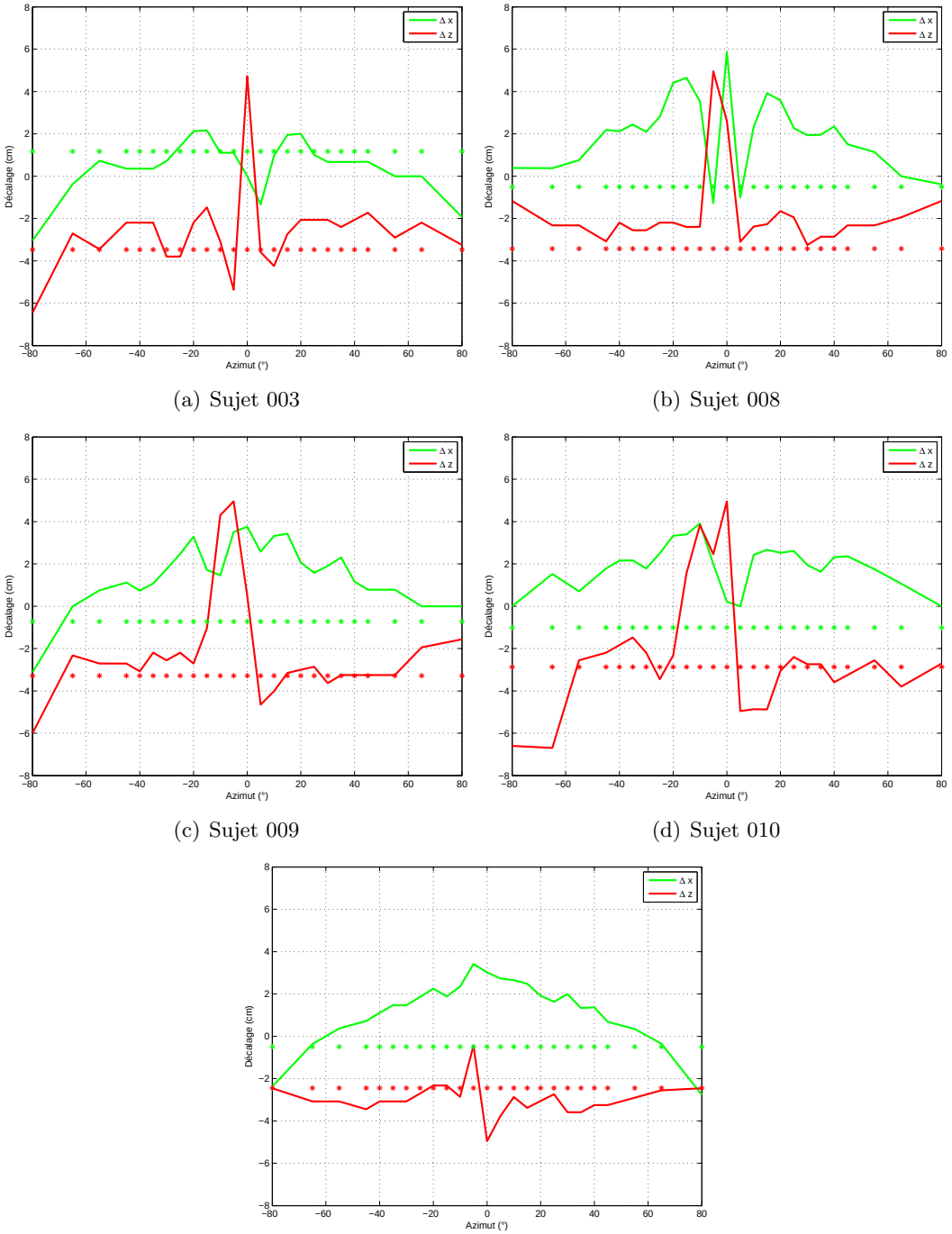


(a) Sujet 003 (b) Sujet 008

(c) Sujet 009 (d) Sujet 010

(e) Sujet 011 (f) Sujet 012

FIG. 3.64 – Suite de la Figure 3.63.



(a) Sujet 003

(b) Sujet 008

(c) Sujet 009

(d) Sujet 010

(e) Sujet 011

FIG. 3.65 – Suite des figures 3.63 et 3.64.

Modèle	$ErreurNorm_{SHM-WOE}$	$P\%_{>1}$
SHM	1.07 [±0.0110]	44.73
$SHM-WOE(R,\Delta\phi,\Delta\theta)_{opt}$	0.34 [±0.0051]	4.69
$SHM-WOE(R,\Delta\phi,\Delta\theta)_{morph}$	1.02 [±0.0109]	41.11
$SHM-WOE(\overline{R},\overline{\Delta\phi},\overline{\Delta\theta})_{[20^o-65^o]}$	0.84 [±0.0092]	32.41

TAB. 3.7 – Performances de modélisation du modèle SHM-WOE : erreur moyenne de modélisation $ErreurNorm_{SHM-WOE}$ (moyenne sur les 17 individus et les 1250 directions) et pourcentage $P\%_{>1}$ de valeurs d'erreurs supérieures à 1 (seuil de discrimination). La configuration du modèle spécifiée par les paramètres $(R,\Delta\phi,\Delta\theta)_{opt}(\phi_i)$ correspond à la mise en œuvre des paramètres optimaux dépendant de l'azimut. Dans les 2 autres configurations, un jeu unique de paramètres est appliqué pour l'ensemble des azimuts.

arbitraire défini comme l'intersection des axes représentant la hauteur et la profondeur de la tête. On peut se demander dans quelle mesure ce point de référence coïncide avec le centre de la sphère par laquelle on souhaite modéliser la tête. Malgré l'échec de la modélisation avec un paramétrage morphologique direct, on constate par ailleurs qu'il existe un jeu de paramètres $(\overline{R},\overline{\Delta\phi},\overline{\Delta\theta})_{[20^o-65^o]}$ avec lequel le modèle SHM-WOE fait bien la preuve de sa capacité à reproduire de façon quasi-transparente une ITD individualisée. Il reste à construire une relation permettant de dériver les paramètres du modèle $(R,\Delta\phi,\Delta\theta)$ à partir des données anthropométriques $x_i, i \in [1,2,3,4,5]$. Il semble cependant que cette relation ne soit pas directe, ni intuitive. Peut-être conviendrait-il même de repenser le protocole de mesure des données morphologiques ou du moins les paramètres descriptifs. Cette étude reste à mener.

3.5 Modélisation des IS

Après la modélisation des indices temporels (ITD), nous allons nous intéresser à celle des indices spectraux (IS), plus exactement à la modélisation du module spectral des HRTF ou encore, de façon équivalente, à la composante à phase minimale des HRIR. Le problème reste inchangé : la méthode la plus directe pour obtenir les HRTF, qu'il s'agisse de leur module ou de leur phase, est la mesure acoustique des fonctions de transfert entre la source sonore et l'entrée des conduits auditifs de l'auditeur. En raison des difficultés et du coût de la mise en œuvre des mesures acoustiques de HRTF (cf. page 133), on cherche à s'en affranchir. Dans l'idéal, on souhaiterait s'en affranchir totalement, mais, plus raisonnablement, on pourrait se satisfaire, du moins dans une première étape ou à un premier niveau, d'une procédure de mesure "allégée" dans laquelle le nombre de directions mesurées est fortement réduit (inférieur à 100 directions mesurées contre 1000 idéalement) sans dégrader la qualité de la spatialisation associée.

Le problème à résoudre est donc le suivant : soit un individu quelconque, on veut construire un modèle de HRTF qui permette, à partir d'un ensemble donné de paramètres décrivant l'**individualité** de cet auditeur en termes de son **encodage binaural** (c'est à dire principalement de ses HRTF), de fournir à cet individu ses HRTF individuelles. La stratégie adoptée ici vise à proposer à l'auditeur une synthèse binaurale dotée de l'encodage spatial adapté à son décodage individuel. L'approche alternative consisterait à miser sur la plasticité du système auditif en aidant l'auditeur à construire un nouveau décodeur correspondant à l'encodeur d'un autre individu ou d'une tête artificielle [Hofman et al., 1998] [Blum, 2003] [Blum et al., 2004] [Savel et al., 2006]. En dépit de tout son intérêt, cette seconde stratégie n'a pas été considérée dans nos travaux.

Au final, la constitution du modèle de HRTF individuelles soulève deux questions fondamen-

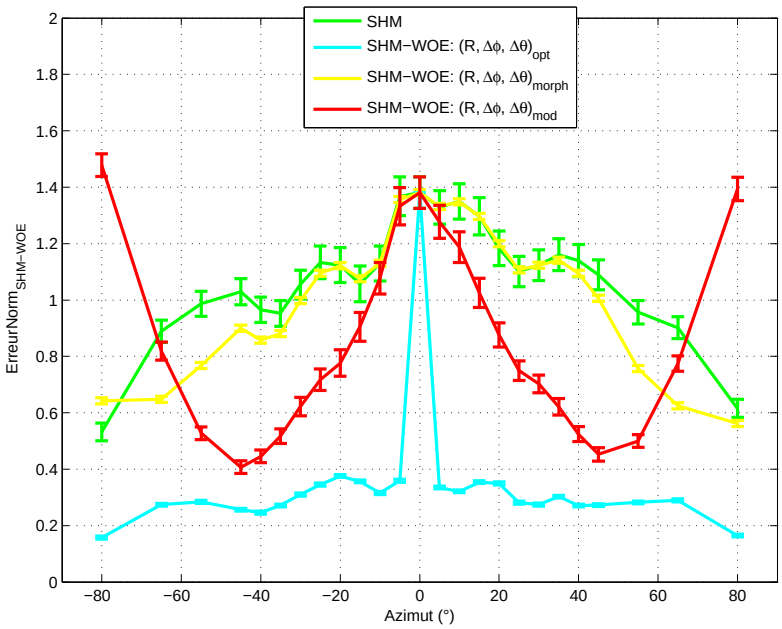


FIG. 3.66 – Erreur moyenne de modélisation des modèles SHM et SHM-WOE en fonction de l'azimut. Pour le modèle SHM-WOE, l'erreur est calculée d'une part avec les paramètres optimisés séparément pour chaque azimut $(R, \Delta\phi, \Delta\theta)_{opt}$ et d'autre part avec 2 jeux de paramètres : $(R, \Delta\phi, \Delta\theta)_{morph}$ (correspondant aux paramètres issus des données de la morphologie) et $(R, \Delta\phi, \Delta\theta)_{mod}$ (correspondant à la moyenne des paramètres optimaux calculée sur la plage d'azimuts [-65°,...,-20°,20°,...,65°]). Moyenne sur les élévations et les individus (17 individus extraits de la base du CIPIC).

tales :

- identifier et sélectionner les **paramètres représentatifs de l'individualité** de l'encodage binaural de l'auditeur, la contrainte étant que ces paramètres doivent être faciles à acquérir dans un contexte grand public,
- identifier et mettre en œuvre une **méthode** permettant d'obtenir les HRTF individuelles à partir des paramètres précédents.

A noter que ces deux questions sont fortement corrélées, ou du moins qu'elles ne sont pas indépendantes : bien souvent le choix de la méthode est le cœur du problème et oriente, voire détermine le choix des paramètres. Il ne faut pas oublier non plus qu'une fois résolus ces deux points, une troisième question sous-jacente émerge : comment valider le modèle et notamment évaluer ses performances d'individualisation ? C'est d'abord à cette question que nous nous intéressons avant de traiter la modélisation des HRTF individuelles où elle prendra alors tout son sens.

3.5.1 Quels outils d'évaluation des modèles ?

Comme pour l'ITD, on ne peut concevoir des modèles de HRTF individuelles sans se poser la question de la validation des modèles, c'est à dire de l'évaluation de leurs performances de modélisation. Cette question peut être abordée :

- d'un point de vue **objectif** : La stratégie la plus naturelle consiste à comparer la HRTF modélisée à la HRTF cible qui est en général la HRTF mesurée. Cette comparaison repose sur une *mesure de distance* ou de *similarité* entre les données.
- d'un point de vue **subjectif** : On cherche en ce cas à comparer la *perception* du sujet. Il reste à définir ce que contient cette perception, car le percept associé à un VAS pris dans sa globalité comporte de nombreuses dimensions qui sont plus ou moins influencées par les filtres binauraux (c'est à dire les HRTF) utilisés pour le générer. D'emblée cette approche soulève davantage de questions qu'elle n'en résout [Martens, 2001]. Mais elle ne peut être éludée : l'objectif de la synthèse binaurale étant de créer un VAS donnant l'illusion à un auditeur de sources sonores spatialisées, il importe de prendre en compte la perception dans le processus de validation.

Mesures de similarité

Pour définir une mesure de similarité des HRTF, la première difficulté concerne la multidimensionalité des HRTF qui dépendent à la fois de la fréquence et de la direction pour un individu donné. Les HRTF (réponses en fréquence) peuvent ainsi être considérées de façon alternative sous la forme de SFRS (fonctions de directivité) (cf. Section 3.1.3). La littérature propose un large choix de mesures de similarité associées aux HRTF[30]. Le plus souvent ces mesures de similarité reposent sur un calcul d'erreur fréquence par fréquence entre les 2 HRTF : la HRTF cible H et la HRTF modélisée $\hat{H}$, se basant sur la différence des modules en valeur linéaire ou logarithmique (auquel cas on se ramène au quotient des modules) [Nicol et al., 2006]. Pour chaque direction, on obtient une erreur correspondant à la moyenne des erreurs sur l'ensemble des bins fréquenciels, une pondération fréquencielle pouvant être appliquée pour prendre en compte la résolution fréquencielle du système auditif (bande critique). Des exemples de ce type de mesure de similarité sont donnés dans le tableau 3.8. Au lieu de considérer la différence des modules spectraux, une mesure originale consiste à calculer l'écart-type [Langendijk & Bronkhorst, 2002] ou la variance [Middlebrooks, 1999b] de cette différence sur l'ensemble des bins fréquenciels [Guillon, 2009] [Hoffmann & Moller, 2008b].

[30] Ces mesure de similarité ne sont pas d'ailleurs utilisées exclusivement pour évaluer les performances de modèles de HRTF, mais aussi pour d'autres opérations relatives aux HRTF telles que la classification [Nicol et al., 2006].

Critère	Définition
MSE (*Mean Square Error*)	$C_{MSE} = \frac{1}{N}\sum_{i=1}^{N}[H(i) - \hat{H}(i)]^2$
CB (*Critical Band*)	$C_{CB} = \frac{1}{N}\sum_{i=1}^{N}\{\alpha(i)[H(i) - \hat{H}(i)]\}^2$ où $\alpha(i)$ désigne la pondération fréquencielle
Fahn [Fahn & Lo, 2003]	$C_F = \frac{\sum_{i=1}^{N}[H(i)-\hat{H}(i)]^2}{\sum_{i=1}^{N}[H(i)]^2}$
Avendano [Avendano et al., 1999]	$C_A = 10log10\left\{\frac{\sum_{i=1}^{N}[H(i)-\hat{H}(i)]^2}{\sum_{i=1}^{N}[H(i)]^2} + 1\right\}$
Durant [Durant & Wakefield, 2002]	$C_D = \frac{1}{N}\sum_{i=1}^{N}\left\{20log10\left[\frac{\hat{H}(i)}{H(i)}\right] - \bar{d}\right\}^2$ où : $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} 20log10\left[\frac{\hat{H}(i)}{H(i)}\right]$

TAB. 3.8 – Exemples de mesures de similarité proposées dans la littérature.

Sur la base de la variance est proposée la mesure de l'ISSD (*Inter-Subject Spectral Difference*) qui a été validée perceptivement par Middlebrooks [Middlebrooks, 1999b], au sens où une réduction de l'ISSD est corrélée à une diminution des confusions avant/arrière et une amélioration de la perception en élévation. Dans tous les cas, on obtient une valeur par direction. Pour disposer d'une mesure **globale**, il convient de calculer la moyenne sur l'ensemble des directions. On peut se poser la question d'appliquer alors une pondération spatiale afin de favoriser les directions où le système auditif offre la résolution maximale (zone frontale) au détriment des directions associées à une faible acuité spatiale (zones latérales, calotte sphérique supérieure). A notre connaissance la pondération spatiale n'a jamais été mise en œuvre. Si l'on considère à présent les SFRS, les mêmes types de mesure de distance (moyenne/ écart-type/ variance de la différence entre 2 SFRS) peuvent être utilisés, cependant une mesure spécifique s'avère plus pertinente : l'**intercorrélation normalisée** [Guillon, 2009] qui évalue la similarité entre 2 fonctions définies sur la sphère. De la même façon que l'intercorrélation entre 2 signaux temporels se calcule en fonction d'un décalage temporel glissant entre les 2 signaux, l'intercorrélation entre 2 SFRS se calcule pour un échantillonnage de l'ensemble des rotations relatives des 2 fonctions sur la sphère. Cette mesure de distance est ainsi capable de détecter la similarité entre 2 SFRS qui ne diffèrent que par une rotation sur la sphère. L'intercorrélation normalisée fournit 2 informations : la valeur du *maximum* d'intercorrélation qui est une mesure quantitative de la distance au mieux entre les 2 SFRS, et la valeur associée de la *rotation* qui est une description qualitative partielle de leur similarité.

Nécessaire ancrage des distances objectives sur la perception

Il reste que ces différentes mesures de similarité ne reflètent que la distance objective entre les HRTF. Cette distance objective n'a de sens qu'à condition qu'elle soit étalonnée sur la base d'une distance perceptive. Une valeur donnée de distance ne signifie rien en soi, si l'on ne sait pas dans quelle mesure cette distance correspond à une différence audible ou non. Si la différence est détectée par le système auditif, il est aussi important de savoir si elle est perçue comme faible ou forte. Cette interprétation nécessite de connaître la JND associée. En ramenant la valeur de la distance en JND, on sait si elle dépasse le seuil de discrimination et si oui dans quelle proportion. Cette démarche a été mise en oeuvre pour la validation des modèles d'ITD dans ce qui précède. Dans le cas des modules spectraux des HRTF, la mesure de JND est une question délicate. Sur la plan perceptif, la "différence" entre 2 HRTF est un percept qui implique de multiples attributs perceptifs. Pour juger de la perception de HRTF, le mieux est de donner à entendre des sources virtuelles synthétisées avec ces HRTF. Toute autre approche comporte le risque de sortir la HRTF de son contexte d'utilisation et de fausser le jugement. Or, dans la perception d'une source virtuelle

en synthèse binaurale, le principal attribut perceptif contrôlé par la HRTF est certes la localisation de la source, mais d'autres attributs sont aussi affectés : son timbre notamment, mais aussi son externalisation, sa largeur apparente, son "réalisme" etc... Dans notre problème, faut-il demander au sujet de se focaliser sur la localisation ? Mais les autres attributs doivent aussi être pris en compte afin de refléter la qualité globale du VAS. Un problème particulièrement critique vient du fait que les différences spectrales tendent à dominer le jugement et par suite sont susceptibles de masquer les autres différences. D'ailleurs fondamentalement les HRTF ne sont autres que des altérations du spectre qui, selon le "contexte", sont interprétées par le système auditif comme une localisation dans l'espace, un élargissement de source ou une simple altération du timbre de la source [Langendijk & Bronkhorst, 2000]. Ce phénomène est frappant dans des expériences psychoacoustiques où le sujet a l'occasion d'écouter des HRTF dont la proximité avec les HRTF individuelles est modifiée selon une échelle progressive [Guillon, 2009]. On se rend compte alors qu'il existe une sorte de continuum perceptif selon lequel une altération spectrale qui n'était pas véritablement perceptible car interprétée en termes de localisation, devient tout à coup audible lorsque les filtres binauraux s'écartent trop de la référence des HRTF individuelles. Quoi qu'il en soit, dans notre problème, il faut considérer que si les différences de timbre sont omniprésentes, car intrinsèquement liées aux mécanismes d'encodage de la synthèse binaurale, elles ne sont pas notre centre d'intérêt. Il convient donc de s'en affranchir, notamment par le choix des stimuli et du protocole expérimental.

L'obstacle majeur à la mesure d'une JND des HRTF est le fait que la HRTF présente autant de degrés de liberté qu'elle comporte de bins fréquenciels. Même si le spectre est considéré avec une résolution réduite, calquée sur une analyse par bande critique, évaluer les JND associées à l'ensemble des différences spectrales qu'on pourraient générer en jouant sur l'ensemble de ces degrés de liberté représente une tâche conséquente qui pour l'instant n'a pas été entreprise. On peut se demander si les différences ainsi obtenues ont toutes un sens physique, c'est à dire qu'elles correspondent à des différences auxquelles un auditeur pourrait être confronté en situation d'écoute naturelle. Le problème physique contraint en effet le domaine des possibilités. Une solution serait de collecter un ensemble des différences "écologiquement" valides à partir des HRTF mesurées sur des individus. On peut imaginer un travail d'analyse et de classification permettant d'extraire les différences les plus représentatives de celles rencontrées sur une large base de données. Néanmoins il faut garder à l'esprit qu'une même différence quantitative (en termes des mesures de similarité précédemment décrites) peut correspondre à plusieurs paires distinctes de HRTF, dans la mesure où chaque paire constitue une distribution particulière des différences spectrales selon l'axe fréquenciel. La question de la discrimination du système auditif en termes de HRTF a été peu traitée jusqu'à présent. Les études récentes de Hoffmann et Møller [Hoffmann & Moller, 2008b] constituent un premier pas. Le seuil de discrimination des HRTF y est évalué pour des paires de HRTF correspondant à un écart angulaire croissant, d'une part en azimut et d'autre part en élévation. Les différences qu'on donne à juger sont donc ancrées sur les variations spatiales des HRTF, ce qui est un choix particulier pour constituer un panel de différences mais constitue une solution aux questions soulevées précédemment. Ce choix est assez naturel et a déjà été adopté par d'autres études [Nicol et al., 2006]. La JND des HRTF est exprimée en termes d'écart angulaire, le principal objectif de l'étude étant de déterminer la résolution minimale requise pour la mesure des HRTF. La portée des résultats de cette étude est cependant amoindrie par le fait que les HRTF utilisées sont celles d'une tête artificielle. Les sujets de l'expérience sont donc placés dans une situation où l'analyse des différences spectrales entre les HRTF est fortement pollué et biaisé par le caractère non individuel (et par la même non écologique) de l'encodage spatial et notamment des IS. Il faut en effet se demander si la discrimination des IS n'est pas modifiée selon que les IS sont perçus comme des attributs de localisation ou des attributs de timbre, comme on vient de le discuter.

Pour étalonner les mesures de similarité sur la base de la perception, la JND est une première

piste, la plus évidente, mais ce n'est pas la seule, ce qui est heureux compte tenu de la difficulté de mesurer la JND des HRTF. Il est aussi possible de se rattacher à d'autres indicateurs perceptifs, comme l'illustre par exemple Middlebrooks pour l'ISSD [Middlebrooks, 1999b]. Il établit le lien entre l'évolution de l'ISSD et celle du taux de confusion avant/arrière et de l'erreur de localisation en élévation. Les méthodes d'évaluation subjective qui vont être décrites fournissent une panoplie de mesures indirectes de la distance entre les HRTF cibles et les HRTF modélisées (mesures indirectes telles que l'erreur de localisation et les taux de confusion) dont les résultats peuvent être corrélés aux distances objectives associées, ce qui constitue un ancrage perceptif à explorer.

Evaluation subjective

Les critères objectifs n'offrant pas des outils pleinement satisfaisants pour l'évaluation des modèles de HRTF individuelles, tournons-nous vers les méthodes d'évaluation subjective. Dans la terminologie des méthodes de tests subjectifs, l'évaluation des HRTF fait appel exclusivement aux méthodes de type *indirect*. Il est en effet impossible de juger de la "qualité" de HRTF en soi, mais indirectement en observant l'effet sur le comportement du sujet.

La méthode de référence est le **test de localisation** [Martens, 2001] [Pernaux, 2003] qui consiste à demander au sujet d'identifier la position des sources sonores virtuelles. Les performances de la spatialisation sont évaluées en termes d'*erreur de localisation* et de *taux de confusion avant/arrière et haut/bas*. Bien que l'usage de cette méthode soit largement répandu et constitue une sorte de standard (y compris en dehors du contexte de la synthèse binaurale), elle est remise en cause par certains auteurs [Martens, 2001] et fait l'objet de nombreuses critiques. Le jugement de localisation est d'abord totalement tributaire de la capacité et de la précision du sujet à localiser des sources sonores, qu'elles soient réelles ou virtuelles. Or tous les individus ne sont pas égaux dans cette tâche, d'une part par leur morphologie : les individus qui grâce à leur morphologie disposent d'indices de localisation particulièrement sailants et lisibles sont dans une certaine mesure avantagés, et d'autre part par leur apprentissage de la localisation auditive : certains individus dans leur "expérience de vie" vont davantage exercer et développer leur aptitude à localiser des sons. De plus, il faut avoir conscience que juger la localisation d'une source virtuelle est une tâche relativement différente de la localisation d'une source réelle. Dans le second cas, l'auditeur dispose du référentiel que constitue l'espace physique où la source sonore possède une réalité multimodale. Pour une source virtuelle, en revanche, elle n'existe que dans l'espace auditif où l'auditeur ne dispose d'aucun référentiel physique. Intervient aussi dans le test de localisation le choix de la méthode de report du jugement de localisation [Pernaux, 2003] : méthodes de pointage, mouvements oculaires, verbalisation des angles d'azimut et d'élévation etc... Dans ces conditions, on peut craindre que le biais introduit à la fois par l'incertitude intrinsèque de localisation du sujet et par la méthode de report du jugement vienne masquer l'effet de la modélisation des HRTF sur la qualité de spatialisation ressentie par le sujet dans le VAS.

Au delà de ce biais d'estimation, il faut se demander si, du point de vue de la qualité de la spatialisation, la mesure de l'erreur de localisation est un indicateur pertinent. En d'autres termes est-il si grave que la source virtuelle soit localisée en $(\phi + \Delta\phi, \theta + \Delta\theta)$ plutôt qu'en (ϕ, θ), dès lors que les écarts $\Delta\phi$ et $\Delta\theta$ restent "raisonnables" ? Il semble plus pertinent d'évaluer si la scène sonore virtuelle est cohérente et naturelle ou si globalement le sujet s'y sent "à l'aise". En ce sens, les taux de confusion avant/arrière et haut/bas sont des mesures plus dignes d'intérêt [Martens, 2001]. Comme le remarque Martens, il convient de ne pas confondre localisation et spatialisation [Martens, 2001]. Plus généralement le test de localisation fait l'hypothèse que la cible à atteindre est la source sonore réelle, mais la copie de la réalité à l'identique est-elle le seul objectif possible ? Créer un VAS réaliste et convainquant, même si la scène sonore virtuelle n'est pas parfaitement identique à une scène réelle

ne suffit-il pas ? La réponse dépend évidemment du contexte applicatif. Dans le cas de simulateurs sonores (pilotes d'avion, conducteurs automobiles, apprentissage d'une machine etc...), il est clair qu'il faut se rapprocher au plus près de réalité. Pour les jeux, on peut commencer à s'en écarter, tandis que pour la création musicale, la spatialisation n'est qu'un outil de composition de l'espace sonore sans référence explicite à une scène réelle...

Afin de dépasser les limites du test de localisation, une étude réalisée à Orange Labs dans le but d'évaluer l'apport de la synthèse binaurale dynamique [Faure, 2005] a proposé deux nouvelles méthodologies d'évaluation subjective :

- une méthodologie **indirecte** où l'on mesure la qualité de la spatialisation non plus sur la base de la capacité du sujet à localiser les sources, mais à décrire la globalité de la scène sonore et à identifier les sources sonores qui la constituent,
- une méthodologie **directe** passant par la définition d'un ensemble d'attributs spatiaux (tels que la précision spatiale, l'enveloppement, la profondeur [Berg & Rumsey, 1999] [Rumsey, 2001] [Zacharov & Koivuniemi, 2001] [Guastavino & Katz, 2004] [Bech & Zacharov, 2006]) décrivant les différentes dimensions de la spatialisation perçue : c'est à travers la grille de ces critères qu'on demande au sujet de juger le rendu spatial.

3.5.2 Etat de l'art des modèles de HRTF individuelles

Dans l'ensemble des modèles de HRTF individuelles proposés dans la littérature [Guillon, 2007], on peut distinguer 4 principales familles de *méthodes* :

- Résolution du problème de diffraction d'une onde acoustique par le corps de l'auditeur (modèle de type 1),
- Reconstruction de HRTF sur une base de vecteurs élémentaires (modèle de type 2),
- Sélection dans une base de données de HRTF (modèle de type 3),
- Transformation de HRTF non individuelles (modèle de type 4).

Ces familles sont illustrées dans les paragraphes qui suivent.

Résolution du problème de diffraction d'une onde acoustique par le corps de l'auditeur

Les HRTF sont obtenues comme les solutions du problème physique de propagation d'une onde acoustique en présence de l'obstacle que constitue l'auditeur. La résolution peut être *analytique* si la géométrie du problème est simple (par exemple si le corps de l'auditeur est modélisé par une sphère [Algazi et al., 2001a]) ou *numérique* dans le cas de géométries complexes où l'on a recours à des méthodes de type BEM (Boundary Element Method) ou FEM (Finite Element Method) [Katz, 1998] [Kahana, 2000]. Le modèle prend en entrée soit une approximation de la morphologie de l'auditeur sur la base de primitives simples de type sphère ou ellipse (par exemple le modèle *snowman* proposé par [Algazi et al., 2002b]), soit un maillage 3D issu d'un scan ou éventuellement d'un jeu de photographies comme le suggèrent de récents travaux [Dellepiane et al., 2008]. Un des intérêts de cette méthode est la possibilité qu'elle offre de modifier la morphologie et d'en évaluer l'impact, par exemple afin de caractériser l'effet d'un élément donné de la morphologie indépendamment des autres (étude des résonances du pavillon [Kahana, 2000]) ou l'effet des paramètres (taille, positionnement, forme...) de cet élément [Iwaya & Suzuki, 2008] [Plaskota & Dobrucki, 2008] [Fels & Vorländer, 2009]. A ce titre, ce type de modélisation est souvent utilisé à des fins d'investigation pour analyser et comprendre comment les IS sont générés en relation avec la morphologie de l'auditeur. En est d'ailleurs issu un *modèle structurel* de HRTF qui consiste à décomposer la fonction de transfert en isolant les contributions de chaque élément morphologique (tête, torse, pavillon) [Brown & Duda, 1998] [Algazi et al., 2001c] [Algazi et al., 2002b].

Reconstruction de HRTF sur une base de vecteurs élémentaires

Il s'agit d'un *modèle par synthèse* consistant à obtenir les HRTF sous la forme d'une combinaison linéaire ou non de fonctions génératrices. La reconstruction peut s'effectuer soit dans le domaine *fréquenciel*, c'est à dire que les fonctions génératrices sont véritablement des HRTF, soit dans le domaine *spatial*, auquel cas les fonctions génératrices sont des fonctions de directivité (selon la terminologie présentée en page 132). Les fonctions génératrices sont le plus souvent communes à tous les individus : dans ce cas, l'individualité d'un nouvel auditeur est donc prise en compte uniquement dans les coefficients de la combinaison. Cependant l'utilisation de fonctions génératrices individuelles n'est pas à exclure [Larcher, 2001].

L'exemple le plus classique se fonde sur une Analyse en Composantes Principales (ACP) ou Indépendantes (ACI) [Larcher, 2001] d'une base de données de HRTF (incluant dans la mesure du possible le plus grand nombre d'individus). L'ACP (ou ACI) vise à extraire l'information responsable de la variance observée dans les données des HRTF (variations à la fois fréquencielles, spatiales et individuelles) : elle fournit au final un ensemble de vecteurs propres qui permettent de reconstruire les données. Une HRTF peut ainsi être exprimée sous la forme d'une somme pondérée (par des coefficients qui définissent les *composantes principales* de l'analyse) de ces vecteurs propres. Le modèle de HRTF individuelles associé consiste à synthétiser les HRTF d'un nouvel auditeur sur la base de vecteurs propres précédemment obtenue, à condition d'identifier les poids de la décomposition adaptés au nouvel individu. On est typiquement dans le cas de fonctions génératrices non individuelles. L'individualisation, c'est à dire en l'occurrence le calcul des coefficients de décomposition associé au nouvel individu, est réalisée à partir d'un ensemble de paramètres anthropométriques décrivant sa morphologie (tête et pavillon principalement) [Rodriguez Soria & Ramirez, 2004] [Rodriguez & Ramirez, 2005a] [Rodriguez & Ramirez, 2005c] [Rodriguez & Ramirez, 2005b] [Inoue et al., 2005] [Jin et al., 2000]. Les coefficients sont en fait prédits à partir des paramètres anthropométriques par un modèle construit par régression linéaire sur une base de données. Cette méthode repose sur l'information acquise par apprentissage statistique sur les bases de données de HRTF initiales. Par suite, le succès de l'individualisation est tributaire de la richesse et de la représentativité de ces bases de données, à la fois en termes d'information fréquencielle, spatiale (finesse de l'échantillonnage spatial de la sphère 3D de mesure) et individuelle (nombre et "variété" des individus). Une étude récente propose d'ajuster les poids des vecteurs propres par un test d'écoute [Hwang et al., 2008].

A la catégorie des modèles de reconstruction appartient aussi un modèle original proposé dans [Hofman & Van Opstal, 2002] : les HRTF sont reconstruites par combinaison linéaire de motifs spectraux élémentaires. L'originalité réside dans la façon dont sont obtenus les coefficients de la combinaison : par une procédure psycho-acoustique consistant en un test de localisation où l'auditeur doit localiser des stimuli correspondant aux motifs spectraux élémentaires et émis par un haut-parleur situé devant lui. Ces motifs spectraux imitent les variations observées sur les HRTF. Pour cette raison ils engendrent des phénomènes d'illusion auditive, c'est à dire qu'ils sont localisés à des positions potentiellement différentes de celle de la source réelle (haut-parleur). Le relevé des localisations perçues associées aux illusions auditives, une fois mises en correspondance avec les spécificités des motifs spectraux qui les engendrent, permet d'exprimer la probabilité que tel motif spectral induise telle localisation. Cette probabilité définit les poids à appliquer pour reconstruire la HRTF dans cette direction par combinaison linéaire des motifs spectraux.

Sélection dans une base de données de HRTF

Cette méthode est sans doute la plus simple : l'auditeur choisit dans une base de données les HRTF (correspondant en général aux HRTF mesurées sur un seul individu) qui lui conviennent

le mieux. La sélection peut se faire sur la base d'un test d'écoute où l'auditeur juge la qualité de spatialisation (notamment l'externalisation, la frontalisation, les confusions avant/arrière...) [Seeber & Fastl, 2003] [Iwaya, 2006]. Un autre critère est la similarité des morphologies : par exemple, pour un nouvel individu, on choisit l'individu de la base de données dont les paramètres anthropométriques relatifs au pavillon sont les plus proches des siens [Zotkin et al., 2002] [Zotkin et al., 2003]. Le choix peut être fait indépendamment pour chaque oreille. Dans ces études, la sélection est réalisée sur les données brutes contenues dans la base, sans qu'aucun tri ne soit effectué au préalable. Des auteurs proposent d'effectuer une classification des HRTF afin de regrouper les données les plus similaires dans des classes et de ne retenir qu'un représentant pour chaque classe. On ne donne ainsi à écouter au nouvel auditeur que les représentants [Shimada et al., 1994].

Transformation de HRTF non individuelles

Comme dans la méthode précédente, on part de HRTF non individuelles sélectionnées dans une base de données, mais ici on ne se contente pas de les sélectionner : les HRTF sont ensuite modifiées par une transformation destinée à les adapter au nouvel individu. Le modèle le plus convaincant de cette catégorie est sans doute celui de *scaling fréquenciel* proposé par Middlebrooks [Middlebrooks, 1999b] [Middlebrooks, 1999a] [Middlebrooks et al., 2000]. Il est basé sur l'observation que, malgré leur spécificité individuelle, les HRTF de deux individus présentent certaines similitudes. On se rend compte notamment que la principale différence entre les HRTF de deux individus réside dans la fréquence des pics et des creux qui glisse sur l'axe fréquenciel d'un individu à l'autre. Ce décalage fréquenciel serait lié à la taille du pavillon. D'où l'idée de Middlebrooks : il suffirait d'appliquer un facteur de dilatation ou de contraction fréquencielle (*scaling*) à des HRTF non individuelles pour les adapter à un nouvel auditeur. Le facteur de dilatation ou contraction est déterminé par le rapport des dimensions caractéristiques des pavillons des deux individus. Middlebrooks montre qu'il est aussi possible d'obtenir ce facteur par une procédure psycho-physique consistant en un test d'écoute où le sujet ajuste le facteur de façon à obtenir la spatialisation la plus fidèle [Middlebrooks, 1999a] [Middlebrooks et al., 2000]. Le jugement porte sur un nombre réduit de directions dans le plan médian. Au final, il propose de déterminer le facteur en deux étapes : d'abord un ajustement grossier sur la base des paramètres anthropométriques, suivi d'un affinement par écoute.

Récemment une étude sur les HRTF de gerbilles de Mongolie [Maki & Furukawa, 2005] suggère d'étendre le modèle de Middlebrooks. L'observation des HRTF dans leur structure spatiale sur la sphère (fonctions de directivités ou SFRS) indique qu'en plus de décalages fréquenciels, les différences individuelles se traduisent par des rotations spatiales. Les différences d'orientation des pavillons d'oreille seraient à l'origine de ces rotations. Les auteurs montrent que la combinaison d'un décalage fréquenciel et d'une rotation spatiale améliore l'individualisation par rapport à un décalage seul ou une rotation seule.

Un autre solution pour ajuster des HRTF non individuelles consiste en une égalisation du spectre par bandes de fréquences (une sorte de *tuning* de HRTF) [Tan & Gan, 1998] [Runkle et al., 2000]. L'égalisation est contrôlée par la perception du sujet qui ajuste les paramètres d'égalisation pour obtenir la spatialisation souhaitée. Une autre méthode originale d'adaptation de HRTF a été proposée par Martens [Martens, 2002]. Au lieu de modifier les HRTF en elles-même c'est leur cartographie spatiale qui est adaptée à la perception du nouvel individu. L'adaptation utilise une procédure psycho-acoustique focalisée sur la localisation dans un plan d'azimut constant (coordonnées polaires interaurales), pour lequel l'ITD est rendue artficiellement constante, afin de reporter toute l'attention de l'auditeur sur les IS. La tâche du sujet consiste à identifier quelle HRTF lui permet de localiser la source virtuelle à une position prédéfinie de référence. Cette identification est réalisée

pour 6 positions de référence comprenant 2 positions dans le plan horizontal (devant et derrière) et 4 élévations ± 45° par rapport au plan horizontal (également devant et derrière), ce qui donne le nom de "*bisection scaling*" à la méthode. L'identification des HRTF pour les positions intermédiaires entre les positions de référence est obtenue par interpolation. Il en résulte une réaffectation spatiale des HRTF (une sorte d'anamorphose de la sphère 3D) adaptée à l'individu.

Paramètres d'individualité

Dans les différents modèles de HRTF individuelles qui viennent d'être décrits, on peut dégager deux principales catégories de paramètres utilisés pour décrire l'individualité de l'auditeur en vue d'obtenir ses HRTF individuelles :

- description **optique** : paramètres anthropométriques (mesurés sur l'individu directement ou sur des photographies), maillage 3D de la morphologie (à partir d'un scan laser, d'une Image à Résonance Magnétique, ou de photographies),
- description **psycho-acoustique** : les paramètres d'entrée sont en quelque sorte la perception de l'auditeur qui vient ajuster le modèle en fonction de ce qu'il perçoit à travers une procédure psycho-acoustique (test de localisation simple ou test d'écoute destiné à juger de la qualité plus ou moins globale de la spatialisation).

A présent que l'état de l'art est posé, nous allons présenter nos contributions sur les modèles de HRTF individuelles.

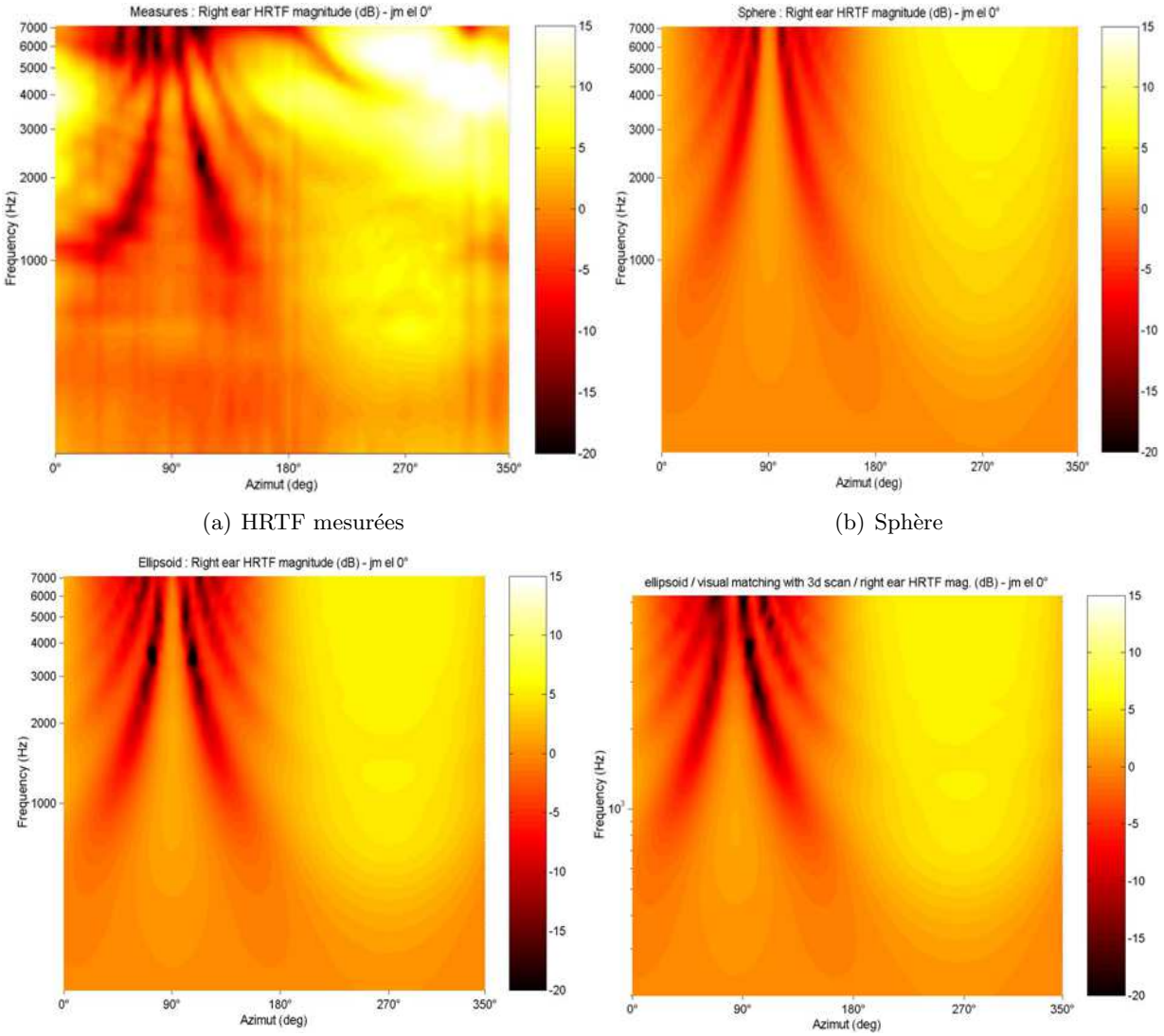
3.5.3 Modèles morphologiques simplifiés pour calcul BEM de HRTF individuelles

Cette étude s'inscrit dans le cadre de la première catégorie de modèles (*Modèle de type 1*). Les modèles BEM ont prouvé leur efficacité pour calculer des HRTF individuelles [Katz, 1998] [Kahana, 2000], dès lors qu'on dispose d'un maillage 3D de la morphologie de l'auditeur. Dans le cadre des travaux de thèse de J.-M. Pernaux, nous avons proposé des modèles morphologiques simplifiés mais individualisés pour des calculs BEM[31] [Pernaux, 2003]. Ces modèles se composent de primitives géométriques élémentaires telles que la sphère, l'ellipsoïde, ou le cylindre. Le premier intérêt de ces modèles est un nombre réduit de paramètres (rayon de la sphère, dimensions et orientation de l'ellipsoïde, ...) à contrôler pour chaque individu. Le second avantage est qu'ils permettent de s'affranchir de l'acquisition d'un maillage 3D de la morphologie de l'auditeur, nécessitant à la fois un matériel spécifique et coûteux de type scan laser, et un protocole de mesure lourd et contraignant. Au contraire les modèles géométriques simplifiés peuvent être ajustés à l'individu à partir d'un simple jeu de photographies. La mise en œuvre de ces modèles morphologiques simplifiés repose sur l'hypothèse fondamentale selon laquelle la ressemblance (voire l'identité) morphologique est le garant d'une synthèse binaurale convaincante.

Modélisation de la tête par un ellipsoïde

Le modèle *snowman* [Algazi et al., 2002a] [Algazi et al., 2002b] a montré que la tête de l'auditeur peut être modélisée avec succès par une sphère. Cependant la forme de la tête évoque plus souvent un ellipsoïde. L'intérêt de l'ellipsoïde pour modéliser la tête a donc été évalué en comparaison de la sphère. La Figure 3.67 illustre les HRTF obtenues dans le plan horizontal pour un modèle sphérique et deux modèles ellipsoïdaux : un ellipsoïde vertical et un ellipsoïde dont à la fois l'orientation et les dimensions des axes sont ajustées pour une correspondance optimale avec

[31] Tous les calculs BEM présentés dans cette section ont été réalisés avec le logiciel *VNoise*™ [VNoise, STS]. Les détails sur la mise en œuvre de ce logiciel sont disponibles dans [Pernaux, 2003] [Busson, 2006].



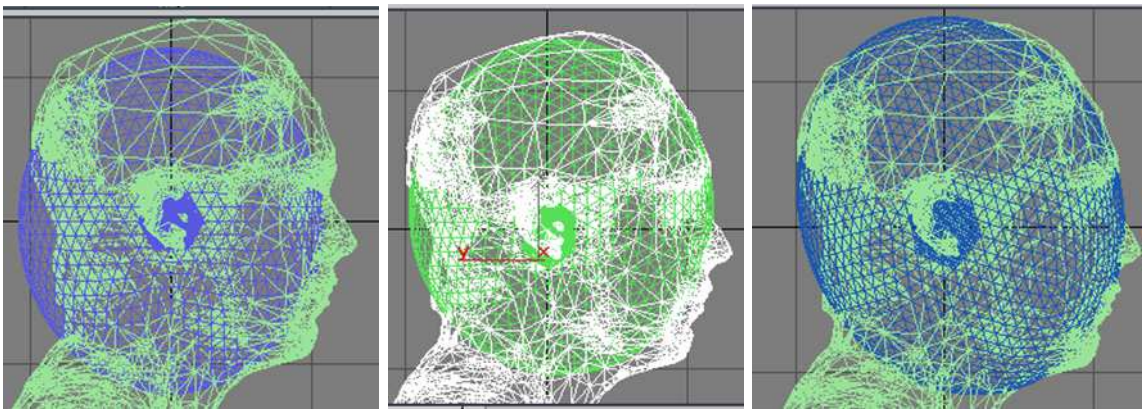
(a) HRTF mesurées

(b) Sphère

(c) Ellispoide vertical

(d) Ellipsoïde individualisé par correspondance visuelle

FIG. 3.67 – HRTF obtenues par calcul BEM pour différents modèles de morphologie (tête seule) : sphère (oreilles diamétralement opposées, rayon individualisé selon l'équation 3.22), ellipsoïde vertical (dont les 3 axes mesurent respectivement $2x_1$, $2x_2$, $2x_3$, cf. Page 174), ellipsoïde individualisé pour une correspondance visuelle avec la morphologie (cf. Fig. 3.68). Module du spectre des HRTF. HRTF obtenues dans le plan horizontal. D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées. Sujet JMP de la base *Jean-Marie Pernaux*.



(a) Sphère (b) Ellipsoïde vertical (c) Ellipsoïde individualisé d'après une correspondance visuelle avec la morphologie

FIG. 3.68 – Correspondance visuelle entre la morphologie d'un sujet et les modèles géométriques. D'après [Pernaux, 2003]. Sujet JMP de la base *Jean-Marie Pernaux*.

la morphologie du sujet. Par rapport à la sphère, l'ellipsoïde apporte des anti-résonances plus marquées et plus proches des HRTF mesurées (cf. Fig. 3.67.a). Quant à l'ellipsoïde individualisé par correspondance visuelle, elle introduit une dissymétrie des anti-résonances entre l'avant et l'arrière, cette dissymétrie étant effectivement présente sur les mesures. Ces résultats se confirment pour d'autres élévations (cf. Fig. 3.69 & 3.70).

Modélisation du torse par un ellipsoïde

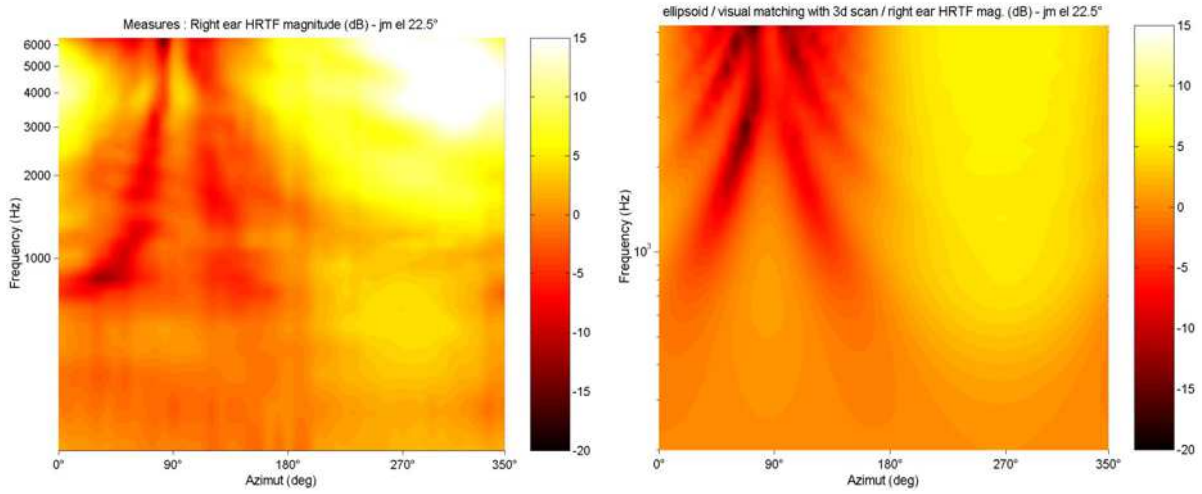
La modélisation de la tête par un ellipsoïde offre une première amélioration par rapport à la sphère, mais il reste encore des caractéristiques spectrales présentes sur les HRTF mesurées qui n'apparaissent pas sur le modèle. D'une part sur la Figure de diffraction du côté controlatéral ($\phi \in$ [0 - 180°]) dans les basses fréquences, la courbure de l'anti-résonance située dans la zone frontale ($\phi \in$ [0 - 90°]) n'est pas reproduite, de même que les oscillations basses fréquences sur la plage [0 - 1 kHz] (cf. Fig. 3.67, 3.69 & 3.70). Du côté ipsilatéral ($\phi \in$ [180 - 360°]), des oscillations hautes fréquences ([1 - 2 kHz] et [4 - 6 kHz]) sont aussi absentes.

Or, il a été montré que les réflexions sur le torse introduisent des IS basses fréquences [Algazi et al., 2001a]. Pour cette raison, nous avons testé l'ajout d'un torse modélisé par un ellipsoïde. Le modèle global se compose d'une tête sphérique et d'un torse ellipsoïdal raccordé à la tête par un cou modélisé par un cylindre (cf. Fig. 3.71). Les figures 3.72, 3.73 & 3.74 illustrent les HRTF obtenues sur la base de ce modèle pour 3 plans d'élévation constante, en comparaison des HRTF mesurées et du modèle de tête sphérique seul. L'ajout du torse modélisé par une ellipse introduit la courbure de l'anti-résonance controlatérale située dans la zone frontale, ainsi que les oscillations ipsilatérales dans les hautes fréquences, ce qui améliore notablement la ressemblance avec les HRTF mesurées.

Proposition d'un modèle morphologique complet : tête + torse

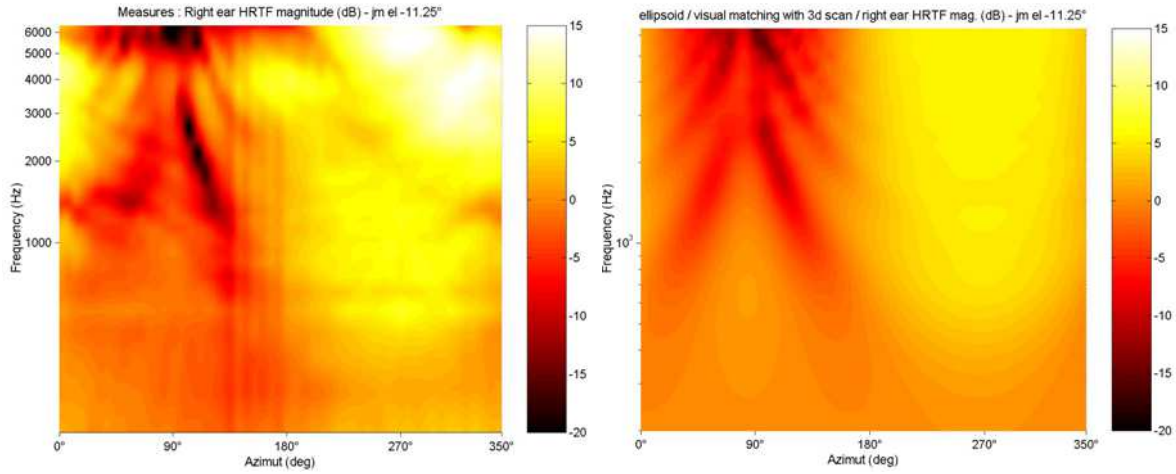
Les résultats qui précèdent conduisent à un modèle morphologique complet constitué de trois primitives géométriques simples (cf. Fig. 3.75) [Pernaux, 2003] :

– un ellipsoïde (tête),



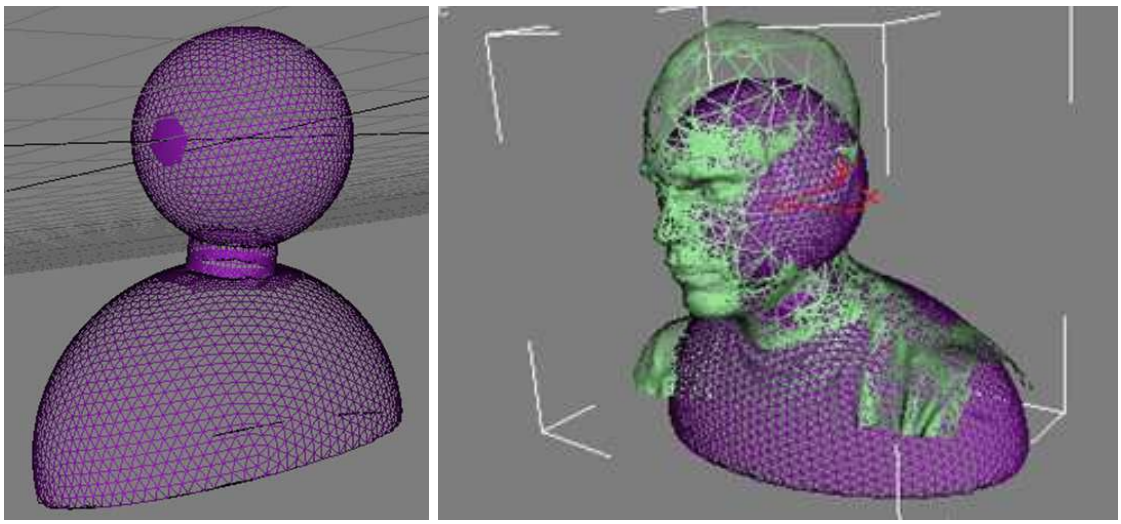
(a) HRTF mesurées (b) Ellipsoïde individualisé par correspondance visuelle

FIG. 3.69 – HRTF obtenues par calcul BEM pour l'ellipsoïde individualisé pour une correspondance visuelle avec la morphologie. Module du spectre des HRTF. HRTF obtenues pour le plan d'élévation 22.5° (coordonnées polaires verticales). D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées. Sujet JMP de la base *Jean-Marie Pernaux*.



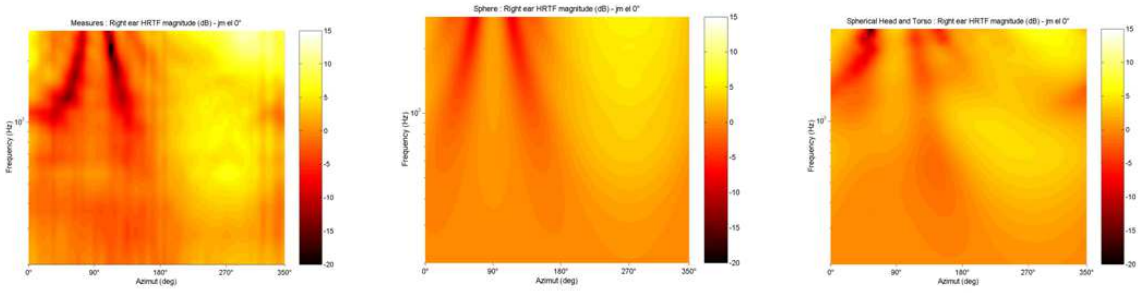
(a) HRTF mesurées (b) Ellipsoïde individualisé par correspondance visuelle

FIG. 3.70 – HRTF obtenues par calcul BEM pour l'ellipsoïde individualisé pour une correspondance visuelle avec la morphologie. Module du spectre des HRTF. HRTF obtenues pour le plan d'élévation -11.5° (coordonnées polaires verticales). D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées. Sujet JMP de la base de données *Jean-Marie Pernaux*.



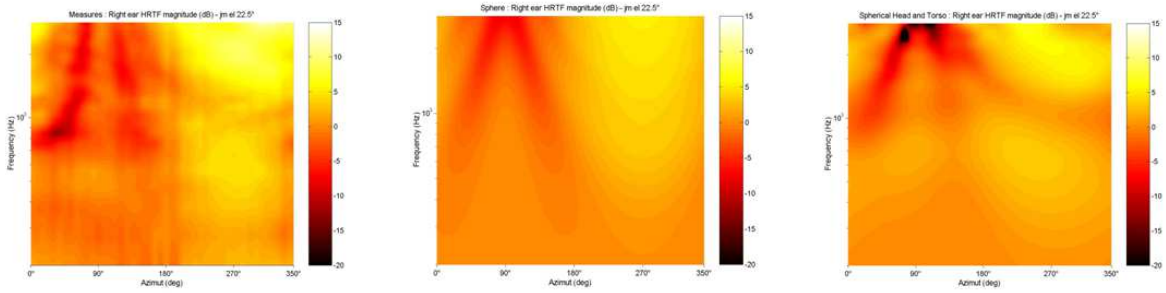
(a) Maillage 3D (b) Correspondance visuelle entre la morphologie et le modèle

FIG. 3.71 – Modèle combinant une tête sphérique et un torse ellipsoïdal raccordé à la tête par un cou modélisé par un cylindre. Sujet JMP de la base *Jean-Marie Pernaux*. D'après [Pernaux, 2003].



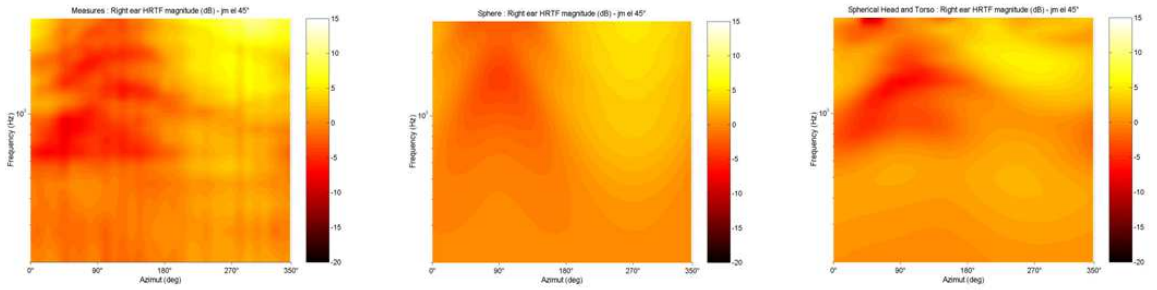
(a) HRTF mesurées (b) Sphère (c) Sphère + Ellipsoïde

FIG. 3.72 – HRTF obtenues par calcul BEM avec le modèle combinant une tête sphérique et un torse ellipsoïdal raccordé à la tête par un cou modélisé par un cylindre (cf. Fig. 3.71). Module du spectre des HRTF. HRTF obtenues pour le plan horizontal. D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées et le modèle de tête sphérique seul. Sujet JMP de la base *Jean-Marie Pernaux*.



(a) HRTF mesurées (b) Sphère (c) Sphère + Ellipsoïde

FIG. 3.73 – HRTF obtenues par calcul BEM avec le modèle combinant une tête sphérique et un torse ellipsoïdal raccordé à la tête par un cou modélisé par un cylindre (cf. Fig. 3.71). Module du spectre des HRTF. HRTF obtenues pour le plan d'élévation 22.5° (coordonnées polaires verticales). D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées et le modèle de tête sphérique seul. Sujet JMP de la base *Jean-Marie Pernaux*.



(a) HRTF mesurées (b) Sphère (c) Sphère + Ellipsoïde

FIG. 3.74 – HRTF obtenues par calcul BEM avec le modèle combinant une tête sphérique et un torse ellipsoïdal raccordé à la tête par un cou modélisé par un cylindre (cf. Fig. 3.71). Module du spectre des HRTF. HRTF obtenues pour le plan d'élévation 45° (coordonnées polaires verticales). D'après [Pernaux, 2003]. Comparaison avec les HRTF mesurées et le modèle de tête sphérique seul. Sujet JMP de la base *Jean-Marie Pernaux*.

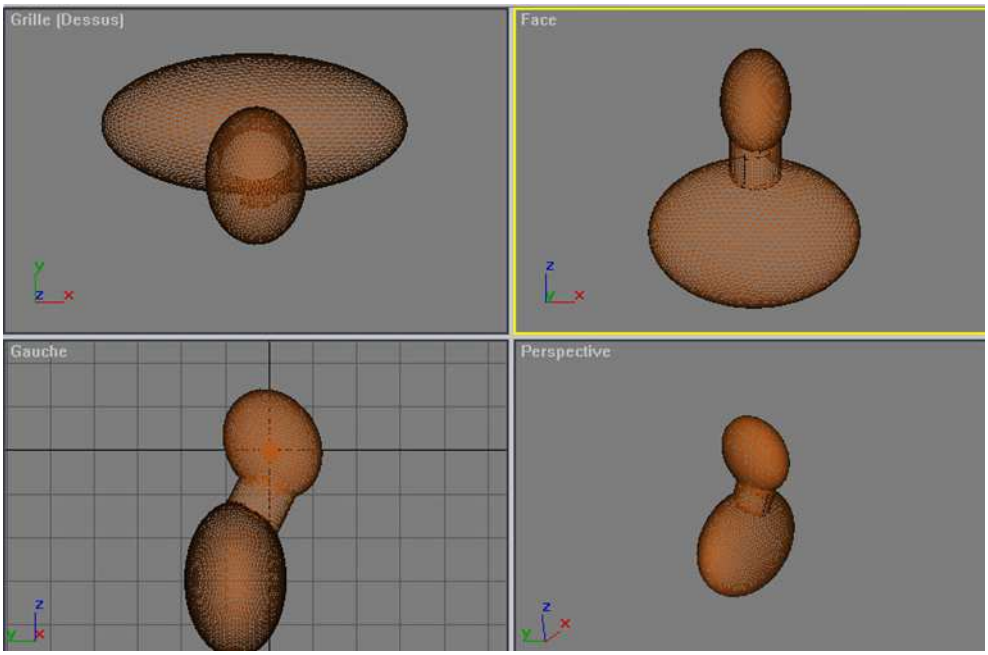


FIG. 3.75 – Modèle morphologique complet combinant une tête et un torse ellipsoïdaux. D'après [Pernaux, 2003].

– un cylindre à section elliptique (cou),
– un ellipsoïde (torse).

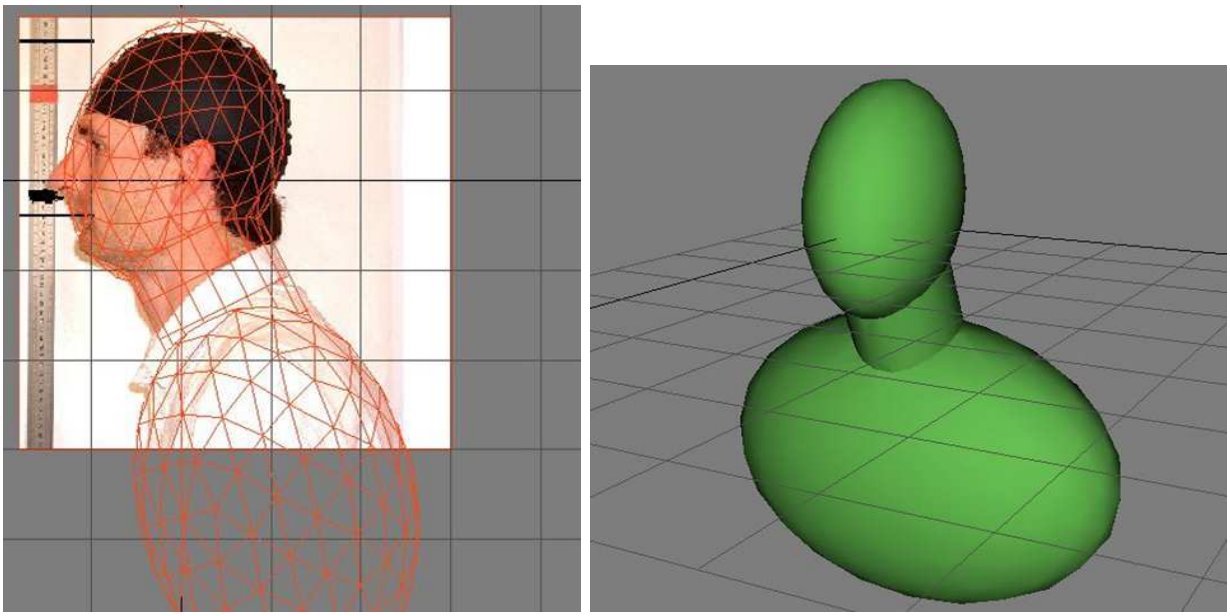
Les paramètres de ce modèle (dimensions et orientation des primitives) sont adaptés à la morphologie de l'individu. La procédure d'individualisation à partir de photographies de face et de profil du sujet (cf. Fig. 3.76) est décrite dans [Pernaux, 2003]. Le modèle, son individualisation et sa mise œuvre ont fait l'objet de 2 brevets [Pernaux et al., 2004a] [Pernaux et al., 2004b].

Le modèle morphologique combinant une tête et un torse ellipsoïdaux a été évalué en comparant les HRTF modélisées aux HRTF mesurées [Busson, 2006]. Les figures 3.77 et 3.78 illustrent les HRTF obtenues pour un sujet de la base *Jean-Marie Pernaux* dans le plan horizontal et le plan médian. Les HRTF ont été calculées sur la plage de fréquences[32] [0 - 4 kHz]. Dans le plan horizontal, la modélisation des figures de diffraction du côté controlatéral est assez fidèle. Du côté ipsilatérale, elle s'avère plus grossière. Dans le plan médian, les motifs d'arche liés à la réflexion sur le torse [Algazi et al., 2001a] prédominent sur les HRTF modélisées, alors qu'elles sont fortement masquées par les résonances du pavillon sur les mesures. L'absence du pavillon s'avère pénalisante pour le modèle, notamment au dessus de 2 kHz.

Conclusions

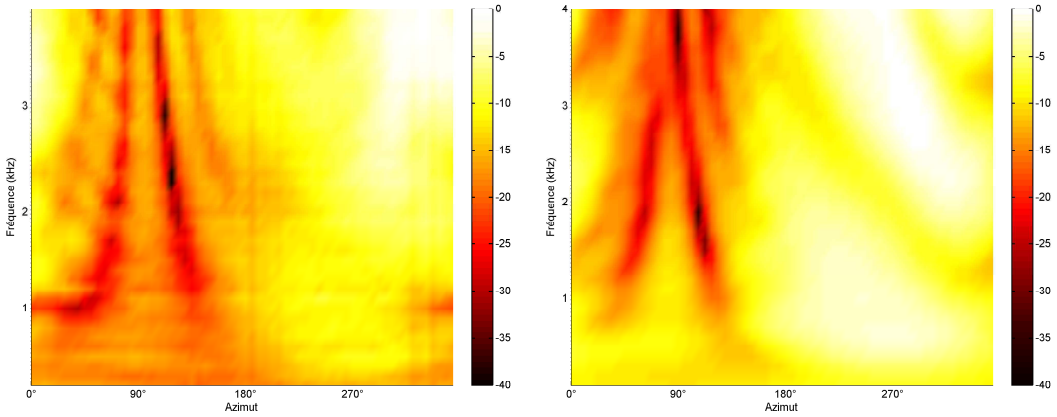
La modélisation BEM des HRTF est une méthode validée, mais dont la mise en œuvre reste problématique du fait de son coût de calcul dès lors qu'on recherche à modéliser les HRTF dans les hautes fréquences. La définition du maillage doit en effet augmenter lorsque la longueur d'onde diminue, ce qui conduit rapidement à un nombre souvent prohibitif d'élements de discrétisation. Or c'est justement dans les hautes fréquences qu'interviennent les IS et que leur individualité est cruciale. De plus l'acquisition de maillage 3D de morphologie reste une opération délicate peu

[32]Le calcul n'a pas pu être étendu aux fréquences supérieures à 4 kHz en raison du nombre de points du maillage résultant, nombre prohibitif compte tenu des capacités de calcul.



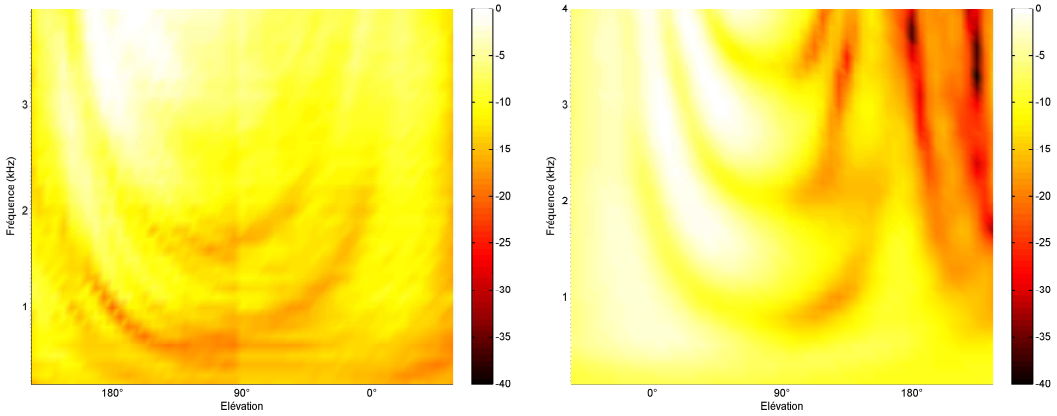
(a) Individualisation à partir d'une photographie (b) Modèle résultant

FIG. 3.76 – Individualisation du modèle morphologique complet combinant une tête et un torse ellipsoïdaux : sujet ME de la base *Jean-Marie Pernaux*. D'après [Pernaux, 2003].



(a) HRTF mesurées (b) HRTF modélisées

FIG. 3.77 – HRTF obtenues par calcul BEM à partir du modèle morphologique complet tête + torse : comparaison avec les HRTF mesurées. HRTF obtenues dans le plan horizontal. Module du spectre des HRTF. Sujet ME de la base *Jean-Marie Pernaux*. D'après [Busson, 2006].



(a) HRTF mesurées (b) HRTF modélisées

FIG. 3.78 – HRTF obtenues par calcul BEM à partir du modèle morphologique complet tête + torse : comparaison avec les HRTF mesurées. HRTF obtenues dans le plan médian. Module du spectre des HRTF. Sujet ME de la base *Jean-Marie Pernaux*. D'après [Busson, 2006].

compatible avec une utilisation grand public. Pour l'ensemble de ces raisons, des pistes alternatives de modélisation des HRTF ont été étudiées et font l'objet des sections qui suivent.

3.5.4 Modélisation des IS par apprentissage statistique basé sur des réseaux de neurones artificiels

La première alternative aux modèles BEM que nous avons examinée est l'**apprentissage statistique**, avec l'idée de collecter les HRTF finement mesurées (à la fois dans le domaine fréquenciel et spatial) d'un grand nombre d'individus afin d'en extraire des informations pertinentes pour "généraliser" à un nouvel individu et exprimer ses HRTF individuelles. Le plus naturellement les nouvelles HRTF sont obtenues par reconstruction sur une base de fonctions génératrices qui sont construites à partir d'une base de données de HRTF (*Modèle de type 2*). La difficulté consiste à déterminer les vecteurs de reconstruction et la procédure pour exprimer les coefficients permettant d'obtenir les HRTF d'un nouvel individu à partir de ces vecteurs. En d'autres termes : comment intégrer la spécificité individuelle du nouvel individu et générer ses HRTF à partir de l'information acquise pour d'autres individus ?

Sur ce point on se propose d'explorer une idée relativement originale : au lieu d'une description optique ou psycho-acoustique de l'individualité du nouvel auditeur (cf. page 211), nous allons utiliser une description **acoustique**, à travers un ensemble de HRTF mesurées sur l'individu pour un nombre réduit de directions (typiquement moins de 100 directions). L'idée d'utiliser des HRTF individuelles comme paramètres d'individualité est somme toute assez naturelle : les HRTF individuelles représentent en effet les données d'individualisation qui sont encore les plus proches des données de sortie du modèle. Ces HRTF individuelles seraient acquises par une procédure allégée de mesure de HRTF individuelles, sous la contrainte d'une session de mesure dont la durée ne devrait pas excéder quelques minutes afin de réduire au maximum l'inconfort du sujet. Il faut aussi avoir présent à l'esprit qu'il ne s'agit pas forcément d'une procédure de mesure de HRTF nécessitant d'obéir rigoureusement au protocole de mesure de HRTF tel qu'il est appliqué pour une mesure exhaustive de HRTF individuelles. L'objectif est d'obtenir une information suffisamment représentative de l'individualité des HRTF, mais cette information peut être sensiblement bruitée par des conditions non idéales de mesure. Le modèle d'individualisation peut parfaitement prendre

en charge un post-traitement des HRTF mesurées afin de corriger les artefacts liés aux éventuelles imperfections de la mesure.

Dans cette section, le concept général du modèle d'individualisation par RNA (Réseau de Neurones Artificiels) issu de ces idées est d'abord présenté. Ce modèle a été développé dans le cadre des travaux de thèse de S. Busson [Busson, 2006] [Lemaire et al., 2005] [Busson et al., 2006] [Nicol et al., 2006] qui se sont appuyés sur les résultats des stages de V. Choqueuse [Choqueuse, 2004] et P. Vovor [Vovor, 2005]. Il a fait l'objet de deux dépôts de brevet [Busson et al., 2005c] [Busson et al., 2005b]. Nous nous intéressons ensuite au choix des directions des HRTF individuelles utilisées comme paramètres d'individualité d'entrée du modèle, avant de donner les résultats d'une première évaluation du modèle.

Modélisation des IS individuels par RNA

Le modèle prend en entrée une sélection de HRTF mesurées sur l'individu considéré, ainsi que la direction pour laquelle on souhaite obtenir les HRTF individuelles (une paire de HRTF pour les oreilles gauche et droite par direction). Il produit en sortie les HRTF de cet individu dans la direction désirée. N'importe quelle direction peut être modélisée. Les HRTF individuelles de sortie sont calculées par un RNA de type perceptron multicouche ou MLP (*Multi Layer Perceptron*) [Busson, 2006]. Un RNA se compose d'un ensemble d'unités élémentaires interconnectées ou neurones (concept de *neurone formel* de McCulloch & Pitts) qui imitent le fonctionnement d'un neurone biologique du cerveau. La propriété fondamentale et caractéristique d'un neurone formel réside dans le fait que sa réponse (c'est à dire ce qui détermine ses variables de sortie) n'est pas simplement le résultat d'une combinaison pondérée de ses variables d'entrée, mais qu'elle dépend aussi de sa *fonction d'activation* qui traduit son activité et introduit potentiellement des non-linéarités. Les RNA présentent un intérêt dans les problèmes comportant un grand nombre de variables explicatives, avec l'avantage qu'ils sont capables d'identifier et d'exploiter les dépendances non-linéaires de haut niveau entre ces variables. Dans notre cas, les variables explicatives se constituent de l'ensemble des bins frequenciels de plusieurs dizaines de HRTF correspondant aux HRTF individuelles mesurées sur l'individu, ce qui représentent effectivement une quantité considérable de données. De plus, le lien entre ces variables et les variables de sortie (c'est à dire les HRTF individuelles dans d'autres directions) est pressenti à la fois comme complexe et sujet à de multiples interdépendances. Toutes ces raisons justifient la mise en œuvre des RNA.

Les RNA ont déjà été utilisés avec succès pour la modélisation de HRTF (ou de leur équivalent temporel les HRIR), mais principalement dans le but d'une représentation alternative des HRTF, éventuellement associée à une interpolation spatiale des données. Ainsi Jenison propose une modélisation par RNA (de type RBF pour *Radial Basis Function*) pour exprimer directement les coefficients du filtre pôle-zéro associé en fonction des coordonnées d'espace, à la fois à des fins d'implémentation dans un moteur de synthèse binaurale et dans un souci d'une représentation compacte des données [Jenison, 1995]. De façon similaire, l'auteur montre qu'un RNA est capable de calculer les composantes principales d'une représentation ACP des HRTF toujours à partir des coordonnées d'espace [Jenison & Fissell, 1996]. D'autres exemples très proches sont donnés dans [Wu et al., 1998] (représentation ACP des HRIR) et [Chu, 2004] (coefficients de filtres tout-pôle). Dans tous les cas, le modèle peut être utilisé pour interpoler les HRTF (plus exactement les coefficients de leur représentation) pour n'importe quelle direction de l'espace ne correspondant pas à une direction de mesure, et réalise ainsi potentiellement une interpolation spatiale. On note cependant que l'ensemble de ces exemples ne s'inscrit pas dans une démarche explicite de calcul de HRTF individuelles, du fait notamment qu'aucun modèle n'intégre dans les variables d'entrée des paramètres d'individualité (qu'il s'agisse d'une description morphologique ou de HRTF individuelles

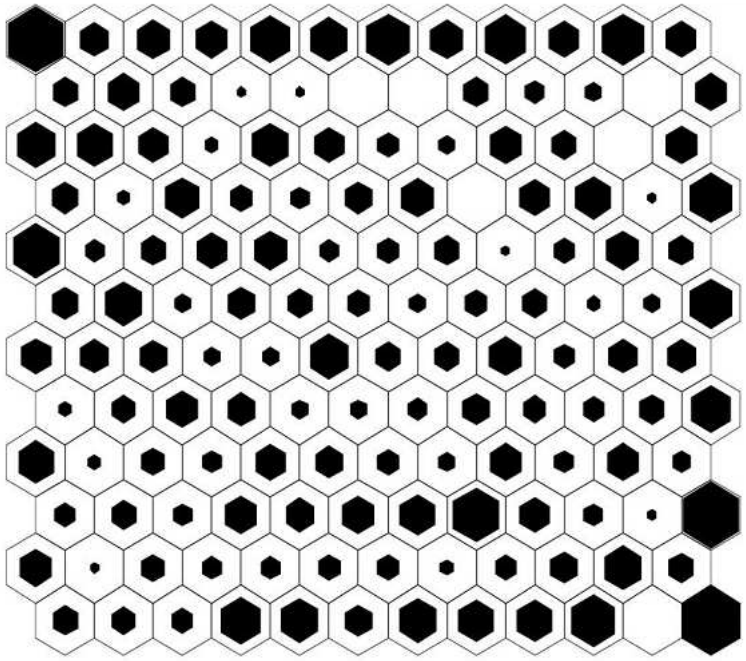


FIG. 3.79 – Carte de Kohonen regroupant les 1250 HRTF d'un individu (Base de données CIPIC). La carte est composée de 12 x 12 neurones, soit 144 classes potentielles. A l'intérieur de chaque neurone, la taille du losange noir est proportionnelle au nombre de HRTF regroupées dans ce neurone.

mesurées), contrairement au modèle que nous proposons.

Notre modèle se construit comme n'importe quel RNA par apprentissage sur une base de données. En l'occurrence l'étude a considéré la base de données de HRTF du CIPIC en raison du nombre important d'individus qu'elle contient (45 individus, cf. Tab. 3.1). L'apprentissage consiste à ajuster les poids des neurones de façon à minimiser une fonction de coût qui traduit la distance entre la sortie attendue et la sortie effective du RNA. Cette opération est menée sur la base d'exemples extraits de la base de données et à partir desquels le RNA élabore sa connaissance des phénomènes (c'est à dire des interactions entre les variables) afin de les modéliser. La base de données est décomposée en trois sous-ensembles : ensemble d'*apprentissage* (correspondant aux données sur lesquelles le RNA apprend et ajuste ses poids), ensemble de *validation* (données sur lesquelles l'erreur de modélisation est évaluée au cours de l'apprentissage et qui permettent de mesurer la capacité du RNA à généraliser la modélisation à des données non apprises) et ensemble de *test* (données non apprises sur lesquelles l'erreur de modélisation est évaluée à l'issue de l'apprentissage et qui permettent de caractériser les performances finales du RNA).

Choix des HRTF individuelles d'entrée du modèle

Les HRTF individuelles appliquées en entrée du modèle au titre de paramètres d'individualité constituent un des éléments clefs du modèle. Deux questions se posent : premièrement quel est le nombre minimum de HRTF nécessaires pour représenter l'individualité de l'encodage binaural d'un auditeur du point de vue de notre modèle, deuxièmement existe-t-il des directions privilégiées, et si oui lesquelles, pour choisir ces HRTF ? Pour y répondre, nous avons décidé d'effectuer une **clas-**

sification (en anglais *clustering*) des HRTF dans le but de trier les HRTF[33] selon leur similarité et de regrouper l'ensemble des HRTF d'un individu en un nombre donné de classes, la règle étant qu'au sein d'une classe les HRTF peuvent être considérées comme semblables (voire indiscriminables), alors que les HRTF de deux classes distinctes sont différentes. L'objectif est de déterminer le nombre de classes par lesquelles on peut représenter la variance spatiale des HRTF d'un individu. L'analyse de HRTF par des méthodes de type ACP ou ACI nous indique déjà qu'il existe une forte redondance spatiale entre les HRTF d'un individu et qu'en conséquence, il est possible de représenter l'ensemble des HRTF d'un individu par moins de 10 composantes principales. Kistler & Wightman ont montré qu'avec seulement 5 composantes on prend en compte 90% de la variance [Kistler & Wightman, 1992], tandis que dans [Jenison & Fissell, 1996], 6 composantes sont utilisées pour atteindre une variance de 98%. Ces résultats suggèrent qu'il est possible de représenter sous une forme compacte (c'est à dire, pour la question qui nous intéresse, avec un nombre réduit de représentants) l'information spatiale contenue dans les HRTF d'un individu. D'ailleurs l'idée de classifier les HRTF n'est pas nouvelle en soi et a déjà été appliquée avec succès [Shimada et al., 1994] [Chuang, 1995] [Lo, 1998] [Fahn & Lo, 2003].

La méthode de classification choisie combine deux algorithmes [Busson, 2006] :

- d'abord une **carte de Kohonen** (ou SOM pour *Self-Organizing Map*) [Kohonen, 1995] pour un premier regroupement des HRTF,
- suivie d'une Classification Hiérarchique Ascendante (CHA) [Lemaire & Clérot, 2002] afin de réduire le nombre de groupes.

La mesure de similarité utilisée est la distance euclidienne définie par :

$$d(H_1, H_2) = \frac{1}{M} \sum_{i=1}^{M} [20 log_{10}(\frac{|H_1(f_i)|}{|H_2(f_i)|})]^2 \tag{3.37}$$

où H_1 et H_2 désignent deux HRTF. Pour l'algorithme de CHA, le regroupement ou agrégation des classes obéit au critère de *Ward* visant à minimiser la perte d'inertie inter-classes à chaque étape de regroupement [Casin, 1999].

La classification est menée sur les HRTF d'un individu de la base de données du CIPIC (cf. Tab. 3.1), soit un ensemble de 1250 HRTF correspondant à l'oreille droite. La Figure 3.79 reproduit la carte de Kohonen obtenue à la première étape. La carte utilisée est constituée de 12 x 12 = 144 classes ou *neurones*[34] qui sont caractérisés par un voisinage hexagonal. Un premier résultat de la carte de Kohonen est le regroupement des HRTF en 144 classes potentielles. Le nombre de HRTF associées à chaque neurone est codé sur la Figure 3.79 par la taille des losanges noirs contenus dans chaque neurone. On observe que certains neurones sont vides, tandis que d'autres contiennent un grand nombre de HRTF. Il est clair que la classification peut être encore affinée, d'où le recours à l'algorithme de CHA en seconde passe. Outre le regroupement des données, la carte de Kohonen visualise un second résultat : du fait qu'il s'agit d'une *carte*, la répartition des neurones a un sens et traduit des relations de voisinage entre les neurones. En d'autres termes, le voisinage topologique de deux neurones sur la carte coïncide avec la similarité des données qu'ils contiennent. Il est alors intéressant de *projeter* sur la carte de Kohonen des informations relatives aux données : par exemple pour chaque neurone on peut afficher l'angle moyen d'azimut des HRTF contenues dans ce neurone, ainsi que l'angle moyen d'élévation. Ainsi sur la Figure 3.80, une couleur est attribuée à chaque neurone pour représenter l'angle moyen d'azimut ou d'élévation des HRTF contenues dans ce neurone. On se rend compte que les neurones situés dans le coin supérieur gauche correspondent

[33] Dans tout ce qui suit, l'étude ne prend pas en compte les indices temporels et se focalisent sur les IS. Aussi, par HRTF, faut-il entendre le module du spectre des HRTF.

[34] Une carte de Kohonen est en effet une catégorie de RNA, c'est pourquoi on parle de neurones.

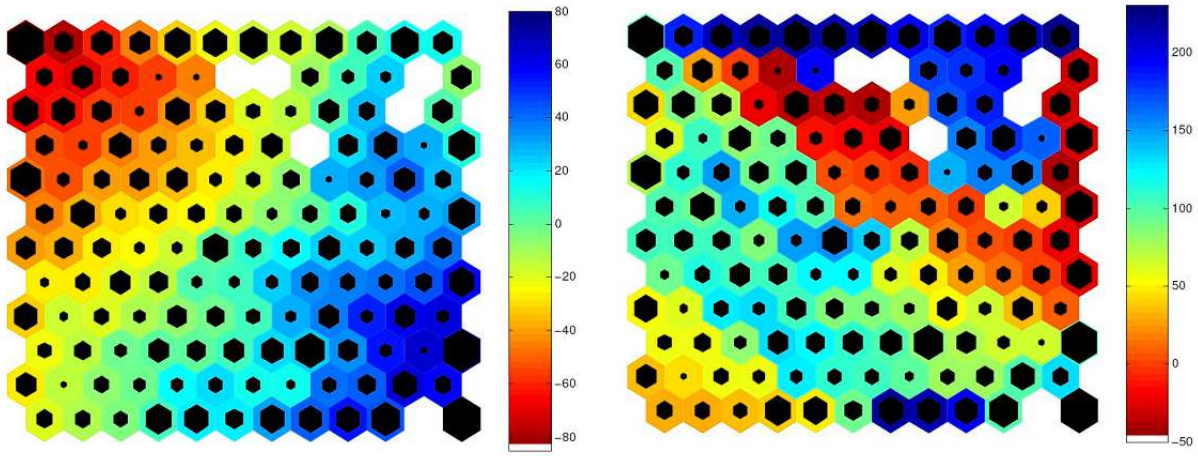


FIG. 3.80 – Carte de Kohonen regroupant les 1250 HRTF d'un individu (Base de données CIPIC) : Projection des angles moyens d'azimut (à gauche) et d'élévation (à droite) sur les neurones. Les valeurs d'angle sont décrites par l'échelle de couleurs figurée sur le côté.

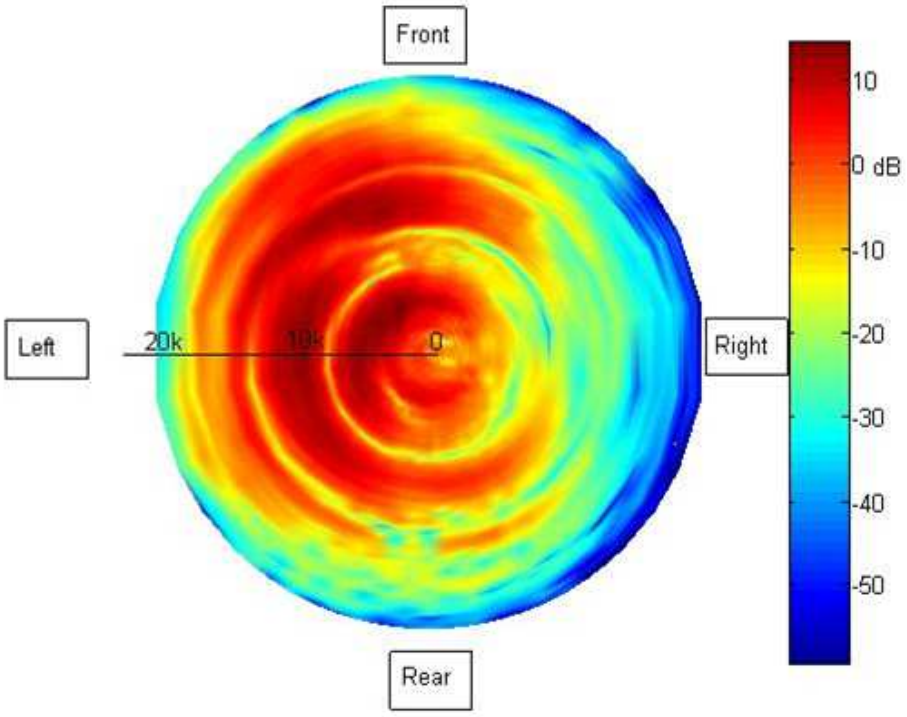


FIG. 3.81 – Evolution des HRTF (module du spectre exprimé en dB) dans le plan horizontal : Représentation polaire, où le rayon correspond à l'axe des fréquences et l'angle à l'angle d'azimut. On observe la relative symétrie entre avant et arrière. Sujet 003 de la base de données CIPIC.

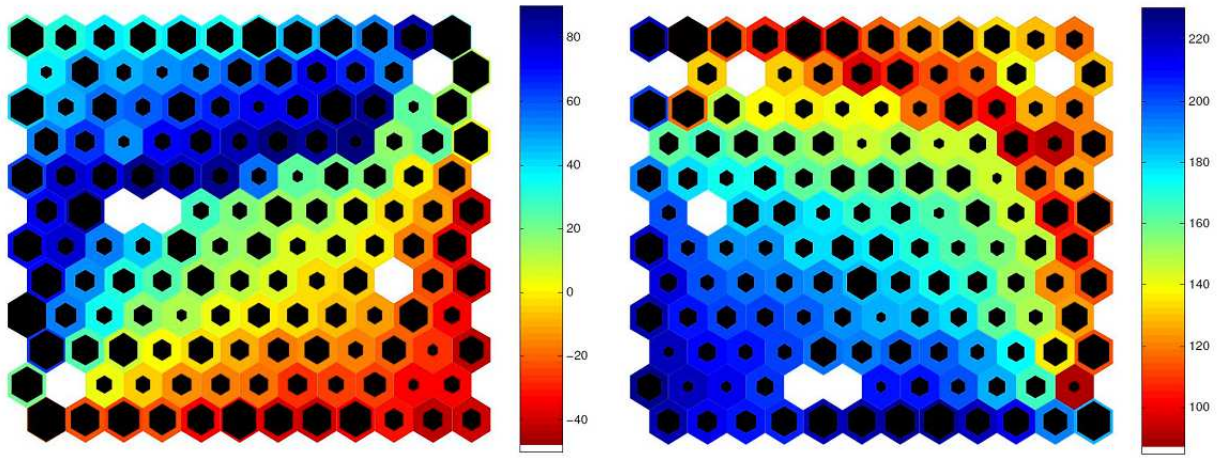


FIG. 3.82 – Carte de Kohonen regroupant les 1250 HRTF d'un individu (Base de données CIPIC) après séparation des HRTF des hémisphères avant et arrière : Projection de l'angle moyen d'élévation pour l'hémisphère avant (à gauche) et l'hémisphère arrière (à droite) sur les neurones.

à des HRTF localisées "en moyenne" sur le côté gauche (azimut moyen proche de -80°), tandis que ceux du coin inférieur droit correspondent à des HRTF localisées sur le côté droit (azimut moyen proche de 80°). De plus on observe qu'entre ces deux positions, les valeurs d'azimut se répartissent quasi-linéairement entre -80° et 80°, avec une progression perpendiculaire à la diagonale indirecte. On a donc une classification qui semble obéir à une distribution des HRTF en fonction de l'angle d'azimut, ce qui n'est pas surprenant étant donné que les HRTF sont générées par des phénomènes acoustiques qui dépendent en partie de l'angle d'azimut. En revanche, si l'on examine l'angle moyen d'élévation pour chaque neurone, la classification ne traduit aucune structure spatiale cohérente avec une distribution relativement aléatoire des valeurs d'élévation sur la carte. On peut remarquer qu'effectivement il existe une forte similarité entre les HRTF mesurées à l'hémisphère avant et l'hémisphère arrière (cf. Fig. 3.81). Cette similarité conduit l'algorithme de classification à regrouper des HRTF situées dans les deux hémisphères, c'est à dire distantes de 180° en termes d'élévation[35], ce qui introduit une forte variance des angles d'élévation associés aux HRTF contenues dans un neurone et, par suite, pollue la lisibilité de l'élévation moyenne sur la carte. Aussi avons-nous reconduit la classification en séparant au préalable les HRTF des deux hémisphères avant et arrière. La projection des angles moyens d'élévation est affichée sur la Figure 3.82 pour les deux hémisphères. On obtient cette fois une progression régulière des valeurs de l'angle moyen le long de la diagonale indirecte, entre le coin supérieur gauche et le coin inférieur droit, pour l'hémisphère avant, et le long de la diagonale directe, entre le coin inférieur gauche et le coin supérieur droit, pour l'hémisphère arrière. La classification semble maintenant cohérente avec la structure spatiale des HRTF.

A l'issue de la première classification opérée par la carte de Kohonen, un algorithme de CHA est appliqué et on obtient au final 13 classes ou *clusters* pour chaque hémisphère[36], correspondant au regroupement des 144 classes identifiées au préalable dans les cartes de Kohonen. Les figures 3.83 & 3.84 illustrent ces 13 classes sur la sphère 3D, en visualisant avec la même couleur les directions des HRTF contenues dans une même classe, ce qui permet de rendre compte à la fois de la localisation et de l'étendue des 13 classes. On observe que les membres d'une même classe

[35] On rappelle que la base de données du CIPIC repose sur le système de coordonnées polaire interaural.

[36] A noter que le fait qu'on obtienne le même nombre de classes pour les deux hémisphères n'est qu'une coïncidence.

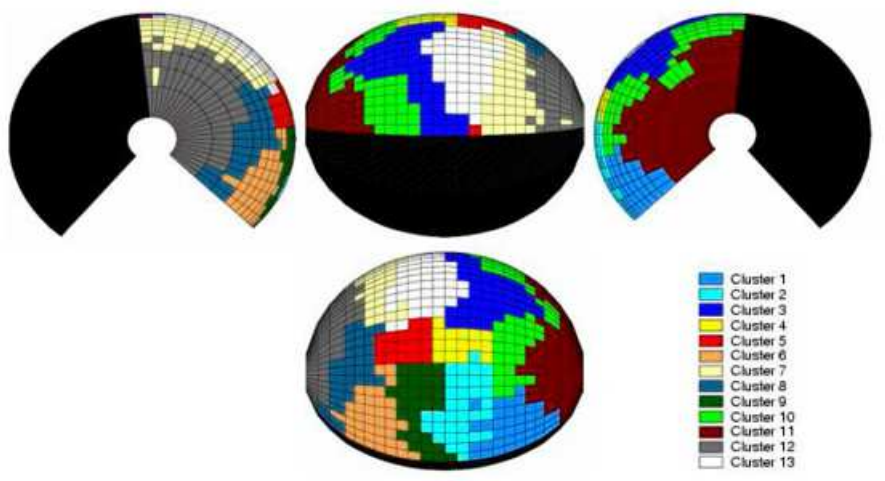


FIG. 3.83 – Visualisation des 13 classes regroupant les 625 HRTF de l'hémisphère avant d'un individu (Base de données CIPIC).

forment des groupes relativement homogènes d'un point de vue spatial, c'est à dire que, dans leur majorité, ils correspondent à des directions adjacentes ou très proches. Par ailleurs, la distribution spatiale des classes renseigne sur la variabilité des HRTF sur la sphère 3D. On remarque que cette variabilité est présente à la fois en azimut et en élévation. Si l'on s'intéresse aux vues frontale et arrière des figures 3.83 & 3.84, il apparaît que la variabilité en azimut semble légèrement plus marquée que celle en élévation, au sens où les classes décrivent pour la plupart des "tâches" formant des bandes verticales étroites, ce qui dénote des variations plus rapides en azimut qu'en élévation. Ce résultat suggère que les IS utilisés pour la localisation en élévation reposent sur des variations spectrales fines de l'ordre de la variance intra-classe.

La dernière étape de la classification consiste à élire une HRTF *représentative* (ou *parangon*) pour chaque classe. Cette HRTF représentative, soit H_{r_n}, est choisie comme le membre de la classe qui minimise la somme des distances (distance au sens de la distance euclidienne définie à l'équation 3.37) avec les autres membres :

$$H_{r_n} = H_{m_r} \quad / \quad m_r = \min_{m \in C_n} \sum_{l=1}^{N_n} d(H_l, H_m) \tag{3.38}$$

où C_n désigne la nième classe regroupant N_n membres : $C_n = \{H_1, H_2, \ldots, H_{N_n}\}$. Les HRTF H_l et H_m sont deux membres de cette classe. Au total, pour les deux hémisphères, 2 x 13 = 26 HRTF représentatives sont élues. Les directions associées à ces HRTF représentatives sont représentées sur la Figure 3.85. L'ensemble de ces HRTF représentatives constituent les paramètres d'individualité à appliquer en entrée du modèle. Afin d'évaluer les performances des HRTF représentatives ainsi identifiées en termes de "représentativité", on définit l'**erreur de quantification** qui mesure l'erreur commise lorsqu'on remplace une HRTF donnée H_m par la HRTF représentative de la classe à laquelle elle appartient, soit H_{r_n} :

$$e_q = \frac{1}{M} \sum_{i=1}^{M} |H_m(f_i) - H_{r_n}(f_i)| \tag{3.39}$$

Pour l'évaluation de leur représentativité, les HRTF représentatives sont comparées à une sélection d'un nombre égal de HRTF choisies de façon à couvrir de façon uniforme la sphère 3D

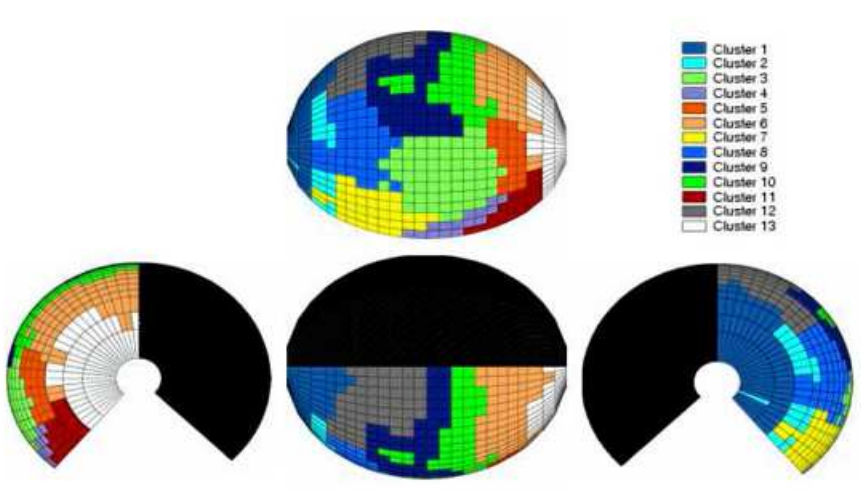


FIG. 3.84 – Visualisation des 13 classes regroupant les 625 HRTF de l'hémisphère arrière d'un individu (Base de données CIPIC).

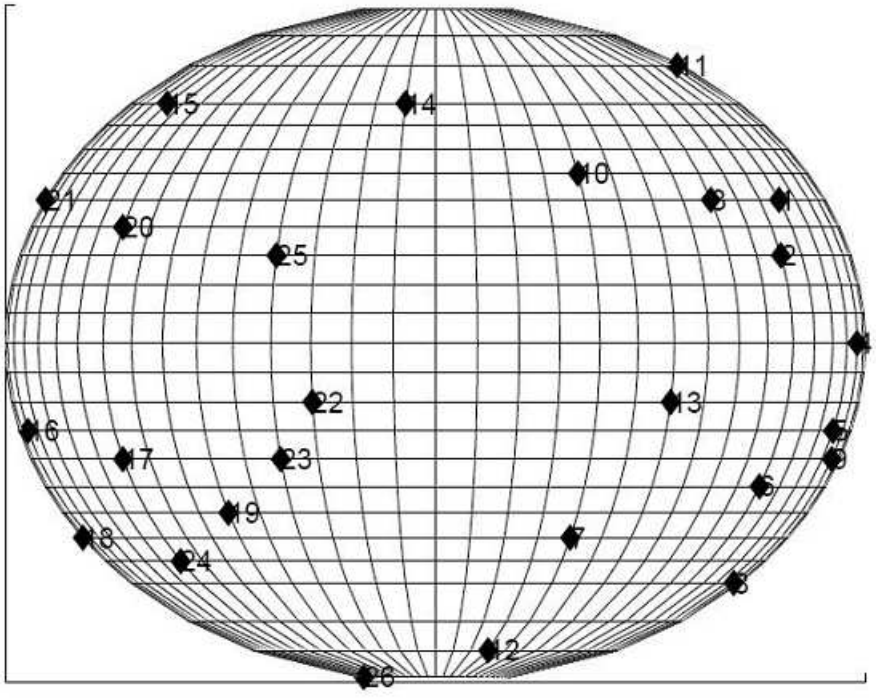


FIG. 3.85 – Visualisation des 26 HRTF représentatives associées aux 26 classes regroupant les 1250 HRTF d'un individu (Base de données CIPIC).

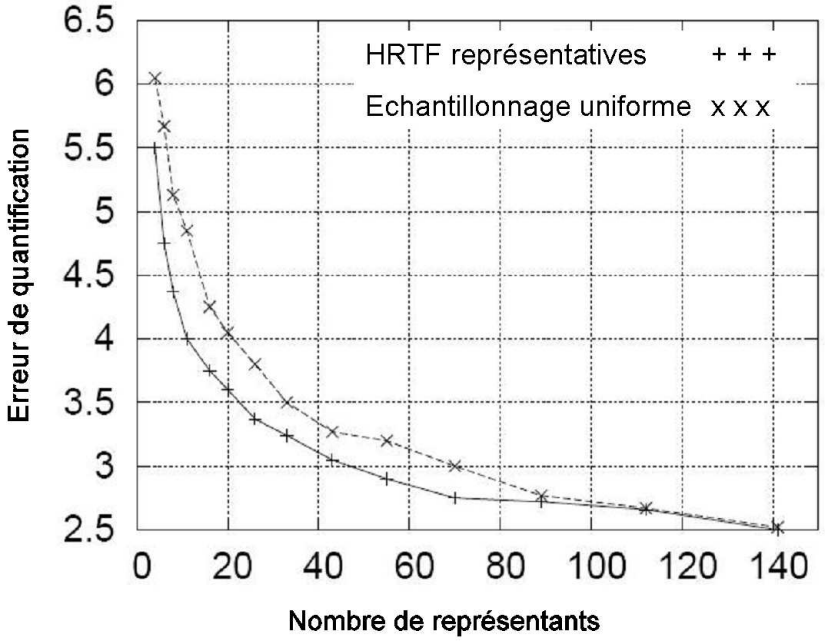


FIG. 3.86 – Evaluation des HRTF représentatives : Erreur de quantification en fonction du nombre de représentants. Comparaison entre les HRTF représentatives et une sélection géométrique uniforme. Cas où seules les 1250 HRTF d'un individu (Base de données CIPIC) sont considérées.

[Busson, 2006]. L'erreur de quantification est reproduite en fonction du nombre de HRTF représentatives sur la Figure 3.86. L'erreur de quantification obtenue avec les HRTF représentatives s'avère inférieure à celle correspondant à une sélection géométrique uniforme lorsque le nombre de HRTF représentatives devient inférieur à 100. Pour atteindre une erreur $e_q = 3$, il suffit de 45 HRTF représentatives, alors qu'une sélection géométrique uniforme requiert 70 HRTF individuelles mesurées. Ces résultats démontrent l'apport des HRTF représentatives pour caractériser l'individualité d'un auditeur, en comparaison d'un échantillonnage uniforme. Cependant, les directions des HRTF représentatives ont été identifiées pour un individu donné. On peut se demander si ces directions sont universelles et permettent de représenter aussi bien n'importe quel individu ? La même évaluation que précédemment est menée en considérant cette fois tous les individus de la base de données du CIPIC. Pour chaque individu, les directions des HRTF représentatives sont celles identifiées pour le premier individu. L'erreur de quantification est représentée sur la Figure 3.87. On note que la sélection géométrique uniforme donne une erreur inférieure à celle obtenue avec les HRTF représentatives, ce qui suggère que les directions des HRTF représentatives du premier individu ne sont pas adaptées aux autres individus. Pour une représentation de l'individualité de l'auditeur plus performante qu'un échantilonnage uniforme, il conviendrait donc d'individualiser les directions des HRTF représentatives, ce qui constitue une sévère limitation au modèle. Une raison possible expliquant ce résultat réside dans le fait que d'un individu à l'autre les différences morphologiques consistent pour partie en des décalages ou des différences d'orientations des éléments morphologiques (pavillon d'oreille principalement). Il n'est donc pas surprenant que les directions représentatives dépendent de l'individu, en fonction des caractéristiques de sa géométrie. Ce point sera rediscuté au cours des sections suivantes.

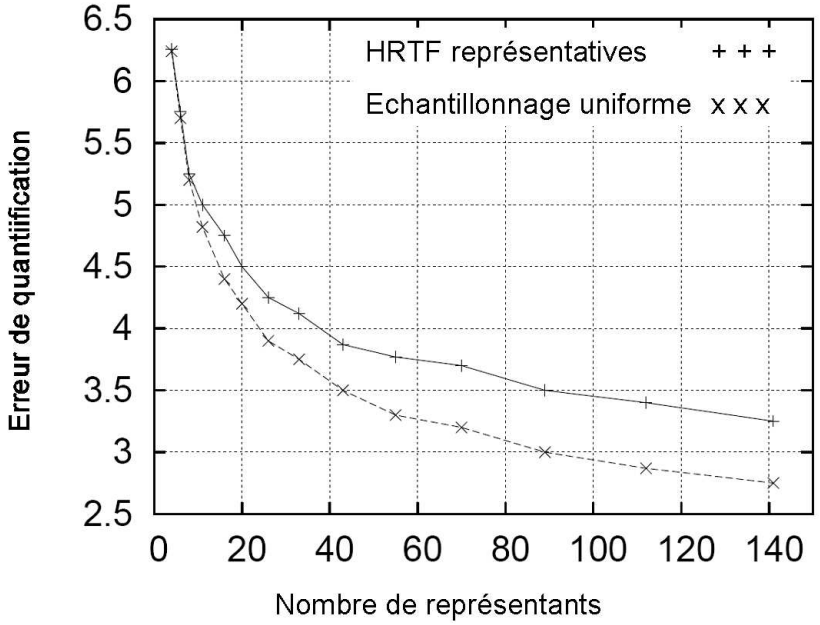


FIG. 3.87 – Evaluation des HRTF représentatives : Erreur de quantification en fonction du nombre de représentants. Comparaison entre les HRTF représentatives et une sélection géométrique uniforme. Cas où tous les individus de la base de données CIPIC sont considérés.

Validation du modèle

La question du choix des directions des HRTF appliquées en entrée du modèle au titre de paramètres d'individualisation vient d'être examinée. On a montré que, pour un individu donné, il existe des directions privilégiées pour choisir ces HRTF d'individualisation, ces directions permettant d'augmenter l'information d'individualité par rapport à un échantillonnage uniforme de la sphère 3D. Mais le problème est que ces directions semblent dépendre de l'individu, ce qui est un obstacle majeur, car faire dépendre de l'individu les directions de mesure des HRTF d'individualisation soulève plusieurs difficultés pratiques. Il faudrait d'abord se doter d'une procédure pour identifier ces directions représentatives pour n'importe quel nouvel individu. Ensuite cela suppose aussi que le système de mesure des HRTF individuelles admette le degré de liberté de choisir n'importe quelle direction de mesure. Pour toutes ces raisons, il est plus judicieux de préférer un échantillonnage uniforme de la sphère 3D. Le modèle est ainsi mis en œuvre sur les 45 sujets de la base de données du CIPIC qui est dans ce but décomposée en trois sous-ensembles : un ensemble d'apprentissage (50% des données), un ensemble de validation (25% des données) et un ensemble de test (25% des données).

En pratique, pour une direction donnée de HRTF à modéliser, le modèle n'exploite pas toutes les HRTF individuelles mesurées appliquées en entrée (paramètres d'individualité), mais uniquement la HRTF correspondant à la classe à laquelle appartient la direction désirée. En l'occurrence, il n'est plus exact de parler de classes ou de HRTF représentatives, étant donné que nous avons opté pour un échantillonnage uniforme de la sphère 3D pour sélectionner les HRTF d'individualisation. Cependant nous conserverons cette terminologie dans un souci de continuité avec ce qui précède, en considérant que l'échantillonnage uniforme est, en quelque sorte, un cas "particulier" de classification. Dans la section précédente, on a évalué l'*erreur de quantification* représentant l'erreur commise lorsqu'on modélise la HRTF dans la direction désirée par la HRTF représentative associée

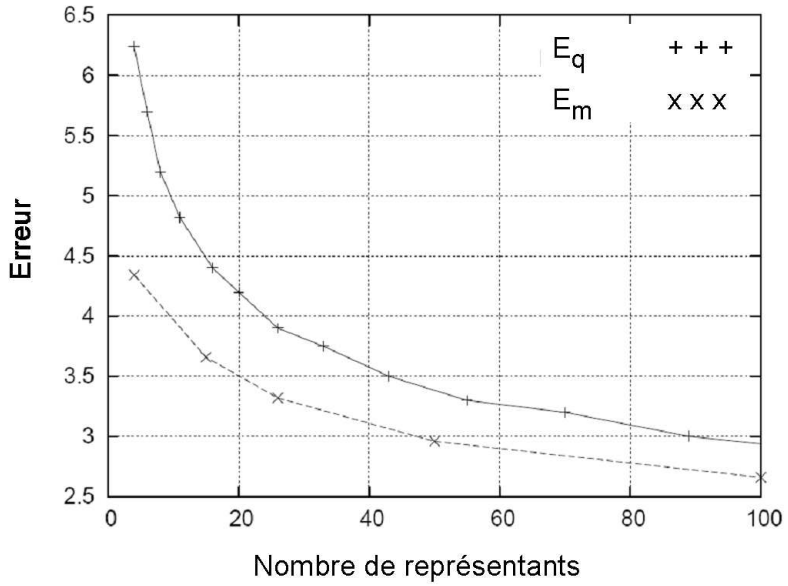


FIG. 3.88 – Validation du modèle de calcul de HRTF individuelles : Comparaison entre l'erreur de quantification E_q et l'erreur de modélisation E_m en fonction du nombre de représentants. Moyenne sur les données de l'ensemble de test.

à la classe de la HRTF désirée. Il s'agit du modèle à l'ordre 0, dans lequel aucun apprentissage n'est mis en œuvre. A présent nous allons comparer cette erreur de quantification à l'*erreur de modélisation* définie comme l'erreur du modèle à l'issue de l'apprentissage :

$$e_m = \frac{1}{M} \sum_{i=1}^{M} |H_m(f_i) - \hat{H}_m(f_i)| \tag{3.40}$$

où $\hat{H}_m$ désigne la HRTF en sortie du modèle. La Figure 3.88 représente les erreurs de quantification et de modélisation. On vérifie que l'erreur de modélisation est bien inférieure à l'erreur de quantification, ce qui dénote l'apport de l'apprentissage réalisé par le RNA et démontre sa capacité à généraliser à partir des données apprises. Le niveau d'erreur de $e_m = 3$ est obtenu pour un peu moins de 50 HRTF représentatives à l'issue de l'apprentissage, alors que 90 HRTF représentatives sont nécessaires pour atteindre le même résultat avec une simple quantification. La Figure 3.89 illustre les HRTF modélisées par le RNA pour 50 et 100 HRTF représentatives. On est frappé par la qualité de la reconstruction, notamment par le degré de finesse et de détail avec lequel sont reproduits les IS, en comparaison par exemple de nos résultats de modélisation BEM (cf. Section 3.5.3). Avec 100 HRTF représentatives, les HRTF modélisées semblent très proches des HRTF mesurées. Lorsqu'on réduit ce nombre à 50 HRTF représentatives, la reconstruction est entachée d'un léger flou.

Conclusion

Le modèle présenté est séduisant par l'originalité de son schéma basé sur la mesure, au titre des paramètres d'individualité, d'un nombre réduit de HRTF dans des directions suffisamment représentatives des phénomènes intrinsèques et de l'individualité de l'encodage binaural. Ces paramètres d'entrée possèdent l'avantage d'être au plus proches des données de sortie du modèle. Dans cette étude de faisabilité, la précision de la reproduction des HRTF individuelles sur l'ensemble de la

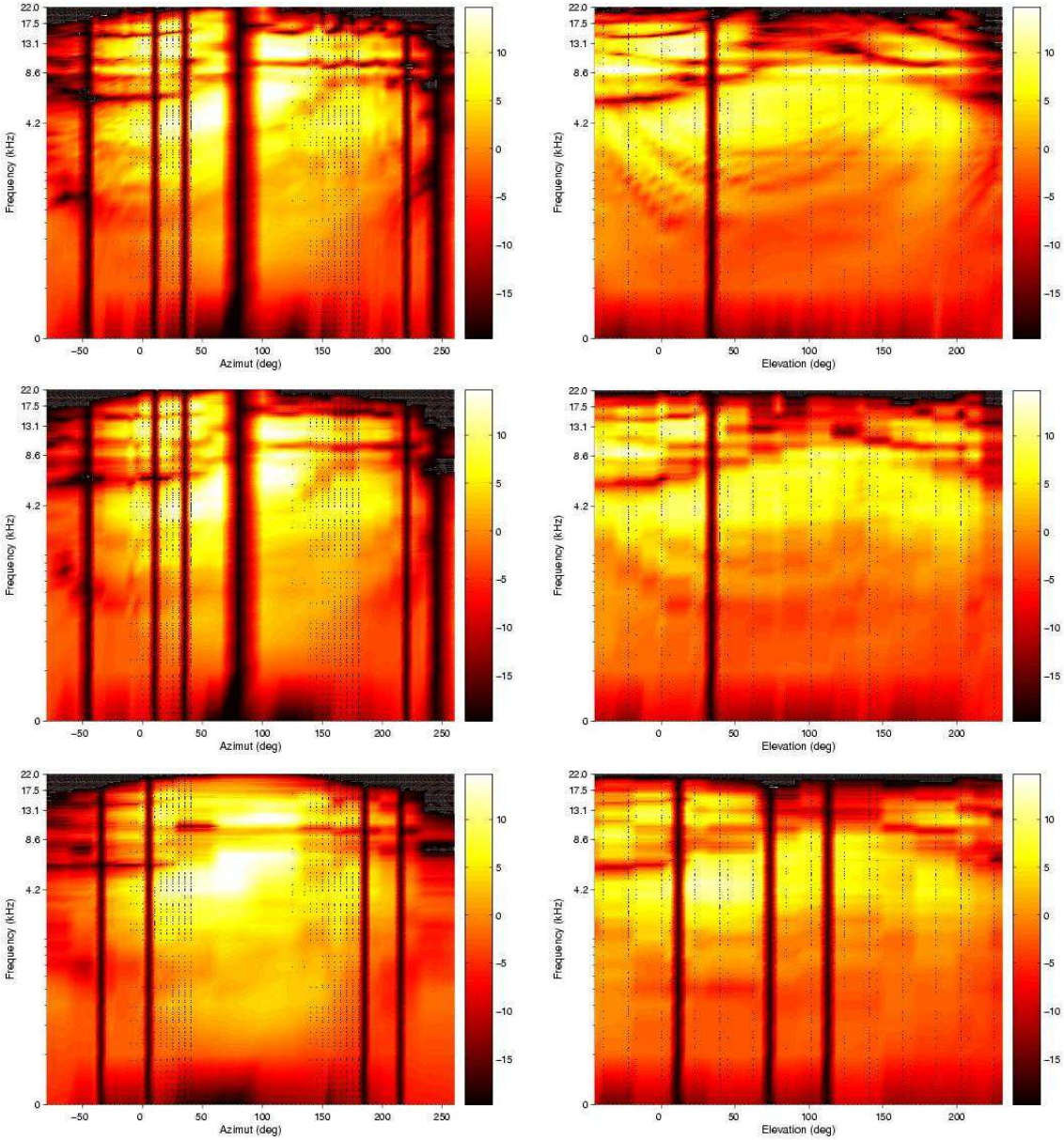


FIG. 3.89 – Validation du modèle de calcul de HRTF individuelles : Illustration des HRTF modélisées avec 100 (ligne du milieu) et 50 (ligne du bas) HRTF représentatives. Comparaison avec les HRTF mesurées (ligne du haut). Représentation dans le plan horizontal (colonne gauche) et dans le plan médian (colonne droite). HRTF d'un individu de la base de données du CIPIC (ensemble de test).

gamme audible [20 Hz - 20 kHz] est assez remarquable, notamment en regard des modèles BEM (cf. Section 3.5.3). Néanmoins la question du choix des directions de mesure des HRTF indviduelles de mesure reste non résolue. On a montré qu'il existe des directions priviligiées permettant d'optimiser les performances de la modélisation. Malheureusement ces directions dépendent de l'individu, ce qui relativise l'intérêt du modèle. En effet l'observation des HRTF révèle que si des phénomènes similaires se produisent d'un individu à l'autre, on remarque que souvent les phénomènes saillants apparaissent dans des directions différentes, en raison de la morphologie de l'individu. Par exemple, une différence morphologique inter-individuelle réside dans l'orientation du pavillon dont on sait qu'il joue un rôle primordial dans la génération des IS [Guillon, 2009] [Maki & Furukawa, 2005]. Il semble donc illusoire de chercher des directions universelles pour les HRTF représentatives.

3.5.5 Modélisation des IS par reconstruction de HRTF individuelles mesurées sur un échantillonnage spatial grossier

Ce modèle repart de la principale innovation de la solution précédente : on considère comme paramètres d'individualisation en entrée du modèle, un ensemble de HRTF individuelles mesurées pour un échantillonnage grossier de la sphère 3D. On cherche alors une méthode pour calculer les HRTF de cet individu dans n'importe quelle direction à partir de ces données d'entrée. Dans son principe, la modélisation s'identifie à un processus de **reconstruction** des HRTF dans les directions non mesurées. Fondamentalement il s'agit donc d'une **interpolation**. L'interpolation des HRTF est un problème qui a déjà été largement traité. Le modèle précédent basé sur un RNA (cf. Section 3.5.4) constitue un cas d'interpolation non linéaire, mais, pour l'essentiel des études, des méthodes d'interpolation linéaire ont été considérées. Pour les HRTF, il semble que la meilleure méthode soit l'interpolation par des fonctions STPS (Spherical Thin Plane Spline) [Hartung et al., 1999] [Guillon, 2009]. Avec cette méthode, Carlile *et al* [Carlile et al., 2000] montrent qu'il faut un minimum de 150 directions mesurées pour ne pas observer de dégradations significatives dans la localisation des sources virtuelles. C'est un ordre de grandeur qu'on retrouve pour d'autres méthodes d'interpolation [Martin & McAnally, 2007]. Même s'il est délicat de définir une limite précise, du fait que la qualité du VAS s'altère *progressivement* au fur et à mesure où le nombre de HRTF individuelles en entrée décroît [Langendijk & Bronkhorst, 2000], nous retiendrons cette valeur de 150 comme valeur de référence qui constitue le défi à relever par les nouveaux modèles de reconstruction.

Aspects novateurs

Qu'elles agissent de façon linéaire ou non, les méthodes d'interpolation opèrent en aveugle, c'est à dire qu'elles restent des méthodes génériques qui traitent les HRTF comme n'importe quelles autres données à interpoler. Afin de dépasser la limite de 150 directions mesurées, il nous semble impossible de faire mieux sans injecter de l'intelligence dans le modèle. En ce sens, le modèle proposé cherche à guider le processus de reconstruction par des informations a priori sur les données à reconstruire. Cette stratégie mise sur les similarités inter-individuelles qu'on a observées sur les SFRS (cf. Section 3.1.3) : le nouvel éclairage qu'apporte l'observation des SFRS sur les différences individuelles permet en effet de dégager l'idée d'un modèle générique (*prototype*) décrivant les évolutions spatiales et fréquencielles des HRTF communes d'un individu à l'autre, et par suite de concevoir la calcul des HRTF individuelles comme un simple ajustement particulier de ce prototype. Cette idée fondamentale impose un double choix méthodologique [Guillon, 2009] :

- les données (HRTF) sont structurées sous forme de **SFRS** pour l'analyse et le traitement,
- la mesure de distance entre HRTF est l'**intercorrélation normalisée** entre SFRS (cf. Section 3.5.1).

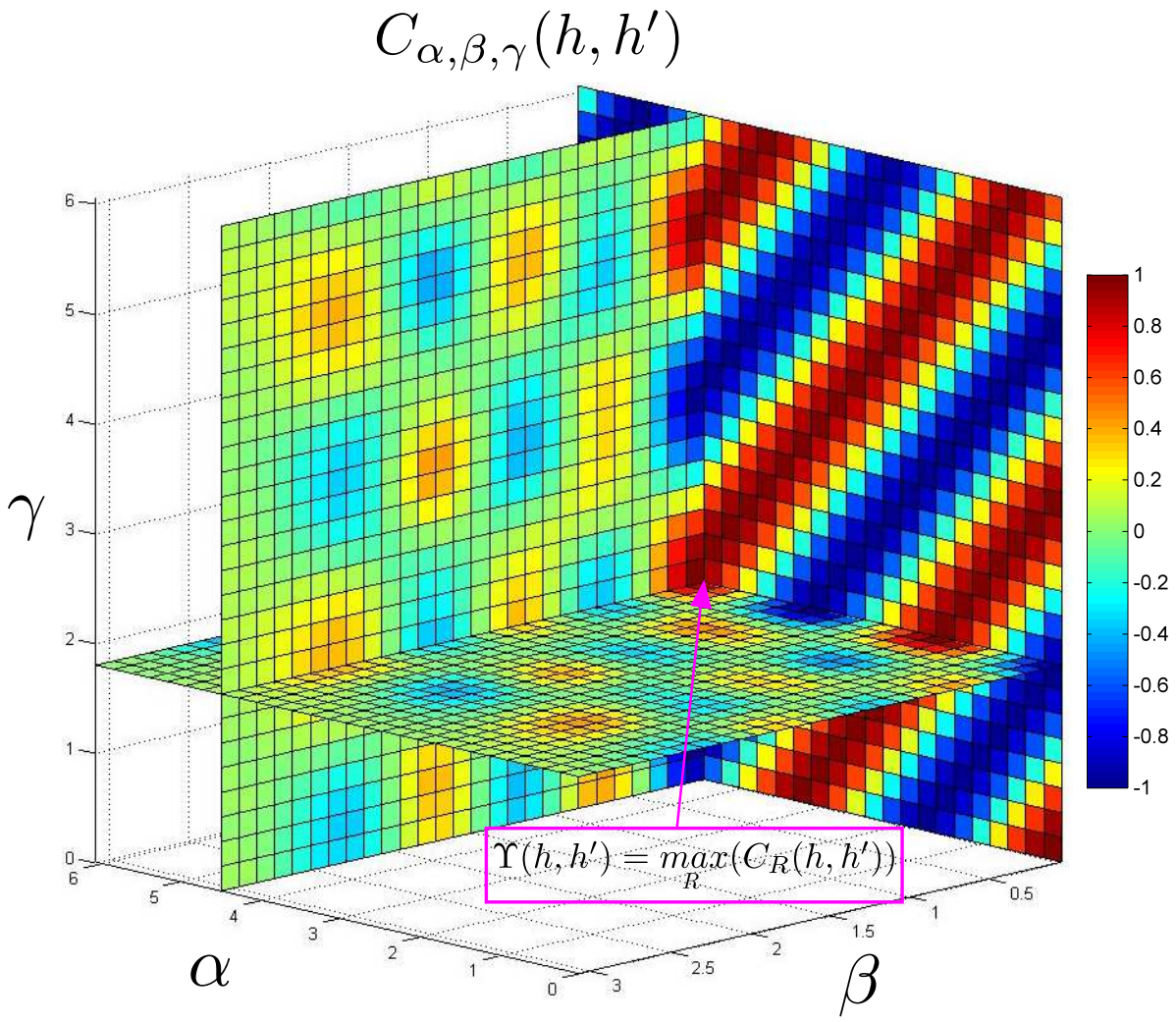


FIG. 3.90 – Illustration du calcul de l'intercorrélation normalisée entre 2 SFRS pour un échantillonnage de toutes les rotations de $SO(3)$. Les rotations sont décrites par le triplet (α, β, γ) des angles d'Euler. $C_{\alpha,\beta,\gamma}$ désigne l'intercorrélation normalisée et Γ son maximum sur l'ensemble des rotations, c'est à dire la mesure de similarité entre les 2 SFRS h et h' (d'après [Guillon, 2009]).

Le second requis de ce modèle est de disposer d'une base de données donnant une information suffisamment détaillée à la fois en termes de variations spatiales, fréquencielles et individuelles. En d'autres termes, il faut collecter les HRTF du plus grand nombre d'individus pour un échantillonnage fin de l'espace et du temps. La base de données considérées pour la mise en œuvre du modèle regroupe les sujets extraits de 4 bases (*Jean-Marie Pernaux*, IRCAM, CIPIC, Université du Maryland) pour constituer une base de 101 sujets avec 1602 directions[37] et 37 bins fréquenciels régulièrement espacés sur la bande [4 - 13 kHz], pour chaque sujet [Guillon, 2009].

La première étape consiste à analyser la base de données de SFRS ainsi collectées afin de vérifier l'hypothèse fondatrice du modèle, à savoir l'existence de prototypes. Cette analyse se fonde sur une

[37] Afin de disposer d'un échantillonnage spatial commun quelle que soit la base, une interpolation STPS est effectuée afin de rééchantillonner les HRTF sur la même grille de directions [Guillon, 2009]. Ce pré-traitement est transparent dans la mesure où pour toutes les bases utilisées, la grille des directions mesurées garantit une reconstruction parfaite par interpolation STPS [Carlile et al., 2000].

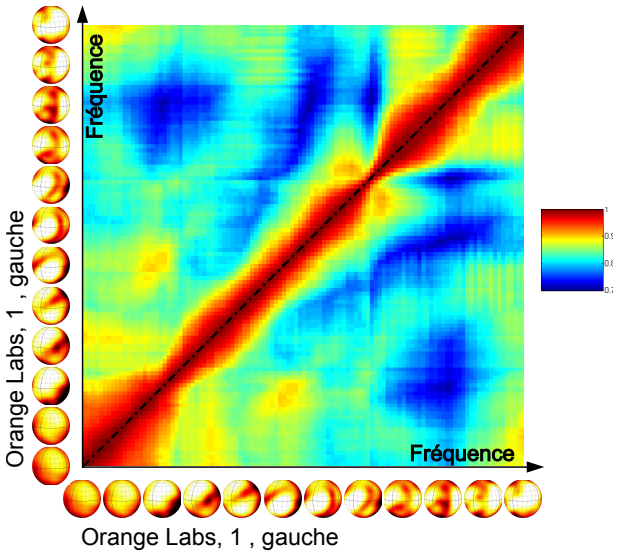


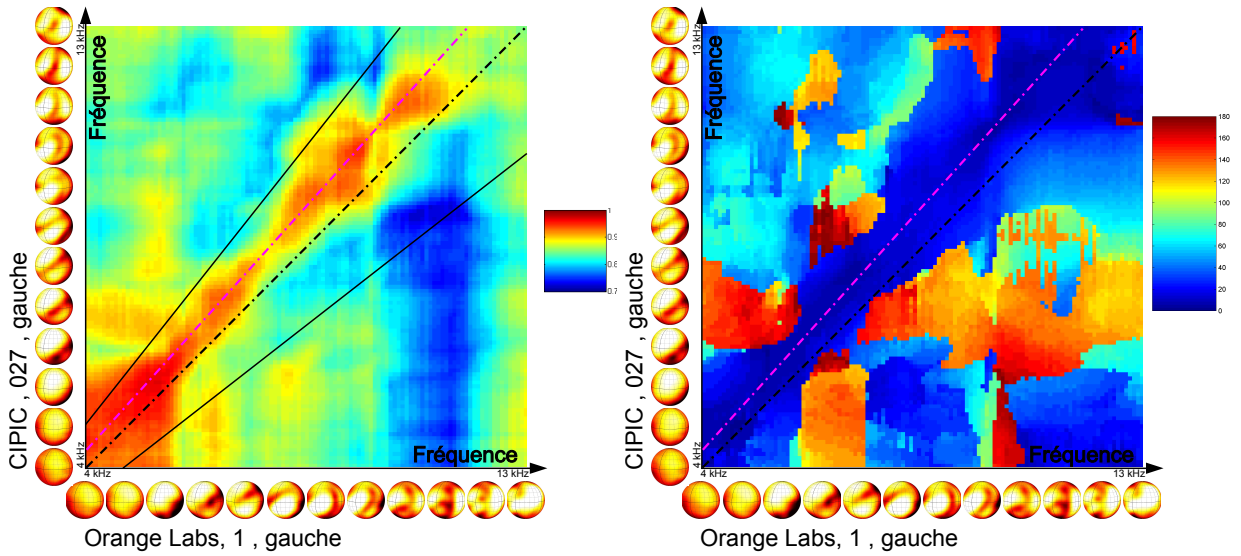
FIG. 3.91 – Mesure de similarité Γ entre les SFRS du même sujet (sujet ME de la base *Jean-Marie Pernaux*, oreille gauche, 37 fréquences, d'après [Guillon, 2009]). On vérifie que la similarité présente sa valeur maximale égale à 1 sur la diagonale, correspondant effectivement à l'autocorrélation entre les SFRS du même bin fréquenciel.

comparaison deux à deux des SFRS, au moyen de la mesure de distance définie par l'intercorrélation normalisée. Chaque sujet fournit 37 SFRS. Les SFRS des oreilles gauches sont symétrisées par rapport au plan médian afin de les rendre comparable aux SFRS des oreilles droites. Dans une première analyse, chaque SFRS est comparée à l'ensemble de toutes les SFRS collectées, incluant celles du même sujet, y compris elle-même, auquel cas on obtient l'auto-corrélation normalisée. Pour chaque paire de SFRS, l'intercorrélation normalisée est calculée pour un échantillonnage de toutes les rotations possibles entre les 2 SFRS (cf. Fig. 3.90) [Guillon, 2009]. Ce calcul de similarité est ainsi capable de mettre en évidence des similarités, abstraction faite d'éventuels décalages sur l'axe des fréquences et/ou de rotations spatiales. Des exemples d'intercorrélation normalisée sont donnés sur les figures 3.91, 3.92, 3.93 et 3.94.

A l'issue de cette première analyse, il apparaît judicieux de restreindre le champ des comparaisons [Guillon, 2009] :

- Dans certains cas, une similarité élevée est obtenue, mais au prix d'une rotation d'un angle très important (parfois proche de 180°), ce qui n'est pas valide d'un point de vue physique. Les rotations supérieures à 40° sont donc éliminées des comparaisons évaluées.
- Au vu des premiers résultats, il ne semble pas nécessaire de comparer les SFRS d'une fréquence donnée aux SFRS de toutes les fréquences, car en général la similarité maximale est obtenue pour une fréquence voisine, les maxima restant en effet proches de la diagonale. La comparaison est ainsi limitée à une bande fréquence $[f - \frac{f_0(f)}{2}; f + \frac{f_0(f)}{2}]$ dont la largeur $f_0(f)$ croît avec la fréquence (cf. Fig. 3.92a).

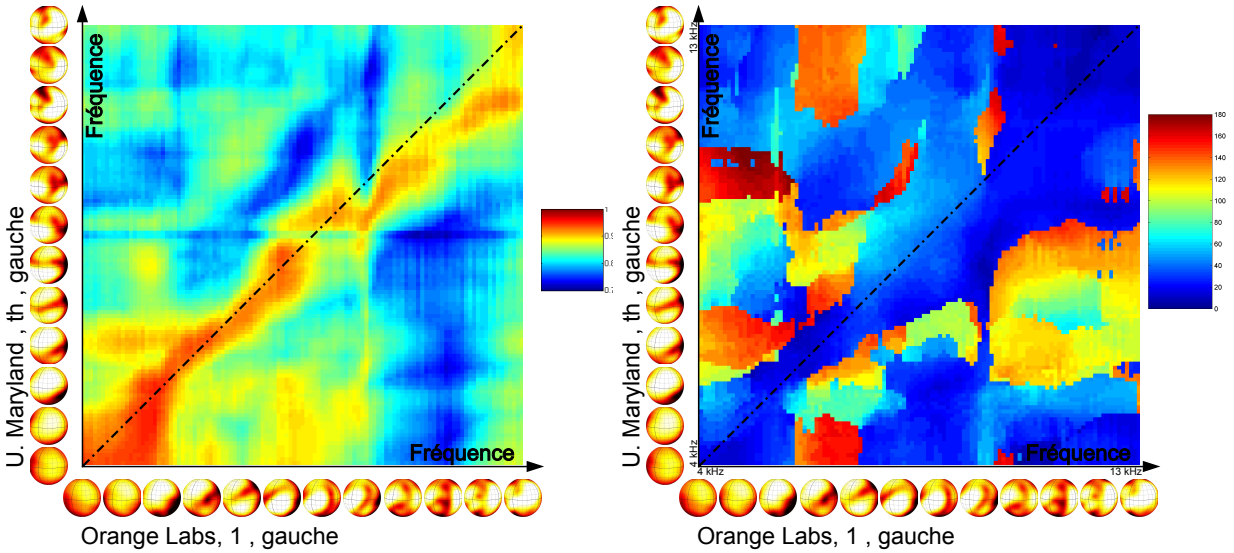
Moyennant ces restrictions, la distance relative entre toutes les SFRS (sujets et fréquences confondus) de la base de données est connue, ce qui permet de classer les SFRS et de les regrouper par similarité. L'outil de classification est la technique de classification spectrale normalisée (*normalized spectral graph clustering*) basée sur une représentation des données sous la forme d'un graphe [Guillon, 2009]. On obtient 300 groupes ou *clusters* de SFRS. Chaque *cluster* définit un prototype de SFRS, c'est à dire une réalisation typique particulière de SFRS. Au sein d'un *cluster*, on trouve



(a) Similarité Γ

(b) Angle global Θ de la rotation associée à la similarité représentée ci-contre

FIG. 3.92 – Mesure de similarité Γ entre les SFRS de 2 sujets (sujet ME de la base *Jean-Marie Pernaux* et sujet 027 de la base CIPIC, oreille gauche, 37 fréquences, d'après [Guillon, 2009]). La ligne en pointillés magenta représente le "rift" de similarité maximale. Le décalage qu'elle présente avec la diagonale (ligne en pointillés noirs) décrit le décalage fréquenciel correspondant à l'homothétie sur l'axe des fréquences. Sur la Figure de droite, on observe que l'angle de la rotation associée est quasiment nul, ce qui signifie que les 2 sujets présentent des SFRS similaires sans appliquer de rotation. Ces résultats sont confirmés par l'observation visuelle des SFRS [Guillon, 2009].



(a) Similarité Γ

(b) Angle global Θ de la rotation associée à la similarité représenté ci-contre

FIG. 3.93 – Mesure de similarité Γ entre les SFRS de 2 sujets (sujet ME de la base *Jean-Marie Pernaux* et sujet TH de la base de l'Université du Maryland, oreille gauche, 37 fréquences, d'après [Guillon, 2009]). Contrairement à la Figure 3.92, les SFRS des 2 sujets ne présentent pas de décalage fréquenciel, mais cette fois l'angle de rotation associée à la similarité maximale n'est pas nul. Ces résultats sont aussi confirmés par l'observation visuelle des SFRS [Guillon, 2009].

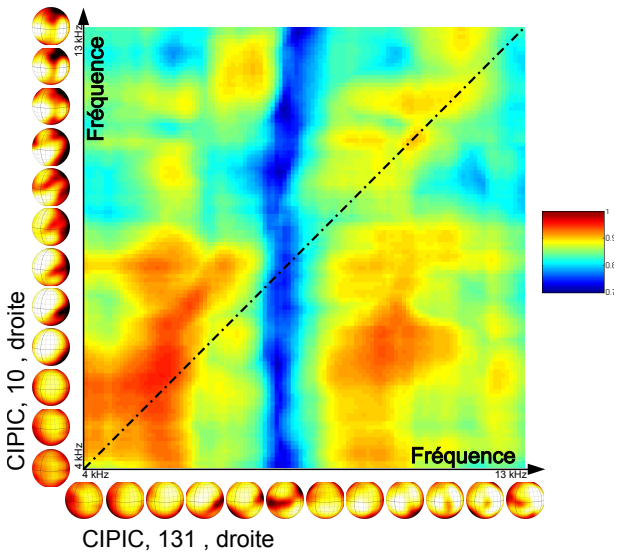


FIG. 3.94 – Mesure de similarité Γ entre les SFRS de 2 sujets (sujets 131 et 10 de la base CIPIC, oreille droite, 37 fréquences, d'après [Guillon, 2009]). Pour ces 2 sujets, la similarité ne présente pas de maximum nettement défini selon un rift caractéristique comme sur les figures 3.92 et 3.92. Les différences entre les 2 sujets ne peuvent donc être réduites par homothétie fréquencielle et rotation spatiale.

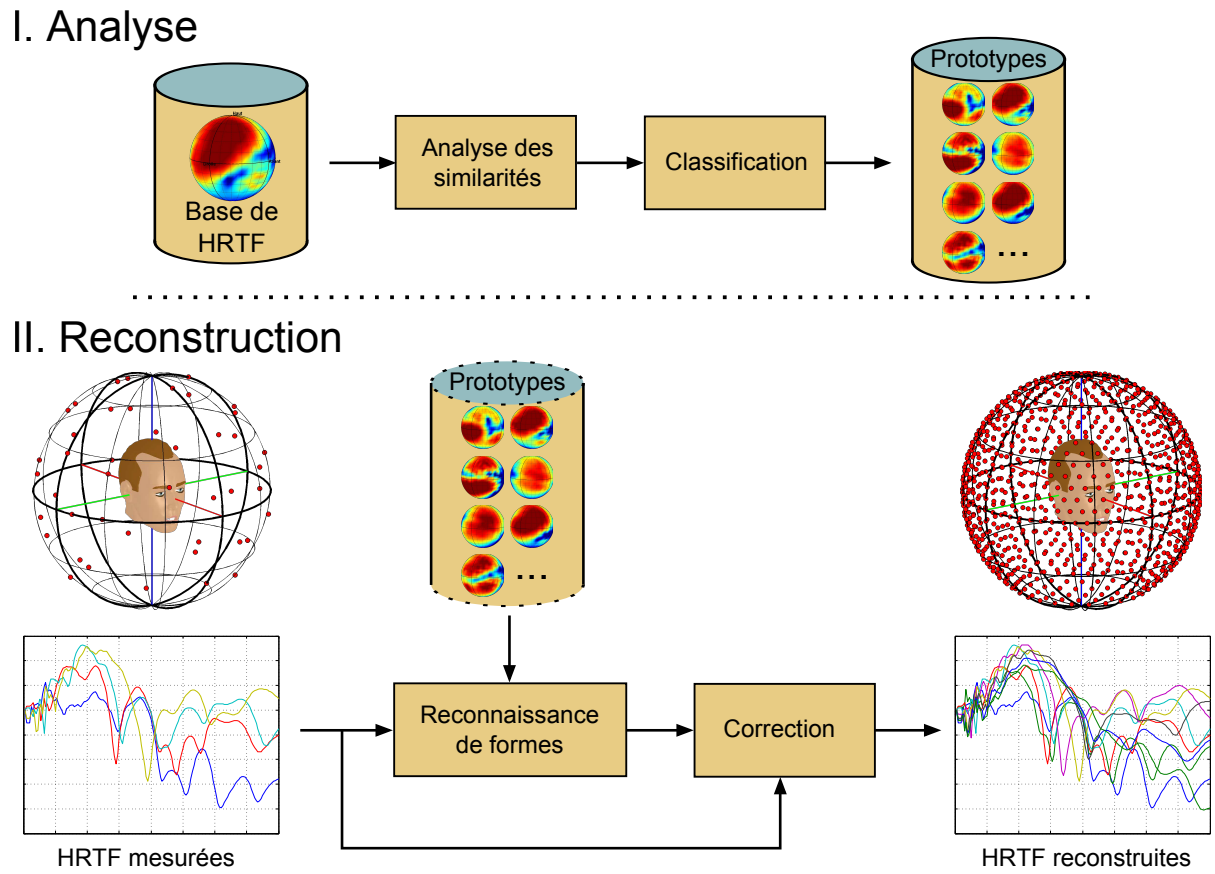


FIG. 3.95 – Schéma de principe du modèle de calcul de HRTF individuelles par reconstruction (d'après [Guillon, 2009]).

des SFRS qui reproduisent ce prototype moyennant une éventuelle rotation. La SFRS représentative d'un cluster, c'est à dire le prototype associé, s'obtient comme la moyenne des SFRS regroupées dans le cluster. L'ensemble de ces *clusters* constitue une sorte de "catalogue" dans lequel le modèle va venir piocher pour construire les HRTF d'un nouvel individu. Ce catalogue est représentatif de la diversité individiduelle et fréquencielle de la base de données appliquée en entrée de l'analyse de similarité. Les clusters contiennent l'information a priori qu'on injecte dans le modèle de calcul de HRTF individuelles.

Description et mise en œuvre du modèle

La mise en œuvre du modèle de calcul de HRTF individuelles comprend les étapes suivantes (cf. Fig. 3.95) [Guillon, 2009] [Guillon & Nicol,] [Guillon & Nicol, 2008] :

- mesure des HRTF du nouvel individu pour un échantillonnage grossier de la sphère 3D,
- comparaison des SFRS issues des HRTF mesurées avec les prototypes et identification du prototype le plus similaire par un processus de reconnaissance de forme,
- première approximation des SFRS du nouvel individu à partir des prototypes sélectionnés,
- correction de cette première modélisation en retranchant l'erreur entre les données modélisées et les données mesurées, l'erreur étant interpolée par STPS sur la sphère 3D,
- au final les HRTF sont obtenues par concaténation fréquencielle des SFRS ainsi corrigées.

Il s'agit d'un modèle de HRTF individuelles de type 2 (cf. Section 3.5.2).

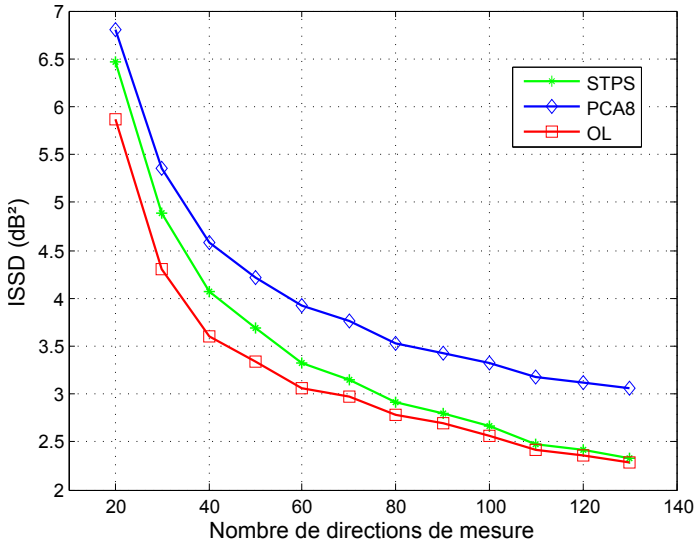


FIG. 3.96 – ISSD en fonction du nombre de directions de mesure des HRTF individuelles d'entrée (d'après [Guillon, 2009]) : comparaison du modèle proposé (désigné par "OL", courbe en rouge, symbole □) avec 2 modèles tirés de la littérature. Le premier modèle est l'interpolation STPS [Hartung et al., 1999] (désigné par "STPS", courbe en vert, symbole *) et le second est la méthode hybride décrite dans [Carlile et al., 2000] et basée sur l'interpolation STPS des poids d'une décomposition ACP (désignée par "PCA8", courbe en bleu, symbole ◇).

La modélisation est évaluée pour 8 individus[38] qu'on a pris soin d'exclure de la base de données utilisée pour la construction des prototypes. L'influence du nombre de directions de mesure des HRTF individuelles appliquées en entrée du modèle est étudiée. Les performances de la reconstruction des HRTF individuelles sont passées au crible de différents critères [Guillon, 2009] :

- erreur quadratique moyenne
- ISSD
- erreur maximale [Langendijk & Bronkhorst, 2000],
- fidélité de reconstruction des CPA (cf. Fig. 3.97).

La Figure 3.96 illustre l'ISSD en fonction du nombre de HRTF individuelles mesurées. Notre modèle de reconstruction démontre sa capacité à modéliser des HRTF individuelles et se démarque des modèles de l'état de l'art (STPS) avec une réduction sensible de l'ISSD. En termes de reconstruction des CPA, le nouveau modèle est aussi plus fidèle, notamment en termes de localisation de la CPA (cf. Fig. 3.98 et 3.99).

Conclusion

Le modèle proposé se distingue par deux idées essentielles :

- injection d'informations a priori sur les données à modéliser dans le processus de reconstruction,
- analyse préalable des informations injectées afin de séparer les comportements similaires (selon les axes frequenciel et individuel) et les comportements proprement spécifiques, cette séparation étant rendue possible par une structuration particulière des données sous forme de SFRS.

[38] Il s'agit des individus suivants : sujets ME et RN de la base *Jean-Marie Pernaux*, sujets DZ, EG et NM de la base de l'Université du Maryland, sujets 012, 040 et 124 de la base CIPIC.

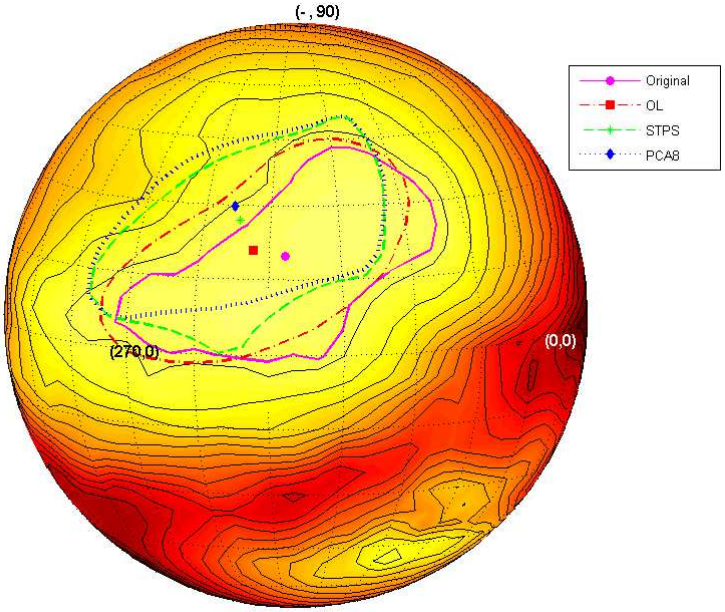


FIG. 3.97 – Illustration de la reconstruction des CPA (d'après [Guillon, 2009]) : comparaison du CPA obtenu sur la SFRS mesurée et les SFRS modélisées (modèles OL, STPS et PCA8). La CPA est définie ici comme la portion de sphère centrée autour du maximum de la SFRS et sur laquelle le spectre d'amplitude de la SFRS reste inférieur au maximum à moins de 1.5 dB. Pour les modèles, 40 HRTF individuelles sont utilisées en entrée. Sujet ME de la base *Jean-Marie Pernaux* (f = 7 kHz).

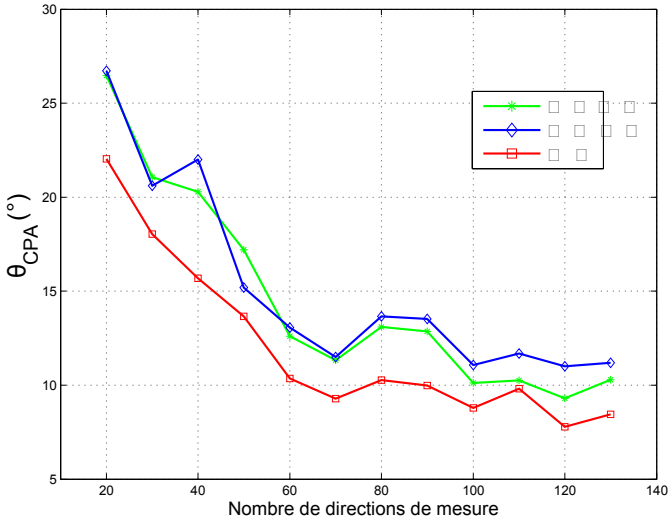


FIG. 3.98 – Reconstruction des CPA (d'après [Guillon, 2009]) : le critère θ_{CPA} représente l'angle entre les centroïdes des SFRS cibles et des SFRS modélisées (modèles OL, STPS et PCA8). Moyenne calculée tous sujets et fréquences confondus.

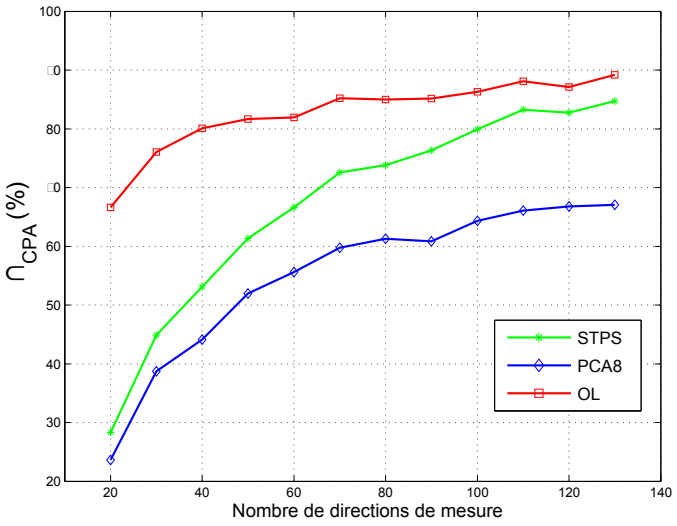


FIG. 3.99 – Reconstruction des CPA (d'après [Guillon, 2009]) : le critère $\cap_{CPA}$ représente le pourcentage de la CPA cible qui est recouvert par la CPA modélisée (modèles OL, STPS et PCA8). Moyenne calculée tous sujets et fréquences confondus.

Une première validation au travers d'un ensemble de critères objectifs montre l'apport du modèle par rapport à des modèles de référence basés sur une interpolation STPS. Il reste à compléter cette validation par une évaluation subjective (test d'écoute) qui est présentée en Section 3.5.7.

3.5.6 Modélisation des IS par adaptation morphologique de HRTF non individuelles

Aller plus loin dans l'exploitation des similarités inter-individuelles

Ce modèle reprend le même point de départ que le modèle de recontruction qui vient d'être présenté : on veut tirer parti des similarités qui existent entre les HRTF de deux individus et qu'on met en évidence lorsqu'on observe les données sous forme de SFRS. Mais, cette fois, cet avantage n'est pas intégré sous la forme d'informations adjacentes qui viennent guider le processus de reconstruction : il intervient au coeur même du processus. L'idée est d'exploiter une autre piste en cherchant à aller plus loin que le modèle de reconstruction dans la prise en compte des similarités inter-individuelles. En effet si l'on approfondit l'observation des similarités inter-individuelles, deux points sont à relever :

- On observe qu'entre deux individus on peut retrouver des SFRS similaires, mais à des **fréquences** distinctes. Plus précisément tout se passe comme si l'évolution des SFRS en fonction de la fréquence subissait une **homothétie selon l'axe fréquenciel** (cf. Fig. 3.100 & 3.101). Il semble alors possible de prédire les SFRS d'un individu à partir de celles d'un autre individu par une simple homothétie permettant une dilatation ou une contraction des variations fréquencielles et dont le rapport est à ajuster en fonction du nouvel individu.
- De même, si d'emblée pour certains individus leurs SFRS ne semblent pas similaires, elles le deviennent à condition de leur appliquer une **rotation des coordonnées d'espace** (cf. Fig. 3.102 & 3.103). Il semble donc possible de déduire les HRTF d'un individu à partir de celles d'un autre auditeur par une rotation des SFRS, l'angle de rotation dépendant de l'individu.

Il en résulte un modèle de calcul de HRTF individuelles consistant à adapter des HRTF d'un individu à un nouvel individu en combinant deux transformations : homothétie sur l'axe des fréquences

et rotation des coordonnées d'espace [Guillon, 2009]. C'est en cela que l'analyse des similarités inter-individuelles ci-dessus définit le principe intrinsèque du modèle. Ce modèle a été proposé initialement pour l'individualisation de HRTF de gerbilles de Mongolie [Maki & Furukawa, 2005]. Il s'agit ici de l'appliquer à des êtres humains. Il reste à déterminer le rapport d'homothétie et l'angle de rotation adaptés à l'individu considéré. D'un point de vue physique le problème est déterminé par la morphologie de l'auditeur, et plus précisément, comme il s'agit des IS, par la morphologie de ses pavillons d'oreille [Algazi et al., 2001a]. L'origine physique du rapport d'homothétie et de l'angle de rotation est donc à chercher dans des différences morphologiques entre les pavillons. L'hypothèse retenue et qui semble la plus plausible est que l'homothétie sur l'axe des fréquences serait liée à des différences de **taille** du pavillon [Middlebrooks, 1999a], tandis que la rotation proviendrait des différences d'**orientation** [Maki & Furukawa, 2005]. Les paramètres de transfomation des HRTF sont donc à déterminer à partir d'une comparaison des morphologies du pavillon, en termes de taille et d'orientation [Guillon, 2009].

L'adaptation de HRTF sur laquelle se fonde le processus de modélisation peut s'interprêter d'une autre façon : comme une *réduction de distance* entre les HRTF de deux individus. Cependant il faut bien avoir conscience que cette réduction des différences inter-individuelles n'est capable de compenser que les différences associées aux différences morphologiques correspondant à la taille et à l'orientation du pavillon. Il reste une part irréductible de différences correspondant à la spécificité morphologique structurelle du pavillon. Ce type de différences ne peut être adapté par le processus de transformation impliquant homothétie et rotation. Néanmoins il n'est pas pour autant éludé par le modèle. On peut penser en effet que la dimension des différences morphologiques structurelles du pavillon est portée par la variété des individus dont les HRTF constituent la base de données où le modèle vient piocher les HRTF à transformer pour calculer les HRTF d'un nouvel auditeur. La question sous-jacente porte sur le choix des HRTF d'entrée du processus d'adaptation : existe-t-il un choix optimal ? En d'autres termes existe-t-il un choix particulier qui permet de minimiser l'erreur de modélisation ? Un point qui mériterait aussi d'être creusé, mais qui ne l'a pas été dans le cadre de cette étude, est le choix des HRTF constituant le "catalogue" des HRTF disponibles comme point de départ de l'adaptation. Une piste serait de partir d'une large base de données d'individus avec une étape de réduction[39] visant à éliminer les redondances liées aux différences de taille et d'orientation du pavillon, afin de ne retenir que les individus présentant des différences morphologiques de type structurel. L'avantage du modèle repose donc sur un catalogue de taille plus compact que dans le cas d'une adaptation de HRTF non individuelles en aveugle, puisque ce catalogue n'est représentatif que des différences morphologiques structurelles.

Description du modèle

Le modèle d'adaptation de HRTF non individuelles comprend les étapes suivantes (cf. Fig. 3.104) [Guillon, 2009] [Guillon et al., 2008] :

- Acquisition de la morphologie des pavillons du nouvel individu (A) par un scan laser (cf. Fig. 3.105) : Dans une première étape, la surface de l'ensemble de la tête est acquise afin de pouvoir évaluer l'orientation du pavillon dans le référentiel de la tête. Le détail du pavillon est obtenu en scannant son moulage, du fait que les surfaces concaves qu'il comporte (notamment la conque) sont difficiles à détecter par le scan lorqu'elles sont en creux, alors que cette détection est facilitée sur le moulage car elles apparaissent en relief.

[39] On pourrait appliquer par exemple une opération de classification comme dans le modèle de reconstruction, avec la propriété de détecter les similarités en dépit d'éventuels décalages fréquenciels et/ou rotations. Cependant, cette fois la classification ne traiterait pas les SFRS indépendamment de l'individu et de la fréquence, mais s'appliquerait à la globalité des HRTF d'un individu.

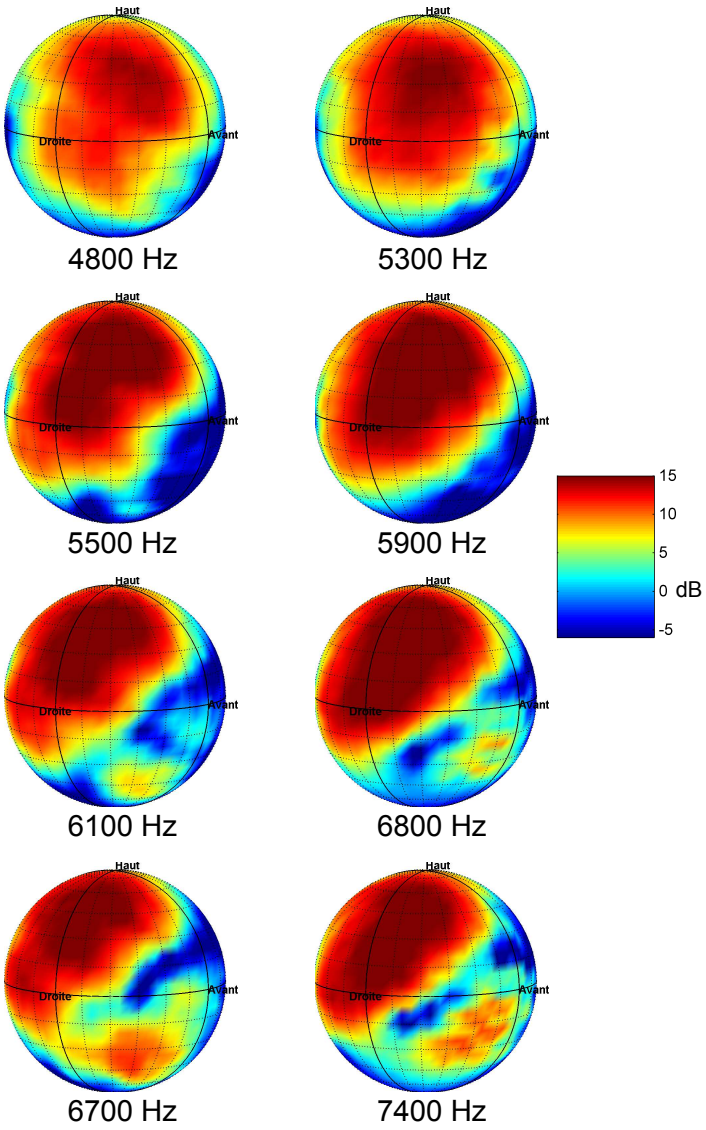


FIG. 3.100 – Comparaison des SFRS de deux individus pour une fréquence croissante (d'après [Guillon, 2009]). A gauche : sujet ME de la base *Jean-Marie Pernaux* (oreille gauche). A droite : sujet 027 de la base CIPIC (oreille gauche). On observe des similarités entre les deux individus, mais avec un décalage fréquenciel.

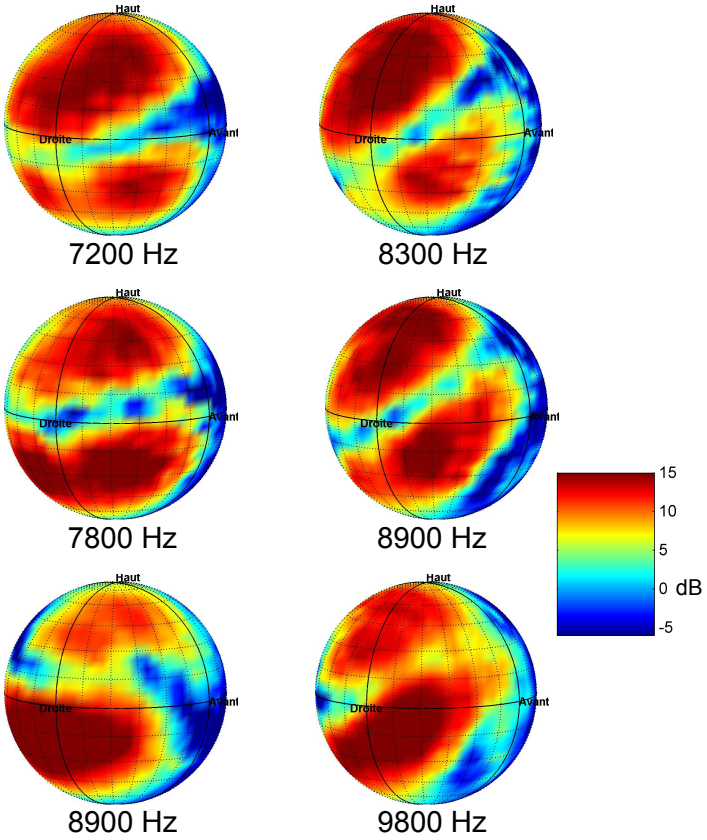


FIG. 3.101 – Suite de la Figure 3.100 pour des fréquences supérieures (d'après [Guillon, 2009]).

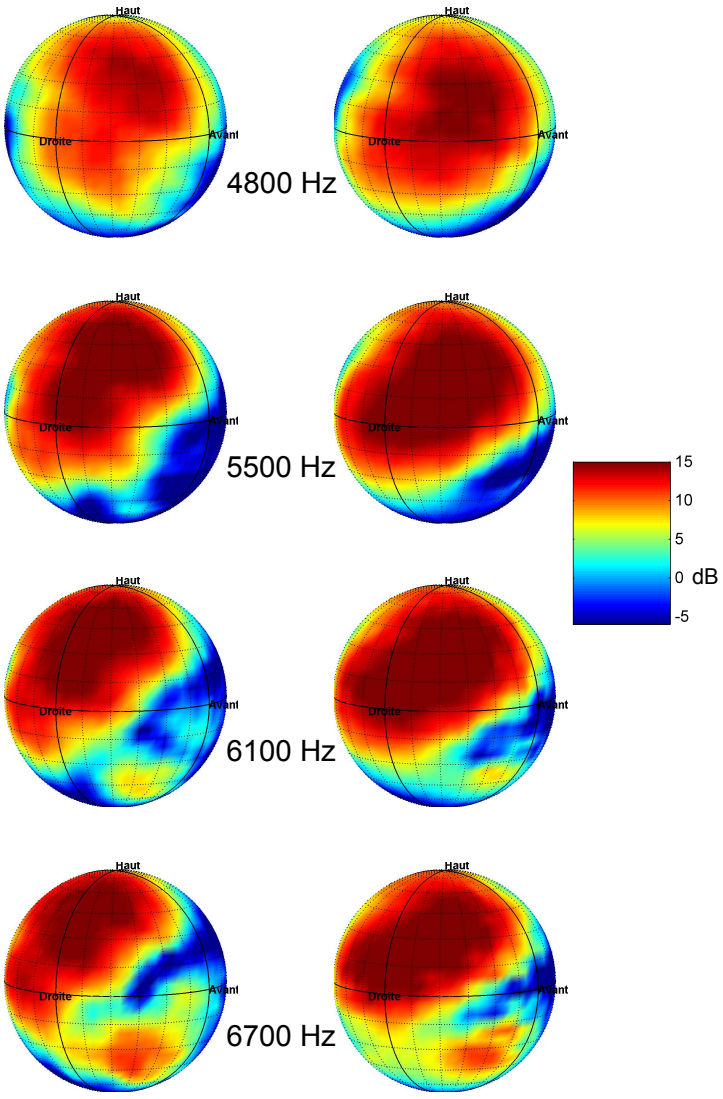


FIG. 3.102 – Comparaison des SFRS de deux individus pour une fréquence croissante (d'après [Guillon, 2009]). A gauche : sujet ME de la base *Jean-Marie Pernaux* (oreille gauche). A droite : sujet TH de la base de l'Université du Maryland (oreille gauche). On observe des similarités entre les deux individus, mais moyennant une rotation du système de coordonnées spatiales.

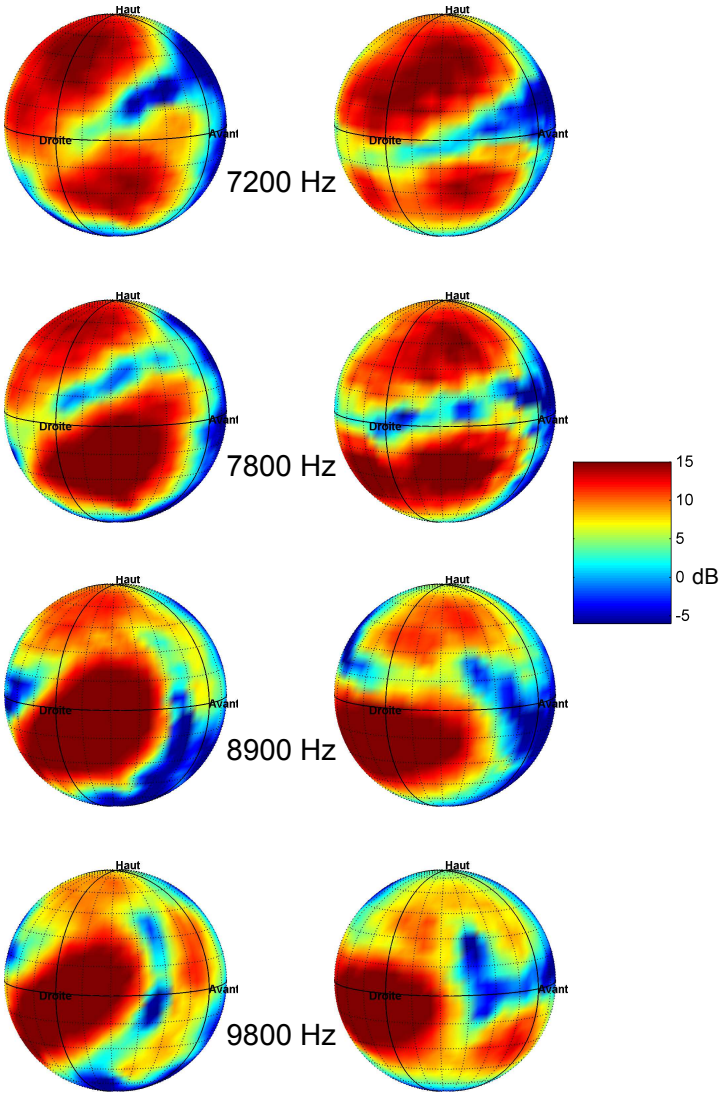


FIG. 3.103 – Suite de la Figure 3.102 pour des fréquences supérieures (d'après [Guillon, 2009]).

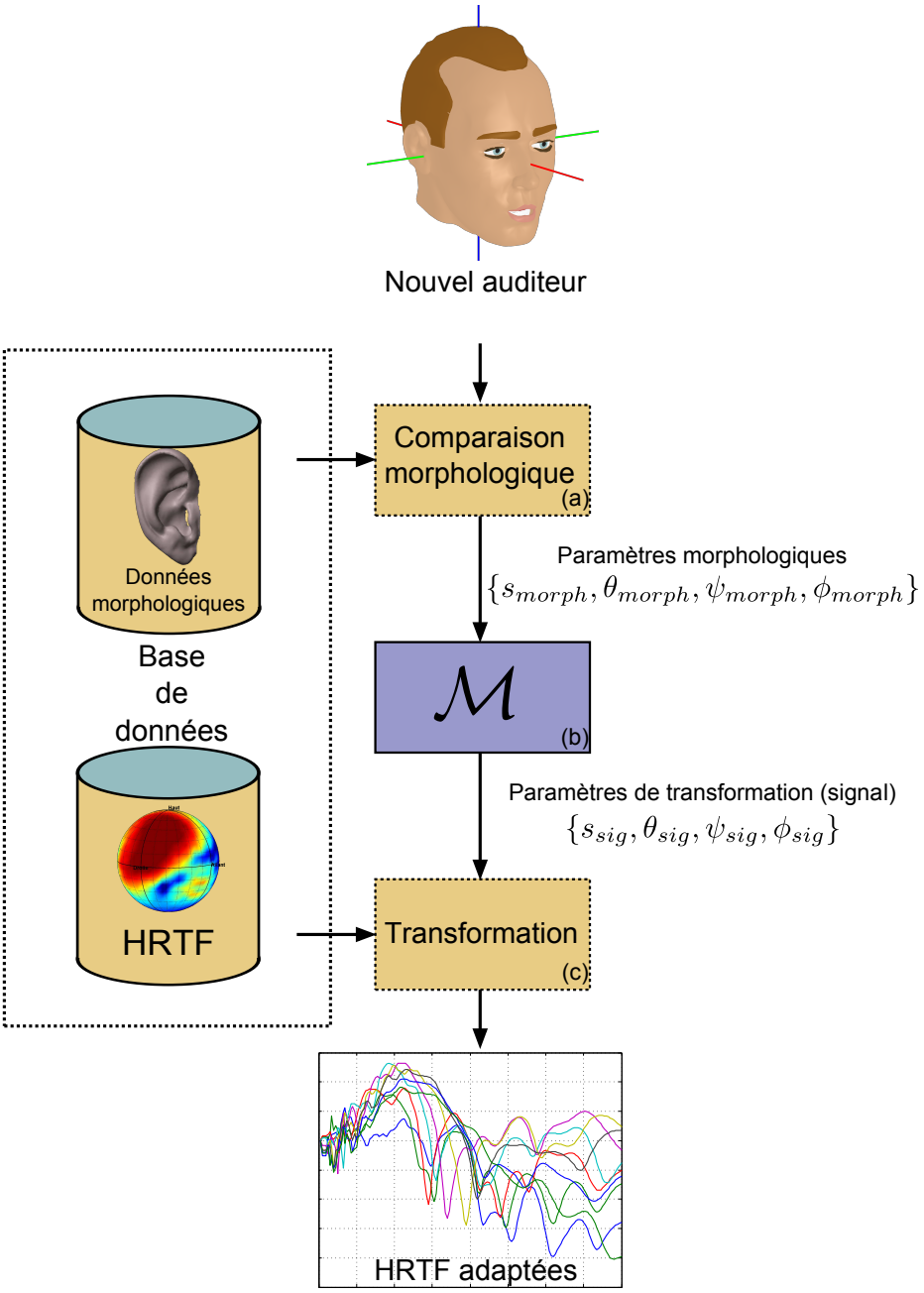


FIG. 3.104 – Schéma de principe du modèle d'adaptation de HRTF non individuelles (d'après [Guillon, 2009]).

- Choix des HRTF non individuelles dans la base de données (HRTF d'un individu B) : Dans cette première mise en œuvre du modèle, ce choix est arbitraire, mais la stratégie de ce choix est une question restant à étudier.
- Comparaison des morphologies du nouvel individu (A) et de l'individu (B) qui est le propriétaire des HRTF qu'on veut adapter au nouvel auditeur : Une technique d'alignement basée sur l'algorithme ICP (*Iterative Closest Point*), qui a déjà été appliquée avec succès pour aligner des surfaces de pavillon [Darkner et al., 2006], est utilisée. Elle consiste à aligner les deux surfaces (définies sous la forme de nuages de points) des pavillons en déterminant la transformation géométrique globale réduisant la distance totale entre les deux nuages de point (cf. Fig. 3.106). Dans la forme classique de l'ICP, les transformations se limitent à la combinaison de rotations et translations. Dans notre mise en œuvre, l'homothétie a donc dû être ajoutée comme degré de liberté supplémentaire afin de rendre compte des différences de taille. Au final, on en retire deux paramètres : le **rapport d'homothétie** et la **rotation** qui décrivent respectivement les différences de taille et d'orientation des deux pavillons.
- Calcul des paramètres de transformation des HRTF à partir des paramètres d'alignement des deux pavillons (A et B) : Les relations permettant d'exprimer les paramètres de transformation (rapport d'homothétie et rotation) des HRTF non individuelles en fonction des paramètres d'alignement des morphologies des deux pavillons (rapport d'homothétie et rotation également), constituent le cœur du modèle. Ces relations sont obtenues par régression linéaire à partir de l'étude des corrélations entre les paramètres d'alignement (pavillon) et les paramètres de transformation (HRTF) dans le cas où les HRTF individuelles sont connues et les paramètres de transformation peuvent être calculés indépendamment de toute comparaison morphologique.
- Transformation des HRTF non individuelles (B) par homothétie et rotation combinées : Les paramètres de transformation précédemment identifiés sont appliqués pour adapter les HRTF non individuelles au nouvel auditeur (A).

Prédiction des paramètres de transformation à partir de la comparaison morphologique

Les relations pour calculer les paramètres de transformation des HRTF à partir des paramètres d'alignement des pavillons sont mises au point par une étude préalable [Guillon, 2009] dans laquelle on considère une base de données de 6 individus (sujets ME, JD, RN, PA, MA, VM de la base *Jean-Marie Pernaux*) constituée d'une part des scans 3D de leurs pavillons et d'autre part de leurs HRTF individuelles. Les paramètres d'alignement des pavillons sont obtenus par la procédure d'alignement décrite précédemment. Pour chaque paire de pavillons, on dispose de 2 paramètres (rapport d'homothétie s_{morph} et rotation) qui représentent leur distance morphologique. Indépendamment, les HRTF associées sont comparées en recherchant, de façon similaire, la transformation optimale permettant d'adapter les HRTF du premier individu à celles du second. L'optimisation vise à minimiser l'ISSD entre les deux jeux de HRTF. Il en ressort deux paramètres de transformation (rapport d'homothétie s_{sig} et rotation) qui décrivent la distance entre les HRTF des deux individus. Dans la suite, la rotation est exprimée sous la forme d'une combinaison de 3 rotations autour des axes X, Y et Z, associées à 3 angles de rotation définis respectivement comme *roll* θ, *pitch* ψ et *yaw* φ.

On observe alors les corrélations (cf. Fig. 3.107) entre les paramètres d'alignement des pavillons (s_{morph}, θ_{morph}, ψ_{morph}, ϕ_{morph}) et les paramètres de transformation des HRTF (s_{sig}, θ_{sig}, ψ_{sig}, ϕ_{sig}). Conformément à ce qu'on pouvait attendre, pour la majorité des paramètres, les paramètres de transformation atteignent la corrélation maximale avec leurs équivalents dans les paramètres d'alignement :

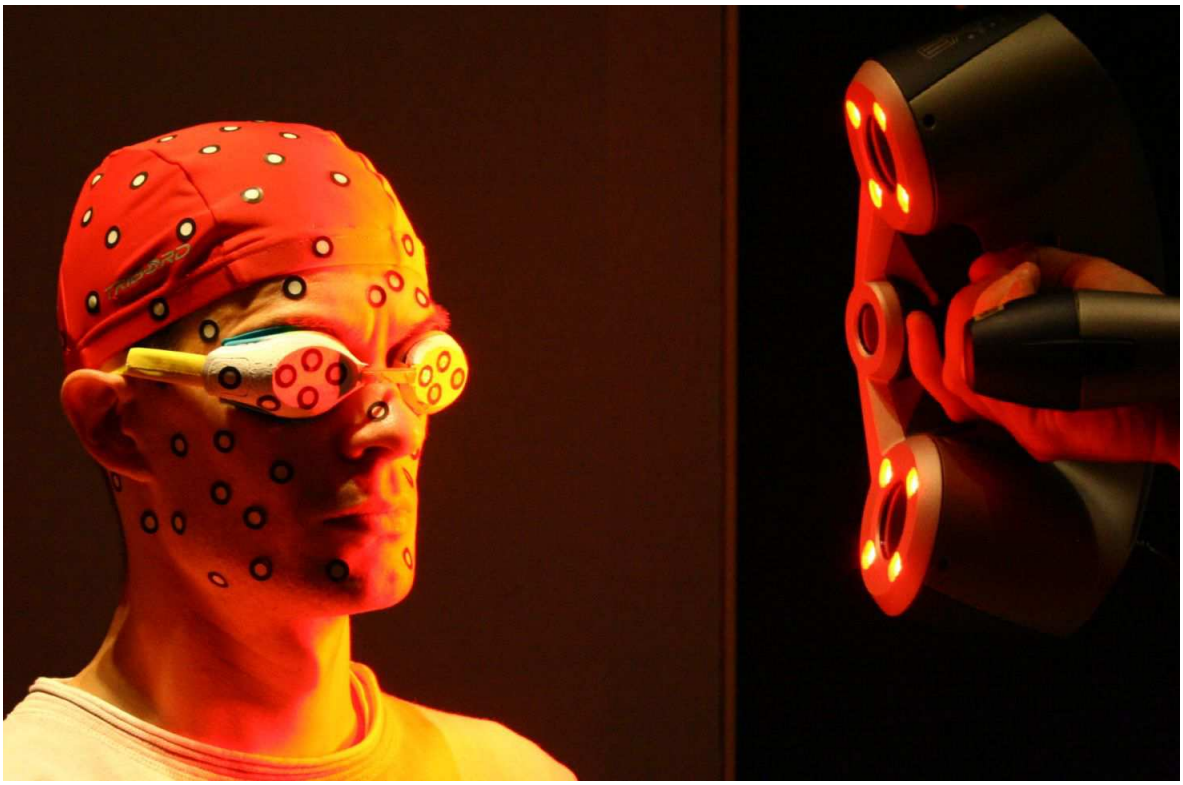


FIG. 3.105 – Acquisition en 3D de la surface de la tête d'un sujet grâce à un scanner laser (d'après [Guillon, 2009]). L'opérateur du scanner le tient dans sa main et tourne autour du sujet.

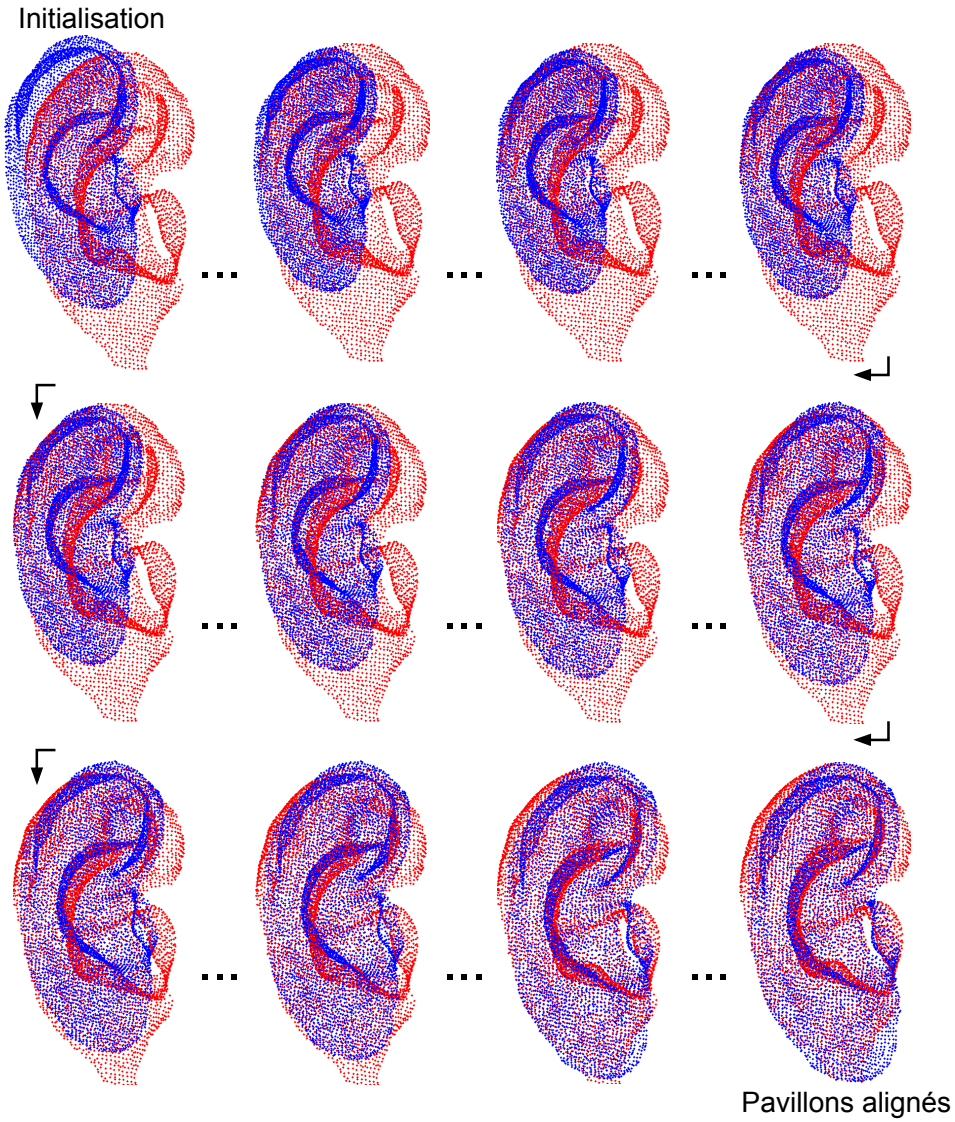


FIG. 3.106 – Illustration du processus d'alignement de deux pavillons par l'algorithme ICP (d'après [Guillon, 2009]). Le pavillon en bleu est transformé pas à pas pour coïncider avec le pavillon rouge qui reste inchangé.

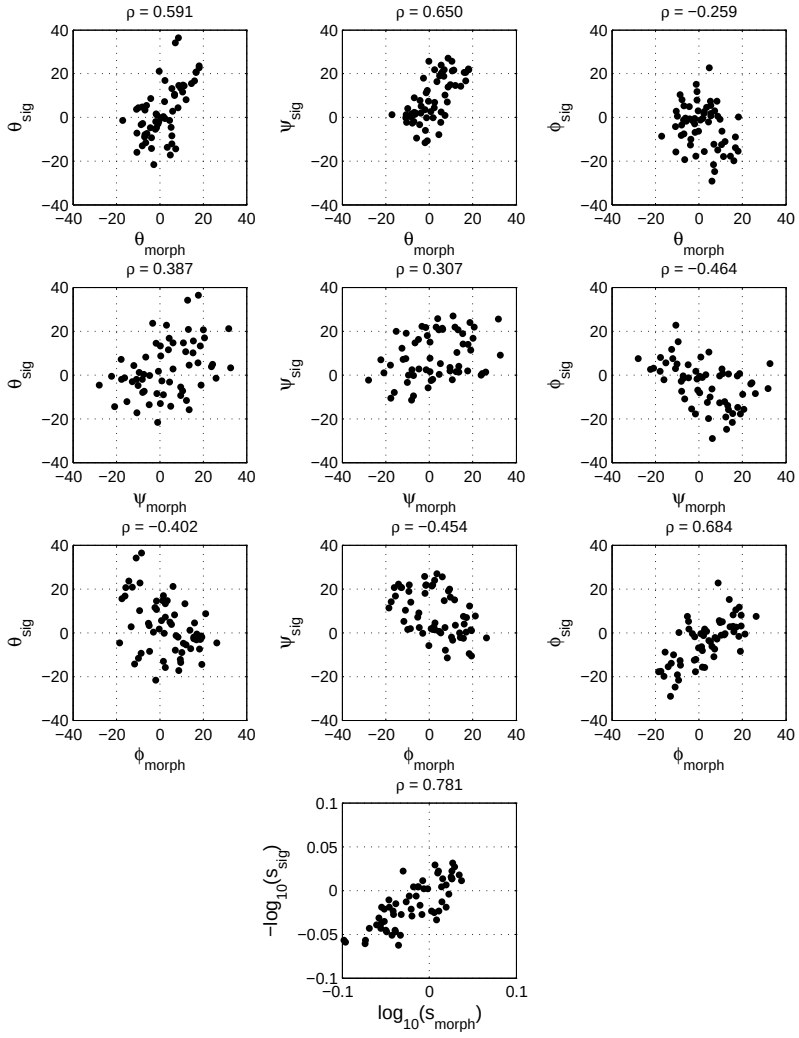


FIG. 3.107 – Corrélations entre les paramètres d'alignement des pavillons et les paramètres de transformations des HRTF (Base *Jean-Marie Pernaux* : sujets ME, JD, RN, PA, MA, VM ; 60 paires ont été considérées au total ; d'après [Guillon, 2009]).

– rapport d'homothétie : s_{sig} est le mieux corrélé à s_{morph},
– angle *roll* : θ_{sig} est le mieux corrélé à θ_{morph},
– angle *yaw* : φ_{sig} est le mieux corrélé à φ_{morph}.

En revanche, pour l'angle *pitch* ψ_{sig}, il s'avère le mieux corrélé avec θ_{morph} et non ψ_{morph}. Ce résultat peut s'expliquer par la difficulté de définir une référence stable pour cet angle et notamment pour faire coïncider les référentiels de mesure des HRTF et des pavillons, étant donné que les deux campagnes de mesure se sont déroulées à quelques années de distance [Guillon, 2009]. Pour ces raisons, on peut craindre une forte incertitude sur les estimations de ψ_{morph} et ψ_{sig}, ce qui pollue toute recherche de corrélation. On note par ailleurs que les valeurs de corrélation maximale ne sont pas très élevées (cf. Fig. 3.107). Une raison possible est le faible nombre de données utilisées pour l'étude (6 sujets). Il conviendrait donc d'étoffer la base de données.

A présent nous allons établir les relations pour prédire les paramètres de transformation à partir des paramètres d'alignement. Des résultats qui précèdent, on conclut qu'on va chercher à exprimer :

– s_{sig} en fonction de s_{morph},
– θ_{sig} en fonction de θ_{morph},
– ψ_{sig} en fonction de θ_{morph},
– φ_{sig} en fonction de φ_{morph}.

Deux modèles ont été testés : modèle linéaire (par exemple : $s_{sig} = \alpha s_{morph}$) et modèle affine (par exemple : $s_{sig} = \alpha s_{morph} + \beta$). Le modèle affine serait a priori le plus approprié, puisqu'on s'attend à ce qu'en l'absence de différence morphologique, les HRTF n'aient à subir aucune transformation. Les résultats des régressions linéaire et affine pour les 4 paramètres sont illustrés sur la Figure 3.108. Contrairement à nos attentes, pour les paramètres ψ_{sig} et φ_{sig}, le modèle affine offre la meilleure prédiction, ce qui signifie la présence d'un biais qui n'a pas d'explication physique. L'hypothèse la plus plausible est que ce biais résulte d'un biais de rotation entre les référentiels de mesure des HRTF et des pavillons, comme on l'a déjà pressenti [Guillon, 2009].

Evaluation du modèle

Les perfomances de modélisation sont évaluées sur la base de l'ISSD. La Figure 3.109 représente l'ISSD après adaptation en fonction de la valeur avant adaptation. La majorité des points se situe en dessous de la diagonale, ce qui signifie que l'adaptation des HRTF permet effectivement de réduire l'ISSD. Quelques points sont localisés au dessus de la diagonale. Il s'agit des points marginaux identifiés comme *outliers* par l'algorithme RANSAC. L'adaptation échoue du fait que le modèle de prédiction des paramètres de transformation n'est pas pertinent dans leurs cas. Pour les autres cas, les performances d'adaptation des HRTF sont proches des performances optimales (correspondant aux paramètres de transformation obtenus par minimisation de l'ISSD), ce qui prouve le succès de la prédiction. Le modèle de prédiction des paramètres de transformation à partir de la morphologie semble bien fonctionner dans la mesure où les performances de modélisation sont proches des performances optimales (calcul des paramètres par minimisation de l'ISSD). On note aussi l'apport de la rotation combinée à l'homothétie : dans la plupart des cas, la réduction de l'ISSD est améliorée par rapport à la transformation par homothétie seule, notamment lorsque l'ISSD initiale est élevée.

Conclusion

Les résultats de cette première étude démontrent, dans le cadre d'une stratégie d'adaptation de HRTF non individuelles, tout l'intérêt, en termes de réduction d'ISSD, de combiner à l'homothétie sur l'axe des fréquences (méthode de *scaling* fréquenciel [Middlebrooks, 1999a]), une tranformation par rotation. On a vu aussi comment exprimer les paramètres de transformation à partir d'une comparaison morphologique des pavillons. Il convient cependant de noter que ces résultats sont

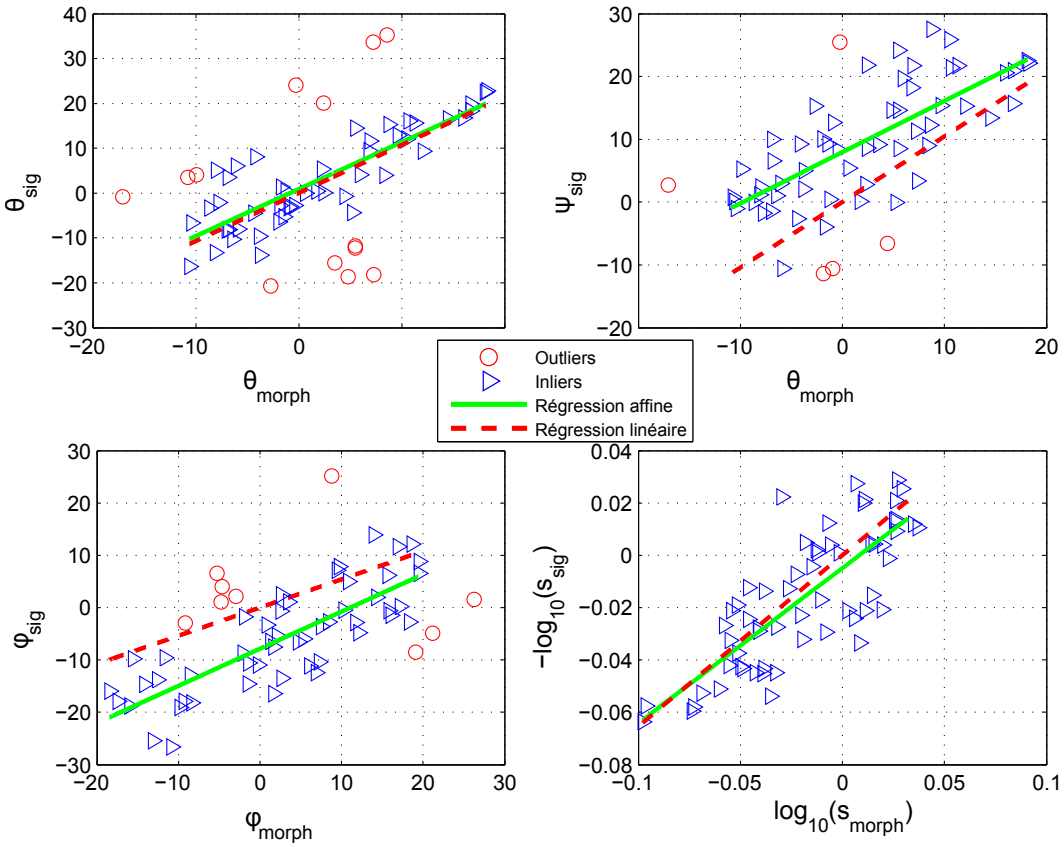


FIG. 3.108 – Régression entre les paramètres d'alignement et les paramètres de transformation (d'après [Guillon, 2009]) : modèle affine et linéaire. En raison des faibles corrélations observées, les données ont été nettoyées par un algorithme RANSAC (*RANdom SAmple Consensus*) visant à éliminer les points marginaux (*outliers*) pour ne conserver que les points conformes au modèle (*inliers*).

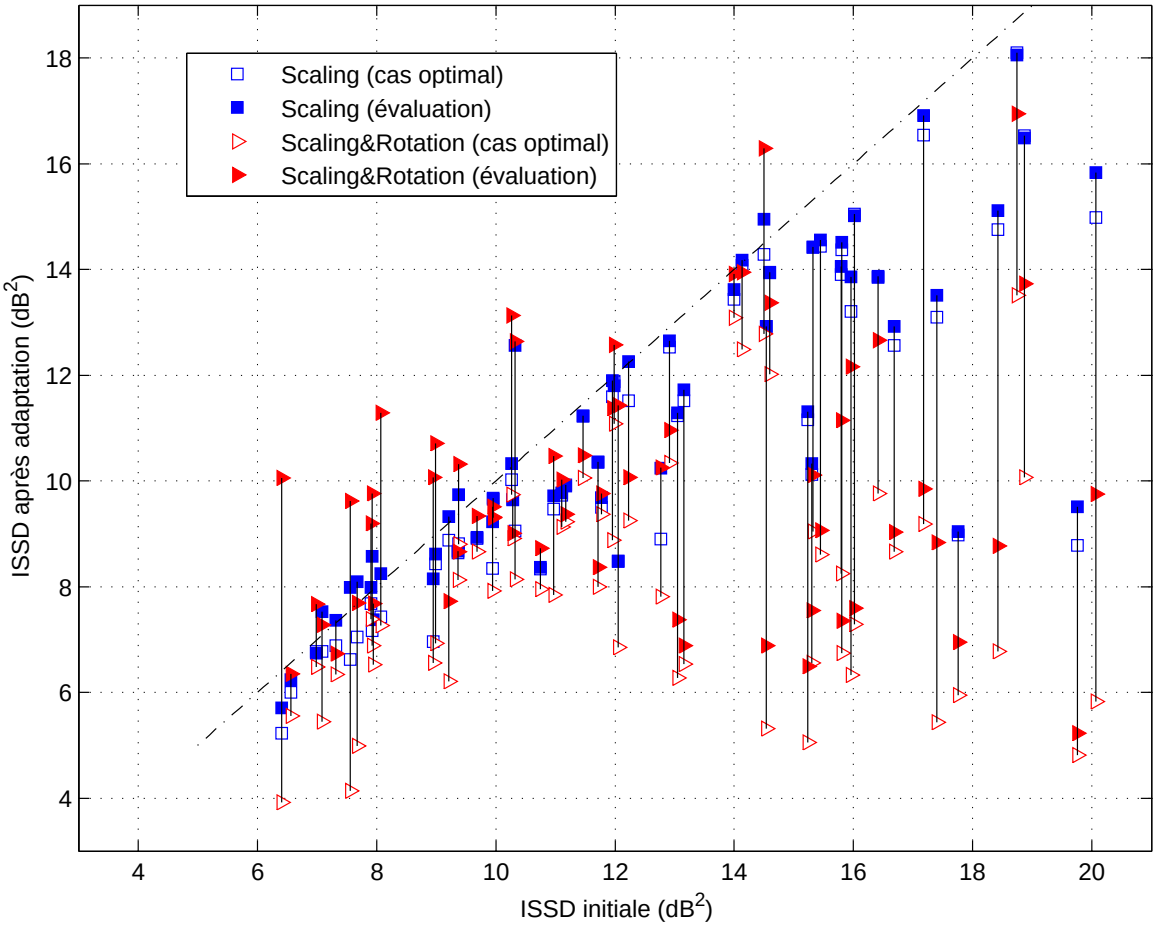


FIG. 3.109 – ISSD obtenue après adaptation des HRTF en fonction de l'ISSD initiale (Base *Jean-Marie Pernaux* : sujets ME, JD, RN, PA, MA, VM ; 60 paires ont été considérées au total ; d'après [Guillon, 2009]). Deux types d'adaptation sont comparés : adaptation par homothétie seule (*scaling* : □ en bleu) ou par homothétie et rotation associées (*Scaling & Rotation* : ▷ en rouge). Sont présentés les résultats de l'adaptation avec les paramètres optimaux qui ont été calculés par minimisation de l'ISSD (*cas optimal* : □ ou ▷ vides) et les paramètres prédits à partir de la moprhologie (*évaluation* : □ ou ▷ pleins).

obtenus sur une base de données ne comportant que 6 sujets et qui mériterait d'être étoffée, tout d'abord pour consolider les relations entre les paramètres de transformation et la morphologie de l'auditeur. Par ailleurs, l'évaluation objective qui a été présentée a repris les données qui avaient été utilisées pour construire le modèle. Il est nécessaire de la compléter en considérant de nouveaux sujets, ce qui nécessite une nouvelle campagne d'acquistion de HRTF et de scans 3D de pavillons. Une évaluation subjective est également à mener.

Concernant le modèle, deux points restent à creuser :

- L'**acquisition des maillages 3D** des pavillons des auditeurs est la seule étape délicate de sa mise en œuvre, principalement dans un contexte grand public d'individualisation de HRTF. Le système utilisé pour notre étude est relativement souple dans la mesure où il repose sur un équipement léger que l'opérateur tient dans sa main. Cependant il impose au sujet une séance dans une situation contraignante et inconfortable (cf. Fig. 3.105). Des travaux récents suggèrent la possibilité de substituer à l'acquisition de scans laser, une séance de prise de **photographies** (cf. Fig. 3.110) à partir desquelles un maillage 3D du pavillon peut être calculé [Dellepiane et al., 2008]. Cette piste prometteuse ouvre la perspective d'appliquer le modèle d'adaptation des HRTF à partir d'un simple jeu de photographies, ce qui élimine le dernier obstacle à sa mise en œuvre.
- Le **choix des HRTF initiales** qu'on va adapter au nouvel individu reste une question ouverte. Il faut d'abord évaluer dans quelle mesure la réduction de l'ISSD dépend du choix de ces HRTF. S'il s'avère qu'il peut exister un choix particulier permettant une adaptation optimale, il reste à déterminer la stratégie et le critère de choix. Une première solution serait un rapprochement morphologique. Une autre piste serait de disposer d'un catalogue suffisamment varié (en termes de différences de type "structurel") de HRTF non individuelles. On pourrait alors proposer plusieurs jeux de HRTF adaptées, en laissant le soin à l'auditeur de choisir les HRTF qui lui conviennent le mieux.

3.5.7 Protocole d'évaluation subjective par mesure des temps de réponse

Proposition d'une nouvelle méthodologie d'évaluation subjective des VAS

Dans le cadre de la validation du modèle de reconstruction de HRTF (cf. Section 3.5.5), une évaluation subjective du modèle a été menée. Elle met en œuvre une nouvelle méthodologie à laquelle est dédiée cette Section et dont les deux principaux aspects innovants sont :

- Synthèse binaurale **dynamique** : La méthode de référence pour évaluer subjectivement la qualité d'un jeu de HRTF est un test de localisation en synthèse binaurale *statique* (cf. page 207) [Wightman & Kistler, 1989b] [Carlile et al., 1997] [Carlile et al., 2000]. Or, la synthèse binaurale statique est une situation d'écoute artificielle, en raison de l'absence des indices dynamiques de localisation, et par la même non *écologique*. Elle risque d'introduire des artefacts de perception susceptibles de perturber l'évaluation des HRTF elles-mêmes. Pour ces raisons nous avons opté pour une synthèse binaurale dynamique basée sur un suivi des mouvements de tête à l'aide de capteurs magnétiques [Guillon, 2009].
- Mesure du **temps de réponse** : Classiquement l'évaluation des HRTF est réalisée en observant les erreurs de localisation des sources virtuelles par les sujets (cf. page 207). Cependant la mesure des erreurs de localisation est entachée de plusieurs biais, liés notamment à la méthode de report du jugement de localisation et aux performances intrinsèques de localisation auditive du sujet. Aussi, même si notre méthodologie se base sur un test de localisation, le jugement de localisation n'est pas notre variable d'observation. Nous nous intéressons au temps de réponse, c'est à dire au temps que met le sujet à identifier correctement la position de la source virtuelle [Chen, 2002] [Yairi et al., 2008]. La **rapidité** de localisation du sujet nous

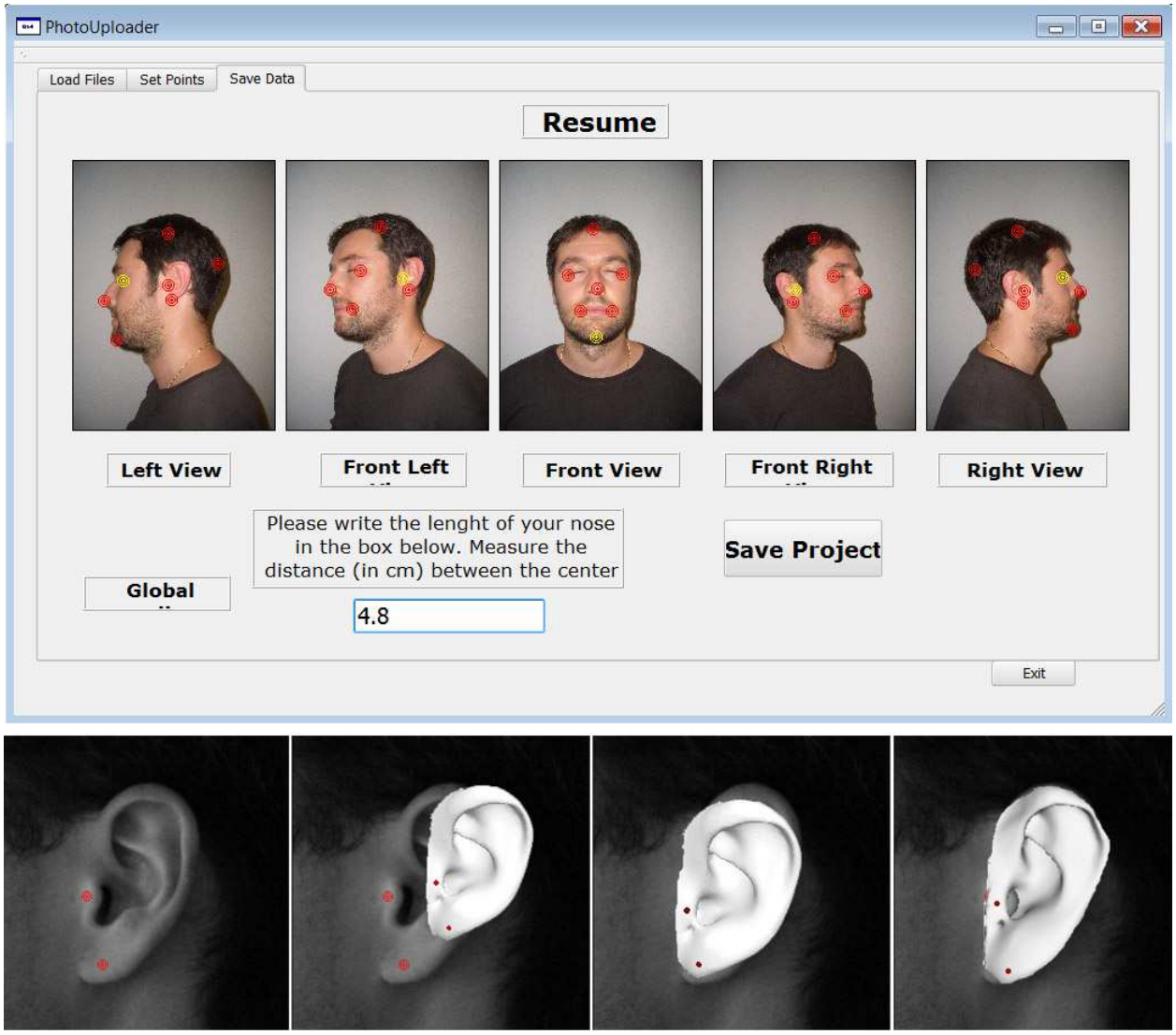


FIG. 3.110 – Obtention de la morphologie 3D d'un individu à partir de photographies (d'après [Dellepiane et al., 2008] [Guillon, 2009]).

semble en effet un indicateur représentatif de la qualité globale, du naturel et de la cohérence du VAS [Guillon, 2009].

Paramètres expérimentaux

Cette nouvelle méthodologie est utilisée ici pour valider le modèle de reconstruction de HRTF. Plus précisément, trois conditions sont évaluées [Guillon, 2009] :

- HRTF individuelles mesurées finement qui définissent la condition idéale de référence où le sujet bénéficie d'un VAS de qualité optimale,
- HRTF individuelles modélisées par reconstruction à partir de HRTF individuelles mesurées pour un nombre limité de directions correspondant à un échantillonnage de plus en plus grossier de la sphère 3D : 6 échantillonnages comprenant de 121 à 19 directions mesurées (cf. Fig. 3.111) sont considérés [Guillon, 2009]. On veut observer l'évolution du comportement du sujet lorsqu'on diminue le nombre de directions mesurées afin de déterminer le nombre minimal de mesures requises, en dessous duquel une dégradation significative du VAS,
- HRTF non individuelles[40] (mesurées finement) : à l'opposé des HRTF individuelles qui constituent l'ancre *haute*, les HRTF non individuelles représentent l'ancre de qualité *basse* reconnue comme affectée par de nombreux artefacts de perception (cf. page 132).

Comme il s'agit d'une nouvelle méthodologie, il aurait été judicieux d'ajouter une troisième ancre correspondant à des sources sonores réelles, ce qui n'a pu être fait pour des raisons pratiques. Néanmoins les HRTF individuelles ont fait l'objet d'une validation spécifique selon la méthologie classique lors d'une étude antérieure [Pernaux, 2003]. Elles constituent donc une ancre haute fiable.

Protocole expérimental

L'expérience se déroule de la façon suivante [Guillon, 2009] : le sujet perçoit une salve de bruits blancs gaussiens et sa tâche consiste à les localiser *le plus rapidement possible* en pointant la tête dans la direction identifiée de la source virtuelle associée (cf. Fig. 3.112 & 3.113). Les stimuli sont diffusés sur un casque circum-auriculaire ouvert (Sennheiser HD600) dont la réponse est compensée par une procédure d'égalisation individuelle, c'est à dire prenant en compte la HPTF de chaque sujet (cf. page 120). Cinq sujets[41] ont participé au test. L'expérience comporte un total de 10 conditions : 1 condition de HRTF individuelles, 6 conditions de HRTF reconstruites (correspondant respectivement à 19, 27, 45, 65, 82, 121 directions de mesure, conditions qui seront désignées ultérieurement par R19, R27, R45, R65, R82, R121), 3 conditions de HRTF non individuelles (correspondant à différents niveaux de distance avec les HRTF individuelles, selon le critère de l'ISSD, cf. Fig. 3.114). Pour chaque condition, un ensemble de 35 positions (cf. Fig. 3.115) de sources virtuelles réparties de façon homogène sur la sphère 3D a été évalué, ceci pour 5 répétitions.

Outils pour l'analyse des résultats

La Figure 3.116 illustre les **trajectoires** mesurées sur un sujet pour une position de source virtuelle. On est frappé par le caractère "direct" très linéaire de la trajectoire. On note que la vitesse angulaire présente un seul maximum en début de la trajectoire. Tous ces éléments indiquent que le sujet a, quasiment dès le début d'émission du stimulus, correctement localisé la source virtuelle et s'est empressé de se diriger vers elle, presque sans hésiter. Les seules hésitations se manifestent en fin de trajectoire, le temps de la validation du jugement de localisation, lorsque le sujet est très

[40] A noter que cette condition n'a été testée que pour 3 sujets, à savoir les sujets ME, JD et RN de la base *Jean-Marie Pernaux*.

[41] Sujets ME, VM, JD, RN et MA de la base *Jean-Marie Pernaux*.

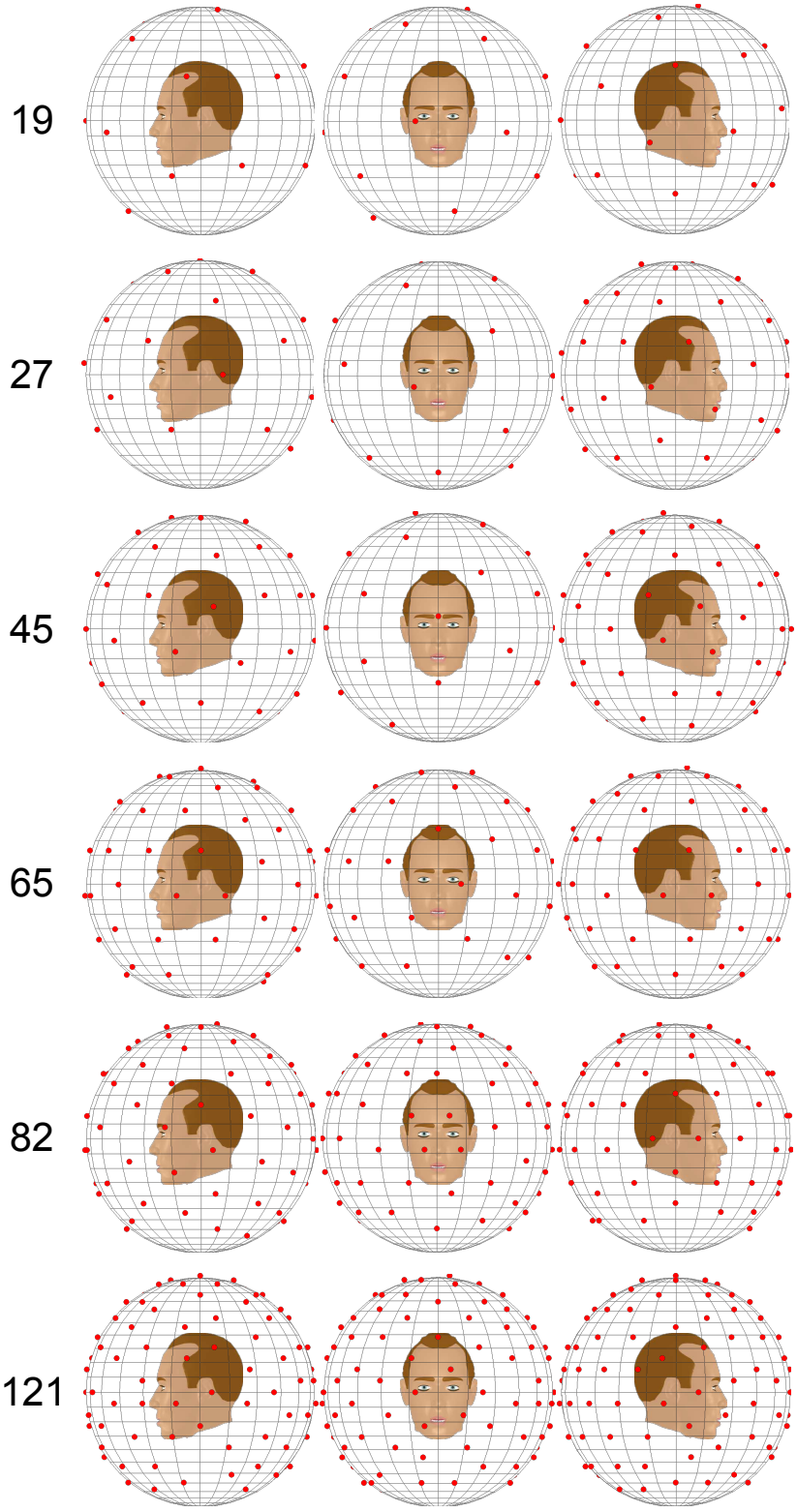


FIG. 3.111 – Echantillonnages spatiaux correspondants aux 19, 27, 45, 65, 82, 121 directions de mesure des HRTF individuelles utilisées en entrée du modèle de reconstruction (d'après [Guillon, 2009]).

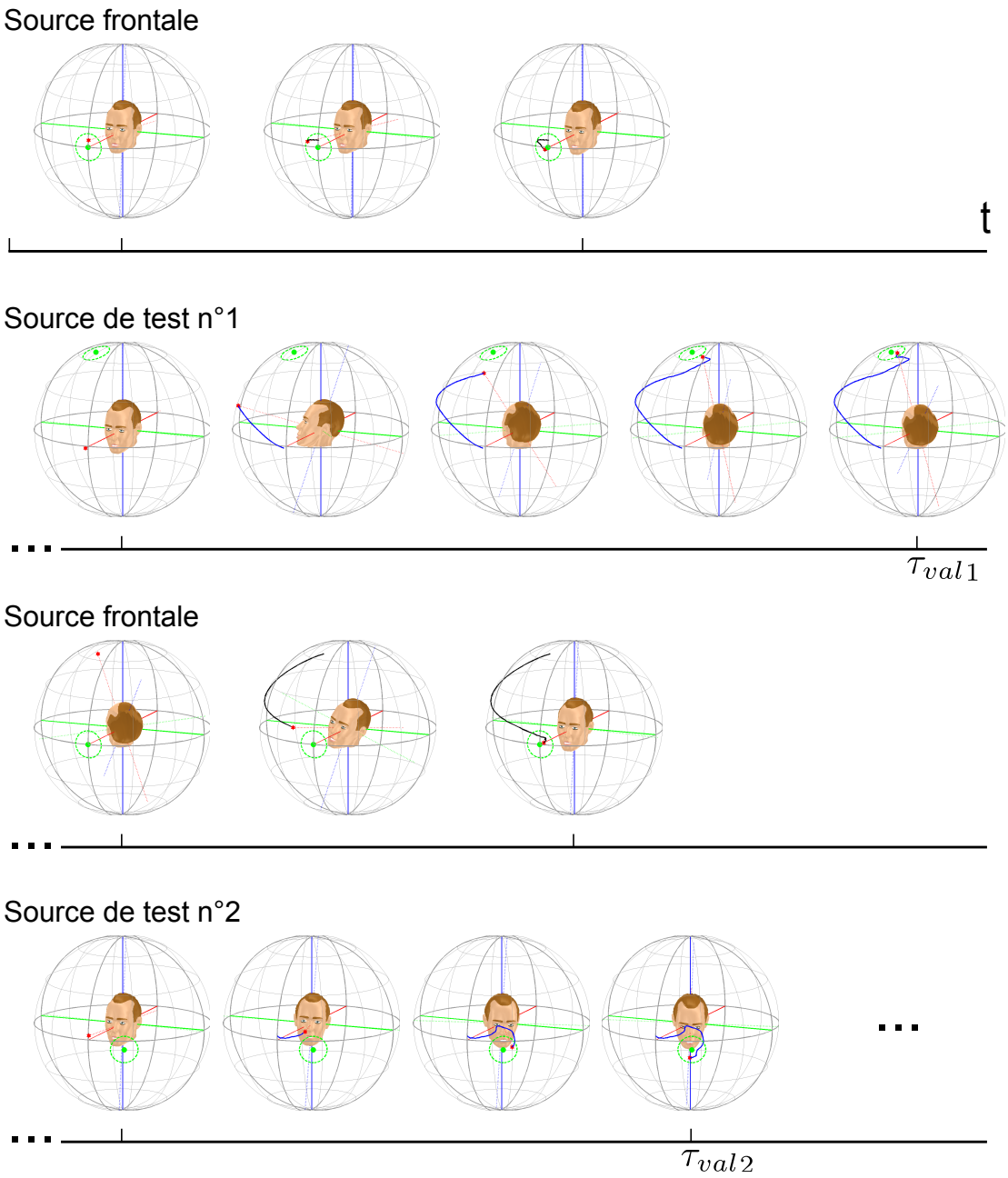


FIG. 3.112 – Illustration du déroulement du test (d'après [Guillon, 2009]) : Une source virtuelle dite "frontale" positionnée en $\phi, \theta = (0°,0°)$ est d'abord émise puis localisée par le sujet. Cette source frontale sert à définir le point de départ (référence identique quelle que soit la position jugée) de la trajectoire de localisation. La source dite "de test" correspondant à une des 35 positions sélectionnées pour l'expérience est ensuite diffusée puis localisée par le sujet. La position de la source de test est symbolisée par le point vert. Le cercle en pointillés verts représente l'aire de validation dans laquelle le sujet doit rester un minimum de 750 ms pour valider son jugement de localisation. Pour chaque source de test, la trajectoire (ligne bleue) suivie par l'axe médian du sujet (droite en pointillés rouges) est enregistrée à l'aide des capteurs de position du système de suivi de mouvements de tête. Le temps τ_{val} nécessaire pour obtenir la validation est aussi relevé.

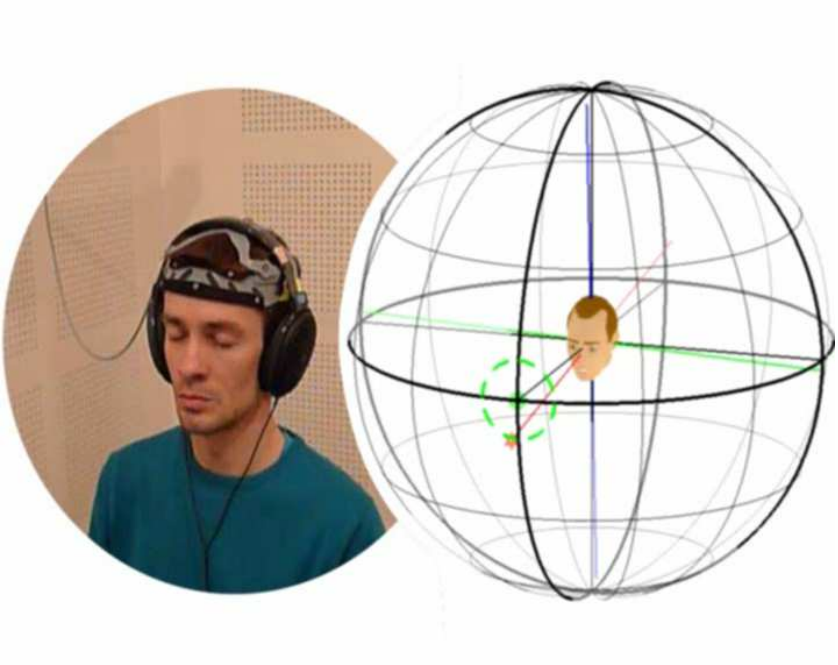


FIG. 3.113 – Illustration d'un sujet passant le test (d'après [Guillon, 2009]).

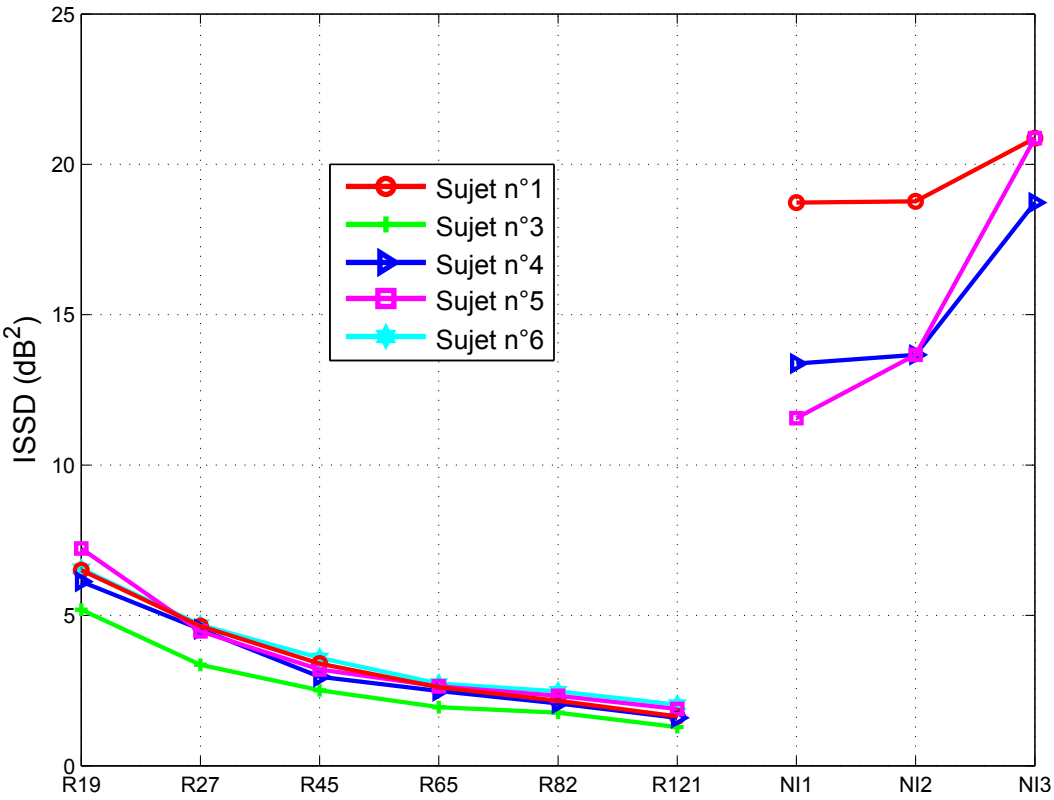


FIG. 3.114 – ISSD calculée entre les HRTF individuelles et les HRTF reconstruites d'une part et les HRTF non individuelles d'autre part (d'après [Guillon, 2009]). Sujets ME (n°1), VM (n°3), JD (n°4), RN (n°5) et MA (n°6) de la base *Jean-Marie Pernaux*.

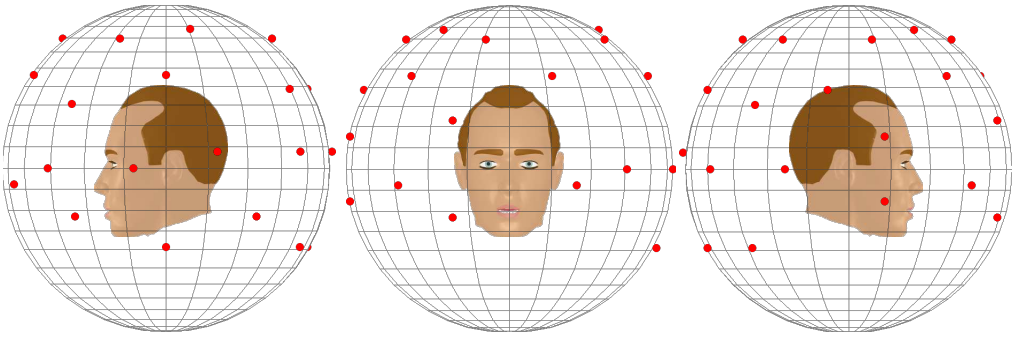


FIG. 3.115 – Illustration des 35 positions de source virtuelle utilisées pour l'expérience (d'après [Guillon, 2009]).

proche de la position cible. On observe un certain effet d'apprentissage[42] dans la mesure où, pour le dernier essai, ces hésitations finales ont presque disparu. Cet exemple est représentatif des réponses collectées pour l'ensemble des sujets et des directions [Guillon, 2009]. Ces observations constituent une première validation du protocole expérimental au sens où elles montrent que le sujet est bien en mesure d'accomplir avec succès la tâche qui lui a été assignée (jugement de localisation), sans difficulté apparente, et en respectant la consigne de rapidité.

Nous allons à présent nous intéresser au **temps de réponse**. Pour chaque jugement de localisation, le temps de réponse τ_{rep} est défini comme le temps séparant l'instant où l'axe médian du sujet sort du cône de 9° entourant la position frontale de référence (0°,0°) de l'instant où il entre dans le cône entourant la source virtuelle à localiser, soit 750 ms avant l'instant de validation (cf. Fig. 3.112 & 3.117) [Guillon, 2009]. Indépendamment des conditions expérimentales, ce temps de réponse dépend de la position de la source virtuelle, puisqu'il est lié à la longueur de la trajectoire à parcourir. Il est aussi influencé par la rapidité intrinsèque du sujet à localiser. Par suite, il est impératif de normaliser chaque temps de réponse afin de se ramener à des valeurs comparables quels que soient la position et le sujet. On choisit de le normaliser sur la base des performances optimales du sujet, en considérant la trajectoire $d_i = \min[d_{rep}(\chi_i)]_{i=1,2,\ldots,35}$ la plus courte qu'il ait accompli pour une direction donnée χ_i, toutes mesures confondues. Le temps de réponse normalisé $\tilde{\tau}_{rep}$ se définit comme [Guillon, 2009] :

$$\tilde{\tau}_{rep}(\chi_i) = \frac{\tau_{rep}(\chi_i)}{d_i} \ . \tag{3.41}$$

Pour chaque condition expérimentale, est collecté un ensemble de 35 (directions) x 5 (sujets) x 3 (essais) = 525 jugements correspondant à des temps de réponse. Ces données se caractérisent par une distribution asymétrique décalée sur la gauche (cf. Fig. 3.118), c'est à dire que la valeur médiane est inférieure à la moyenne, ce qui est typique d'une distribution de temps de réponse. Les descripteurs statistiques *classiques* tels que la *moyenne* ou la *médiane* et qui sont associés à des variables aléatoires présentant une loi normale (c'est à dire gaussienne) s'avèrent donc inappropriés [Guillon, 2009]. Il convient, conformément à ce qui a été proposé dans [Hohle, 1965] [Ratcliff, 1978], de modéliser les temps de réponse comme la somme d'une variable aléatoire normale et d'une variable aléatoire exponentielle, ce qui définit une variable **ex-gaussienne** (cf. Fig. 3.119) [Lacouture & Cousineau, 2008]. Les descripteurs statistiques associés à une variable ex-gaussienne sont au nombre de 3 (cf. Fig. 3.119) :

[42] Une observation attentive de l'évolution des réponses des sujets au fur et à mesure des répétitions indique une stabilisation à partir de la troisième répétition, ce qui nous a conduit, pour la suite de l'étude, à éliminer les résultats des deux premiers essais et à regrouper ceux des trois derniers essais pour constituer les données analysées.

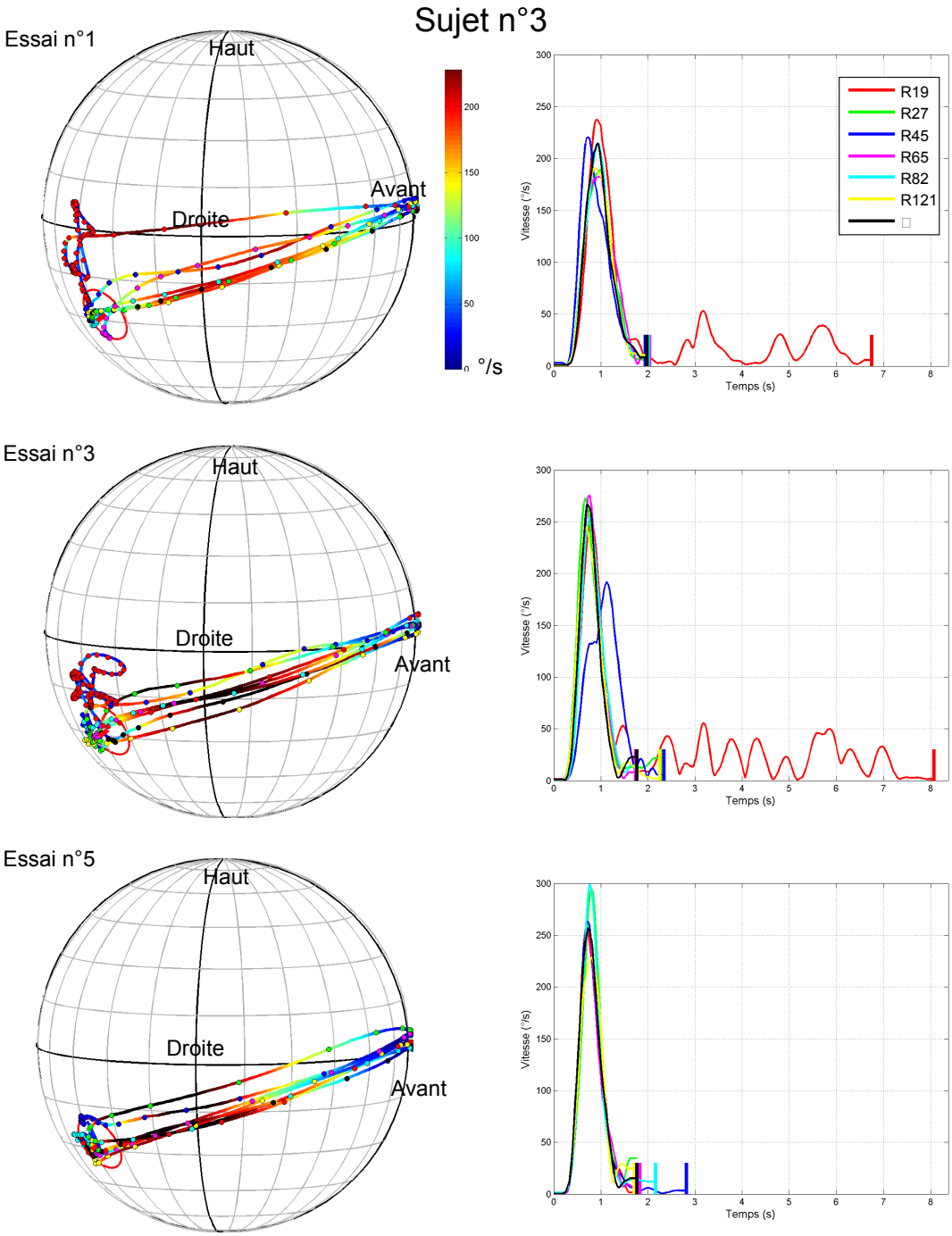


FIG. 3.116 – Trajectoire suivie par le sujet VM (base *Jean-Marie Pernaux*) pour la source située en $\phi, \theta = (229°,\text{-}28.15°)$ (système polaire vertical, d'après [Guillon, 2009]) : Chaque condition a été mesurée 5 fois, ce qui donne 5 essais. Sont présentées ici les réponses mesurées pour les essais n°1, 3 et 5. La position de la source virtuelle à localiser est entourée d'un cercle rouge. Les trajectoires entre la source frontale et la position identifiée comme celle de la source virtuelle sont visualisées à gauche. La vitesse angulaire, qui est aussi représentée à droite, est codée en couleur le long de chaque trajectoire.

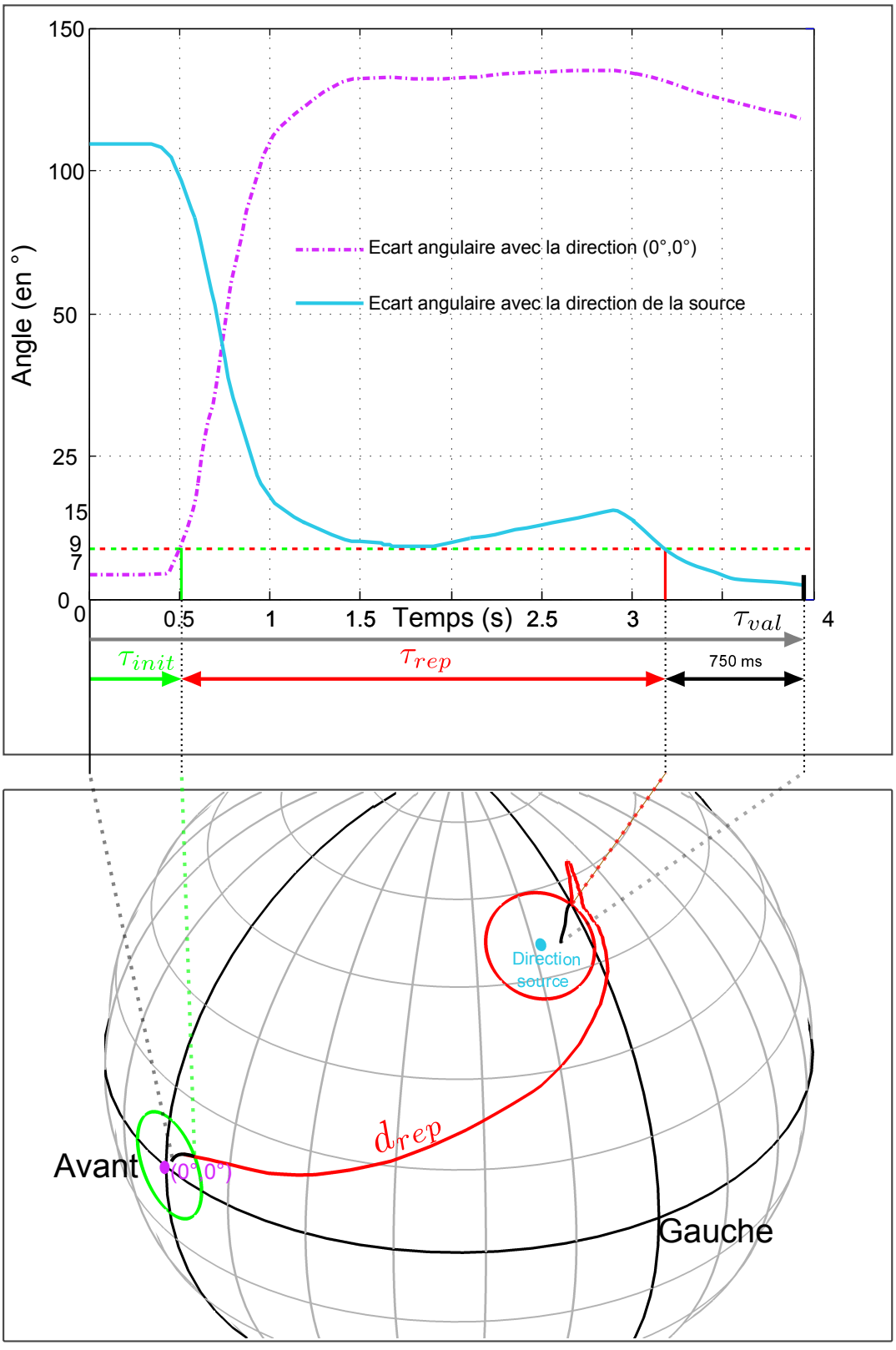


FIG. 3.117 – Définition du temps de réponse (d'après [Guillon, 2009]).

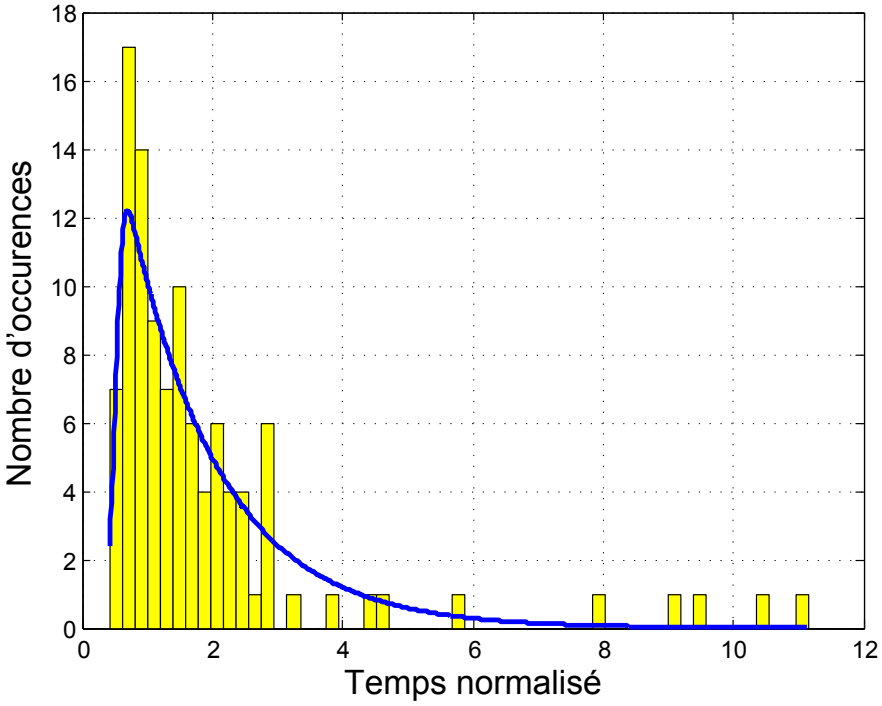


FIG. 3.118 – Exemple de distribution des temps de réponse réalisés par un sujet pour une condition expérimentale (d'après [Guillon, 2009]).

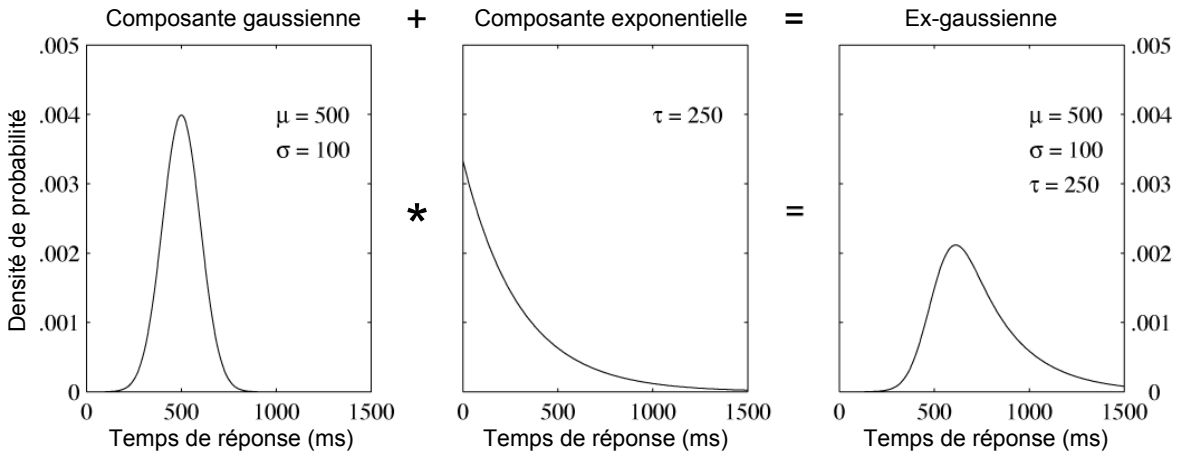


FIG. 3.119 – Définition et densité de probabilité d'une variable ex-gaussienne (d'après [Guillon, 2009]).

– μ et σ : respectivement la moyenne et l'écart-type de la composante normale,
– τ : moyenne de la composante exponentielle.

Pour une distribution donnée, les paramètres μ, σ et τ sont estimés en cherchant le jeu de paramètres permettant de la modéliser au mieux par une distribution ex-gaussienne. Dans notre cas, cet ajustement a été réalisé par maximisation de la vraisemblance (*Maximum Likelihood Estimation* ou MLE) [Guillon, 2009]. Cependant, pour chaque distribution considérée (correspondant à une condition expérimentale), la qualité de l'estimation n'est pas connue, notamment en termes de *dispersion*, du fait qu'on ne dispose que d'un échantillon (correspondant à la distribution considérée qui n'est pas disponible sous la forme de plusieurs réalisations). La technique d'inférence statistique appelée *bootstrapping* permet de créer artificiellement plusieurs pseudo-réalisations de cet échantillon [Efron & Tibshirani, 1993]. Il est alors possible de calculer la valeur moyenne et l'intervalle de confiance associés aux paramètres μ, σ et τ estimés pour une distribution particulière décrivant une condition expérimentale particulière [Guillon, 2009]. L'analyse de l'ensemble des résultats indique que, dans le cas de nos données, des 3 descripteurs statistiques, seul le paramètre τ apporte une information lisible sur le comportement du sujet [Guillon, 2009]. Par suite c'est sur ce paramètre que se focalise notre analyse.

HRTF reconstruites vs HRTF mesurées

La Figure 3.120 reproduit le paramètre τ pour les HRTF individuelles mesurées et les HRTF reconstruites avec un nombre décroissant de directions mesurées. On observe que ce paramètre présente une évolution relativement caractéristique et cohérente d'un sujet à l'autre : τ est minimal dans le cas des HRTF mesurées et croît au fur et à mesure que le nombre de directions mesurées diminue dans le cas des HRTF reconstruites. Cette augmentation dénote un allongement du temps de réponse des sujets que nous interprétons comme une dégradation de la qualité du VAS correspondant à une détérioration de la reconstruction des HRTF. Le paramètre τ apparaît donc comme un indicateur fiable et apte à rendre compte de l'impact de la qualité de modélisation des filtres binauraux sur la perception audio spatialisée du sujet. A l'exception d'un sujet, les HRTF reconstruites avec 121 directions mesurées ne sont pas perçues comme significativement différentes des HRTF individuelles mesurées [Guillon, 2009], ce qui constitue une première validation du modèle de reconstruction. Il reste à déterminer le nombre de directions mesurées à partir duquel les différences avec les HRTF mesurées sont significatives. Ce nombre dépend sensiblement du sujet : il vaut 45 pour les sujets JD et RN, 27 pour le sujet ME et 19 pour le sujet MA. Le sujet VM présente un comportement marginal : aucun des jeux de HRTF reconstruites n'est jugé différent des HRTF mesurées. Il semble que pour ce sujet, le seuil soit inférieur à 19 directions de mesure. Si l'on prend le maximum des seuils sur l'ensemble des sujets, un minimum de **65 directions de mesure** est requis pour garantir une reconstruction transparente quel que soit l'individu. En comparaison des 965 directions initialement mesurées, ce résultat représente un gain d'un rapport de près de 15, ce qui est considérable.

HRTF non individuelles vs HRTF individuelles

Une dernière question à étudier concerne le positionnement de HRTF non individuelles par rapport aux HRTF individuelles reconstruites. Quel est l'apport des HRTF reconstruites en comparaison d'un choix arbitraire de HRTF non individuelles dans une base de données ? Pour illustrer un choix arbitraire de HRTF non individuelles, on considère trois jeux de HRTF non individuelles correspondant à des valeurs croissantes d'ISSD vis à vis des HRTF individuelles. Le paramètre τ estimé à partir des réponses de 3 sujets est représenté sur la Figure 3.121 pour l'ensemble des conditions des HRTF individuelles et non individuelles. Dans la majorité des cas, les HRTF non

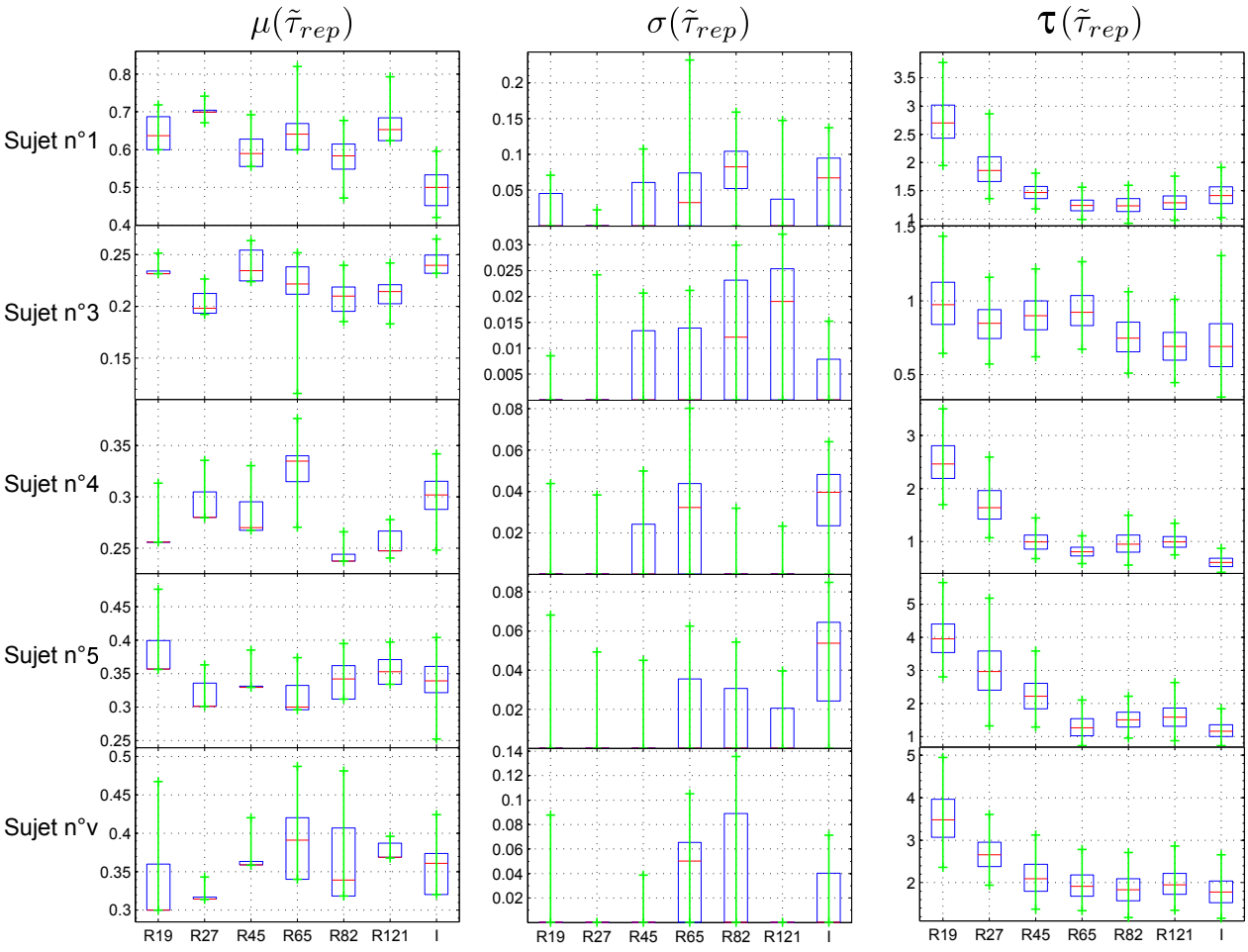


FIG. 3.120 – Paramètres μ, σ et τ décrivant la distribution des temps de réponse $\tilde{\tau}_{rep}$ de chaque sujet en fonction des différentes conditions expérimentales (d'après [Guillon, 2009]). Seules les HRTF individuelles sont considérées : HRTF mesurées (condition I) ou reconstruites (conditions R19, R27, R45, R65, R82, R121). A noter que l'échelle en ordonnée diffère selon le sujet.

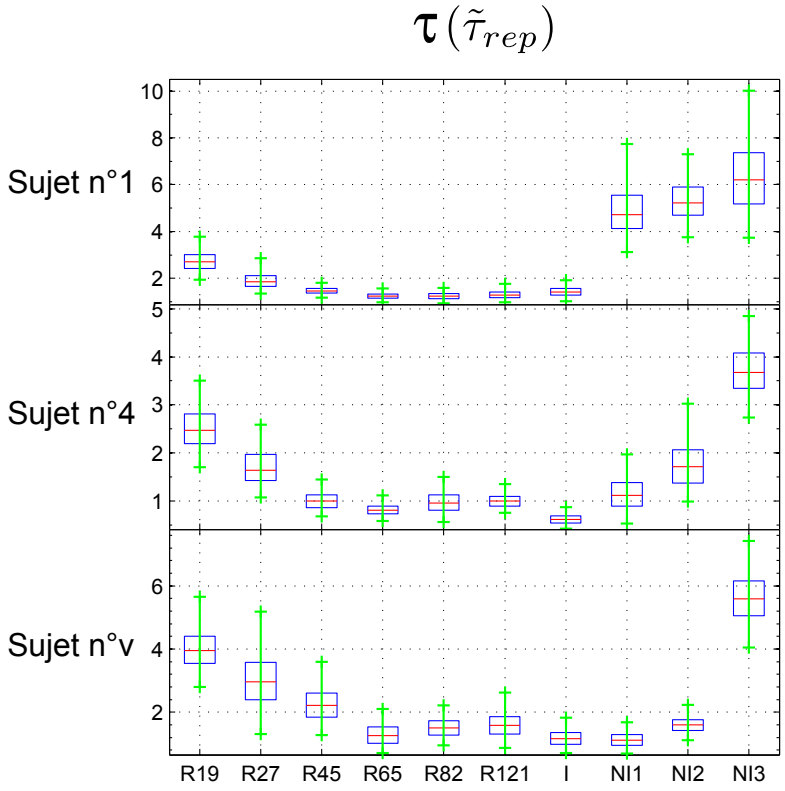


FIG. 3.121 – Paramètre τ décrivant la distribution des temps de réponse $\tilde{\tau}_{rep}$ de chaque sujet en fonction des différentes conditions expérimentales (d'après [Guillon, 2009]). Comparaison des HRTF individuelles (HRTF mesurées ou reconstruites) et non individuelles. Trois jeux de HRTF non individuelles sont considérés, correspondant à trois niveaux croissants d'ISSD (conditions NI1, NI2 et NI3, cf. Fig. 3.114). L'étude est restreinte aux sujets ME (n°1), JD (n°4) et RN (n°5). A noter que l'échelle en ordonnée diffère selon le sujet.

individuelles introduisent une augmentation très significative du paramètre τ. Cette augmentation est fortement corrélée à la valeur de l'ISSD associée au jeu de HRTF non individuelles. On note que, pour tous les sujets, l'allongement du temps de réponse pour le jeu présentant l'ISSD maximale avec les HRTF individuelles (condition NI3) excède largement les pires résultats obtenus avec les HRTF reconstruites (condition R19). Un choix arbitraire de HRTF non individuelles peut donc avoir des conséquences dramatiques, bien pires qu'une mauvaise modélisation individualisée. Dans le détail, les résultats dépendent du sujet. La variabilité est d'autant plus marquée que l'étude ne comporte que 3 sujets. Le sujet ME témoigne d'une gêne importante et quasiment équivalente avec les jeux de HRTF non individuelles, quelle que soit l'ISSD. Il semble, en quelque sorte, très "résonant" sur ses HRTF. A l'opposé, le sujet RN "s'accommode" des HRTF NI1 et NI2 pour lesquelles son temps de réponse n'est pas significativement différent des HRTF individuelles (condition I), mais "rejète" fortement les HRTF NI3. Le sujet JD offre un comportement intermédiaire avec un accroissement du paramètre τ sensiblement proportionnel à l'ISSD.

Conclusion

En alternative aux tests de localisation, une nouvelle méthodologie d'évaluation subjective des VAS a été proposée et mise en œuvre. Elle se fonde sur l'observation des temps de réponse des sujets dans une tâche de localisation des sources virtuelles. Elle requiert des outils dédiés d'analyse statistique (modélisation de la densité de probabilité par une fonction ex-gaussienne et ses descripteurs associés). L'analyse met en évidence un indicateur (paramètre τ), représentatif du temps de réponse, dont l'évolution dénote une corrélation certaine avec la qualité des HRTF ("qualité" au sens des différentes conditions I, R19, R27, R45, R65, R82, R121, NI1, NI2 et NI3 données à entendre aux sujets lors du test d'écoute), et qui, par suite, semble capable de détecter l'effet de la modélisation des filtres binauraux sur la perception du sujet. Au delà des descripteurs statistiques, la méthodologie proposée fait aussi ses preuves dans le ressenti des sujets et la richesse de leurs commentaires qui permettent d'affiner et de nuancer les résultats quantitatifs [Guillon, 2009].

3.6 Quel(s) usage(s) de la synthèse binaurale ?

Dans ce qui précède, des modèles de calcul de filtres binauraux individualisés à l'auditeur, à la fois en termes d'ITD et d'IS, ont été proposés. La finalité de ces travaux vise l'intégration d' unmoteur de spatialisation binaurale dans des applications grand public dans le contexte étendu des télécommunications d'aujourd'hui. En guise de conclusion, des exemples de telles mises en œuvre vont être décrits. De nos jours, les technologies binaurales demeurent principalement appliquées dans le contexte des **laboratoires de recherche**, notamment pour explorer et comprendre les mécanismes de la perception auditive. Cependant des applications commerciales commencent à émerger. Dans le domaine des contenus, la première utilisation concerne la **synthèse de haut-parleurs virtuels** : avec une paire de filtres binauraux on est capable de simuler pour une écoute au casque un haut-parleur dans n'importe quelle direction. Le principe peut être étendu à tout système multi haut-parleurs (stéréophonique, multicanal de type 5.1 et dérivés 6.1/7.1/10.2/22.2, HOA). L'intérêt est de proposer une alternative d'écoute sur un équipement léger et discret (casque ou paire d'écouteurs qui présente notamment l'avantage d'offrir une diffusion limitée à l'auditeur sans déranger ses voisins) de contenus audio spatialisés dédiés à des équipements lourds impliquant des réseaux multi haut-parleurs. Le procédé de synthèse binaurale de haut-parleurs virtuels s'apparente à une *conversion de formats audio 3D* dans la mesure où il consiste à convertir un flux multicanal (de type 5.1 ou HOA par exemple) en un flux bicanal au format binaural. On parle d'ailleurs de **downmix binaural**. Dans le cas spécifique de contenus 5.1, on rencontre le terme de *virtual*

surround qu'on trouve dans des produits tels que le *Dolby® Headphone* ou le concept de *Headzone 5.1* de Beyerdynamic®. La technologie de *virtual surround* n'est d'ailleurs pas limitée à une restitution sur casque mais s'applique également à une paire de haut-parleurs (cf. produit *Dolby® Virtual Speaker*). Dans le contexte des services conversationnels, la **conférence audio spatialisée** constitue une application en voie d'émergence où les technologies binaurales démontrent tous leurs avantages, que ce soit avec une prise de son naturelle par tête artificielle d'une salle de réunion où sont présents plusieurs locuteurs, ou avec la synthèse binaurale pour spatialiser différents locuteurs distants répartis sur des lieux distincts [Nagle, 2008]. Dans le cadre de ces réunions téléphoniques, la spatialisation des locuteurs distants améliore notablement l'intelligibilité des voix, ainsi que le confort et le naturel des interactions. Sans oublier l'intérêt en termes de développement durable, étant donné que ce type de service, dès lors qu'il offre une qualité proche ou équivalente d'une réunion physique des personnes, permet de réduire les transports de personnes.

Ce qui séduit dans les technologies binaurales c'est la simplicité de leur mise en œuvre : une paire de microphones placés dans les oreilles de n'importe quel individu suffit pour une prise de son binaurale. En dépit de cette simplicité, la spatialisation présente des qualités impressionnantes de naturel et d'immersion permettant de créer des illusions auditives véritablement capables de leurrer l'auditeur. Une première application est la *carte postale sonore* consistant à enregistrer une scène sonore (concert, spectacle, moment de convivialité, paysage sonore...) avec une paire de microphones binauraux et à la transmettre en temps réel ou différé à des personnes distantes. Une autre déclinaison est le *webmike* dans lequel on utiliserait une tête artificielle pour enregistrer l'évolution en temps réel de la scène sonore associée à un lieu donné, en vue par exemple d'une diffusion en continu sur internet. La tête artificielle constitue depuis longtemps un outil de référence pour la captation de paysages sonores ; notamment pour l'observation de données environnementales telles que le bruit. Jusqu'à l'apparition des réseaux multimicrophoniques pour des captations HOA, la tête artificielle représentait d'ailleurs le seul système de prise de son 3D. En termes de coût de transmission, le flux binaural représente un compromis optimal entre les performances de spatialisation et le débit des données audio : pour le coût d'un flux stéréophonique, il offre en effet une spatialisation bien supérieure couvrant l'ensemble de la sphère 3D. Il ne semble même pas nécessaire de développer des techniques de compression audio dédiée.

La simplicité de mise en œuvre se retrouve au niveau de la restitution puisqu'une paire d'écouteurs (casque ou écouteur intra-auriculaire, voire paire de haut-parleurs) suffit, ce qui désigne la technologie binaurale comme l'outil de spatialisation privilégié pour tous les équipements mobiles de type téléphone mobile ou ordinateur portable, ou plus généralement la catégorie des *handheld devices*. Ce type d'équipements impose alors des contraintes en termes d'implémentation en raison de leur miniaturisation : il requiert en particulier le développement spécifique d'un moteur de synthèse binaurale optimisé pour offrir un coût de calcul minimal compatible avec leurs performances limitées. Le schéma d'*implémentation binaurale multicanale* [Larcher, 2001] est une solution à ce problème : elle consiste à décomposer chaque filtre binaural sous la forme d'une combinaison linéaire d'un nombre réduit de filtres qui sont communs à toutes les directions. Ainsi dans le moteur de synthèse binaurale, un seul jeu de filtres est implémenté une fois pour toutes. Les filtres binauraux sont ensuite synthétisés par composition de ces filtres élémentaires, en jouant sur les pondérations pour simuler une direction donnée. On note que, quel que soit le nombre de sources sonores virtuelles, le nombre de filtres est identique, ce qui constitue un avantage considérable par rapport à un schéma classique d'implémentation dès que la scène sonore est complexe et comporte un grand nombre de sources.

Malgré ses nombreux avantages, la technologie binaurale n'est pas encore aujourd'hui reconnue comme une technologie de prise de son à part entière, principalement par la communauté des ingénieurs du son. Elle reste en effet associée dans les esprits au contexte des laboratoires de

FIG. 3.122 – Tête artificielle KU 100 de Neumann®.

recherche (perception auditive ou métrologie). Le principal obstacle réside dans les colorations spectrales qui dénaturent le timbre des sources et disqualifient la technologie binaurale en termes de qualité audio et de transparence aux oreilles des professionnels audio, en dépit de la qualité de sa spatialisation. Des premiers signes d'évolution émergent cependant, avec par exemple la proposition par Neumann® de la tête artificielle KU 100 (cf. Fig. 3.122) dédiée explicitement à la prise de son, ou les travaux de l'Association Omnihead [Rueff & Blum, 2003], poursuivis par P. Rueff [Rueff, 2010] visant à explorer le potentiel des technologies binaurales pour la prise de son (cf. Fig. 3.5).

Chapitre 4

Conclusion

Ce document a présenté un panorama des travaux de recherche que j'ai menés au sein de l'équipe *Audio 3D* à Orange Labs à l'issue de ma thèse. Il s'agit de travaux de recherche amont visant à l'amélioration et l'enrichissement des briques technologiques de spatialisation sonore en vue d'intégration dans des futurs services de communication (conférence audio 3D, MMS, contenus audio 3D...). Pour une large part les études présentées s'appuient sur les travaux des thèses que j'ai (co-)encadrées [Pernaux, 2003, Busson, 2006, Guillon, 2009], la frontière entre mes travaux en propre et ceux des thèses étant ténue.

Les thèmes abordés couvrent l'ensemble des aspects de la chaîne de (re)création d'un VAS, à savoir :

- captation ou synthèse d'une scène audio 3D,
- transmission du flux audio 3D associé, incluant une éventuelle compression des données,
- restitution de la scène,
- perception par un auditeur.

De façon transverse, les concepts de format(s) audio 3D et de conversion(s) associée(s) sont aussi présents. Toutes les études n'ont pas été détaillées. J'ai choisi de focaliser le contenu du document sur mes contributions majeures (et notamment pour lesquelles je suis reconnue par la communauté scientifique internationale), concernant d'une part le lien entre les technologies WFS et HOA et d'autre part la modélisation individualisée des filtres binauraux pour la synthèse binaurale. Pour chaque étude présentée, le contexte et l'état des travaux antérieurs est rappelé, à la fois pour donner une vision complète de la technologie, mais surtout pour bien mettre en évidence la portée de mes apports.

Aujourd'hui l'état des lieux des technologies de spatialisation sonore pour les VAS indique qu'il existe une panoplie relativement matures de solutions pour la captation (arbre multicanal, tête artificielle, réseau microphonique de type HOA) et la restitution (casque, réseau multi haut-parleurs), avec la possibilité de s'adapter aux contraintes de contextes applicatifs variés (écoute multi auditeurs/ mono auditeur, mobilité de l'utilisateur, miniaturisation des équipements, communication...). Les prochains défis semblent s'orienter sur les questions suivantes :

- définir un (de) format(s) audio 3D, alternatif au format 5.1, offrant une spatialisation enrichie, et reconnu notamment comme standard audio professionnel auprès des instances de normalisation,
- développer des schémas de codage audio 3D[1] pour la compression des flux audio multicanal (notamment les flux HOA dans la perspective où cette dernière technologie émerge),
- se doter d'outils de conversion de format permettant avec un système d'écoute donné d'avoir

[1] Cette question est traitée avec le travail de thèse d'A. Daniel que je co-encadre depuis décembre 2007.

accès à l'ensemble des contenus audio 3D disponibles quels que soient leurs formats,
- relâcher les contraintes sur les spécificités des équipements audio, notamment les réseaux de microphones et de haut-parleurs utilisés pour la prise et la restitution du son, à la fois en vue d'être capable de tirer parti des transducteurs existants dans le contexte (téléphones, ordinateurs, équipements domestiques de type chaîne HIFI...) pour s'affranchir de l'ajout d'équipements spécifiques et ainsi avoir une mise en œuvre "discrète", et s'adapter aux contraintes (physiques, esthétiques, financières ou autres) de l'utilisateur[2].

Au delà de ces aspects technologiques, un point essentiel à traiter est d'"exorciser" les technologies audio 3D aux yeux non seulement de la communauté audio professionnelle, mais aussi du grand public. Ces technologies sont trop souvent perçues comme cantonnées aux laboratoires de recherche, en raison de leur complexité ou du coût matériel, ce qui n'est pas toujours le cas comme en témoigne la technologie binaurale. Il s'agit donc de démontrer l'intérêt et l'apport de l'audio 3D afin de sensibiliser les auditeurs non experts à ces nouvelles technologies. Sans oublier qu'il est tout aussi important de travailler à en simplifier la mise en œuvre et l'usage pour "l'homme de la rue".

[2]Cette question est traitée avec le travail de thèse de R. Deprez que j'encadre depuis décembre 2008.

Bibliographie

[Algazi & Duda, 2008] Algazi, V. & Duda, R. (2008). Effective use of psychoacoustics in motion-tracked binaural audio. In *Tenth IEEE International Symposium on Multimedia, ISM 2008.*

[Algazi et al., 2004] Algazi, V., Duda, R., & Thompson, D. (2004). Motion tracked-binaural sound. *J. Audio Eng. Soc.*, 52(11), pp. 1142–1156.

[Algazi et al., 2001a] Algazi, V. R., Avendano, C., & Duda, R. O. (2001a). Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.*, 109(3), pp. 1110–1122.

[Algazi et al., 2001b] Algazi, V. R., Avendano, C., & Duda, R. O. (2001b). Estimation of a spherical-head model from anthropometry. *J. Audio Eng. Soc.*, 49(6), pp. 472–479.

[Algazi et al., 2002a] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., & Tang, Z. (2002a). Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.*, 112(5), pp. 2053–2064.

[Algazi et al., 2001c] Algazi, V. R., Duda, R. O., Morrison, R. P., & Thompson, D. M. (2001c). Structural composition and decomposition of HRTFs. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2001.*

[Algazi et al., 2002b] Algazi, V. R., Duda, R. O., & Thompson, D. M. (2002b). The use of head-and-torso models for improved spatial sound synthesis. In *AES 113th Convention.*

[Algazi et al., 2001d] Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001d). The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

[Atal & Schroeder, 1966] Atal, B. S. & Schroeder, M. (1966). Apparent sound source translator. United States Patent 3,236,949.

[Avendano et al., 1999] Avendano, C., Duda, R. O., & Algazi, V. R. (1999). Modeling the contralateral HRTF. In *AES 16th International Conference on Spatial Sound Reproduction* (pp. 313–318).

[Bamford, 1995] Bamford, J. S. (1995). *An analysis of Ambisonic sound system of first and second order.* PhD thesis, University of Waterloo.

[Baskind, 2003] Baskind, A. (2003). *Modèles et méthodes de description spatiale de scènes sonores : Application aux enregistrements binauraux.* PhD thesis, Université Paris 6.

[Batteau, 1967] Batteau, D. W. (1967). The role of pinna in human localization. *Proc. Royal Society London*, 168, pp. 158–180.

[Bauck & Cooper, 1996] Bauck, J. & Cooper, D. H. (1996). Generalized transaural stereo and applications. *J. Audio Eng. Soc.*, 44(9), pp. 683–705.

[Bech & Zacharov, 2006] Bech, S. & Zacharov, N. (2006). *Perceptual Audio Evaluation : Theory, Method And Application.*

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

[Berg & Rumsey, 1999] Berg, J. & Rumsey, F. (1999). Spatial attribute identification and scaling by repertory grod technique and other methods. In *AES 16th International Conference on Spatial Sound Reproduction*.

[Berkhout, 1988] Berkhout, A. J. (1988). A holographic approach to acoustic control. *J. Audio Eng. Soc.*, 36, pp. 977–995.

[Berkhout et al., 1993] Berkhout, A. J., de Vries, D., & Vogel, P. (1993). Acoustic control by wave field synthesis. *J. Acoust. Soc. Am.*, 93(5), pp. 2764–2778.

[Bertet, 2009] Bertet, S. (2009). *Formats audio 3D hiérarchiques : caractérisation objective et perceptive des systèmes Ambisonics d'ordres supérieurs*. PhD thesis, INSA Lyon.

[Blauert, 1970] Blauert, J. (1969-1970). Sound localization in the median plane. *Acustica*, 22, pp. 205–213.

[Blauert, 1983] Blauert, J. (1983). *Spatial hearing : The psychophysics of human sound localization*. The MIT Press, Cambridge, Massachusetts.

[Bloom, 1977] Bloom, P. (1977). Creating source elevation illusions by spectral manipulations. *J. Audio Eng. Soc.*, 25(9), pp. 561–565.

[Blum, 2003] Blum, A. (2003). *Etude de la plasticité du système auditif en localisation sonore. Application au problème de l'individualisation en synthèse binaurale.* Technical report, Université d'Aix-Marseille II.

[Blum et al., 2004] Blum, A., Katz, B., & Warusfel, O. (2004). Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training. In *Proc. CFA/DADA'04*.

[Braasch & Hartung, 2002] Braasch, J. & Hartung, K. (2002). Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I psychoacoustical data. *Acta Acustica*, 88, pp. 942–955.

[Breebaart et al., 2005a] Breebaart, J., Disch, S., Faller, C., Herre, J., Hotho, G., Kjörling, K., Myburg, F., Neusinger, M., Oomen, W., Purnhagen, H., & Rödén, J. (2005a). MPEG spatial audio coding / MPEG surround : overview and current status. In *AES 119th Convention, New York, USA*.

[Breebaart et al., 2005b] Breebaart, J., Par, S. V. D., Kohlrausch, A., & Schuijers, E. (2005b). Parametric coding of stereo audio. *EURASIP Journal on Applied Signal Processing*, 9, pp. 1305–1322.

[Bregman, 1990] Bregman, A. S. (1990). *Auditory scene analysis : The perceptual organization of sound*. The MIT Press.

[Bronkhorst, 1995] Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98(5), pp. 2542–2553.

[Brookes & Treble, 2005] Brookes, T. & Treble, C. (2005). The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings. In *Proceedings of the 118th Convention of the Audio Engineering Society*.

[Brown & Duda, 1998] Brown, C. P. & Duda, R. O. (1998). A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5), pp. 476–488.

[Bruneau, 1983] Bruneau, M. (1983). *Introduction aux théories de l'acoustique*. Université du Maine, Le Mans.

[Brungart et al., 2004] Brungart, D., Simpson, B., McKinley, R., Kordik, A., Dallman, R., & Ovenshire, D. (2004). The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. In *Proc. Int. Conf. on Auditory Dispaly (ICAD)*.

[Busson, 2006] Busson, S. (2006). *Individualisation d'indices acoustiques pour la synthèse binaurale*. PhD thesis, Université de la Méditerranée Aix-Marseille II.

[Busson et al., 2006] Busson, S., Nicol, R., Choqueuse, V., & Lemaire, V. (2006). Non-linear interpolation of head related transfer function. In *CFA06, 8ème Congrès Français d'Acoustique (Société Française d'Acoustique)*.

[Busson et al., 2005a] Busson, S., Nicol, R., & Katz, B. (2005a). Subjective investigations of the interaural time difference in the horizontal plan. In *118th A.E.S. Convention, Barcelona, 2005 May 28-31*.

[Busson et al., 2005b] Busson, S., Nicol, R., & Lemaire, V. (2005b). Individualisation de HRTF utilisant une modélisation par éléments finis couplée à un modèle correctif. Brevet EP1946612 (WO2007048900, US2008306720).

[Busson et al., 2005c] Busson, S., Nicol, R., & Lemaire, V. (2005c). Procédé de modélisation de HRTF pour l'interpolation et l'individualisation des HRTF. Brevet FR2880755 (EP1836876 , WO2006075077).

[Butler, 1987] Butler, R. (1987). An analysis of the monaural displacement of sounds in space. *Percept. Psychophysics*, 41(1), pp. 1–7.

[Butler & Helwig, 1983] Butler, R. & Helwig, C. (1983). The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. *Am. J. Otolaryngol.*, 4, pp. 165–173.

[Butler et al., 1990] Butler, R., Humanski, R., & Musicant, A. (1990). Binaural and monaural localization of sound in two-dimensional space. *Perception*, 19, pp. 241–256.

[Campbell et al., 2006] Campbell, R., Doubell, T., Nodal, F., Schnupp, J., & King, A. (2006). Interaural timing cues do not contribute to map of space in the superior colliculus : a virtual acoustic space study. *J. Neurophysiol.*, 95, pp. 242–254.

[Capra et al., 2007] Capra, A., Fontana, S., Adiaensen, F., Farina, A., & Grenier, Y. (2007). Listening tests of the localization performance of stereodipole and ambisonic systems. In *123rd AES Convention*.

[Carlile et al., 2000] Carlile, S., Jin, C., & van Raad, V. (2000). Continuous virtual auditory space using HRTF interpolation : Acoustic and psychophysical errors. In *International Symposium on Multimedia Information Processing* Sydney, NSW, Australia.

[Carlile et al., 1997] Carlile, S., Leong, P., & Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hearing Research*, 114(1-2), pp. 179–196.

[Carlile & Pralong, 1994] Carlile, S. & Pralong, D. (1994). The location-dependent nature of perceptually salient features of the human head-related transfer functions. *J. Acous. Soc. Am.*, 95(6), pp. 3445–3459.

[Casin, 1999] Casin, P. (1999). *Analyse des données et des panels de données*. De Boeck Université.

[Chen, 2002] Chen, F. (2002). The reaction time for subjects to localize 3d sounds via headphones. In *AES 22nd International Conference*.

[Cheng & Wakefield, 2000] Cheng, C. & Wakefield, G. (2000). A tool for volumetric visualization and sonification of Head Related Transfer Functions (HRTFs). In *International Conference on Auditory Display 2000, Atlanta, GA*.

[Cheng & Wakefield, 1999] Cheng, C. I. & Wakefield, G. H. (1999). Spatial frequency response surfaces : An alternative vizualization tool for head-related transfer functions (HRTF's). In *ICASSP*.

[Choqueuse, 2004] Choqueuse, V. (2004). *Utilisation d'outils statistiques pour l'individualisation des HRTF*. Rapport de stage ingénieur, Université de Technologie de Troie.

[Chu, 2004] Chu, W. (2004). Deployment of head-related transfer function using all-pole filters and neural network-based storage devices. In *2004 IEEE International Joint Conference on Neural Networks*.

[Chuang, 1995] Chuang, C. (1995). *Study on HRTF clustering and synthesis with 3D sound applications*. PhD thesis, National Chiao Tung University.

[Constan & Hartmann, 2003] Constan, Z. A. & Hartmann, W. H. (2003). On the detection of dispersion in the head-related transfer function. *J. Acoust. Soc. Am.*, 114(4), pp. 998–1008.

[Cooper & Bauck, 1989] Cooper, D. H. & Bauck, J. L. (1989). Prospects for transaural recording. *J. Audio Eng. Soc.*, 37(1/2), pp. 3–19.

[Corteel & Nicol, 2003] Corteel, E. & Nicol, R. (2003). Listening room compensation for wave field synthesis. what can be done ? In *AES 23rd International Conference on Signal Processing in Audio Recording and Reproduction, Helsingor, 23-25 may*.

[Craven & Gerzon, 1977] Craven, P. G. & Gerzon, M. A. (1977). U. K. Patent 394,325.

[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*. PhD thesis, Université Paris 6.

[Daniel, 2009] Daniel, J. (2009). Evolving views on HOA : From technological to pragmatic concerns. In *Ambisonics Symposium 2009 (Graz)*.

[Daniel et al., 2003] Daniel, J., Nicol, R., & Moreau, S. (2003). Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. In *114th AES Convention*, number 5788.

[Darkner et al., 2006] Darkner, S., Vester-Christensen, M., Larsen, R., Nielsen, C., & Paulsen, R. (2006). Automated 3D Rigid Registration of Open 2D Manifolds. In *MICCAI 2006 Workshop From Statistical Atlases to Personalized Models*.

[de Bruijn, 2004] de Bruijn, W. (2004). *Application of Weve Field Synthesis in videoconferencing*. PhD thesis, Delft University of Technology, Delft, The Netherlands.

[Dellepiane et al., 2008] Dellepiane, M., Pietroni, N., Tsingos, N., Asselot, M., & Scopigno, R. (2008). Reconstructing head models from photographs for individualized 3D-audio processing. In *Pacific Graphics 2008*.

[Deutsch, 2009] Deutsch, D. (2009). *Musical Illusions (Encyclopedia of Neuroscience)*, chapter 5, (pp. 1159–1167). Academic Press, Oxford.

[Domnitz, 1973] Domnitz, R. H. (1973). The interaural time JND as a simultaneous function of interaural time and interaural amplitude. *J. Acoust. Soc. Am.*, 53, pp. 1549–1552.

[Domnitz & Colburn, 1977] Domnitz, R. H. & Colburn, H. S. (1977). Lateral position and interaural discrimination. *J. Acoust. Soc. Am.*, 61, pp. 1586–1598.

[Driscoll & Healy, 1994] Driscoll, J. R. & Healy, D. M. (1994). Computing Fourier Transforms and convolutions on the 2-sphere. *Adv. Appl. Math.*, 15, pp. 202–250.

[Duda et al., 1999] Duda, R. O., Avendano, C., & Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *ICASSP, Proceedings of the Acoustics, Speech, and Signal Processing*.

[Duda & Martens, 1998] Duda, R. O. & Martens, W. L. (1998). Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5), pp. 3048–3058.

[Durant & Wakefield, 2002] Durant, E. & Wakefield, G. (2002). Efficient model fitting using a genetic algorithm : Pole-zero approximations of HRTFs. *IEEE Transactions on speech and audio processing*, 10(1).

[Efron & Tibshirani, 1993] Efron, B. & Tibshirani, R. (1993). *An introduction to the Bootstrap*. NY Monographs on Statistics and Applied Probablilty Chapman and Hall.

[Engdegard et al., 2008] Engdegard, J., B. Resch a, d. C. F., Helmuth, O., Hilpert, J., Hoelzer, A., Terentiev, L., Breebaart, J., Koppens, J., Schuijers, E., & Oomen, W. (2008). Spatial audio coding object (saoc) - the upcoming mpeg standard on parametric object based audio coding. In *124th AES Convention, 2008 May 17-20, Amsterdam, The Netherlands*.

[Fahn & Lo, 2003] Fahn, C. & Lo, Y. (2003). On the clustering of head-related transfer functions used for 3d sound localization. *Journal of Information Science and Engineering*, 19, pp. 141–157.

[Faller & Baumgarte, 2002] Faller, C. & Baumgarte, F. (2002). Binaural cue coding : a novel and efficient representation of spatial audio. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP'02), Orlando, USA*.

[Farina & Ugolotti, 1999] Farina, A. & Ugolotti, E. (1999). Subjective comparison between stereo dipole and 3D ambisonic surround systems for automotive applications. In *AES 16th International Conference. on Spatial Sound Reproduction (Rovaniemi, Finland)*.

[Farrar, 1979a] Farrar, K. (1979a). Soundfield microphone. *Wireless World*, (pp. pp. 48–50).

[Farrar, 1979b] Farrar, K. (1979b). Soundfield microphone - 2. *Wireless World*, (pp. pp. 99–103).

[Faure, 2004] Faure, J. (2004). *Les systèmes de Head Tracking*. Technical report, France Telecom.

[Faure, 2005] Faure, J. (2005). *Evaluation de la synthèse binaurale dynamique*. Technical report, France Telecom.

[Fazi et al., 2008] Fazi, F. M., Nelson, P. A., Christensen, J. E. N., & Seo, J. (2008). Surround system based on three dimensional sound field reconstruction. In *Proceedings of the 125th AES Convention*.

[Fazi et al., 2009] Fazi, F. M., Nelson, P. A., & Potthast, R. (2009). Analogies and differences between three methods for sound field reproduction. In *Ambisonics Symposium*.

[Fels & Vorländer, 2009] Fels, J. & Vorländer, M. (2009). Antropometric parameters influencing head-related transfer functions. *Acta Acustica United With Acustica*, 95, pp. 331–342.

[Gardner, 1997] Gardner, W. G. (1997). *3-D audio using loudspeakers*. Kluwer Academic Publishers.

[Genuit, 1992] Genuit, K. (1992). Standardization of binaural measurement technique. In *French Congress on Acoustics (SFA), 1992, April 14-17, Arcachon*, volume 2 (pp. 405–407). : Journal de Physique IV, Colloque Ce, supplément au Journal de Physique III.

[Gerzon, 1980] Gerzon, M. A. (1980). Practical periphony : The reproduction of full-sphere sound. In *Proceedings of the A.E.S. 65th Convention*.

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.*, 33(11), pp. 859–871.

[Gerzon, 1992a] Gerzon, M. A. (1992a). General metatheory of auditory localisation. In *Proceedings of the A.E.S. 92nd Convention*.

[Gerzon, 1992b] Gerzon, M. A. (1992b). Optimum reproduction matrices for multispeaker stereo. *J. Audio Eng. Soc.*, 40(7/8), pp. 571–589.

[Greff & Katz, 2007] Greff, R. & Katz, B. (2007). Round robin comparison of HRTF simulation results : Preliminary results. In *123rd AES Convention, 2007 October 5-8, New York, NY, USA*.

[Guastavino & Katz, 2004] Guastavino, C. & Katz, B. (2004). Perceptual evaluation of multi-dimensional spatial audio reproduction. *J. Acoust. Soc. Am.*, 116, pp. 1105–11115.

[Guastavino et al., 2007] Guastavino, C., Larcher, V., Catusseau, G., & Boussard, P. (2007). Spatial audio quality evaluation : Comparing transaural, ambisonics and stereo. In *Proceedings of the 13th International Conference on Auditory Display*.

[Guillon, 2007] Guillon, P. (2007). *Rapport d'avancement de thèse : synthèse binaurale, indices spectraux, individualisation des HRTF*. Technical report, France Telecom.

[Guillon, 2009] Guillon, P. (2009). *Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*. PhD thesis, Université du Maine, Le Mans, France.

[Guillon et al., 2008] Guillon, P., Guignard, T., & Nicol, R. (2008). Head-Related Transfer Function customization by frequency scaling and rotation shift based on a new morphological matching method. In *Proc. 125th Convention of the Audio Eng. Soc.*

[Guillon & Nicol,] Guillon, P. & Nicol, R. Procédé et dispositif pour la détermination de fonctions de transfert de type HRTF. Brevet WO2009106783.

[Guillon & Nicol, 2008] Guillon, P. & Nicol, R. (2008). Head-related transfer function reconstruction from sparse measurements considering a priori knowledge from database analysis : a pattern recognition approach. In *125th AES Convention, 2008 October 2-5, San Fransisco, CA, USA*.

[Hafter & Maio, 1975] Hafter, E. R. & Maio, J. D. (1975). Difference thresholds for interaural delay. *J. Acoust. Soc. Am.*, 57, pp. 181–187.

[Han, 1994] Han, H. L. (1994). Measuring a dummy head in search of pinna cues. *J. Audio. Eng. Soc.*, 42(1/2), pp. 15–36.

[Hartmann & Wittenberg, 1996] Hartmann, W. M. & Wittenberg, A. (1996). On the externalization of sound images. *J. Acous. Soc. Am.*, 99, pp. 3678–3688.

[Hartung et al., 1999] Hartung, K., Braasch, J., & Sterbing, S. (1999). Comparison of different interpolation methods for the interpolation of head-related transfer functions. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction. Rovaniemi, Finland*.

[Hebrank & Wright, 1974] Hebrank, J. & Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *J. Acous. Soc. Am.*, 56(6), pp. 1829–1834.

[Hershkowitz & Durlach, 1969] Hershkowitz, R. M. & Durlach, N. I. (1969). Interaural time and amplitude JNDs for a 500-hz tone. *J. Acoust. Soc. Am.*, 46, pp. 1464–1468.

[Hetherington et al., 2003] Hetherington, C., Tew, A., & Tao, Y. (2003). Three-dimensional elliptic fourier methods for the parameterization of human pinna shape. In *ICASSP*.

[Hiranaka & Yamasaki, 1983] Hiranaka, Y. & Yamasaki, H. (1983). Envelope representations of pinna impulse response relating to three-dimensional localization of sound sources. *J. Acoust. Soc. Am.*, 73(1), pp. 291–296.

[Hoffmann & Moller, 2008a] Hoffmann, P. F. & Moller, H. (2008a). Audibility of differences in adjacent head-related transfer function. *Acta Acustica United With Acustica*, 94, pp. 45–54.

[Hoffmann & Moller, 2008b] Hoffmann, P. F. & Moller, H. (2008b). Some observations on sensitivity to HRTF magnitude. *J. Audio Eng. Soc.*, 56(11), pp. 972–982.

[Hofman et al., 1998] Hofman, P., Riswick, J. V., & Opstal, A. V. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, 1(5), pp. 417–421.

[Hofman & Van Opstal, 2002] Hofman, P. M. & Van Opstal, A. J. (2002). Bayesian reconstruction of sound localization cues from responses to random spectra. *Biol. Cybern.*, 86, pp. 305–316.

[Hofman & Van Opstal, 2003] Hofman, P. M. & Van Opstal, A. J. (2003). Binaural weighting of pinna cues in human sound localization. *Exp. Brain. Res.*, 148, pp. 458–470.

[Hohle, 1965] Hohle, R. (1965). Inferred components of reaction times as functions of foreperiod duration. *J. Exp. Psychol.*, 69.

[Humanski & Butler, 1988] Humanski, R. & Butler, R. (1988). The contribution of the near and far ear toward localization of sound in the sagittal plane. *J. Acous. Soc. Am.*, 83(6), pp. 2300–2310.

[Huopaniemi & Smith, 1999] Huopaniemi, J. & Smith, J. O. (1999). Spectral and time-domain preprocessing and the choice of modelling error criteria for binaural digital filters. In *AES 16th International Conference*.

[Hwang et al., 2008] Hwang, S., Park, Y., & Park, Y.-S. (2008). Modeling and customization of head-related impulse response based on general basis functions in time domain. *Acta Acustica United With Acustica*, 94, pp. 965–980.

[Ianaga et al., 1995] Ianaga, K., Yamada, Y., & Koizumi, H. (1995). Headphone system with out-of-head localisation applying dynamic HRTF (Head Related Transfer Function). In *Proceedings of the 98th Convention of the Audio Engeering Society*.

[Iida & Itoh, 2006] Iida, K. & Itoh, M. (2006). A novel head-related transfer function model based on spectral and interaural difference cues. In *The 9th Western Pacific Acoustics Conference, WESPAC IX 2006*.

[Inoue et al., 2005] Inoue, N., Nishino, T., Itou, K., & Takeda, K. (2005). HRTF modeling using physical features. In *Forum Acusticum 2005* Budapest, Hungary.

[Iwaya, 2006] Iwaya, Y. (2006). Individualization of head-related transfer functions with tournament-style listening test : Listening with other's ears. *Acoust. Sci. & Tech.*, 6, pp. 340–343.

[Iwaya & Suzuki, 2008] Iwaya, Y. & Suzuki, Y. (2008). Numerical analysis of the effects of pinna shape and position on the characteristics of head-related transfer functions. *J. Acoust. Soc. Am.*, 123, pp. 3297.

[Jenison, 1995] Jenison, R. (1995). A spherical basis function neural network for pole-zero modeling of head-related transfer functions. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Jenison & Fissell, 1996] Jenison, R. & Fissell, K. (1996). A spherical basis function neural network for modeling auditory space. *Neural Computation*, 8, pp. 115–128.

[Jessel, 1973] Jessel, M. (1973). *Acoustique théorique, propagation et holophonie*. Masson, Paris.

[Jin et al., 2000] Jin, C., Leong, P., Leung, J., Corderoy, A., & Carlile, S. (2000). Enabling individualized virtual auditory space using morphological measurements. In *IEEE Pacific-Rim Conference on Multimedia - International Symposium on Multimedia Information Processing* (pp. 220–223).

[Jot et al., 1995] Jot, J. M., Larcher, V., & Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. In *AES 98th Convention*.

[Jouhaneau, 1994] Jouhaneau, J. (1994). *Notions élémentaires d'acoustique - Electroacoustique : Les microphones et les haut-parleurs*. Collection "'Acoustique Appliquée". Lavoisier.

[Kahana, 2000] Kahana, Y. (2000). *Numerical modelling of the Head-Related Transfer Function*. PhD thesis, University of Southampton.

[Katz, 1998] Katz, B. F. G. (1998). *Measurement and calculation of individual Head-Related Transfer Functions using a Boundary Element Model Including the measurement and effect of skin and air impedance*. PhD thesis, Pennsylvania State University.

[Kendall, 1995] Kendall, G. S. (1995). The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19, pp. 71–87.

[Kim & Choi, 2005] Kim, S.-M. & Choi, W. (2005). On the externalisation of virtual sound images in headphone reproduction : A wiener filter approach. *J. Acoust. Soc. Am.*, 117(6), pp. 3657–3665.

[King & Oldfield, 1997] King, R. & Oldfield, S. (1997). The impact of signal bandwidth on auditory localization : Implications for the design of three-dimensional audio displays. *Human Factors*, 39(2), pp. 287–295.

[Kirkeby et al., 1997] Kirkeby, O., Nelson, P. A., & Hamada, H. (1997). The stereo dipole – a virtual source imaging system using two closely spaced loudspeakers. In *102nd AES Convention*.

[Kistler & Wightman, 1992] Kistler, D. J. & Wightman, F. L. (1992). A model of head related transfer function based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91, pp. 1637–1647.

[Klumpp & Eady, 1956] Klumpp, R. G. & Eady, H. R. (1956). Some measurement of interaural time difference thresholds. *J. Acous. Soc. Am.*, 28(5), pp. 859–860.

[Kohonen, 1995] Kohonen, T. (1995). *Self-organizing maps*, volume 30 of *Springer Series in Information Science*. Springer.

[Kuhn, 1977] Kuhn, G. (1977). Model for the interaural time difference in the azimuthal plane. *J. Acoust. Soc. Am.*, 62(1), pp. 157–167.

[Kulkarni & Colburn, 1998] Kulkarni, A. & Colburn, H. (1998). Role of spectral detail in sound-source localization. *Nature*, (pp. pp. 747–749).

[Kulkarni & Colburn, 2000] Kulkarni, A. & Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *J. Acoust. Soc. Am.*, 107(2), pp. 1071–1074.

[Kulkarni et al., 1999] Kulkarni, A., Isabell, S., & Colburn, H. (1999). Sensitivity of human subjects to head-related transfer function phase spectra. *J. Acous. Soc. Am.*, 105(5), pp. 2821–2840.

[Kulkarni et al., 1995] Kulkarni, A., Isabelle, S. K., & Colburn, H. S. (1995). On the minimum-phase approximation of head-related transfer functions. In *IEEE WASPAA*.

[Lacouture & Cousineau, 2008] Lacouture, Y. & Cousineau, D. (2008). How to use matlab to fit the ex-gaussian and other probability functions to a distribution of response times. Tutorials in Quantitative Methods for Psychology.

[Langendijk & Bronkhorst, 2002] Langendijk, E. & Bronkhorst, A. (2002). Contribution of spectral cues to human sound localization. *J. Acous. Soc. Am.*, 112(4), pp. 1583–1596.

[Langendijk & Bronkhorst, 2000] Langendijk, E. H. A. & Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *J. Acoust. Soc. Am.*, 107(1), pp. 528–537.

[Larcher, 2001] Larcher, V. (2001). *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Université de Paris VI.

[Leipp, 1997] Leipp, E. (1997). *La machine à écouter : Essai de psycho-acoustique*. Dunod.

[Lemaire & Clérot, 2002] Lemaire, V. & Clérot, F. (2002). SOM-based clustering for on-line fraud behaviour classification : a case study. In *Fuzzy Systems and Knowledge Discovery (FSKD)*.

[Lemaire et al., 2005] Lemaire, V., Clérot, F., Busson, S., Nicol, R., & Choqueuse, V. (2005). Individualized HRTFs from few measurements : A statistical learning approach. In *International Joint Conference on Neural Networks IJCNN 2005*.

[Levitt, 1971] Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acous. Soc. Am.*, 49(2), pp. 467–477.

[Lo, 1998] Lo, Y. (1998). *A clustering and synthesis method for the head-related transfer functions in the minimum-phase approximation*. PhD thesis, National Taiwan University of Science and Technology.

[Mackensen, 2004] Mackensen, P. (2004). *Auditive Localization. Head movements, an additional cue in Localization*. Technical report, TU Berlin.

[Macpherson, 1996] Macpherson, E. (1996). Effect of source spectrum iregularity and uncertainty on sound localization. *J. Acous. Soc. Am.*, 99, pp. 2515–2529.

[Macpherson, 1998] Macpherson, E. (1998). *Spectral cue processing in the auditory localization of sounds with wideband non-flat spectra*. PhD thesis, University of Wisconsin-Madison, USA.

[Macpherson & Middlebrooks, 2002] Macpherson, E. & Middlebrooks, J. (2002). Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited. *J. Acoust. Soc. Am.*, 111(5), pp. 2219–2236.

[Majdak et al., 2007] Majdak, P., Balazs, P., & Laback, B. (2007). Multiple exponential sweep method for fast measurement of head related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8), pp. 623–637.

[Maki & Furukawa, 2005] Maki, K. & Furukawa, S. (2005). Reducing individual differences in the external-ear transfer functions of the mongolian gerbil. *J. Acous. Soc. Am.*, 118(4), pp. 2392–2404.

[Martens, 2001] Martens, W. L. (2001). Uses and misuses of psychophysical methods in the evaluation of spatial sound reproduction. In *Audio. Eng. Soc. 110th Convention* Amsterdam, The Netherlands.

[Martens, 2002] Martens, W. L. (2002). Rapid psychophysical calibration using bisection scaling for individualized control of source elevation in auditory display. In *Proc. Int. Conf. on Auditory Display, ICAD 2002* Kyoto, Japan.

[Martin, 2006] Martin, G. (2006). Microphone techniques for stereo and multichannel. In *Tutorial Seminar* : Audio Eng. Soc.

[Martin & McAnally, 2007] Martin, R. & McAnally, K. (2007). *Interpolation of Head-Related Transfer Functions*. Technical Report DSTO-RR-0323, Australian Government - Department of Defence.

[Marvit et al., 2003] Marvit, P., Florentine, M., & Buus, S. (2003). A comparative of psychophysical procedures for level-discrimination thresholds. *J. Acoust. Soc. Am.*, 113, pp. 3348–3361.

[HEAD acoustics, 2003] HEAD acoustics (2003). *Notes on performing and playing back binaural measurements*. Technical report, HEAD Application Note.

[HEAD acoustics, 2005] HEAD acoustics (2005). *Artificial head equalization*. Technical report, Brochure.

[HEAD acoustics, 2006] HEAD acoustics (2006). *Binaural measurement, analysis and playback*. Technical report, Application Note.

[IOSONO Sound, 2010] IOSONO Sound (2010). Iosono (http ://www.iosono-sound.com/).

[Møller, 1992] Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, 36, pp. 171–218.

[McAnally & Martin, 2002] McAnally, K. I. & Martin, R. L. (2002). Variability in the headphone-to-ear-canal transfer function. *J. Audio Eng. Soc.*, 50(4), pp. 263–266.

[Middlebrooks, 1992] Middlebrooks, J. (1992). Narrow-band sound localization related external ears acoustics. *J. Acoust. Soc. Am.*, 92(5), pp. 2607–2624.

[Middlebrooks, 1999a] Middlebrooks, J. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acous. Soc. Am.*, 106(3), pp. 1480–1492.

[Middlebrooks, 1999b] Middlebrooks, J. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J. Acous. Soc. Am.*, 106(3), pp. 1493–1510.

[Middlebrooks et al., 1989] Middlebrooks, J., Makous, J., & Green, D. (1989). Directional sensivity of sound pressure levels in the human ear canal. *J. Acous. Soc. Am.*, 86(1), pp. 89–108.

[Middlebrooks et al., 2000] Middlebrooks, J. C., MacPherson, E. A., & Onsan, Z. A. (2000). Psychophysical customization of directional transfer functions for virtual sound localization. *J. Acous. Soc. Am.*, 108(6), pp. 3088–3091.

[Mills, 1972] Mills, A. (1972). *Foundations of modern auditory theory*, chapter Auditory localisation. New York : Academic Press.

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *J. Acoust. Soc. Am.*, 30, pp. 237–248.

[Minnaar et al., 1999] Minnaar, P., Christensen, F., Mø ller, H., Olesen, S., & Plogsties, J. (1999). Audibility of all-pass components in binaural synthesis. In *Proceedings of the 106th Convention of the Audio Engineering Society*, number 4911.

[Minnaar et al., 2000] Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., & Möller, H. (2000). The interaural time difference in binaural synthesis. In *AES 108th Convention, preprint n° 5133.*

[Moore, 2009] Moore, A. H. (2009). *Towards the perception of externalised auditory images using binaural technology.* PhD thesis, University of York (Department of Electronics).

[Moore et al., 1989] Moore, B., Oldfield, S., & Dooley, G. (1989). Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *J. Acoust. Soc. Am.*, 85(2), pp. 820–835.

[Moreau, 2006] Moreau, S. (2006). *Etude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de la distance.* PhD thesis, Université du Maine, Le Mans, France.

[Morimoto, 2001] Morimoto, M. (2001). The contribution of two ears to the perception of vertical angle in sagittal planes. *J. Acous. Soc. Am.*, 109(4), pp. 1596–1603.

[Morimoto et al., 2003] Morimoto, M., Yairi, M., Iida, K., & Itoh, M. (2003). The role of low frequency components in median plane localization. *Acoust. Sci. & Tech.*, 24(2), pp. 76–82.

[Musicant & Butler, 1984] Musicant, A. & Butler, R. (1984). The influence of pinnae-based spectral cues on sound localization. *J. Acoust. Soc. Am.*, 75(4), pp. 1195–1200.

[Musicant et al., 1990] Musicant, A., Chan, J., & Hind, J. (1990). Direction-dependent spectral properties of cat external ear : New data and cross-species comparisons. *J. Acous. Soc. Am.*, 87(2), pp. 757–781.

[Nagle, 2008] Nagle, A. (2008). *Enrichissement de la conférence audio en voix sur IP au travers de l'amélioration de la qualité et de la spatialisation sonore.* PhD thesis, Telecom Paris.

[Nam et al., 2008] Nam, J., Abel, J., & Smith III, J. (2008). A method for estimating interaural time difference for binaural synthesis. In *Proceedings of the 125th Convention of the Audio Engineering Society*, number 7612.

[Nicol, 1999] Nicol, R. (1999). *Restitution sonore spatialisée sur une zone étendue : Application à la téléprésence*. PhD thesis, Université du Maine.

[Nicol et al., 2008] Nicol, R., Daniel, J., Emerit, M., Pallone, G., Virette, D., Chetry, N., Guillon, P., & Bertet, S. (2008). Le son 3d dans toutes ses dimensions. *Acoustique & Technique*, 52.

[Nicol et al., 2006] Nicol, R., Lemaire, V., Bondu, A., & Busson, S. (2006). Looking for a relevant similarity criterion for HRTF clustering : a comparative study. In *AES 120ème Convention*.

[Opstal & Esch, 2003] Opstal, A. V. & Esch, T. V. (2003). Estimating spectral cues underlying human sound localization. *NAG-journaal*, 168, pp. 1–10.

[Palmer & Russell, 1986] Palmer, A. & Russell, I. (1986). Phase locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, 24, pp. 1–15.

[Pernaux et al., 1998] Pernaux, J., Boussard, P., & Jot, J. (1998). Virtual sound source positioning and mixing in 5.1. implementation on the real-time system genesis. In *DAFX98, November 19-21, Barcelona, Spain*.

[Pernaux et al., 2004a] Pernaux, J., Nicol, R., & Emerit, M. (2004a). Procédé et système d'obtention automatisée de fonctions de transfert acoustiques associées à la morphologie d'un individu. Brevet FR2851878.

[Pernaux et al., 2004b] Pernaux, J., Nost, G. L., & Bouguet, A. (2004b). Procédé de mesure de fonctions de transfert acoustiques associées à la morphologie d'un individu. Brevet FR2851877.

[Pernaux, 2003] Pernaux, J. M. (2003). *Spatialisation du son par les techniques binaurales : Application aux services de télécommunications*. PhD thesis, Institut National Polytechnique de Grenoble.

[Plaskota & Dobrucki, 2008] Plaskota, P. & Dobrucki, A. (2008). The influence of pinna position on head-related transfer function. *J. Acoust. Soc. Am.*, 123(5), pp. 3724.

[Plogsties et al., 2000] Plogsties, J., Minnaar, P., Olesen, S. K., Christensen, F., & Möller, H. (2000). Audibility of all-pass components in head-related transfer functions. In *AES 108th Convention*.

[Poletti, 2005] Poletti, A. (2005). Three-dimensional surround sound systems based on spherical harmonics. *J. Audio Eng. Soc.*, 53(11), pp. 1004–1024.

[Pralong & Carlile, 1996] Pralong, D. & Carlile, S. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *J. Acoust. Soc. Am.*, 100(6), pp. 3785–3793.

[Pulkki & Hirvonen, 2005] Pulkki, V. & Hirvonen, T. (2005). Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions onSpeech and Audio Processing*, 13, pp. 105–119.

[Pulkki & Lokki, 1998] Pulkki, V. & Lokki, T. (1998). Creating auditory displays with multiple loudspeakers using VBAP : A case study with DIVA project. In *ICAD'98*.

[Raatgever, 1980] Raatgever, J. (1980). *On the binaural processing of stimuli with different interaural phase relations*. PhD thesis, Dutch Efficiency Bureau, Pijnacker, Delft, The Netherlands.

[Rakerd, 1999] Rakerd, B. (1999). Identification and localization of sound sources in the median sagittal plane. *J. Acous. Soc. Am.*, 106(5), pp. 2812–2820.

[Ratcliff, 1978] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), pp. 59–108.

[Renard, 2000] Renard, C. (2000). *Analyse objective et subjective d'une technique de rendu sonore 2D sur une zone d'écoute étendue, l'holophonie, en vue de réaliser un mur de téléprésence.* Technical report, Université du Maine.

[Rodriguez & Ramirez, 2005a] Rodriguez, S. G. & Ramirez, M. A. (2005a). Extracting and modeling approximated pinna-based transfer functions from HRTF data. In *Proc. Int. Conf. on Auditory Display, ICAD 2005* Limerick, Ireland.

[Rodriguez & Ramirez, 2005b] Rodriguez, S. G. & Ramirez, M. A. (2005b). Linear relationships between spectral characteristics and anthropometry of the external ear. In *Proc. Int. Conf. on Auditory Display, ICAD 2005* (pp. 336–339). Limerick, Ireland.

[Rodriguez & Ramirez, 2005c] Rodriguez, S. G. & Ramirez, M. A. (2005c). HRTF individualization by solving the least squares problem. In *Audio. Eng. Soc. 118th Convention* Barcelona, Spain.

[Rodriguez Soria & Ramirez, 2004] Rodriguez Soria, S. G. & Ramirez, M. A. (2004). Feature identification techniques for HRTF individualization. In *Proc. Int. Workshop on Telecommunications, INATEL 2004* (pp. 188–193). Santa Rita do Sapucaí, Brasil.

[Rueff, 2010] Rueff, P. (2010). www.binaural.fr.

[Rueff & Blum, 2003] Rueff, P. & Blum, A. (2003). Association Omnihead.

[Rumsey, 2001] Rumsey, F. (2001). *Spatial audio.* Focal Press.

[Runkle et al., 2000] Runkle, P., Yendiki, A., & Wakefield, G. H. (2000). Active sensory tuning for immersive spatialized audio. In *Proc. Int. Conf. on Auditory Display, ICAD 2000* Atlanta, Georgia, USA.

[Savel et al., 2006] Savel, S., Meunier, S., & Rabau, G. (2006). Individual differences and systematic errors in human sound localization of real sources. In *8ème Congrès Français d'Acoustique, Tours, France.*

[Schobben & Aarts, 2005] Schobben, D. & Aarts, R. (2005). Perzonalized multi-channel headphone sound reproduction based on active noise cancellation. *Acta Acustica united with Acustica*, 91, pp. 440–450.

[Seeber & Fastl, 2003] Seeber, B. U. & Fastl, H. (2003). Subjective selection of non-individual head-related transfer function. In *Proc. 2003 International Conference on Auditory Display* (pp. 259–262). Boston, MA, USA.

[Shaw & Teranishi, 1968] Shaw, E. A. G. & Teranishi, R. (1968). Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *J. Acoust. Soc. Am.*, 44(1), pp. 240–249.

[Shimada et al., 1994] Shimada, S., Hayashi, N., & Hayashi, S. (1994). A clustering method for sound localization transfer functions. *J. Audio. Eng. Soc.*, 42(7/8), pp. 577–584.

[Solvang, 2009] Solvang, A. (2009). *Representation of High Quality Spatial Audio.* PhD thesis, Norwegian University of Science and Technology.

[Sonoda et al., 2001] Sonoda, S., Mori, M., & Goishi, A. (2001). Pattern of localisation error in patients with stroke to sound processed by a binaural sound space processor. *J. Neurol. Neurosurg. Psychiatry*, 70, pp. 43–49.

[Sontacchi et al., 2002] Sontacchi, A., Noisternig, M., Majdak, P., & Höldrich, R. (2002). Subjective validation of perception properties in binaural sound reproduction systems. In *21st Int. Audio Eng. Soc. Conference (St Petersburg, Russia).*

[Spors, 2009] Spors, S. (2009). Comparison of wave field synthesis and higher-order ambisonics. In *Ambisonics Symposium.*

[Start, 1997] Start, E. (1997). *Direct sound enhancement by Wave Field Synthesis.* PhD thesis, Delft University of Technology, Delft, The Netherlands.

[Sterbing et al., 2003] Sterbing, S., Hartung, K., & Hoffman, K.-P. (2003). Spatial tuning to virtual sounds in the inferior colliculus of the guinea pig. *J. Neurophysiol.*, 90, pp. 2648–2659.

[Tan & Gan, 1998] Tan, C.-J. & Gan, W.-S. (1998). User-defined spectral manipulation of HRTF for improved localisation in 3D sound systems. *Electronic Letters*, 34(25), pp. 2387–2389.

[Tao et al., 2002] Tao, Y., Tew, A., & Porter, S. (2002). The differential pressure synthesis method for estimating acoustic pressures on human heads. In *112th AES Convention, 2002 May 10-13, Munich, Germany.*

[Theile, 2001] Theile, G. (2001). *Multichannel Natural Music Recoding Based on Psychoacoustic Principles.* Technical report, IRT.

[Toole, 1970] Toole, F. E. (1970). In-head locatedness of acoustic images. *J. Acous. Soc. Am.*, 48, pp. 943–949.

[Verheijen, 1998] Verheijen, E. (1998). *Sound reproduction by Wave Field Synthesis.* PhD thesis, Delft University of Technology, Delft, The Netherlands.

[VNoise, STS] VNoise (STS). http ://www.sts-soft.com/vnoise.aspx.

[Vogel, 1993] Vogel, P. (1993). *Application of Wave Field Synthesis in room acoustics.* PhD thesis, Delft University of Technology, Delft, The Netherlands.

[Vovor, 2005] Vovor, P. (2005). *Utilisation d'outils statistiques pour l'individualisation de HRTF.* Rapport de stage master ATIAM, Université Paris IV.

[Wanrooij & Opstal, 2006] Wanrooij, M. V. & Opstal, A. J. V. (2006). Sound localization under perturbed binaural hearing. *J. Neurophysiol.*, 97, pp. 715–726.

[Watkins, 1978] Watkins, A. (1978). Psychoacoustical aspects of synthesized vertical locale cues. *J. Acous. Soc. Am.*, 63(4), pp. 1152–1165.

[Wenzel, 1999] Wenzel, E. (1999). Effects of increasing system latency on localization of virtual sounds. In *Proc. 16th Conference of the Audio Eng. Soc. (Rovaniemi, Finland).*

[Wenzel, 1995] Wenzel, E. M. (1995). The relative contribution of interaural time and magnitude cues to dynamic sound localization. In *Proceedings of the IEEE Workshop on Applications of Signal Procesing to Audio & Acoustics (WASPAA).*

[Wenzel et al., 1993] Wenzel, E. M., Kistler, D. J. ., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94(1), pp. 111–123.

[Wightman & Kistler, 1989a] Wightman, F. & Kistler, D. (1989a). Headphone simulation of free-field listening. I : Stimulus synthesis. *J. Acoust. Soc. Am.*, 85, pp. 858–867.

[Wightman & Kistler, 1989b] Wightman, F. & Kistler, D. (1989b). Headphone simulation of free-field listening. II : Psychophysical validation. *J. Acoust. Soc. Am.*, 85, pp. 868–878.

[Wightman & Kistler, 1992] Wightman, F. L. & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3), pp. 2149–2162.

[Wightman & Kistler, 1997] Wightman, F. L. & Kistler, D. J. (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter Factors affecting the relative salience of sound localization cues, (pp. 1–23). Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

[Woodworth & Schlosberg, 1954] Woodworth, R. & Schlosberg, H. (1954). *Experimental Psychology*. New York : Holt.

[Wu et al., 1998] Wu, Z., Weng, T., & Wang, W. (1998). Neural network model of binaural hearing based on spatial feature extraction of the head related transfer function. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 20.

[Xiang et al., 1993] Xiang, N., Genuit, K., & Gierlich, H. W. (1993). Investigations on a new reproduction procedure for binaural recordings. In *Proceedings of the 95th AES Convention, 1993 October 7-10, New York*.

[Yairi et al., 2008] Yairi, S., Iwaya, Y., & Suzuki, Y. (2008). Influence of large system latency of virtual display on behavior of head movement in sound localization task. *Acta Acustica united with Acustica*, 94, pp. 1016–1023.

[Zacharov & Koivuniemi, 2001] Zacharov, N. & Koivuniemi, K. (2001). Audio descriptive analysis and mapping of spatial sound displays. In *Proceedings of the 2001 International Conference on Auditory Display*.

[Zotkin et al., 2002] Zotkin, D., Duraiswami, R., & Davis, L. (2002). Customizable auditory displays. In *Proc. Int. Conf. on Auditory Display, ICAD 2002* Kyoto, Japan.

[Zotkin et al., 2003] Zotkin, D., Hwang, J., Duraiswami, R., & Davis, L. (2003). HRTF personalization using anthropometric measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2003* New Patlz, NY, USA.

[Zotkin et al., 2004] Zotkin, D. N., Duraiswami, R., & Gumerov, E. G. N. (2004). *Fast head related transfer function measurement via reciprocity*. Technical Report Technical report CS-4620 and UMIACS-20004-62, University of Maryland, Computer Science and UMIACS.

[Zwislocki & Feldman, 1956] Zwislocki, J. & Feldman, R. (1956). Just noticeable differences in dichotic phase. *J. Acous. Soc. Am*, 28, pp. 860.

Abstract

Representation and perception of Virtual Auditory Space
Rozenn Nicol

A Virtual Auditory Space (VAS) is a virtual sound scene which is composed of several sound sources which only exist in the perceptive space of the listener. This space is created by technologies of sound spatialization (such as : stereophony, binaural technology, Wave Field Synthesis or Higher Order Ambisonics) which relies on models for representing the sound scene. Modelling is the first issue to be investigated : it concerns the steps of recording and rendering the spatial information. The concept of spatial audio format (as well as the related topics concerning format adaptation and spatial audio coding) is implicit. The opposite issue is the perception of the VAS, i.e. how the listener perceives the virtual sound sources. This document provides food for thought about all these issues. In addition to an overview of current knowledge, two questions are examined in details. The first question concerns spatialization technologies for multi loudspeaker array, focussing on Wave Field Synthesis (WFS) and Higher Order Ambisonics (HOA). It is shown how to derive feasable systems from the theoretical equations. A unified description allows one to point out the convergence between the two technologies and opens a comparative study. The second question deals with the adaptation of sound spatialization to individual (i.e. mono listener) and handheld devices, which implyies rendering over headphones. It is based on binaural technology which consists in reproducing the acoustic signals at the entrance of the listener's ear. This technology relies on the reproduction of the localization cues which result from the interaction of the acoustic wave with the listener's body and are therefore strongly individual. It is presented how to model these localization cues, considering the temporal information (i.e. Interaural Time Difference or ITD) and the spectral information (i.e. the Spectral Cues or SC), and how to customize them for one particular individual.