Thèse de doctorat de l'Université Paris 6

Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia

Jérôme DANIEL

31 juillet 2001

Résumé

Ce travail de thèse s'intéresse à la représentation de champ acoustique pour la restitution spatialisée sur haut-parleurs ou au casque, appliquée au large domaine des applications multimédias, dont les nouvelles applications de navigation 3D dans des scènes virtuelles composites sur Internet. Ce domaine allie la *reproduction* spatialisée de champs sonores complexes préexistants (*e.g.* sous forme multi-canal "5.1") et un travail de *spatialisation* à part entière (pan-pot 3D et effet de salle). De plus en plus, les spécifications de ce type d'application sont caractérisées par la variabilité d'un grand nombre de paramètres: débit de transmission, ressources de l'utilisateur (CPU et dispositif de restitution), conditions d'écoute (individuelle ou collective), diversité des matériels sonores et audio-visuels brassés, point de vue et position des objets dans la scène virtuelle (interactivité). La question de la représentation – ensemble de signaux à restituer directement ou bien à décoder au préalable – intéresse à la fois la transmission (objectif de concision), et l'étape intermédiaire de spatialisation (encodage global pour une factorisation des opérations en aval).

Nous avons choisi d'approfondir l'approche ambisonique, basée sur une *décomposition du champ acoustique en harmoniques sphériques* centrée sur le point de vue de l'auditeur. Elle est restée longtemps connue sous une forme restreinte au premier ordre, qui réalise un encodage directionnel minimal du champ sonore à travers quatre composantes (*B-format*): W (pression) et X, Y, Z (gradient de pression), offrant une manipulation aisée du champ (rotations). Moyennant un décodage optimisé en fonction des conditions d'écoute (idéale-centrée ou collective-excentrée), un *rendu cohérent et homogène* de l'espace sonore peut être obtenu sur divers dispositifs panoramiques (2D) ou périphoniques (3D). Cette *restitution "à géométrie variable*" s'étend à la paire d'écouteurs ou de haut-parleurs grâce aux techniques binaurales (haut-parleurs virtuels). Avec la prise en compte de composantes d'ordres supérieurs – à laquelle encore peu de recherche avait été consacrée –, s'ajoute à toutes ces propriétés la notion de *représentation "à résolution variable*" (*scalabilité*), exploitable en fonction du nombre de haut-parleurs et/ou de la capacité de transmission.

Nous présentons d'abord les fondements acoustiques et psychoacoustiques, et une revue critique des stratégies de spatialisation (stéréo, *surround*, binaural, transaural et nouvelles variantes). Nous explicitons le lien intrinsèque entre représentation ambisonique et caractéristiques locale (vecteur vélocité $\vec{V}$) et globale (vecteur énergie $\vec{E}$) de propagation du champ restitué – compte-tenu du dispositif –, ainsi que les relations de prédiction entre ces dernières et l'effet de localisation selon la mobilité de la tête. Les théories de localisation impliquées dans le décodage ambisonique (Gerzon) sont ainsi approfondies. Cette démarche étendue à l'ensemble des approches recensées souligne l'intérêt d'Ambisonic.

La généralisation d'Ambisonic à tous les ordres touche tous les aspects évoqués à l'ordre 1, notamment le formalisme d'encodage et les principes de décodage (2D et 3D). Après développement de la notion d'échantillonnage directionnel de la base d'harmoniques sphériques (applicable aussi à la prise de son), les trois formes de décodage originelles (Gerzon, Malham) sont généralisées en trois familles de solutions, à appliquer selon les conditions d'écoute. Des évaluations objectives étayées par des écoutes informelles confirment l'apport des ordres supérieurs et des solutions optimisées. Cet apport se manifeste sur le plan acoustique à travers l'expansion radiale de la reconstruction du champ et la propagation globale ($\vec{E}$), et sur le plan perceptif à travers la précision et la robustesse des images sonores même en conditions non-idéales, ainsi que la préservation des impressions spatiales (séparation latérale, défectueuse à l'ordre 1).

La partie mise en oeuvre et expérimentation couvre, outre Ambisonic (expérimenté jusqu'à l'ordre 2), d'autres techniques de spatialisation (pan-pot, binaural, transaural, plus réverbération artificielle), intégrées à une interface sur PC. Ambisonic a pu ainsi être appliqué avec succès à la manipulation temps-réel et le mélange des sources (mono, multi-canal, B-format), et comparé ou combiné avec les autres techniques, en mode binaural comme sur haut-parleurs. Cet outil pourrait servir à une validation subjective complète de l'approche ambisonique et des théories sous-jacentes.

L'approche ambisonique apporte une réponse globale très satisfaisante aux enjeux de départ, bien que pour certains problèmes spécifiques – matriçage/décodage *surround*, synthèse binaurale performante d'une scène complexe – elle ne rivalise pas avec d'autres stratégies évoluées. Son extension aux ordres supérieurs intéresse de nombreux domaines et devrait connaître un essor grâce aux recherches et projets en cours.

Mots-clés

Spatialisation - Son 3D - *Surround* - Multimédia - Navigation 3D - Ambisonic(s) - *B-format* - Théorie de la localisation - Vecteur vélocité - Vecteur énergie - Décodage psychoacoustique - Représentation de champ acoustique - Décomposition en harmoniques sphériques - Echantillonnage directionnel - *Scalabilité*

Abstract

This thesis deals with acoustic field representation for spatial reproduction over loudspeakers or headphones, as applied to the large multimedia domain, including new applications for browsing in virtual composite 3D-scenes on the Internet. This domain combines *spatial reproduction* of pre-existent complex sound fields (*e.g.* in the "5.1" multi-channel format) and *constructive spatialisation tasks* (3D pan-pot and room effects). This kind of application is increasingly characterised by the variability of a number of parameters: transmission bit-rate, user resources (CPU, reproduction layout), listening conditions (individual or collective), diversity of sound or audio-visual material handled, view-point and object positions in the virtual environment (interactivity). The problem of the representation – seen as a set of signals to be directly diffused or to be decoded before – concerns the transmission (purpose of conciseness) as well as the intermediary spatialisation step (global encoding to factorise the next process).

We have chosen to study thoroughly the ambisonic approach, which is based on *spherical harmonic decomposition of the acoustic field*, centred on the listener view-point. It has been known for a long time as a first order restricted form, which processes a minimal, directional sound field encoding through four components (B-format): W (pressure) and X, Y, Z (pressure gradient), offering easy sound field manipulations, such as rotations. With a decoder optimised in terms of the listening conditions (ideal/centred or collective/off-centred), a *coherent and homogeneous sound space rendering* can be obtained for various panoramic (2D) or periphonic (3D) loudspeaker rigs. This *"variable geometry" rendering* extends to headphones or a pair of loudspeakers *via* binaural techniques (virtual loudspeakers). The consideration of higher order components, which have just begun to be studied, introduces the concept of *"variable resolution" representation* (scalability), used as a function of the number of loudspeakers and/or the transmission capability.

We present acoustic and psychoacoustic foundations and a critical review of spatialisation strategies (stereo, surround, binaural, transaural and new derived forms). We explain the intrinsic link between ambisonic representation and local (velocity vector $\vec{V}$) and global (energy vector $\vec{E}$) propagation characteristics of the reproduced sound field (considering also the loudspeaker geometry), and the prediction laws between the latter and the localisation effect according to the head moving. Thus, the localisation theories implied in ambisonic decoding (Gerzon) are thoroughly justified. While being extended to other reviewed approaches, this kind of analysis highlights Ambisonic.

The generalisation of Ambisonic to all higher orders involves all the aspects presented for the first order systems, especially the encoding formalism and the decoding principles (2D and 3D). We develop the notion of directional sampling of the spherical harmonic base (also usable for the sound field pick-up problem), then the three original decoding forms (Gerzon, Malham) are generalised into three solution families, to be applied according to the listening conditions. Objective evaluations supported by informal listening experiments confirm the contribution of the higher orders and the optimised solutions. The improvement appears through the radial expansion of the acoustic field reconstruction and the global propagation ($\vec{E}$), or with regard to the perceptive aspects, through the sound image precision and robustness – even in non-ideal conditions – and the preservation of spatial impressions (lateral separation, a bit weak with the first order).

In addition to Ambisonic (tested up to the second order), other spatialisation techniques (pan-pot, binaural, transaural, plus artificial reverberation) are implemented, incorporated in an interface on a PC, and then experimented. This way, Ambisonic has been successfully applied to real-time source manipulation and mixing (with mono, multi-channel and B-format as input sources), and also compared or combined with other techniques, over headphones (binaural mode) as well as over loudspeakers. This tool could be used for a complete subjective validation of the ambisonic approach and the underlying strategies.

The ambisonic approach gives a very satisfying, global solution to the initial issues, although other strategies are better suited for specific problems – surround matrix systems, efficient binaural synthesis of complex scenes. Its extension to higher orders concerns many application fields and should develop in the near future thanks to the current studies and projects.

Keywords

Spatialisation - 3D-Sound - Surround - Multimedia - 3D-browsing - Ambisonic(s) - B-format - Localisation theory - Velocity vector - Energy vector - Psychoacoustic decoding - Sound field representation - Spherical harmonic decomposition - Directionnal sampling - Scalability

Table des matières

Introduction

Contexte et enjeux

La *(re-)création d'un espace sonore subjectif* est une préoccupation qui intéresse des domaines aussi variés que le cinéma, la diffusion musicale, la navigation dans les scènes virtuelles 3D, les jeux et bien d'autres applications multimédias ou de type "réalité virtuelle". Parce qu'elle permet d'élargir le champ des informations captées et des sensations perçues, la reproduction sonore spatialisée peut apporter une plus-value considérable à une application, quand elle n'en est pas l'objet principal. Elle y joue d'une part un *rôle informatif* – complémentaire des données visuelles le cas échéant – sur la présence des objets ou acteurs dans la scène représentée et sur leur localisation, y compris hors du champ visuel. Mais elle a aussi un *pouvoir suggestif* et un *potentiel immersif* qui font fréquemment défaut aux seules informations visuelles présentées sur un écran frontal.

Dans un cadre de diffusion sonore ou audio-visuelle, les formes les plus répandues de *représentation du champ sonore* sont dédiées à un dispositif de restitution propre, et consistent en un ensemble de signaux associés un à un aux haut-parleurs: c'est le cas du format stéréo conventionnel (pour deux haut-parleurs frontaux) et du format multi-canal "5.1" (stéréophonie panoramique ou "surround" sur cinq haut-parleurs). En s'intéressant à la reconstitution d'un espace sonore subjectif, on peut dire que la représentation se définit également par les mécanismes de création d'illusion sonore qui y conduisent, eux-mêmes propres aux techniques de prise de son ou de mixage employées à l'étape de production. Dans un contexte plus large de *spatialisation sonore*, incluant les applications de type réalité virtuelle, la scène sonore peut être au départ spécifiée par un certain nombre d'objets sonores (sources monophoniques) et l'acoustique du lieu virtuel. Pour la création d'images sonores localisées autour de l'auditeur (i.e. le travail de spatialisation), on dispose désormais d'une variété de techniques de positionnement 3D – auxquelles se joignent des techniques de réverbération artificielle –: des simples pan-pots traditionnels sur haut-parleurs aux techniques binaurales (sur écouteurs) et transaurales (sur paire de haut-parleurs) offrant des conditions d'illusion sonore plus performantes au prix de conditions d'écoute plus strictes et d'un coût de traitement plus élevé. Au stade intermédiaire de spatialisation, entre la description "décomposée" initiale de la scène et la restitution, la représentation globale et compacte du champ – liée ou non au dispositif de restitution selon les approches – peut permettre de réduire le coût de calcul par factorisation du traitement en aval. La diversité des approches s'exprime donc en termes de formats de représentation intermédiaire, de qualité d'image sonore, de dispositifs requis ou envisageables, de conditions d'écoute (dont l'étendue de l'auditoire)...

Les applications multimédias "faisant feu de tout bois", sont ainsi amenées à brasser une variété de matériels sonores et audio-visuels, mais aussi à concilier leur restitution avec des ressources CPU, des dispositifs et des conditions d'écoute variables selon les utilisateurs. Ce brassage atteint son paroxysme avec les nouvelles applications de navigation 3D dans des scènes virtuelles composites sur Internet: plaquage de matériel audio-visuel 2D sur des surfaces 3D, superposition d'objets sonores individuels spatialisés avec des champs sonores précomposés (enregistrements stéréo ou multi-canal), etc... Ce type d'application a par ailleurs une

dimension interactive qui suppose la manipulation des sources sonores, et plus globalement du champ sonore restitué, en fonction du point de vue de l'utilisateur dans la scène virtuelle.

La navigation 3D sur Internet fait partie – avec l'application de téléprésence [Nic99] – des domaines dans lesquels France-Télécom Recherche et Développement (F-T R&D)[1] est investi, et où se marient restitution sonore spatialisée et transmission numérique des informations. A ce titre, F-T R&D participe à la définition de la norme MPEG4 (Moving Picture Expert Group) concernant la description et la transmission *via* Internet de scènes audiovisuelles composites. La composante "transmission" impose une nouvelle contrainte à cette application: celle du débit, qui dépend à la fois du serveur, de la capacité du réseau et du client-utilisateur. Le problème de la réduction des informations sonores – que l'on rencontre également en télévision et radio numériques[2] – est jusqu'ici habituellement résolu par la compression audio-numérique de chaque canal individuellement (codage MPEG2 par exemple). La transmission d'un champ sonore complexe sous la forme "multi-canal" reste cependant relativement coûteuse avec une telle stratégie, qui ne tient aucun compte d'éventuelles redondances des informations spatiales entre les canaux.

Problématique et orientation de la thèse

L'orientation de ce travail de thèse sur la question de la représentation du champ sonore est ainsi motivée par plusieurs enjeux: sa transmission sous une forme concise et minimale pour répondre aux contraintes de débit; son usage au stade intermédiaire de traitement (spatialisation) pour en factoriser les opérations et en diminuer le coût; la possibilité d'une restitution sur des dispositifs variés, et adaptée à différentes envergures d'auditoire; la possibilité d'une manipulation du champ acoustique.

Bien qu'elle fut restée longtemps marginalisée, *Ambisonics*[3] se présente d'emblée comme une approche rationnelle dont les multiples propriétés apparaissent comme autant d'éléments de réponse à ces enjeux. Mathématiquement, elle repose sur une décomposition du champ acoustique en harmoniques sphériques autour d'un point, assimilable au point de vue de l'auditeur. Cette approche est plus connue sous une forme restreinte au premier ordre, qui réalise un encodage directionnel minimal du champ sonore à travers une composante omnidirectionnelle W (pression) et trois composantes bidirectionnelles X, Y, Z (gradient de pression), la composante Z étant omise pour une restitution sur un dispositif horizontal. Le *format B* ainsi constitué transporte de façon explicite et concise les informations directionnelles de la scène sonore et peut donner lieu, au moyen d'un décodage peu coûteux, à une restitution sur des dispositifs *surround* de géométries variées (quatre ou plus de haut-parleurs). Plusieurs formes de décodage ont déjà été proposées ([Ger92a], [Mal92]), s'adaptant à des conditions d'écoute différentes: écoute individuelle centrée, auditoire modérément ou très élargi. Combinée avec les techniques binaurales et transaurales (principe des haut-parleurs virtuels), une restitution sur paire d'écouteurs ou de haut-parleurs vient compléter les possibilités d'écoute individuelle. De par sa compacité, la représentation ambisonique peut être avantageusement envisagée à un stade intermédiaire de spatialisation pour le mélange des sources composant la scène sonore, rendant le coût de traitement en aval (décodage) indépendant de la complexité initiale de la scène. Enfin, elle se prête aisément à des manipulations du champ de type rotation ou distorsion de perspective.

Le développement et la vulgarisation des systèmes d'ordre 1 sont en grande partie attribués à Gerzon [Ger92e] [Ger92a]. On lui doit notamment les principes du décodage dit "psychoacoustique", basés sur une théorie de la localisation séparée entre des domaines basse-fréquence (vecteur vélocité $\vec{V}$) et haute-fréquence (vecteur énergie $\vec{E}$) [Ger92b]. L'extension de l'approche aux ordres supérieurs est quant à elle relativement

1. Anciennement: le CNET (Centre National d'Études des Télécommunications).

2. La radio numérique (ou DAB: *Digital Audio Broadcasting*) faisait également partie des thèmes de travail du CCETT (France-Télécom+Télédiffusion de France) – désormais centre de F-T R&D – au début de cette thèse.

3. ... ou encore *Ambisonic*, avec ou sans majuscule, dont on tire l'adjectif francisé "ambisonique".

récente ([BV95] pour l'ordre 2 et [Pol96a], de façon plus déguisée). Elle repose sur la prise en compte des composantes harmoniques sphériques d'ordres supérieurs – donc de nouveaux canaux – qui, en augmentant la résolution spatiale de la représentation, permettent une image sonore restituée plus précise, requérant toutefois plus de haut-parleurs. Les premières études de la restitution d'ordre supérieur se sont cependant restreintes à l'extension du domaine (basse-fréquence et/ou zone d'écoute) de reconstruction du champ acoustique, laissant dans l'inconnu la caractérisation des artefacts perceptibles hors de ce domaine. Quoiqu'il en soit, on voit apparaître l'idée éminemment prometteuse d'une *représentation "à résolution variable"*[4], transmise et exploitée en fonction des ressources de l'utilisateur et des capacités du réseau.

Les objectifs poursuivis au cours de cette thèse sont donc les suivants: la mise en oeuvre d'Ambisonic, et l'extension de ses principes (encodage et décodage pour toutes conditions d'écoute) aux ordres supérieurs; une validation de l'approche – d'abord par des arguments objectifs et théoriques, qu'il restera à compléter par une validation subjective formelle –; l'exploitation d'Ambisonic pour le mélange et la manipulation de sources; son utilisation parmi d'autres techniques de spatialisation, qu'il faut donc implémenter.

Plus fondamentalement, nous avons eu le souci de montrer le lien entre la représentation ambisonique – donc la caractérisation de la propagation acoustique qu'elle donne intrinsèquement – et les qualités spatiales perçues à la restitution. Il s'agit non seulement d'interpréter et justifier les vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ introduits par Gerzon et couramment utilisés pour prédire la localisation, mais aussi d'observer les conséquences sur les impressions spatiales et la qualité du champ diffus, par exemple. Cette étude est menée sur un double-front: extension de la reconstruction acoustique *et* caractérisation des artefacts.

Plan du document

La partie I peut être vue superficiellement comme un rappel des notions de base sur l'acoustique et la perception auditive spatiale (chapitre 1), et un tour d'horizon – ou état de l'art – des techniques de reproduction sonore spatialisée (chapitre 2)[5]. Une vocation plus profonde du chapitre 1 est d'expliciter les relations entre caractérisation acoustique du champ et localisation auditive[6], c'est-à-dire finalement donner des fondations rigoureuses aux théories de la localisation basées sur les vecteurs vélocité et énergie, qui sont des piliers de l'approche ambisonique. On y détermine également, d'après la description de la perception en "champs sonores naturels", les propriétés attendues des systèmes de reproduction.

La revue critique donnée au chapitre 2 couvre les techniques stéréophoniques et multi-canal, Ambisonic, ainsi que les techniques binaurales, transaurales et leurs nouvelles extensions, s'intéressant conjointement à la qualité de restitution, aux contraintes d'écoute, et au coût de calcul. En complément des éléments de la littérature existante, nous y tentons une analyse "basée acoustique" et unifiée des mécanismes de création d'illusion sonore et cherchons ainsi à montrer la dépendance des propriétés potentielles de restitution vis-à-vis de la représentation intermédiaire du champ acoustique, mais aussi vis-à-vis du dispositif de restitution. De cette manière, les spécificités d'*Ambisonic* – surtout décrite à travers les systèmes traditionnels d'ordre 1 – sont mises en lumière au milieu des autres approches, qu'elles soient déjà éprouvées de longue date ou bien plus prospectives et prometteuses.

La partie II est entièrement consacrée à la généralisation de l'approche ambisonique aux ordres supérieurs. Le chapitre 3 décrit l'extension des différentes étapes du système ambisonique: de l'encodage à la généralisation des solutions de décodage en fonction de la géométrie du dispositif (2D et 3D) et des conditions d'écoute, en passant par les transformations du champ. Le décodage s'appuie sur les outils de prédiction

4. On emploiera aussi le néologisme "*scalable*", déjà utilisé en codage d'image.

5. Avis au lecteur: ces premiers chapitres peuvent être dans un premier temps survolés, le lecteur pouvant être ensuite invité à s'y référer pour y trouver des justifications approfondies des notions manipulées dans le reste du document.

6. La démarche est expliquée en détail en introduction de ce chapitre 1.

$\vec{V}$ et $\vec{E}$ justifiés en 1.5. Nous avons également cherché à préciser certaines propriétés mathématiques – liées à la décomposition du champ en harmoniques sphériques (3D) ou cylindriques (2D) – qui sont sous-jacentes à l'approche ambisonique: la notion d'échantillonnage de la base de décomposition intervient au niveau du décodage et de la prise de son (abordée de façon plus marginale); l'étude des propriétés fondamentales liées à la troncature de la décomposition donne une idée de la qualité de restitution.

Dans le chapitre 4, on évalue l'apport des ordres supérieurs et des solutions de décodage développées sur la qualité de restitution, pour les différentes conditions d'écoute envisagées. Pour la position d'écoute centrée, une évaluation des indices de localisation se base sur une simulation binaurale du rendu ambisonique, et est étayée par des tests d'écoute informels. Les relations de prédiction d'après les vecteurs vélocité et énergie, établies en 1.5, sont à l'occasion validées ou précisées. Le cas de zone d'écoute élargie (ou de position excentrée) est abordé aussi bien *dans* et *en dehors* des conditions de reconstruction étendue (ordre élevé). Un bilan des développements de la partie II est enfin dressé.

La partie III fait état des développements pratiques réalisés au cours de cette thèse, ainsi que des applications et perspectives promises à *Ambisonic*. Le chapitre 5 présente notamment un outil d'expérimentation et d'évaluation où sont incorporées les techniques de positionnement 3D (pan-pot, ambisonic, binaural, transaural) et de synthèse d'effet de salle implémentées au cours de cette thèse, et rapporte quelques expériences de composition de scène sonore et de mélange/manipulation de sources. Après une étude sur la transmission/restitution du son multi-canal moyennant un encodage/décodage *via* le format B horizontal (W,X,Y), le chapitre 6 expose d'autres potentialités de l'approche ambisonique aux ordres supérieurs, applicables au multimédia parmi d'autres contextes.

Une conclusion générale dresse le bilan des développements effectués et des validations à poursuivre.

Première partie

Fondements théoriques et analyses préliminaires

Chapitre 1

Phénomènes acoustiques et perception auditive spatiale

1.1 Objectifs et démarche: de la caractérisation acoustique à l'évaluation et la prédiction de l'effet auditif spatial

1.1.1 Rôle et objectifs du chapitre

En se plaçant en préambule de l'étude et du développement de systèmes de restitution spatialisée, ce chapitre a pour vocation d'introduire et d'approfondir les notions fondamentales qui leur sont sous-jacentes, qui concernent à la fois la caractérisation ou la modélisation des phénomènes acoustiques, et les mécanismes de la perception auditive spatiale.

En premier lieu, puisque *l'observation et la caractérisation des phénomènes acoustiques* produits sont des points incontournables de notre étude, la section 1.2 en présente quelques *notions essentielles*. Un effort particulier est dédié à la question de la propagation et de sa caractérisation par le vecteur vélocité $\vec{V}$, dont on montre plus loin les relations avec l'effet de localisation auditive. Bien que d'abord exposés dans un cadre très général, ces aspects ont une place déterminante dans la description du fonctionnement de la restitution spatialisée sur haut-parleurs. En complément, l'évocation des phénomènes d'interface sert entre autre à la modélisation de la diffraction autour d'une tête – dont il est fait usage par la suite – et s'ouvre sur une description schématique des phénomènes acoustiques associés à un lieu d'écoute (acoustique d'une salle, par exemple). Ces phénomènes sont par la suite interprétés sur le plan de la perception auditive spatiale, et leurs procédés de synthèse (réverbération artificielle) sont partiellement appliqués au chapitre 5.

La section 1.3 poursuit avec la description des *mécanismes objectifs de la perception auditive spatiale* habituellement reconnus: de l'exploitation des indices de localisation liés au premier front d'onde (en 1.3.2 et 1.3.3), à l'effet des réflexions et de la réverbération sur l'image sonore, l'impression spatiale et l'enveloppement (en 1.3.4). A l'occasion de cette description basée sur les effets d'*expériences auditives naturelles*, quelques spécifications sur les informations de localisation à reproduire aux oreilles se dégagent pour définir l'aptitude des systèmes de restitution étudiés au chapitre suivant, à créer l'illusion d'images sonores convaincantes et plus globalement, à reproduire ou préserver des qualités spatiales et impressions spatiales naturelles.

Dans l'optique de *fournir des outils objectifs à l'analyse critique* des performances de restitution des techniques étudiées (en 4.1 par exemple), des *méthodes classiques et originales* sont ensuite proposées *pour l'évaluation des indices de localisation* à partir de réponses binaurales (section 1.4). Signalons que pour

traiter de façon précise ce problème de l'évaluation, et celui de la prédiction abordé juste après (1.5), il a été utile d'approfondir la modélisation du retard interaural (ITD) en 1.3.2, en correction aux modèles simplifiés traditionnels.

Enfin, la dernière section (1.5) *met en relation les deux domaines* – acoustique et perception auditive – pour proposer des outils de *prédiction de l'effet de localisation à partir d'une caractérisation de la propagation acoustique*. Ces outils seront exploités avec avantage pour *la caractérisation de la qualité de la restitution spatialisée*. Nous utilisons comme grandeurs synthétiques pour cette caractérisation, les *vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$*, traditionnellement connus sous la qualité d'indices "psychoacoustiques" depuis Gerzon. C'est lui qui les a en effet définis comme tels: il y fonde sa théorie de la localisation [Ger92b] et en dérive des "critères psychoacoustiques" pour l'optimisation des systèmes de restitution, et en premier les systèmes ambisoniques, qui sont l'objet principal de notre étude. Si l'on doit à Gerzon la vulgarisation et l'exploitation pratique de ces notions, on peut cependant regretter que les arguments fondateurs de ses théories soient pour le moins peu explicites, ainsi que les propriétés exactes attachées aux vecteurs vélocité et énergie – sa principale référence étant [Mak62], concernant le vecteur vélocité. Il en résulte une certaine méconnaissance des conditions exactes d'application et d'interprétation de ces critères lorsqu'on cherche à les utiliser. C'est donc à un important "travail d'éclaircissement" qu'il a paru utile de se livrer ici. Comme un point de départ solide pour la définition de ces vecteurs, il s'avère naturel de les ancrer *tout d'abord* dans une réalité physique, en commençant par *les identifier comme grandeurs acoustiques caractérisant la propagation* (en 1.2.2 et 1.5.1), dans un cadre général (champ acoustique quelconque) et avec le souci de bien préciser leur domaine de définition sur les plans spatial, temporel et fréquentiel. *Ensuite*, en appliquant cette caractérisation au lieu où doit se placer l'auditeur, nous en montrons les *implications sur les indices de localisation*, puis nous en interprétons les conséquences en termes d'*effet supposé (ou plausible) de localisation* et en fonction d'éventuelles rotations de la tête. Cette étude s'attarde au passage sur des cas de figures particuliers, comme des caractéristiques de propagation manifestement artificielles, propres à la reproduction stéréophonique sur haut-parleurs.

Le lecteur ne manquera pas de remarquer que les développements techniques présentés dans cette étude sont orientés suivant des *arguments essentiellement objectifs*: mécanismes de localisation, caractérisation acoustique, évaluation des indices de localisation... Il était donc au moins nécessaire d'en fournir des fondations théoriques solides. Mais c'est en définitive aux *effets subjectifs* perçus par l'auditeur qu'il importe de s'intéresser! Après un rappel préliminaire de notions fondamentales de la perception auditive, la discussion qui suit présente notre démarche, où l'adoption d'outils ou de critères objectifs – basés par exemple sur une caractérisation acoustique – se justifie par la préoccupation permanente d'une interprétation sur le plan perceptif, sans nier la nécessité de validations subjectives ou, en dernier lieu, écologiques.

1.1.2 Démarche et notions fondamentales: une approche objective et une fin subjective

Place de la cognition dans la perception auditive spatiale

L'ouïe a un rôle essentiel dans l'appréhension de l'environnement. Du fait qu'elle opère une collecte omnidirective des informations, elle a – à la manière de l'odorat chez d'autres espèces animales, et en tous cas plus que la vue – une fonction d'alerte primordiale pour la survie d'une espèce, depuis toujours et jusqu'à nos jours (dans un environnement urbain, par exemple!). Ainsi, *l'identification de la source sonore*[1] intervient comme une *opération précoce* de la perception, qui précède et – en association avec la connaissance[2] – *in-*

1. "Graou...? C'est un tigre!"
2. "C'est Hobbes..." (processus cognitif)

fléchit sa caractérisation subjective[3] [Cas94], y compris en terme de situation dans l'espace[4]. Cela n'exclue évidemment pas que des informations partielles de localisation ("A ma droite!" / "A ma gauche!") soient prises en compte précocement, comme en témoigne le réflexe de tourner la tête ("audition-réflexe" ou effet d'alarme). Il semble d'ailleurs délicat de discerner des relations d'antériorité entre le traitement de ces informations dans le processus de perception. Il reste néanmoins que la perception spatiale et la constitution d'un univers sonore subjectif sont la plupart du temps assujetties à des éléments contextuels forts – et à l'ensemble des informations sensorielles – qui définissent une *attitude d'écoute* [Cas94], qu'il s'agisse d'un contexte environnemental (attitude exploratoire, veille) ou d'un contexte culturel plus policé (écoute musicale, suivi d'un scénario théâtral).

Interprétation objective des informations sonores

Cependant, si l'on s'attache à la localisation et/ou la caractérisation spatiale des sources sonores perçues, on peut définir un certain nombre de mécanismes de détection qui exploitent les informations sonores parvenant aux oreilles, et ceci *en toute objectivité*, c'est-à-dire en mettant à l'écart toute donnée cognitive relative à l'identité de la source. Par exemple, la section 1.3.2 recense, quantifie et illustre les *indices objectifs de localisation* d'un son transporté par une onde plane (modélisant l'onde directe venant de la source sonore), basés essentiellement sur les différences des signaux perçus entre les deux oreilles. En utilisant les variations de ces indices lors de mouvements de la tête, nous montrons en 1.3.3 l'*émergence "naturelle" de mécanismes* conduisant à la *détection* de la direction de provenance *à la manière d'une machine* (!) [DRP99], qu'elle soit complète ou ambiguë.

Replaçant cette "expérience d'onde plane" (pour le premier front d'onde et les réflexions) au coeur d'une situation ordinaire d'écoute (1.3.1), on peut penser que le système perceptif exploite ces mêmes informations, utilise des mécanismes similaires, et qu'il les incorpore au processus de localisation et de caractérisation spatiale – par corrélation avec les autres sens et la connaissance acquise – au gré des expériences sensorielles d'événements acoustiques "ordinaires" et "naturels". *Cette analyse n'a recours à d'autre processus cognitif que l'effet d'un apprentissage plausible de ces mécanismes objectifs de détection*. L'analyse peut en outre être élargie, avec plus de limites cependant, aux effets des réflexions et de la réverbération sur l'appréhension d'une source en largeur et en profondeur, ainsi que sur l'impression spatiale et l'enveloppement.

Prédiction d'après caractérisation acoustique et évaluation: un effet perceptif supposé ou plausible

Lorsque nous observons le comportement des indices de localisation mesurés en situation "artificielle" (reproduction stéréophonique: 1.5, 2.6.1, 4.1), et que nous le confrontons aux lois et mécanismes issus d'un apprentissage en "situation naturelle" (1.3), nous nous livrons à une *interprétation objective* de l'effet de localisation, voire de la qualité de l'image sonore en termes de "naturel", précision, stabilité. Notons bien qu'il s'agit alors d'un *effet perceptif supposé ou plausible, en l'absence de toute présomption sur d'éventuelles données cognitives supplémentaires* (identification, contexte, vision associée).

Par extension, lorsque la *caractérisation de la propagation acoustique* (vecteurs $\vec{V}$ et $\vec{E}$) au lieu de l'auditeur conduit à la prédiction des indices de localisation en fonction de l'orientation de la tête (1.5), on accède dans une certaine mesure à la *prédiction des mêmes effets perceptifs supposés, selon les degrés de liberté de rotation de la tête*. C'est à l'issue de ce cheminement que les vecteurs vélocité et énergie méritent finalement leurs dénominations de "prédicteurs" ou d'"indices psychoacoustiques" de l'effet de localisation. Si Gerzon

3. "... qu'il est bête!"

4. "Il attend derrière la porte pour me sauter dessus." [Wat90]

avait d'emblée réduit ce fossé sémantique, ce cheminement nous permet d'associer à ces grandeurs $\vec{V}$ et $\vec{E}$ une interprétation plus complète et d'en préciser les conditions d'utilisation.

Validations subjectives, correspondances entre échelles subjective et objectives

Des expériences d'écoute validant le rôle des indices de localisation (différences interaurales en particulier) jalonnent la littérature. Des évaluations subjectives systématiques des qualités prédictives des vecteurs vélocité et énergie semblent en revanche y faire défaut. Les expériences d'écoute effectués durant cette thèse, qui semblent corroborer les prédictions théoriques et les mesures objectives [DRP98], restent de nature informelle et mériteraient d'être soutenus par des tests de plus grande envergure. Ces tests subjectifs doivent par ailleurs s'étendre à l'évaluation comparative des qualités de restitution des systèmes ambisoniques, suivant les conditions d'écoute et les solutions de décodage que nous avons généralisées (3.3). Au-delà de validations qualitatives, il s'avérerait instructif de mener des expériences qui permettent de définir quantitativement une échelle de sensibilité subjective aux paramètres (critères, indices) objectifs dont nous nous servons. Ces indices objectifs sont en particulier les normes r_V et r_E des vecteurs vélocité $\vec{V}$ et $\vec{E}$, ainsi que la variance directionnelle associée au vecteur énergie et sa projection sur l'axe interaural, que nous introduisons en 1.5. Il s'agirait en somme d'établir des correspondances entre une échelle objective et une échelle subjective. A défaut de temps pour avoir réalisé des séries de tests, nous en suggérons des méthodologies dans la partie III de ce document.

Validation écologique, efficacité et qualité de l'illusion sonore

La validation subjective des théories de la localisation passe par des expériences "de laboratoire" au conditionnement très strict. Mais considérant les techniques de restitution qui peuvent en être issues, c'est dans un contexte final d'application qu'il importe de juger la qualité de l'illusion sonore. A ce stade, c'est une *validation écologique* qui semble prévaloir: l'illusion peut être jugée satisfaisante si elle provoque chez le sujet *les mêmes réactions* que provoquerait la situation "réelle" à laquelle est est censée se substituer[5]. (CF THÈSES LAM)

Force est d'admettre cependant, que depuis le temps que l'homme invente et améliore les moyens de créer des illusions, la perception humaine a montré sa *capacité à ne pas "s'y laisser berner" de façon durable*. Par expérience renouvelée, par comparaison, ou une fois passé l'effet de surprise[5] [6], l'individu parvient à distinguer ces illusions d'événements réels ou originaux, voire à identifier leurs procédés. Finalement, il s'est familiarisé avec bon nombre de ces procédés (cinéma, télévision, photo, téléphone, etc...), a acquis un sens critique vis-à-vis de leurs artefacts, ainsi que l'aptitude à éventuellement se blaser[6] quand ils sont trop grossiers et quand l'argument narratif est insuffisant. Ces observations imposent plusieurs conclusions:

- C'est au scénariste, à l'artiste, au créateur, de créer la surprise, de fournir un contenu narratif satisfaisant...
- C'est aussi au preneur de son ou à l'ingénieur du son, de "jouer avec" les artefacts propres à la technique utilisée et au contexte de restitution, et d'exploiter avec pertinence les éléments cognitifs complémentaires pour "les imposer" à la perception (auditive) et à la reconstitution de l'espace (sonore) subjectif.
- Dans un contexte général de restitution du son, l'oreille critique du spectateur/utilisateur "audiophile" s'aiguise et justifie la recherche d'une qualité accrue de l'illusion, pour des raisons de confort et d'im-

5. La projection du "train rentrant en gare", des frères Lumière, avait mis les spectateurs en déroute...

6. ... de nos jours, il en faut beaucoup plus pour faire déguerpir un gamin!

mersion, ce qui se traduit par l'exigence d'une *haute-fidélité* y compris en termes de restitution spatiale [Mal99b].

Cette notion de "haute-fidélité" – qu'il est prudent de préférer à celle de "réalisme" – est parfois poussée jusqu'à un point où elle est confondue avec une reproduction des événements acoustiques qui soit conforme à la situation originale représentée (téléprésence, enregistrements musicaux) ou à une situation physiquement réaliste. En pratique, la plupart des systèmes sont loin d'atteindre cet "idéal", du moins en laissant à l'auditeur une certaine liberté de mouvement. A défaut, quelques propriétés particulières doivent être recherchées pour une qualité d'illusion satisfaisante: par exemple, la dématérialisation ou l'oubli du dispositif de restitution (haut-parleurs), corrélée avec la propriété de restitution homogène des images sonores (*coherency* pour Malham [Mal99b]), ou encore (pour Malham) l'homogénéité 3D de la scène reproduite, par une "représentation" ou une restitution équivalente des événements sonores quelles que soient leurs directions. Ces propriétés sont parmi les qualités les plus caractéristiques de l'approche ambisonique, très largement développée dans ce document.

Ambitions et nature de l'étude: technicité et généricité des développements

Les développements présentés dans cette thèse revêtent un caractère relativement technique, et sont dédiés à un contexte d'application encore très générique. Un des objectifs est l'utilisation des techniques ambisoniques dans des applications à caractère interactif, comme la navigation dans des scènes virtuelles 3D. Dans ce cadre particulier, mais aussi de manière générale, il nous est donc difficile d'incorporer le point de vue d'un preneur de son, qui agit sur la production sonore en projetant son écoute dans l'espace de restitution. Au contraire, nous devons faire face à une production "automatique" de l'espace sonore reproduit, donc *a priori non supervisée*. C'est pourquoi nous nous intéressons principalement à des performances de restitution qui soient significatives sur le plan perceptif de par les informations objectives perçues, mais sans autre présomption d'ordre cognitif sur l'expérience d'écoute. Nous jugeons donc un "potentiel d'implication" sur la perception sonore, en termes de localisation, précision, "naturel" des images sonores, ainsi qu'en termes d'enveloppement et d'immersion. Si les développements techniques sont guidés par des critères objectifs, des validations subjectives doivent bien-sûr parachever leur appréciation, voire également les orienter.

1.2 Notions d'acoustique

1.2.1 Equation des ondes et définitions

Les aspects acoustiques qui intéressent la présente étude concernent essentiellement la transmission de l'information sonore dans l'air (c'est le *médium*), de sa source jusqu'à sa réception par les oreilles ou d'autres transducteurs (microphones). Cette transmission a lieu sous forme d'une vibration se propageant par le jeu de compressions et raréfactions de l'air. Le *signal de pression p* mesurable en un point correspond à des variations autour de la pression atmosphérique moyenne.

Acoustique linéaire, équation des ondes

Tout au long de ce document, les phénomènes acoustiques sont décrits dans le cadre de l'approximation de l'*acoustique linéaire*, c'est-à-dire que les variations p sont petites par rapport à la pression atmosphérique, les phénomènes dissipatifs (absorption par l'air) étant quant à eux négligés, sauf mention spéciale. Sous cette

hypothèse, ils obéissent à l'équation des ondes, écrite ici dans le *domaine temporel* [Bru98]:

$$\left(\triangle - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\phi(\vec{r},t) = -q(\vec{r},t), \tag{1.1}$$

où c désigne la vitesse du son[7], $\phi(\vec{r},t)$ le *potentiel des vitesses* et $q(\vec{r},t)$ le débit volumique des sources au point $\vec{r}$ et à l'instant t. La *pression* p et la *vitesse particulaire* $\vec{v}$ sont reliées au potentiel des vitesses par les relations[8]:

$$p(\vec{r},t) = \rho\frac{\partial\phi}{\partial t}(\vec{r},t), \tag{1.2}$$

où ρ représente la densité de l'air, et:

$$\vec{v}(\vec{r},t) = -\vec{\nabla}\phi(\vec{r},t) \tag{1.3}$$

Ces grandeurs vérifient elles-mêmes l'équation des ondes (1.1), au terme de source près, qui devient $\rho\partial q/\partial t$ avec p, et $-\nabla q$ avec $\vec{v}$. Sur les domaines où ce terme de sources est nul, les trois quantités ϕ, p et $\vec{v}$ vérifient effectivement la même équation.

Ecriture dans le domaine fréquentiel

Il est pratique d'exprimer ces équations dans le *domaine fréquentiel*. A chaque fonction temporelle $\underline{\mathcal{F}}(\vec{r},t)$ est alors associée une fonction fréquentielle complexe $\mathcal{F}(\vec{r},\omega)$ obtenue par transformée de Fourier de $\underline{\mathcal{F}}$:

$$\mathcal{F}(\vec{r},\omega) = \int_{-\infty}^{\infty} \underline{\mathcal{F}}(\vec{r},t)e^{-j\omega t}\,dt \tag{1.4}$$

où la pulsation ω est liée à la fréquence f par $\omega = 2\pi f$. A partir de la spécification $\mathcal{F}(\vec{r},\omega)$ du champ dans le domaine fréquentiel, son expression $\underline{\mathcal{F}}(\vec{r},t)$ dans le domaine temporel est obtenue par la transformée de Fourier inverse:

$$\underline{\mathcal{F}}(\vec{r},t) = \int_{-\infty}^{\infty} \mathcal{F}(\vec{r},\omega)e^{j\omega t}\,d\omega \tag{1.5}$$

Dans le cas particulier d'un *régime harmonique* de pulsation ω, un champ $\underline{\mathcal{F}}(\vec{r},t)$ est défini par l'amplitude complexe $\mathcal{F}(\vec{r},\omega)$ de la façon suivante[9]:

$$\underline{\mathcal{F}}(\vec{r},t) = \Re(\mathcal{F}(\vec{r},\omega)e^{j\omega t}) \tag{1.6}$$

Pour simplifier, nous confondrons les notations dans le domaine temporel et dans le domaine fréquentiel pour décrire les grandeurs ϕ, p et $\vec{v}$, et les phénomènes acoustiques seront la plupart du temps décrits en régime harmonique, sans qu'il y ait nécessairement restriction à un champ monochromatique (une seule fréquence). L'équation (1.1) devient alors:

$$(\triangle + k^2)\phi(\vec{r},\omega) = -q(\vec{r},\omega), \tag{1.7}$$

où apparaît le *nombre d'onde* $k = \omega/c$. Et par ailleurs, (1.2) devient:

$$p(\vec{r},\omega) = j\omega\rho\phi(\vec{r},\omega), \tag{1.8}$$

d'où:

$$\vec{v}(\vec{r},\omega) = \frac{j}{\omega\rho}\vec{\nabla}p(\vec{r},\omega) \tag{1.9}$$

7. Typiquement, $c = 340m/s$ dans des conditions atmosphériques ordinaires.

8. Les conventions de signe pour le membres droits des équations (1.2) et (1.3) peuvent être inversées, comme on peut le rencontrer dans la littérature, sans conséquence sur les résultats qui en dérivent, dès lors que l'expression du potentiel ϕ disparaît.

9. Il faut savoir que cette convention adoptée (souvent implicitement) dans la majorité des références mentionnées dans cette thèse, y compris [Bru98], n'est pas celle utilisée dans [MI68], où le terme $e^{j\omega t}$ est remplacé par $e^{-j\omega t}$. Il peut en résulter une confusion quant au sens de la propagation à la lecture des équations.

Résolution des problèmes acoustiques

La résolution des équations pour la description ou la prédiction des phénomènes acoustiques peut schématiquement être dissociée en *deux problèmes*: celui de la *propagation en champ libre* (domaine exempt de parois) et le *problème dit "aux limites"* pour les phénomènes d'interface (réflexions, diffraction, absorption, transmission au niveau des objets et parois) et de rayonnement (en cas de source d'énergie acoustique à la frontière).

1.2.2 Propagation en champ libre, caractérisation par le vecteur vélocité $\vec{V}$

En champ libre et hors des sources, le terme q est nul et (1.7) devient:

$$(\triangle + k^2)\phi(\vec{r},\omega) = 0, \tag{1.10}$$

soit, dans le domaine temporel:

$$\left(\triangle - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\phi(\vec{r},t) = 0 \tag{1.11}$$

Solutions particulières et essentielles

Deux formes particulières de solutions retiennent notre attention: l'une à dépendance uni-axiale – les ondes planes, d'incidence $\vec{u}$ dans (1.12) – et l'autre à caractère sphérique, écrite en (1.13) comme se propageant "vers l'extérieur".

$$p_{\vec{u}}(\vec{r},t) = Ae^{jk\vec{u}.\vec{r}}e^{j\omega t} \tag{1.12}$$

$$p_{\vec{\rho}}(\vec{r},t) = V\frac{e^{-jk|\vec{r}-\vec{\rho}|}}{|\vec{r}-\vec{\rho}|}e^{j\omega t} \qquad \text{avec} \quad V = \frac{j\omega\rho_0 q}{4\pi} \tag{1.13}$$

$$p_{\vec{\rho}}(\vec{r},t) = A.d\frac{e^{-jk|\vec{r}-\vec{\rho}|}}{|\vec{r}-\vec{\rho}|}e^{j\omega t} \tag{1.14}$$

où le point $\vec{\rho}$ désigne le centre de symétrie de l'onde sphérique, soit encore le lieu de sa source d'émission. Dans l'écriture de la propagation reliée à l'émission par une source ponctuelle (1.13), q désigne la force de la source et ρ_0 la densité du médium (à ne pas confondre avec la distance $\vec{\rho}|$). Pour faire apparaître une quantité A homogène à une amplitude, comme dans 1.14, on introduit la notion de distance de référence d, à laquelle est le champ de pression est mesuré avec l'amplitude A.

Caractérisation locale de la propagation

Avant de commenter davantage ces solutions, la notion de *propagation acoustique* mérite d'être précisée et développée. Tout en introduisant un certain nombre d'outils essentiels, nous allons montrer qu'elle est indissociable de la notion de transport d'énergie. On définit pour cela le *flux d'énergie instantané* $\vec{I}_i$ (intensité acoustique) [Bru98] par:

$$\vec{I}_i(\vec{r},t) = p(\vec{r},t)\,\vec{v}(\vec{r},t) \tag{1.15}$$

En écrivant par ailleurs le vecteur intensité acoustique complexe $\vec{\Pi}$ (domaine fréquentiel):

$$\vec{\Pi} = \frac{1}{2}p\vec{v}^* = -\frac{j}{2\omega\rho}p\vec{\nabla}p^* = \vec{I} + j\vec{J}, \qquad \text{i.e.} \quad \left\{\begin{array}{rcl}\vec{I} & = & \Re(\vec{\Pi}) \\ \vec{J} & = & \Im(\vec{\Pi})\end{array}\right. , \tag{1.16}$$

il advient [DRP99] [SS94] (à des différences de conventions près), dans le cas d'un champ monochromatique, que:

$$\vec{I}_i(\vec{r},t) = 2\cos^2(\omega t + \varphi_p)\vec{I} - \sin(2(\omega t + \varphi_p))\vec{J} \tag{1.17}$$

L'intensité instantanée s'exprime comme somme d'un terme à moyenne temporelle $\vec{I}$ non-nulle, et d'un terme à moyenne nulle (puissance fluctuante), associé au vecteur $\vec{J}$. Le vecteur $\vec{I} = \langle \vec{I}_i \rangle$, appelé *intensité active*, traduit donc les transferts d'énergie, c'est-à-dire véritablement les *phénomènes propagatifs*, alors que la quantité $\vec{J}$, dite *intensité réactive*, ne traduit que des échanges locaux [Bru98] et révèle le *caractère stationnaire* du champ.

Admittance spécifique normalisée

Pour aller plus loin dans la caractérisation du champ acoustique, il est utile d'introduire l'*admittance spécifique normalisée*, définie comme le vecteur:

$$\vec{\beta}(\vec{r}) = \frac{1}{\rho c}\frac{\vec{v}(\vec{r})}{p(\vec{r})} \tag{1.18}$$

En utilisant (1.9) et en posant $p = |p|e^{j\varphi_p}$, il ressort rapidement [DRP99] que:

$$\vec{\beta} = \frac{j}{k}\frac{\vec{\nabla} p}{p} = -\frac{1}{k}\vec{\nabla}\varphi_p + j\frac{1}{2k}\vec{\nabla}\ln|p|^2 \tag{1.19}$$

d'une part, et d'autre part:

$$\vec{\Pi} = \frac{1}{2\rho c}|p|^2\vec{\beta}^* = \frac{\omega^2\rho}{2c}|\phi|^2\vec{\beta}^*, \qquad \text{soit} \quad \left\{ \begin{array}{rcl} \vec{I} & = & \frac{1}{2\rho c}|p|^2\Re(\vec{\beta}) \\ \vec{J} & = - & \frac{1}{2\rho c}|p|^2\Im(\vec{\beta}) \end{array} \right. \tag{1.20}$$

Au regard des équations (1.19) et (1.20), il apparaît clairement que:

- La direction locale de propagation du *front d'onde*[10], c'est-à-dire la *propagation de phase*, indiquée par $\Re(\vec{\beta})$ (gradient de phase, orthogonal au front d'onde), est aussi celle du flux d'énergie $\vec{I}$. De surcroît, en faisant l'approximation au premier ordre de la fonction de phase $\varphi_p(\vec{r} + d\vec{r}) = \varphi(\vec{r}) - k_{loc}\, d\vec{r}.\vec{n}$, $\vec{n}$ étant la normale au front d'onde de même direction que $\Re(\vec{\beta})$, l'équation (1.19) permet de déduire la vitesse apparente $c_{loc} = \omega/k_{loc}$:

$$c_{loc} = c/|\Re(\vec{\beta})| \qquad \text{puisque} \qquad k_{loc} = k\,|\Re(\vec{\beta})| \tag{1.21}$$

- $\Im(\vec{\beta})$ (gradient d'énergie, échelle logarithmique) relate le caractère réactif du champ (intensité réactive) à travers les *variations spatiales du champ d'énergie* (Cf [Bru98] pour une définition plus complète de l'intensité réactive).

Notons que l'admittance spécifique normalisée est définie dans le domaine fréquentiel et ne dépend en outre que du lieu $\vec{r}$. Les différentes grandeurs caractéristiques introduites sont illustrées Figure 1.1.

10. Un front d'onde est défini comme une ligne ou une surface iso-phase, c'est-à-dire sur laquelle φ_p est constant.

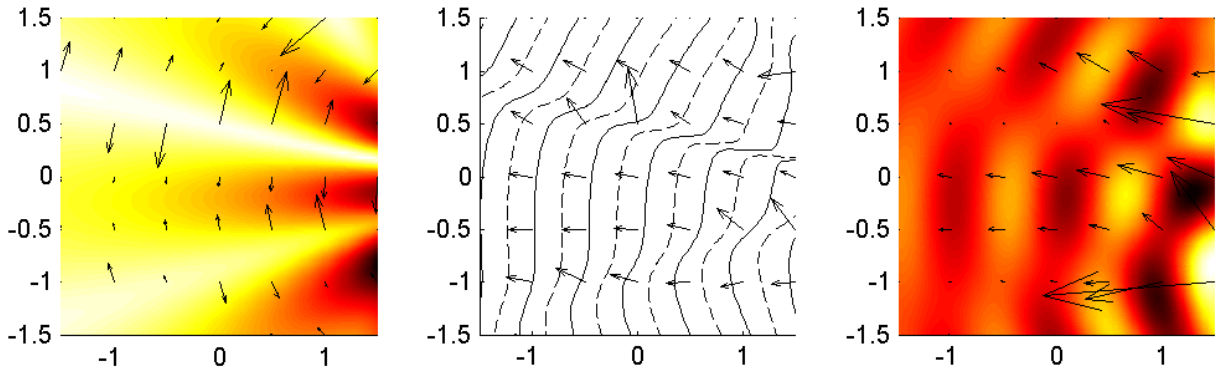

FIG. 1.1 – *Visualisation et caractérisation d'un champ acoustique produit par interférence de deux ondes sphériques monochromatiques. De gauche à droite: champ d'énergie (ou plus précisément d'amplitude $|p|$: coloris foncé pour forte amplitude), avec représentation du gradient d'énergie $\Im(\vec{\beta})$; champ de phase φ_p (courbes iso-phase $0[2\pi]$ et $\pi[2\pi]$) avec son gradient $\Re(\vec{\beta})$; vue instantanée $p(\vec{r},t)$ (coloris foncé pour les plus grandes valeurs) et intensité active $\vec{I}$. Noter que l'intensité active a bien la même direction que $\Re(\vec{\beta})$, mais qu'elle traduit en plus la pondération énergétique.*

Introduction du vecteur vélocité $\vec{V}$ pour la caractérisation locale

Pour des raisons de compatibilité avec un formalisme largement rencontré dans la littérature que nous citerons (par exemple, [Ger92b]), nous introduisons le *vecteur vélocité* $\vec{V}$ qui n'est autre que l'opposé du vecteur $\vec{\beta}$:

$$\vec{V}(\vec{r}) = -\vec{\beta}(\vec{r}) = -\frac{1}{\rho c}\frac{\vec{v}(\vec{r})}{p(\vec{r})} = -\frac{j}{k}\frac{\vec{\nabla} p(\vec{r})}{p(\vec{r})}, \tag{1.22}$$

dont on nomme les parties réelle et imaginaire $\vec{\Omega}$ et $\vec{\Phi}$:

$$\left\{ \begin{array}{rclcll} \vec{\Omega}(\vec{r}) & = & \Re(\vec{V}(\vec{r})) & = & r_V\,\vec{u}_V, & |\vec{u}_V| = 1 \\ \vec{\Phi}(\vec{r}) & = & \Im(\vec{V}(\vec{r})) & = & r_\Phi\,\vec{u}_\Phi, & |\vec{u}_\Phi| = 1 \end{array} \right. \tag{1.23}$$

Le vecteur unitaire $\vec{u}_V$, orthogonal au front d'onde, pointe dans la *direction de provenance de l'onde*, dont la *vitesse apparente de propagation* c_{loc} est donnée – d'après (1.21) ou [DRP98] [DRP99] – par:

$$c_{loc} = c/r_V \tag{1.24}$$

C'est le terme $\vec{\Omega}$ qui est révélateur du transport d'énergie (intensité active $\vec{I}$), alors que $\vec{\Phi}$ – habituellement appelé "vecteur *phasiness*" – mesure le caractère stationnaire (local) du champ. [Dans le cas particulier d'une onde plane progressive – "meilleur transporteur d'énergie" –, le vecteur vélocité est réel ($\vec{\Phi} = 0$) et $r_V = 1$.]

Caractéristiques de l'onde plane et de l'onde sphérique

Les deux cas particuliers de propagation évoqués plus haut méritent d'être illustrés ici, puisqu'ils modélisent les formes de contributions élémentaires d'une source sonore que nous utiliserons pour la description des phénomènes acoustiques. Il s'agit donc de l'onde sphérique progressive (1.14), typiquement émise par

une source ponctuelle, et de l'onde plane progressive (1.12), dont l'onde sphérique approche les caractéristiques de propagation en champ lointain. Il est facile de déduire de l'équation de l'onde plane (1.12) l'admittance spécifique normalisée et le vecteur vélocité qui lui sont associés:

$$\vec{V}_{\vec{u}}(\vec{r}) = -\vec{\beta}_{\vec{u}}(\vec{r}) = \vec{u} \tag{1.25}$$

Le vecteur vélocité est donc uniforme spatialement, réel (donc pas de déphasage entre p et $\vec{v}$), de norme 1 et... tout simplement égal au vecteur incidence $\vec{u}$. Il s'agit donc de l'onde "idéalement progressive" par excellence, et qui présente la "meilleure intensité acoustique globale". Considérant maintenant le cas de l'onde sphérique (1.14), il advient que:

$$\vec{\beta}_{\vec{\rho}}(\vec{r}) = -\vec{V}_{\vec{\rho}}(\vec{r}) = \left(1 - \frac{j}{k|\vec{r}-\vec{\rho}|}\right)\frac{\vec{r}-\vec{\rho}}{|\vec{r}-\vec{\rho}|} \tag{1.26}$$

Il apparaît clairement que $\vec{V}_{\vec{\rho}}(\vec{r})$ tend vers un vecteur réel et unitaire – donc la caractéristique de propagation d'onde plane – quand $k|\vec{r}-\vec{\rho}| >> 1$, c'est-à-dire en hautes-fréquences pour une distance $|\vec{r}-\vec{\rho}|$ donnée, ou à grande distance pour une limite basse-fréquence donnée. En champ proche en revanche, le déphasage est dû à la décroissance radiale exponentielle (gradient d'énergie).

Principe de superposition

Le principe de superposition est une propriété intrinsèque de l'acoustique linéaire, et peut être appliqué à la caractérisation locale par le vecteur vélocité. Superposons donc des champs acoustiques i, chacun étant décrit en un point $\vec{r}$ (et dans un domaine temps/fréquence donné) par une amplitude complexe $a(\vec{r})$ et un vecteur vélocité $\vec{V}_i(\vec{r})$. Un calcul simple permet de déduire le vecteur vélocité $\vec{V}(\vec{r})$ au point considéré du champ résultant:

$$\vec{V}(\vec{r}) = \frac{\sum_i a_i(\vec{r})\vec{V}_i(\vec{r})}{\sum_i a_i(\vec{r})} \tag{1.27}$$

Cas de l'interférence d'ondes planes

Dans le cas de la superposition d'ondes planes d'incidences $\vec{u}_i$, le remplacement $\vec{V}_i(\vec{r}) = \vec{u}_i$ s'opère dans l'équation (1.27), où les $\vec{u}_i$ sont réels, unitaires et uniformes. Considérée en un point central $\vec{r} = \vec{0}$, elle peut donc s'écrire:

$$\vec{V} = \frac{\sum_i a_i \vec{u}_i}{\sum_i a_i} \tag{1.28}$$

C'est cette expression qui est donnée par Gerzon pour la définition du vecteur vélocité [Ger92b], dans le cadre de la caractérisation et l'optimisation de procédés de restitution sur haut-parleurs (comme nous le développons également en 2.6.1), les $\vec{u}_i$ décrivant les directions des haut-parleurs vus du centre du dispositif.

Il est d'ores et déjà intéressant d'aborder le cas particulier d'amplitudes a_i réelles au point considéré (plus généralement: ondes en phase ou en opposition de phase). Le vecteur vélocité résultant est alors réel, de norme r_V non-nécessairement unitaire et pouvant pointer dans une direction différente des $\vec{u}_i$.

Nous cherchons ici à sensibiliser le lecteur au fait que l'*on peut observer localement une propagation qui ressemble à celle d'une onde plane*, dans le sens où le vecteur vélocité$\vec{V}$ est réel (pas de gradient d'énergie, pas de fluctuations locales), mais *qui n'en a pas la vitesse de propagation naturelle c* (i.e. $r_V \neq 1$). Bien-sûr, de telles caractéristiques ne sont pas "stables" – c'est-à-dire uniformes – spatialement: leur expansion est limitée au moins suivant un axe. Il faut signaler en revanche que si une telle caractéristique de propagation est observée (en un point) *uniformément sur une large bande de fréquence*, cela signifie qu'elle résulte de

l'interférence d'ondes planes portant un même signal et *convergeant de façon synchrone* au point considéré: sauf curiosité de la nature[11], il s'agit d'un *phénomène essentiellement artificiel!* Cette anticipation de l'étude des phénomènes stéréophoniques (2.6.1) met déjà en évidence la *notion de "naturel"* des événements acoustiques, qui sera abordés sur le plan de la perception en 1.5.

Expansion radiale d'une caractérisation locale: tenseurs/harmoniques sphériques d'ordres supérieurs

La pression p et son gradient $\vec{\nabla}p$ (ou bien la vitesse particulaire $\vec{v}$) constituent les tenseurs d'ordre 0 et d'ordre 1 du champ, et c'est donc leur rapport – en substance, le vecteur vélocité $\vec{V}$ – qui caractérise localement la propagation. Il semble logique que l'*expansion de cette caractérisation* à partir et autour d'un point soit alors spécifiée par les *tenseurs d'ordres 2 et supérieurs* au point considéré. S'intéressant à une expansion à caractère radial (donc sans privilégier de direction pour une distance donnée), un formalisme basé sur un système de coordonnées sphériques apparaît plus judicieux que l'expression de ces tenseurs en coordonnées cartésiennes. En accord avec l'équation des ondes, cela conduit naturellement ([MI68] [Bru98] et Annexe A) à l'émergence d'harmoniques sphériques (ou cylindriques, si l'on s'attache à une expansion radiale parallèlement au plan horizontal) et à l'écriture du champ en série de Fourier-Bessel.

Il est d'ailleurs assez facile de vérifier [MI68] que les cinq harmoniques sphériques d'ordre 2, assimilables aux fonctions de directivité de quadrupôles, sont contenues de façon redondante dans le tenseur d'ordre 2, matrice 3×3 symétrique de trace constante (équation des ondes), ne dépendant donc que de cinq paramètres indépendants. De façon plus générale, l'étude présentée dans [DK99] illustre la redondance de l'écriture de l'expansion du champ en série de Taylor (coordonnées cartésiennes) par rapport à l'expansion en série de Fourier-Bessel utilisant les harmoniques sphériques.

Au cours des chapitres suivants, ces arguments, augmentés de propriétés supplémentaires[12], donnent à l'approche ambisonique une place privilégiée.

Propagation et transport d'énergie en général: intérêt de la caractérisation

Note: la digression qui suit fait écho à certains débats concernant la localisation, mais il n'est pas indispensable de retenir les grandeurs qui y sont évoquées.

Parmi les discussions qui concernent la perception directionnelle des flux sonores, il n'est pas rare de voir émerger des questions qui pourraient se formuler ainsi: «En quoi le système auditif s'apparente-t-il à un appareil de mesure du transport acoustique des informations sonores?», et vu sous cet angle: «Quelle(s) grandeur(s) acoustique(s) mesure-t-il alors, qu'on puisse associer à l'effet directionnel perçu?». En prélude à l'étude 1.5, mentionnons déjà les deux propositions qui viennent le plus couramment argumenter une telle approche: l'oreille (pour dire "le système auditif dans son ensemble") *serait* un "capteur de la propagation de phase" dans un domaine basse-fréquence, et un "détecteur du transport d'énergie" dans un domaine haute-fréquence. La première proposition est relativement facile à admettre, de nombreuses théories lui donnent des arguments sérieux, et c'est pour lui apporter une interprétation complète et rigoureuse (en 1.5.2) que nous avons détaillé l'étude du vecteur vélocité. La seconde proposition reste au contraire confuse – d'autant qu'il existe plusieurs définitions du transport d'énergie – et ne semble pas avoir trouvé à ce jour d'argumentation explicite.

Ces questions méritent au moins une petite digression sur les notions de propagation et de transport d'énergie dans un contexte général de champ acoustique, et sur ce que l'on peut exploiter ou non des dé-

11. Par exemple, conditions de symétrie des parois réfléchissantes par rapport à la source et au point de mesure.

12. On verra que l'ordre de grandeur de l'expansion est également indiqué par la similarité du vecteur vélocité et du vecteur énergie, introduit en 1.5.1.

finitions usuelles du transport d'énergie, en observant notamment les propositions de Stanzial *et al* [SS94] relatives à ce domaine[13].

Nous venons de montrer que pour caractériser les phénomènes de propagation à une fréquence donnée, il suffit d'observer le rapport entre la vitesse particulaire et la pression considérées dans le domaine fréquentiel. La *propagation de phase* ainsi décrite reflète alors, à un facteur d'énergie près et dans le cas d'un *champ monochromatique*, l'intensité active, c'est-à-dire le transport moyen d'énergie en un point. Mais dans le cas d'un champ non-monochromatique, cette correspondance n'a plus de sens car c'est bien un rapport entre grandeurs temporelles qu'il faut considérer. Par analogie avec le vecteur vélocité dans le domaine fréquentiel, on peut s'intéresser à la *vélocité du transport d'énergie acoustique* [SS94] – "pleine-bande", pourrait-on ajouter. La vélocité d'énergie instantanée peut être définie comme le rapport entre l'intensité acoustique $\vec{I}_i(\vec{r},t)$ (densité de flux d'énergie) et la densité d'énergie $w(\vec{r},t)$:

$$\vec{U}_E = \frac{\vec{I}_i(\vec{r},t)}{w(\vec{r},t)}, \tag{1.29}$$

avec

$$w(\vec{r},t) = \frac{1}{2}\rho\left(\vec{v}^2(\vec{r},t) + \frac{1}{\rho^2 c^2}p^2(\vec{r},t)\right) \tag{1.30}$$

S'intéressant plus particulièrement à dégager une grandeur moyenne, Schiffrer et Stanzial [SS94] préconisent, plutôt que de s'attacher à la moyenne $\langle\vec{U}_E\rangle$, de considérer la *vélocité de l'énergie transférée par l'intensité moyenne*, dite *u-velocity*:

$$\vec{U} = \frac{\langle\vec{I}_i\rangle}{\langle w\rangle} \tag{1.31}$$

Une propriété remarquable attachée à ces grandeurs est que leurs versions normalisées$\vec{U}_E/c$, $\langle\vec{U}_E/c\rangle/c$ et $\vec{U}/c$ sont de modules systématiquement inférieurs à 1: elles traduisent une vitesse obligatoirement inférieure ou égale à la vitesse du son c. Par comparaison, il est frappant de remarquer que le vecteur vélocité peut quant à lui décrire une vitesse locale de propagation de phase qui soit *supérieure* à la vitesse du son, lorsque le module r_V de sa partie réelle est *inférieur* à 1! On verra que cette propriété acoustique a des implications effectives sur la perception directionnelle dans la mesure où la caractérisation de la propagation de phase s'étend au moins à l'échelle de la tête. Il n'y a par ailleurs aucune contre-indication pour que cette propagation "supersonique" locale puisse être observée sur une large bande de fréquence. Dans ce cas[14], le vecteur vélocité caractérise également une *propagation de groupe*. Bien que la vélocité d'énergie soit assimilable à un vecteur vitesse, ce n'est donc pas un transport d'énergie au sens d'une de ses définitions, que "mesure" le système auditif pour la localisation dans le domaine basse-fréquence. Une raison pour laquelle la caractérisation locale par un vecteur vélocité fréquentiellement uniforme, et celle donnée par l'une des grandeurs – $\vec{U}_E/c$, $\langle\vec{U}_E/c\rangle/c$ ou $\vec{U}/c$ – se différencient sur le plan des propriétés traduites, vient du choix de normalisation (par la densité d'énergie w ou sa moyenne $\langle w\rangle$) de (1.29) et (1.31): c'est le terme $\vec{v}^2$ du dénominateur w qui borne la vélocité d'énergie par c, alors que sans lui, on retombe sur une définition semblable à celle du vecteur vélocité (tout au moins dans le cas monochromatique): $\Re(\vec{V}(\vec{r})) = -\frac{\langle I_i(\vec{r},t)\rangle}{\langle p^2(\vec{r},t)/\rho c\rangle}$, d'après (1.20). De fait, le vecteur *u-velocity* $\vec{U}$ a plus pour vocation [SS94] de traduire le degré d'activité ($|\vec{U}|$ proche de 1) ou de réactivité ($|\vec{U}|$ proche de 0) du champ au point de mesure, que de traduire une vitesse locale de propagation.

D'un autre côté, il n'est pas raisonnable de considérer que c'est une caractérisation locale en champ libre par la vélocité d'énergie que le système auditif "mesure" pour la localisation haute-fréquence: l'instrument de

13. Ces références sont régulièrement mises en avant par Angelo Farina lors des discussions de la liste [Sur], et c'est surtout pour cette raison que nous nous y sommes intéressés.

14. Quelle que soit la valeur de r_V, pourvu qu'elle soit uniforme.

mesure, en l'occurrence *la tête, perturbe le champ* de manière significative par le volume qu'il occupe, et ses "capteurs" (les oreilles) sont espacés l'un de l'autre. Dans un domaine haute-fréquence (courtes longueurs d'onde), la caractérisation locale est insuffisante car de courte portée, donc *une caractérisation globale à l'échelle de la tête* est nécessaire pour se faire une idée de ce que peut détecter le système auditif[15]. Si l'on cherche à ne retenir qu'une grandeur synthétique, c'est donc une *caractérisation statistique de la propagation sur la zone d'écoute* qu'il faut définir. Le vecteur énergie introduit par Gerzon (Cf 1.5.1) est en ce sens un bon candidat. Nous prendrons la peine d'expliciter quelques aspects de sa nature en 1.5.1, puis nous mettrons à jour sa manifestation dans les indices de localisation et la façon dont il peut prédire l'effet de localisation.

1.2.3 Problèmes aux limites

Intérêt pour l'étude présente

La question des phénomènes d'interface concerne notre étude de façon plus ou moins indirecte: elle est évidemment présente lorsqu'on s'intéresse à la production d'un champ complexe par une source dans un lieu (réflexions, diffusion et réverbération dans une salle, par exemple); mais elle apparaît aussi au dernier stade de la transmission acoustique des informations sonores jusqu'aux oreilles d'un auditeur, au moins sous forme de diffraction des ondes incidentes autour de la tête et du corps. Cependant il n'est pas indispensable d'en retenir les détails pour une compréhension générale, c'est pourquoi nous n'en présentons ici qu'une formulation très générale et les principes de base de résolution. Le développement du cas de la diffraction autour d'une tête – modélisée par une sphère – est reporté en annexe A.3. Quelques exemples particuliers classiques comme la réflexion plane et des modèles de diffraction, sont ensuite brièvement évoqués en prélude à une modélisation sommaire des phénomènes acoustiques dans une salle, et à la présentation de procédés de spatialisation complémentaires (chapitre 5).

Formulation générale

Les phénomènes d'interface surviennent *aux limites de l'espace où l'on change de milieu élastique*[Val95], soit un fluide avec des caractéristiques de propagation (ρ,c) différentes (l'eau, par exemple) ou bien un solide (paroi, objet). En général, des conditions dites *aux limites* [Bru98] (voir également en annexe de [Nic99]) peuvent être spécifiées portant sur des grandeurs définies à l'interface entre les deux milieux, à savoir la pression p:

$$p(\vec{r}) = p_S(\vec{r}) \qquad \text{(condition de Dirichlet)} \tag{1.32}$$

ou bien la vitesse particulaire $\vec{v}$:

$$\vec{v}(\vec{r}).\vec{n} = v_n(\vec{r}) \qquad \text{(condition de Neumann)} \tag{1.33}$$

ou bien sur le rapport des deux (impédance ou admittance):

$$\vec{v}(\vec{r}).\vec{n} + Y^{\perp}(\vec{r}).p(\vec{r}) = v_n(\vec{r}) \qquad \text{(condition de Churchill)} \tag{1.34}$$

où, selon le type de condition: p_S est la pression imposée sur la surface S, v_n est la vitesse normale à S également imposée, $\vec{n}$ désignant la normale à S, unitaire et extérieure au domaine considéré.

15. On suggère implicitement que la tête est libre de mouvements de rotation.

Interprétation et principe de la résolution

La résolution des équations qui dérivent des conditions énoncées implique la “création" d’ondes réfléchie et/ou transmise ou d’un champ diffracté, s’il s’agit de traiter les conséquences d’une onde incidente, ou encore d’un champ rayonné s’il y a production d’énergie acoustique au niveau de l’interface, voire éventuellement les deux[16]. Il est possible et même avantageux de séparer les deux problèmes – perturbation d’une onde incidente et émission –, et c’est au premier des deux que nous nous intéressons. A partir du problème global décrit dans le domaine fréquentiel par l’une des conditions générales (1.32) (1.33) ou (1.34), et en éliminant les termes de source, les équations deviennent homogènes, et peuvent se résumer à la relation d’impédance:

$$\frac{p(\vec{r})}{\vec{v}(\vec{r}).\vec{n}} = Z^{\perp} = \frac{1}{Y^{\perp}}, \tag{1.35}$$

où $Z^{\perp}$ est dite *impédance normale*, éventuellement infinie si $v_n = 0$ est imposé (surface rigide).

Cas particuliers

Quelques cas se prêtent bien à une résolution mathématique rigoureuse de l’équation (1.35), surtout s’il existe un système de coordonnées adapté à la géométrie de la surface (l’interface): coordonnées cartésiennes pour une surface plane, sphériques pour une sphère (annexe A.3), cylindriques pour un cylindre [Bru98] [MI68].

Le schéma descriptif (en 1.2.4) du champ créé par une source dans un salle fait intervenir en particulier le cas très classique de la réflexion d’une onde plane – et par extension, d’une onde sphérique – sur une surface plane (de grande dimension par rapport à la longueur d’onde considérée): l’effet de cette *réflexion plane* ou *réflexion spéculaire* s’interprète comme celui d’une *source miroir*, par analogie au domaine visuel (Figure 1.2). De par sa simplicité, ce principe est largement utilisé en acoustique prévisionnelle et pour la synthèse temps-réel d’effet de salle (Chapitre 5).

D’autres cas de figure également très courants méritent d’être cités, qui requièrent des traitements plus délicats: il s’agit par exemple de la *diffraction au bord d’un demi-plan ou par une fenêtre*, dont différentes approximations peuvent être recensées [Bru98] et appliquées [Gas98]; les irrégularités de surfaces “globalement" planes ont quant à elles un *effet diffusant* d’un point de vue macroscopique (au moins pour les longueurs d’onde pas très grandes par rapport aux irrégularités), qui peut s’interpréter grossièrement par le fait qu’une onde incidente est réfléchie statistiquement dans toutes les directions. Par opposition à la réflexion spéculaire, on parle alors de *réflexion diffuse*, dont [DKS94] présente un éventail des différentes modélisations et techniques de simulation associées dans le domaine de l’acoustique prévisionnelle.

1.2.4 Champ complexe généré par une source: exemple de l’acoustique des salles

La description succincte de l’effet de salle associé à une source, présentée maintenant, a une double vocation:

- montrer qu’il se dégage une structure temporelle et spatiale typique qui conditionne l’organisation des mécanismes de perception auditive spatiale et l’interprétation subjective qui en découle en termes de localisation et d’impression spatiale (1.3.1 et 1.3.4);
- présenter des modèles physiques et statistiques qui sont utiles dans un cadre de synthèse artificielle d’effet de salle, et que nous exploiterons concrètement pour la réalisation d’un module de réverbération artificielle au chapitre 5 (d’après [Jot92]).

16. Sans parler des phénomènes dissipatifs: absorption, couche visco-thermique

Pour contenter le premier objectif, il est d'abord proposé une description qualitative des événements acoustiques qui touchent un auditeur, suite à l'émission d'un son dans une salle. Les caractéristiques physiques de l'effet de salle sont ensuite quantifiées en fonction de la géométrie de la salle et des qualités acoustiques des revêtements, à l'aide des formules de prédiction et des modèles classiques. Des descriptions plus approfondies peuvent être trouvées dans [Jot92] ou [Mar96], par exemple.

Description qualitative

La figure 1.2 expose le schéma typique de la production du champ acoustique en un point, par une source dans un lieu clos: une salle. Le modèle de réflexion spéculaire induit la notion de trajet acoustique de la source vers un point de mesure, ce qui offre une description très commode du processus de création du champ, à la fois géométrique et temporelle. La structure temporelle très caractéristique des événements acoustiques parvenant à l'auditeur ou au point de mesure peut être schématisée en trois groupes (Figure 1.3): l'onde directe, le groupe des réflexions précoces et la réverbération tardive. En un point d'observation particulier dans la salle, chacune des réflexions précoces est caractérisée par un temps d'arrivée, une intensité et un contenu fréquentiel qui dépendent de la position de la source dans la salle, en plus des qualités réfléchissantes des parois. Au bout d'un certain temps, la densité temporelle des réflexions est telle qu'elles ne sont plus dissociables: c'est la réverbération tardive, dont les propriétés ne dépendent plus de la position de la source de façon significative, mais seulement des caractéristiques de la salle (volume, absorption moyenne). On accorde généralement à la réverbération tardive des propriétés de champ diffus, c'est-à-dire d'un ensemble d'ondes temporellement dense, d'incidences équiréparties dans toutes les directions, cette propriété d'isotropie étant observée en tout point de la salle. L'établissement du champ diffus est favorisé par les réflexions diffuses (plus que spéculaires), c'est-à-dire la multiplication des surfaces réfléchissantes dans la salle.

La portion de la réponse impulsionnelle correspondant à la réverbération tardive est caractérisée par une décroissance exponentielle – soit à une décroissance linéaire en dB (Figure 1.3) –, dans la mesure où il s'agit d'une réverbération dans un lieu clos. Cette décroissance peut être elle-même caractérisée par un temps de réverbération T_{60}, au cours duquel l'énergie de la réponse diminue de 60 dB.

Caractérisation quantitative, modèle statistique

Pour une modélisation plus quantitative de l'effet de salle et de la réverbération tardive en particulier, il est utile d'observer les phénomènes à la fois sur le plan fréquentiel et sur le plan temporel.

D'un point de vue fréquentiel, la salle peut être considérée comme un résonateur aux modes propres amortis, l'amortissement étant dû à l'absorption par les parois et par l'air. Deux principaux aspects rentrent en jeu pour caractériser la réverbération diffuse: la densité modale, avec le souci d'un recouvrement fréquentiel suffisant, et l'amortissement moyen, auquel est directement lié le temps de réverbération. Pour une salle parallélépipédique de dimensions $L_x \times L_y \times L_z$ et à parois rigides, les fréquences des modes propres non amortis sont données par:

$$f_{n_x,n_y,n_z} = \frac{c}{2}\sqrt{\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2} \qquad \text{avec } n_x, n_y \text{ et } n_z \text{ entiers} \tag{1.36}$$

En généralisant la formule aux salles de formes quelconques [Kut79], on peut déduire l'expression de la densité modale $D_m(f)$ en fonction du volume V de la salle:

$$D_m(f) = 4\pi V \frac{f^2}{c^3} \qquad \text{(Nombre de modes propres par Hz autour de } f\text{)} \tag{1.37}$$

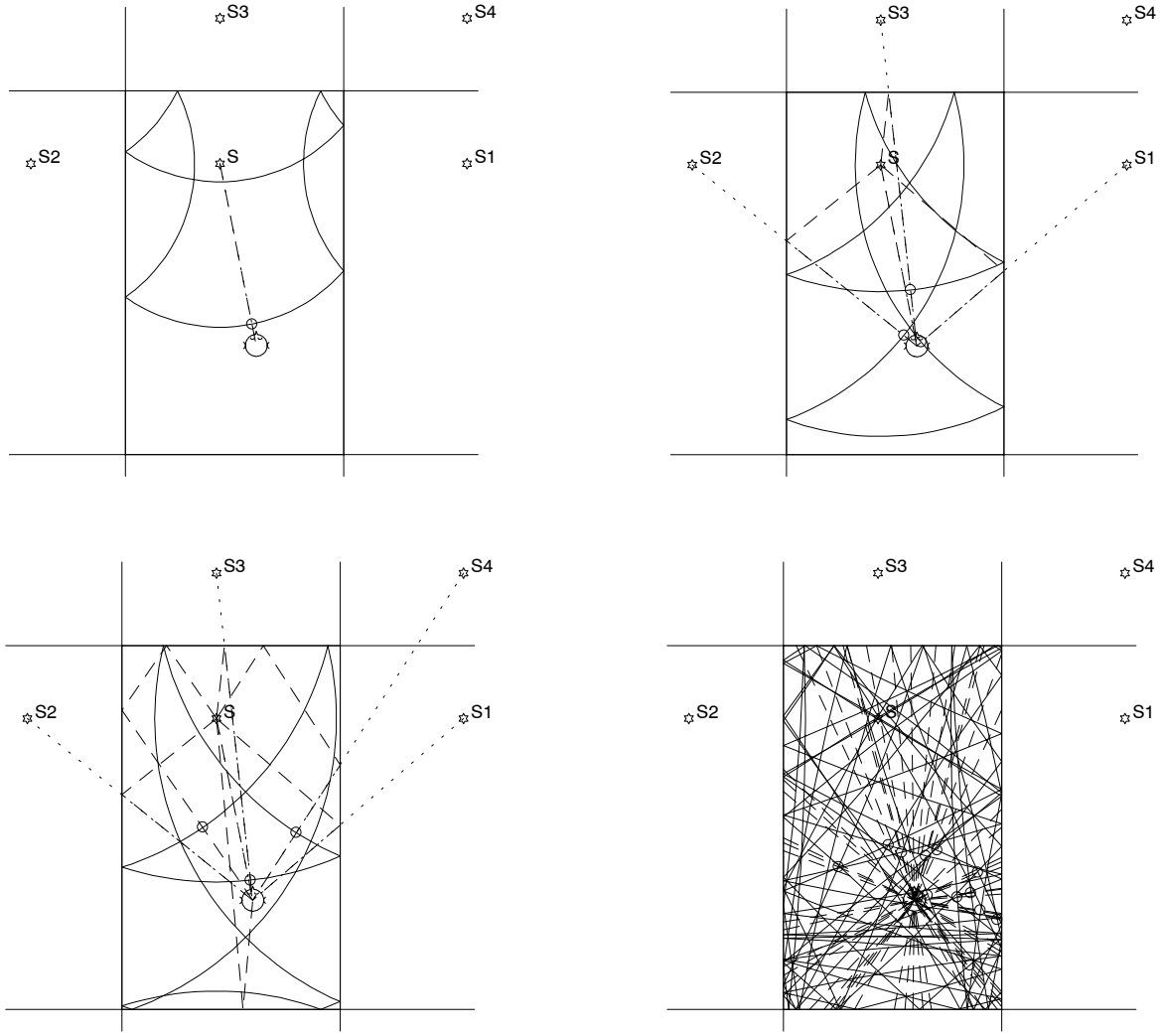

FIG. 1.2 – *Description temporelle et géométrique des événements acoustiques perçus en un point (auditeur), suite à l'émission d'un front d'onde par une source S et de par les interactions avec les parois de la salle. Chaque petit cercle indique le point de front d'onde (arc de cercle) qui va atteindre l'auditeur (trajet en tirets). Exemple d'une salle rectangulaire (ici, de 3 m sur 5 m). De gauche à droite et de haut en bas: (a) A $t = +6$ ms, un premier front d'onde (onde directe) parvient à l'auditeur. (b) A $t = +11$ ms, les premières réflexions (dites précoces) d'incidences diverses parviennent ensuite, assimilables à des ondes émises par des sources-miroir S_i, images de la source primaire par rapport aux plans de réflexion (murs). (c) Plus tard ($t = +100$ ms), les réflexions se densifient temporellement et proviennent de toutes les directions (ce qui rend la représentation graphique illisible!). Vu de l'auditeur, le champ est isotrope. Il en est de même en tout lieu de la salle: le champ est diffus. Ce schéma ne fait intervenir que des réflexions spéculaires. La plupart du temps, des phénomènes de diffusion entrent également en jeu, et les réflecteurs sont plus nombreux et variés, accélérant l'établissement du champ diffus.*

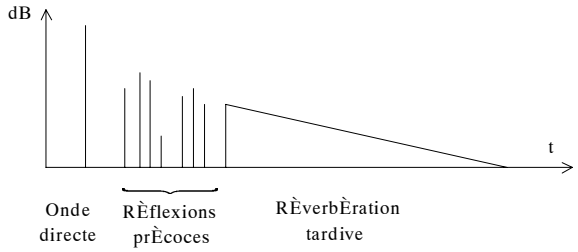

FIG. 1.3 – *Structure typique d'une réponse impulsionnelle dans une salle. C'est plus précisément la réponse temporelle d'énergie qui est représentée, en dB.*

En présence d'absorption, chaque mode propre amorti i possède une largeur de bande Δf lié à son coefficient d'amortissement σ_i par $\Delta f = \frac{\sigma_i}{\pi}$. Schroeder [Sch62] a montré que lorsque la densité modale $D_m(f)$ est suffisante par rapport à la largeur de bande des modes, la réverbération peut être considérée comme un processus aléatoire gaussien. Le taux de recouvrement modal requis pour cela doit être au moins de $\Delta f.D_m \approx 3$, ce qui est réalisé au-delà de la fréquence dite "de Schroeder":

$$f_s \approx 2000\sqrt{\frac{T_{60}}{V}} \quad \text{(Hz)} \qquad \text{avec } V \text{ en m}^3 \text{ et } T_{60} \text{ en Hz,} \tag{1.38}$$

le temps de réverbération T_{60} étant lié à l'amortissement moyen σ par:

$$20\log_{10}\left(e^{-\sigma T_{60}}\right) = -60 \text{ dB} \quad \Rightarrow \quad T_{60} = \frac{3\ln 10}{\sigma} \tag{1.39}$$

Il est normalement fonction de la fréquence.

Connaissant les caractéristiques de la salle, le temps de réverbération peut être approché par la *formule de Eyring* [CM82]:

$$T_{60} = \frac{-0{,}163V}{S\ln(1-\bar{\alpha}) - 4\mu V} \qquad \text{avec:} \quad \begin{cases} V & \text{Volume de la salle (m}^3\text{)} \\ S & \text{Surface totale des parois (m}^2\text{)} \\ \bar{\alpha} = \frac{\sum \alpha_i S_i}{S} & \text{Coefficient moyen d'absorption} \\ \mu & \text{Coefficient d'absorption de l'air (m}^{-1}\text{)} \end{cases} \tag{1.40}$$

ou bien encore la *formule de Sabine*, applicable lorsque $\bar{\alpha} \ll 1$:

$$T_{60} = \frac{0{,}163V}{\sum \alpha_i S_i + 4\mu V}, \tag{1.41}$$

où l'on peut négliger le terme $4\mu V$ dans le cas de petites salles et de fréquences relativement basses (< 8 kHz).

Observée sur le plan temporel, il est montré [Pol88] que la portion de réponse temporelle correspondant à la réverbération diffuse peut également être considéré comme un processus aléatoire gaussien non-stationnaire, dès lors que la densité temporelle d'échos est suffisante. Alors la réponse impulsionnelle peut être modélisée par un bruit gaussien centré $b(t)$ pondéré par une enveloppe exponentielle [Pol88]:

$$h(t) = b(t)\, e^{-\sigma t}, \quad t \geq 0 \tag{1.42}$$

Le temps de mélange à partir duquel cette modélisation stochastique peut être considérée comme valide est tel que la densité d'échos $D_e(t)$ traduit la superposition d'au moins 10 échos pendant un intervalle d'intégration

de l'oreille ($\Delta t \approx 24$ ms). Le modèle de salle parallélépipédique offre une estimation aisée de la *densité d'échos* $D_e(t)$, c'est-à-dire du nombre de réflexions par seconde au voisinage d'un instant t:

$$D_e(t) \approx 4\pi c^3 \frac{t^2}{V}, \tag{1.43}$$

où V désigne le volume de la salle. Par suite, cette expression peut se généraliser aux salles de formes quelconques [Pol88]. Le temps de mélange peut alors être estimé approximativement par la formule:

$$t_{\text{mélange}} \approx \sqrt{V} \quad \text{en ms, avec } V \text{ en m}^3 \tag{1.44}$$

La modélisation stochastique (1.42), indépendante de la position de la source et du récepteur, peut donc être appliquée dans une région du plan temps-fréquence bornée inférieurement par la fréquence de Schroeder f_s et le temps de mélange $t_{\text{mélange}}$. En pratique, on cherche à privilégier le caractère diffus de la réverbération, ce qui revient à rendre ces grandeurs aussi basses que possible, que ce soit pour la conception de salles de concert ou la définition de réverbérateur artificiel (Chapitre 5).

1.3 Perception auditive spatiale

1.3.1 Objectifs et données préliminaires

Objectifs

Comme annoncé en introduction du chapitre, un des objectifs de cette section est de *décrire les mécanismes de la perception auditive spatiale* – et plus particulièrement de la localisation – *et de les interpréter en relation avec des expériences d'écoute "ordinaires"*. C'est en effet par ce type d'expérience que le système de perception apprend à exploiter les informations sonores pour appréhender son environnement et reconstituer un espace sonore subjectif. Par ailleurs, en introduisant les *indices objectifs de localisation*, nous chercherons à préciser les paramètres et le domaine de validité de certains modèles simplifiés habituels, notamment au sujet de l'ITD (retard interaural). La présentation formelle des lois d'ITD révèle en effet des propriétés qu'il est important de prendre en compte, d'une part pour définir des méthodes efficaces d'estimation de l'ITD (en 1.4), et d'autre part pour l'établissement de théories de prédiction de la localisation (en 1.5).

Données physiologiques et propriétés psychoacoustiques de l'oreille humaine

Avant de s'abandonner à l'analyse de la localisation auditive – et de la perception spatiale en général – à partir des seules données acoustiques parvenant au niveau des oreilles, il est important de fournir quelques précisions sur la manière dont les signaux de pression acoustique sont transformés au sein de l'organe auditif, et sur les propriétés qui en sont retenues[17].

On dissocie dans l'organe en question trois parties: l'oreille externe (pavillon et conduit auditif), l'oreille moyenne (chaîne d'osselets pour l'adaptation d'impédance acoustique) et l'oreille interne (dont la cochlée, siège de la transduction mécano-nerveuse du signal auditif). Les oreilles externe et moyenne ont grossièrement l'effet d'un filtre passe-bas coupant à 20 kHz et résonnant à environ 3 kHz. De fait, la *bande des fréquences audibles* est typiquement comprise entre 20 Hz et 20 kHz, et le *seuil d'audition* (au repos) est plus important pour les fréquences aiguës et graves[18]. En prenant en compte l'ensemble de la chaîne, il advient

17. Pour lecture, ces aspects physiologiques et psychoacoustiques sont développés de façon plus approfondie dans les références suivantes: [Dur98] [ZF81] [BCDS89b] [Bla83], etc...

18. Cela peut avoir des implications subtiles sur les sensations auditives spatiales: l'effet d'enveloppement dû à la réverbération basse-fréquence dépend donc du niveau de la source sonore (Cf Griesinger).

que *la perception des informations sonores obéit à des échelles ou à des comportements* **non-linéaires**: perception différentielle du niveau sonore (échelle des *sonies* non-conforme à celle des dB), seuils d'audition et phénomènes de masquage. Les deux dernières propriétés sont en fait des conséquences d'une même propriété d'origine physiologique, traduite par la notion de *bande critique*. Dans une définition "idéalisée", la bande critique correspondrait à la largeur fréquentielle sur laquelle il faut intégrer l'énergie pour rendre compte de l'excitation en un point de la cochlée (ou au niveau d'une cellule sensorielle), c'est-à-dire finalement pour traduire l'information transmise dans le nerf auditif. Deux échelles ont été définies pour caractériser cette propriété: les Barks et les Erbs. La largeur des bandes critiques suit une échelle approximativement logarithmique en hautes fréquences (largeur proportionnelle à la fréquence centrale) et tend à être constante en basse-fréquence (1 Bark correspond à environ 100 Hz en-dessous de 500 Hz). Pour simplifier, on pourra interpréter ces échelles comme des "échelles de résolution spectrale perceptive", et utiliser la bande critique comme une fenêtre fréquentielle glissante sur laquelle il faut intégrer les informations sonores pour en retenir un poids perceptif pertinent. Concernant la *perception binaurale* et les mécanismes de la localisation, il a été prouvé ([Bla83], p.313 et suivantes) que la notion de bande critique se manifeste de façon quasi-identique à la définition des Barks (expériences monaurales).

Ces propriétés psychoacoustiques – particulièrement les notions de courbe de masquage et de bande critique – sont largement exploitées en matière de codage audio-numérique, et leur modélisation constitue un point clé de l'efficacité des systèmes de codage [Dur98]. Etendues aux interactions binaurales, elles semblent par contre être encore très peu prises en compte dans un contexte de définition de systèmes de restitution spatiale[19] ou d'évaluation objective de ces systèmes. Nous discuterons brièvement, dans la section 1.4, l'intérêt éventuel de prendre en compte les propriétés de bande critique pour l'estimation de l'ITD et de l'ILD. Signalons qu'une implémentation du filtrage par bande critique (échelle des Erbs), au moyen des *filtres de Patterson*, est disponible au sein d'une boîte à outils fonctionnant sous `matlab` [Sla].

Description d'une situation d'écoute ordinaire

Plaçons l'auditeur dans une situation ordinaire – un espace qu'il partage avec une source sonore (Figure 1.2) – et observons le phénomène acoustique associé à l'émission sonore par la source.

Bien souvent, si la source est "visible" pour l'auditeur, le *premier front d'onde* qui atteint l'auditeur – celui qui apporte *la primeur* de l'information sonore – est l'*onde directe*, qui suit le plus court trajet entre la source et l'auditeur. C'est donc cette onde directe qui fournit à l'auditeur les informations objectives principales pour la *localisation* de la source en terme de direction, lorsque les conditions sont bonnes, c'est-à-dire quand ces informations ne sont pas noyées par des réflexions trop présentes ou trop précoces (*loi du premier front d'onde*, ou *effet d'antériorité* [Bla83] [Beg94]). La section 1.3.2 se propose de quantifier l'information perçue par les oreilles, associée à ce premier front d'onde, et l'interpréter en terme de *localisation directionnelle*, en faisant émerger et en illustrant les indices de localisation habituellement utilisés.

Les réflexions et la réverbération tardive révèlent, parce qu'elles en sont la conséquence, la présence des objets et parois réfléchissantes, et les dimensions de l'espace qu'elles délimitent (plus le fait qu'il soit clos ou ouvert), en même temps qu'elles fournissent des indices sur la position (distance, profondeur) de la source dans cet espace. Les réflexions précoces en particulier modifient la perception de l'image sonore associée à l'onde directe, et sont en grande partie responsables de la largeur apparente de la source. De manière plus générale, l'ensemble des réflexions et de la réverbération est traduit comme une impression d'espace. Un effet d'enveloppement découle en particulier des réflexions latérales [Bar70], du fait qu'elles révèlent au mieux,

19. Mentionnons quand même la tentative d'utilisation de l'échelle des Barks (ou des Erbs) pour une modélisation "économe" et efficace des filtres binauraux [JLW95] (ou comme rappelé dans [HK97]).

grâce à une décorrélation interaurale maximale, l'enveloppement physique de l'auditeur par les éléments environnementaux (typiquement une salle). Ces points sont développés en 1.3.4.

Repérage spatial relatif à l'auditeur

C'est un système de coordonnées sphériques qui est utilisé pour décrire le repérage spatial relatif à la tête. L'axe vertical de la tête définissant l'axe polaire, l'angle de la projection de la source dans le plan horizontal est appelé *azimut* (*azimuth* en anglais), qui sera souvent noté θ, et l'information de hauteur est traduite par le *site*[20] (*elevation*) noté δ (Figure 1.4).

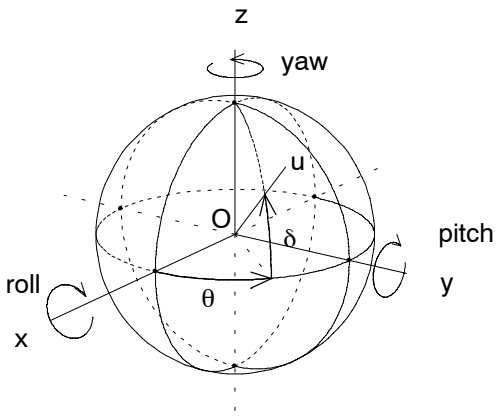

FIG. 1.4 – *Repérage directionnel par un vecteur incidence $\vec{u}$, d'azimut* θ *et de site* δ*, dans un référentiel $(O,\vec{x},\vec{y},\vec{z})$ relatif à la tête, O étant le centre de la tête, $\vec{x}$ sa direction frontale, et $\vec{y}$ orientant son axe interaural. Illustration des degrés de liberté de rotation de la tête (*yaw, pitch, roll*).*

Le repère associé à la tête est dirigé par trois vecteurs unitaires orthogonaux: $\vec{x}$, qui pointe vers l'avant (vu de la tête!); $\vec{y}$, dirigeant l'*axe interaural* de la droite vers la gauche; $\vec{z}$, dirigeant l'axe vertical vers le haut. La direction ou *incidence* de la source considérée est elle-même indiquée par le vecteur unitaire $\vec{u}$ de coordonnées sphériques (θ,δ) et de coordonnées cartésiennes $x = \cos\theta\cos\delta$, $y = \sin\theta\cos\delta$, $z = \sin\delta$.

En désignant par O le centre de la tête, le plan $(O,\vec{x},\vec{z})$ est appelé *plan médian*, ou encore plan sagittal.

Les rotations de la tête, dont nous montrerons plus loin l'importance pour la localisation, sont généralement décomposées en *trois rotations élémentaires successives* (si nécessaires), d'axes respectifs: $(O,\vec{z})$ (*yaw*/*rotate*), $(O,\vec{y})$ (*pitch*/*tumble*), et $(O,\vec{x})$ (*roll*/*tilt*), auquel on a associé les dénominations anglophones usuelles (*yaw, pitch, roll*, d'après [Beg94], et *rotate, tilte, tumble* d'après [Mal93]). La figure 1.4 en précise les conventions de sens.

Dans les cas ultérieurs où l'on aura besoin de tenir compte des rotations de la tête, on pourra parler d'azimut et site *relatifs ou bien absolus* pour préciser à quel référentiel ils se rapportent, à savoir un *référentiel lié à la position courante de la tête, ou bien lié à sa position initiale*.

1.3.2 Premier front d'onde: analyse des différences interaurales et effet de latéralisation

Bien que les différences interaurales – de temps (ITD) et d'intensité (ILD) – soient parmi les indices de localisation les plus communément évoqués et utilisés, et de ce fait les plus connus, il a semblé utile de leur

20. Attention, ne pas confondre le site δ avec l'angle polaire, lui-même noté θ en annexe A, qui sont complémentaires à $\pi/2$: $\delta = \pi/2 - \theta$.

consacrer ici une analyse relativement approfondie. En effet, puisque notre étude en fait un usage essentiel – qu'il s'agisse d'estimations objectives (en 1.4 puis en 4.1) ou de théories de prédiction de la localisation (en 1.5) –, il importait de présenter des lois "exactes" (plus particulièrement celles de l'ITD, suivant la fréquence) issues de simulations ou de mesures, afin de corriger les modèles simplifiés habituellement employés ou d'en préciser les domaines de validité. A l'occasion, il est également instructif pour le lecteur de visualiser le phénomène acoustique de diffraction par la tête.

Nous rappelons que le front d'onde considéré est ici *modélisé par une onde plane*, approximation valide lorsque la source n'est pas trop proche de l'auditeur, et hypothèse généralement – sinon unanimement – adoptée comme *référence* pour l'étude de la localisation sonore.

Les conventions utilisées dans la suite pour le repérage spatial sont celles introduites dans la section précédente .

Latéralisation d'après les différences interaurales: ITD et ILD

Les figures 1.6, 1.7, et plus schématiquement 1.5, décrivent le cas d'onde plane arrivant au niveau de la tête avec une incidence latérale. Des informations de localisation viennent du fait que le signal porté par l'onde arrive aux oreilles en suivant deux trajets acoustiques différents: un trajet *ipsilatéral* – onde "directe" pour l'oreille du côté de la source – et un trajet *contralatéral* – onde de "contournement" – dont la figure 1.5a donne un schéma idéal. La différence de marche entre ces deux trajets[21] induit un *retard interaural* – i.e. retard dans l'arrivée du signal à une oreille par rapport l'autre – (ITD: *Interaural Time Difference*), auquel s'ajoute une *atténuation interaurale* due au masquage par la tête (ILD: *Interaural Level Difference* ou IID: *Interaural Intensity Difference*).

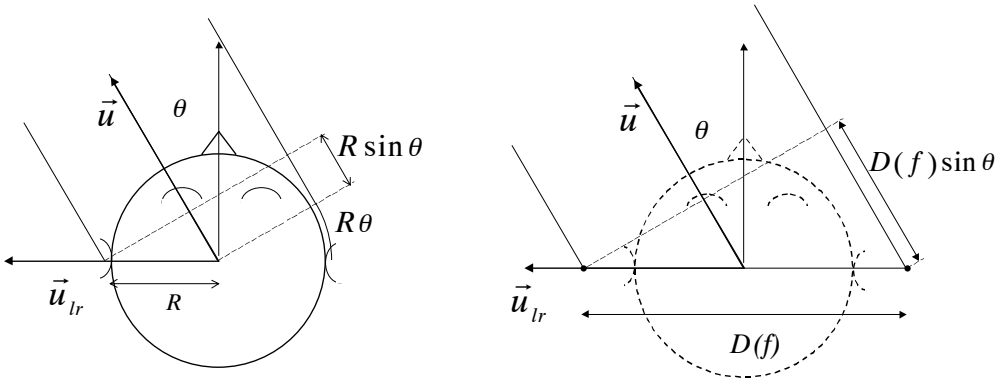

FIG. 1.5 – *A gauche (a): Différence de marche entre les contributions ipsi-latérale et contra-latérale d'une onde plane incidente (ici,* $\delta = 0$*,* $\sin\theta = \vec{u}.\vec{y}$*, avec* $\vec{y} = \vec{u}_{lr}$*). Schéma idéal, tendance asymptotique pour les hautes fréquences. Tête modélisée par une sphère de rayon R. A droite (b): Modèle équivalent pour la définition de l'ITD dans le domaine des basses fréquences: tête acoustiquement transparente et capteurs fictifs plus écartés que les oreilles (distants de* $D(f)$ *au lieu de* $2R$ *et en fonction de la fréquence* f*, avec en particulier* $D(f \to 0) = 3R$ *pour un modèle purement sphérique).*

Nous nous proposons maintenant de quantifier ces indices interauraux, préciser les lois qu'ils suivent, et illustrer les conditions de détection (en particulier celle de l'ITD). Nous utilisons pour cela tout d'abord un

21. Attention: il n'est pas dit que les trajets acoustiques soient "parcourus" par le signal avec une vitesse identique ou uniforme!

modèle de tête sphérique, qui permet une simulation rigoureuse du phénomène de diffraction et sa visualisation. Ce modèle est ensuite (un peu plus loin) confronté à un modèle plus réaliste (mesures effectuées sur un mannequin KEMAR), afin de compléter les données, mais aussi pour constater que les lois observées sont semblables et que les principes de détection à partir des différences interaurales sont identiques.

Modèle sphérique: conditions de détection (ITD) et lois basse-fréquence

La figure 1.6 décrit le cas d'une onde plane monochromatique "basse-fréquence" dans le sens où sa longueur d'onde excède la taille de la tête. On constate que l'effet de masquage, donc l'*ILD*, est *très peu significatif*. En revanche, le retard étant inférieur à une période, il est détectable sans ambiguïté d'après le *déphasage interaural* (IPD: *Interaural Phase Difference*).

On note également que *la loi de l'ITD est d'allure sinusoïdale dans le domaine basse-fréquence* où nous nous trouvons (fréquence inférieure à 1,5 ou 2 kHz). En considérant en plus la symétrie de la tête autour de l'axe interaural, *cette loi a la propriété très intéressante – pour les développements mathématiques ultérieurs – d'être proportionnelle à la projection linéaire du vecteur incidence $\vec{u}$ sur l'axe interaural* (dirigé suivant $\vec{y}$):

$$\begin{aligned} \mathrm{ITD}^{LF}(\vec{u}) &= \frac{D(f)}{c}\vec{u}.\vec{y} \\ \mathrm{ITD}^{LF}(\theta,\delta) &= \frac{D(f)}{c}\cos\delta\sin\theta \end{aligned} \tag{1.45}$$

où l'on a introduit une quantité $D(f)$ homogène à une distance. Celle-ci correspondrait au diamètre qu'une "tête" acoustiquement transparente devrait avoir pour produire la même loi d'ITD à la fréquence f considérée (Figure 1.5b). Il faut noter que ce diamètre équivalent – ou encore *l'amplitude de la loi – varie suivant la fréquence f*, comme nous l'illustrons plus loin (Figure 1.9). Signalons enfin, comme résultat particulier dû au modèle sphérique, que ce diamètre équivalent vaut, en très basses fréquences, trois fois le rayon effectif R de la sphère:

$$D(f \to 0\mathrm{Hz}) = 3R \tag{1.46}$$

Ce résultat peut être obtenu en utilisant les développements de l'annexe A.3[22], et est également présenté dans [Kuh77].

Modèle sphérique: conditions de détection et lois haute-fréquence

En considérant un domaine haute-fréquence (Figure 1.7), le phénomène de diffraction induit *une loi d'ILD beaucoup plus marquée, quoique non-monotone* (creux d'amplitude pour les incidences purement latérales). Bien que significatif, l'ILD n'apparaît donc pas comme un indice très précis de la position latérale.

Quant au retard interaural (ITD), il équivaut désormais, à une fréquence donnée, à plusieurs périodes du signal pour les incidences latérales (et dans un secteur angulaire de plus en plus large à mesure que la fréquence augmente): il n'est donc plus détectable à partir d'un simple déphasage interaural (ambigu)[23]. De fait, il est reconnu – et des expériences le montrent [Bla83] – que *le système auditif est peu sensible à l'information de phase en hautes-fréquences*, et d'autant moins que que des informations de localisation satisfaisantes (retard de phase cohérent) sont fournies en basse-fréquence [KIH99].

22. Pour le démontrer rigoureusement, il faut utiliser l'équation (A.50) et effectuer un développement en série en $1/(ka) = 1/(kR)$ de $\frac{p_t(\pi/2,0)}{p_t(-\pi/2,0)}$, limité au premier ordre. La phase de ce rapport complexe tend vers la fonction $\varphi_{lr} = 3kR\sin\theta\cos\delta$.

23. Dans une région fréquentielle intermédiaire, cette limitation n'est plus aussi catégorique: on peut supposer que les rotations de la tête peuvent permettre de résoudre cette ambiguïté là.

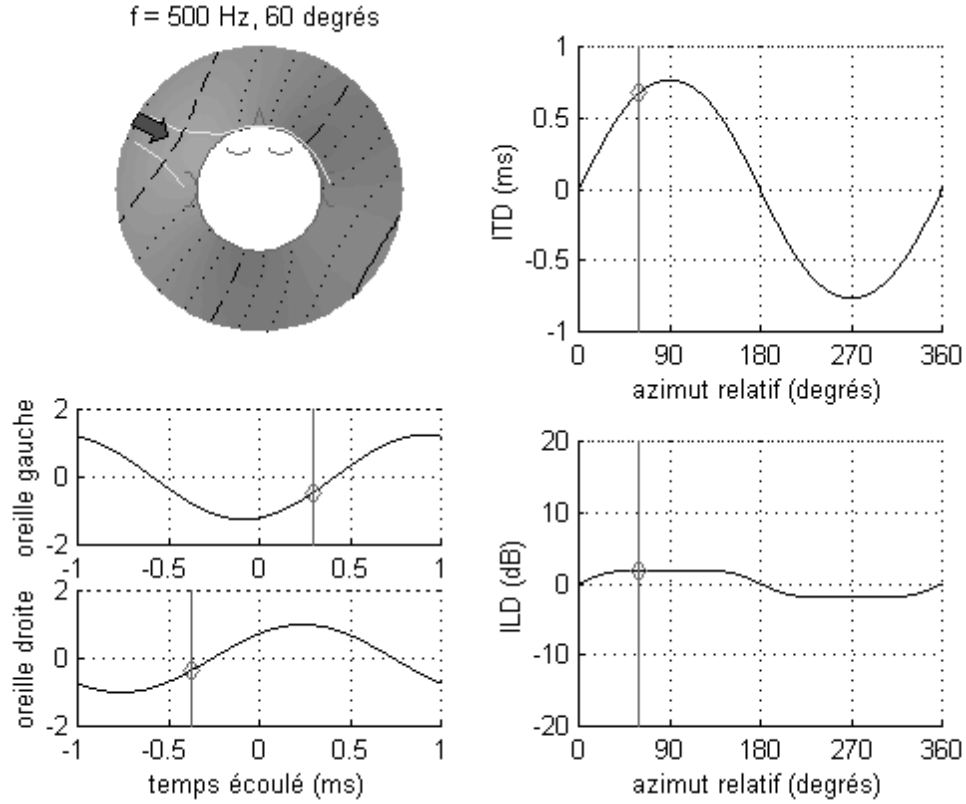

FIG. 1.6 – *En haut à gauche: effet de diffraction d'une onde plane monochromatique "basse-fréquence" (500 Hz) et d'incidence latérale (60°) autour d'une tête modélisée par une sphère (de diamètre 17,5 cm). Le niveau de gris représente le champ d'amplitude (donc lié à l'énergie). La propagation de phase est représentée par des fronts d'onde orthogonaux à la direction de propagation (courbes iso-phase: la distance séparant la ligne continue de la ligne tiretée correspond à une demi-longueur d'onde ou une demi-période). Deux lignes de courant parvenant à proximité des oreilles sont également tracées. En bas à gauche: signaux résultants au niveau des oreilles en fonction du temps écoulé par rapport à l'instant présent: les marques verticales (et les losanges) datent l'arrivée d'un même front d'onde à chacune des oreilles. Le retard interaural (ITD), plus petit que la période du signal, est donc détectable sous forme de différence de phase. En revanche, l'atténuation (ILD) est faible d'une oreille à l'autre. A droite: lois de l'ITD (d'allure sinusoïdale) et de l'ILD (peu significative) en fonction de l'azimut de l'onde incidente avec marquage du cas de gauche (60°).*

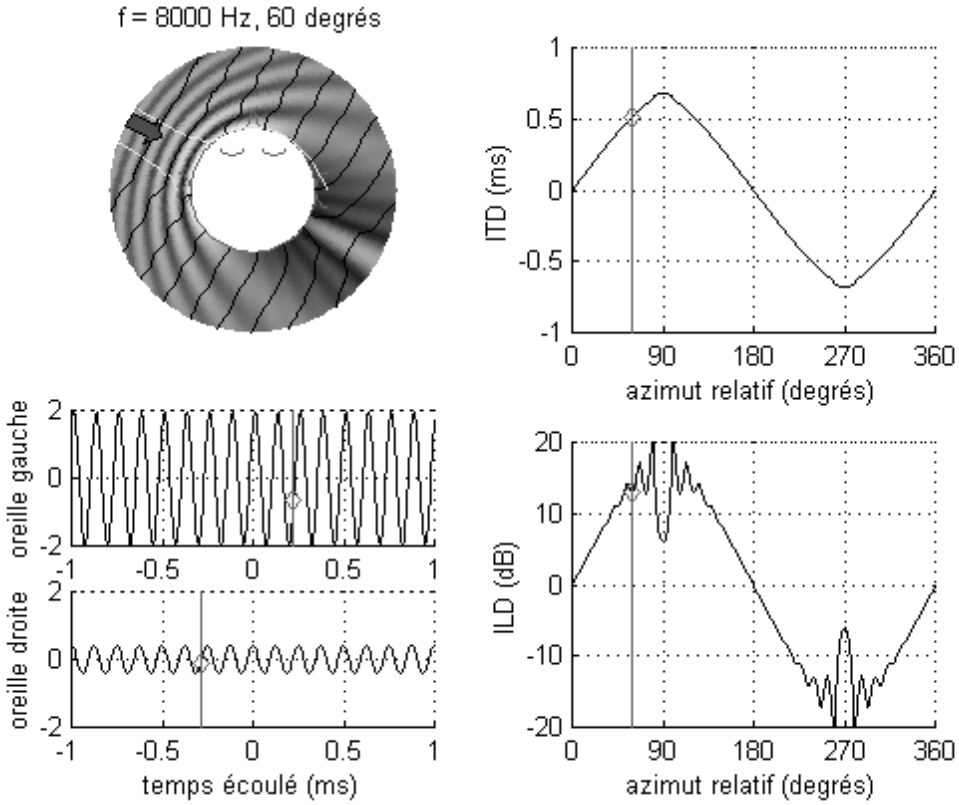

FIG. 1.7 – *Cas comparable à la figure 1.6, mais avec une fréquence élevée (8 kHz). Même légende, sinon que seuls des fronts d'onde de phase identique (modulo* 2π*) sont représentés (traits continus). Commentaires: le retard interaural (équivalent à plusieurs périodes du signal) n'est plus détectable sans ambiguïté à partir du déphasage. Le rapport d'amplitude est beaucoup plus marqué qu'en basse-fréquence. La loi de l'ITD n'est plus de forme sinusoïdale et est d'amplitude (ITD maximal) moins prononcée qu'en basse-fréquence. La loi de l'ILD, bien marquée, n'est en revanche pas monotone entre* 90° *(minimum local) et* 270° *(maximum local), ce qui s'explique par le fait que les "contributions diffractées" (signaux contournant la sphère) s'additionnent en phase au point diamétralement opposé à la direction d'incidence (zone parfaitement contralatérale), y créant donc un pic d'énergie (voir l'image du champ d'énergie diffracté, en haut à gauche de cette figure).*

Avant de montrer avec quel mécanisme et sous quelles conditions le retard peut être détecté *en hautes-fréquences*, il est intéressant de remarquer (Figure 1.9) que *la loi d'ITD converge vers une loi basée sur un modèle idéal*, d'après lequel les trajectoires directe et de contournement seraient "minimales" avec vitesse uniforme c (modèle schématisé Figure 1.5(a), et dont l'inexactitude est commentée plus loin) [Bla83] [DRP99]:

$$\begin{aligned} \mathrm{ITD}^{HF}(\vec{u}) &= \frac{R}{c}\left(\arcsin(\vec{u}.\vec{y}) + \vec{u}.\vec{y}\right) \\ \mathrm{ITD}^{HF}(\theta,\delta) &= \frac{R}{c}\left(\arcsin(\cos\delta\sin\theta) + \cos\delta\sin\theta\right) \end{aligned} \tag{1.47}$$

Du fait de cette convergence, *la loi (1.47) reste vérifiée* avec une approximation acceptable lorsqu'elle est transposée au cas du transport d'un *signal à bande non-nulle essentiellement haute-fréquence* (décomposable en signaux élémentaires mono-fréquence).

Considérons en particulier les signaux dont l'enveloppe d'amplitude varie au cours du temps, et avec des pentes suffisantes pour que la variation soit sensible sur la durée du retard d'arrivée entre les deux oreilles. C'est alors un *retard de l'enveloppe d'amplitude* qui est détecté, à condition que le caractère modulant ne présente pas une périodicité trop "serrée" (i.e. la période ne doit pas être plus courte que l'ITD maximal). Ces conditions sont remplies par les signaux de type transitoire, ou impulsionnel (impulsions de Gauss, par exemple), ou bien générés par battement entre fréquences voisines (modulation d'amplitude[24]), au contraire des signaux monochromatiques ou polychromatiques stationnaires qui n'induisent pas de battements sensibles (sons harmoniques continus, par exemple). La figure 1.8 illustre le cas d'un signal modulé en amplitude (battement) dont la période d'enveloppe est assez grande par rapport au retard interaural.

Confrontation avec la "référence" *KEMAR* – Domaines de validité des lois et des mécanismes

Les simulations qui viennent d'être présentées doivent être confrontées à un modèle plus conforme à la morphologie humaine. Nous utilisons pour cela les mesures effectuées par Martin et Gardner sur un mannequin KEMAR (tête+buste) et mises à disposition par ces auteurs [GM94]. Ce sont particulièrement les mesures d'ITD des deux modèles – sphérique et KEMAR – qu'il est intéressant de comparer. La figure 1.9 présente les lois estimées pour chaque modèle et pour un jeu de fréquences réparties en octaves.

Si les lois sont très semblables dans une région moyenne-fréquence, celles du modèle KEMAR sont plus perturbées en haute-fréquence, et d'amplitude plus grande pour les basses ou très basses fréquences – mais toujours d'allure sinusoïdale – même en corrigeant le diamètre équivalent. Ces divergences ne sont pas très surprenantes et peuvent facilement s'interpréter de la manière suivante:

1. En haute fréquence, le pavillon d'oreille devient un détail non-neutre du relief de la tête vis-a-vis de la diffraction. Cela explique les perturbations et la dissymétrie (avant-arrière, Figure 1.9), qui par ailleurs disqualifie l'axe interaural comme meilleur axe de symétrie[25].
2. Dans la région basse fréquence où les longueurs d'onde sont au moins de la dimension du corps, c'est l'ensemble du corps (tête, cou, buste...) qui doit être pris en compte pour estimer l'effet de diffraction, et il faut sans doute préférer au modèle sphérique celui d'un cylindre tronqué. C'est probablement ce qui explique l'amplitude d'ITD excessive par rapport au modèle purement sphérique.
3. Enfin, il faut peut-être considérer d'autres facteurs directement liés à la mesure des réponses binaurales sur KEMAR: le champ proche en basse fréquence, la distance du haut-parleur à la tête étant de 1,4 m seulement; les caractéristiques du lieu de mesure (proximité des parois), etc...

24. Le principe peut également être étendu aux signaux modulés en fréquence ("retard d'enveloppe spectrale").

25. Il est d'ailleurs souvent suggéré, comme correction du modèle sphérique, de déplacer le lieu des oreilles jusqu'à 110° par rapport à la direction frontale.

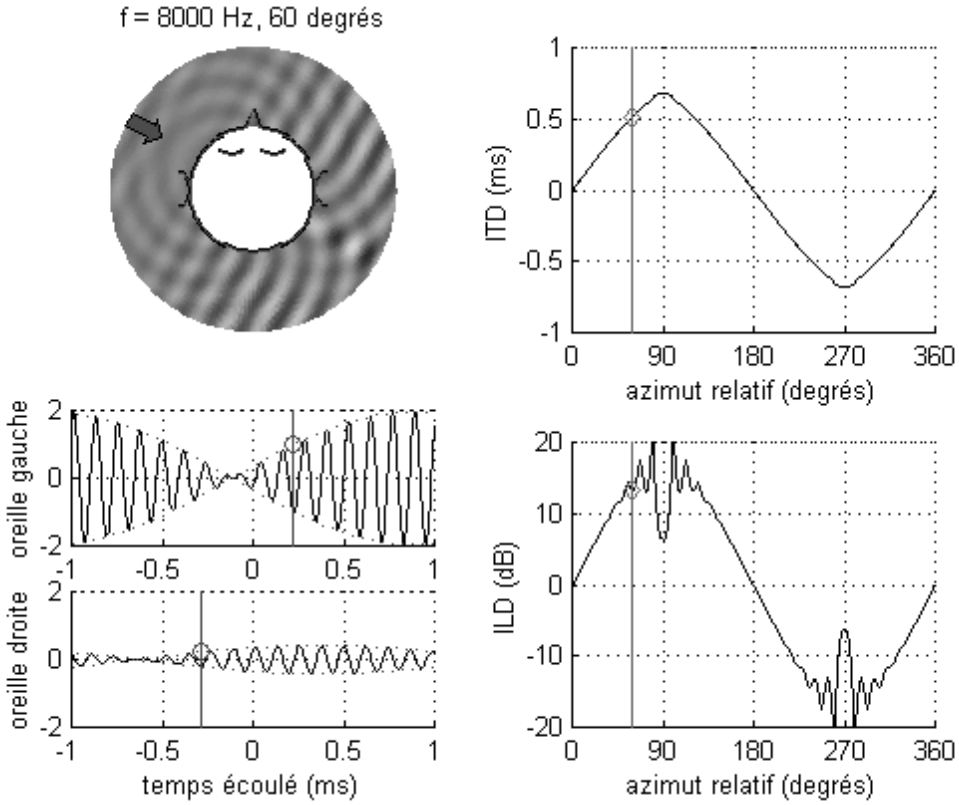

FIG. 1.8 – *Présentation similaire aux figures 1.6 et 1.7, sinon que c'est le champ de pression instantanée qui est représenté en niveau de gris. Cas d'un battement (deux fréquences voisines 8 kHz et 8,5 kHz), soit encore d'un sinus (*porteuse *à 8,25 kHz) modulé en amplitude (*modulante *à 250 Hz), la périodicité de l'enveloppe en valeur absolue étant donc de 500 Hz. Le retard est détectable non pas à travers la phase du signal (porteuse) mais son enveloppe (modulante), dont la période est supérieure au retard interaural. Pour le calcul de l'ILD, l'énergie perçue à chaque oreille est la somme de l'énergie associée à chaque son pur (chaque fréquence).*

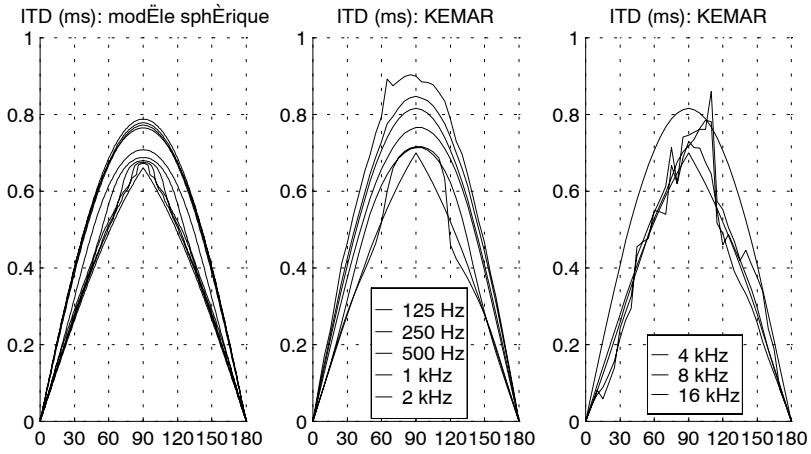

FIG. 1.9 – *Lois d'ITD (retard de phase) obtenues avec les modèles* sphère *et* KEMAR*, pour un jeu de fréquences. Les lois théoriques très basse-fréquence (la plus ample des deux) et très haute-fréquence pour un modèle sphérique sont ajoutées en gras. Pour la comparaison avec le modèle KEMAR, le diamètre de sphère équivalent est légèrement surévalué:* 18,5 *cm au lieu de* 17,5 *cm (à gauche).*

Nous retiendrons surtout que l'allure sinusoïdale de la loi d'ITD basse- et moyenne-fréquence subsiste jusqu'à environ 1,5 kHz[26].

Commentaires sur les modèles usuels

Les lois (1.45) et (1.47) rejoignent celles fréquemment rencontrées dans la littérature de référence [Bla83] [Beg94], mais il faut avoir conscience qu'elles correspondent à des modèles idéaux ou approximatifs, et qu'il est en toute rigueur abusif d'appliquer l'un ou l'autre sur la totalité des fréquences audibles.

Lois basse-fréquence et lois empiriques Les mesures et les simulations ont établi que les lois basse- et moyenne-fréquence pour l'ITD sont de forme sinusoïdale (jusqu'à environ 2 kHz), bien que de plus en plus approximative quand la fréquence augmente. Cette loi sinusoïdale (1.45) ne manque pas de rappeler les lois les plus couramment reportées et utilisées dans la littérature, la plupart du temps présentées comme résultant d'une approche empirique [Bla83], comme corrections au modèle de tête acoustiquement transparente (Figure 1.5b), et adoptées par commodité de calcul. Il est d'ailleurs intéressant de noter que Blauert indique comme correction un rayon équivalent de l'ordre de $1{,}2R$ à $1{,}3R$ [Bla83], ce que corrobore le rapport $(1+\pi/2)/2 \simeq 1{,}285$ des ITD maximaux entre la tendance haute-fréquence (1.47) et le modèle transparent (équ. 1.45 avec $D(f) = 2R$). D'autres auteurs (Griesinger par exemple) préconisent, dans un contexte de prise de son "sans tête", un espacement de l'ordre de 25 cm entre les microphones (contre un diamètre de tête de 17 cm, soit environ trois fois le rayon) qui permet d'obtenir un effet de latéralisation plus adéquat compte-tenu des basses fréquences, et par voie de conséquence un meilleur enveloppement (voir 1.3.4).

Digression sur l'onde de contournement et application aux sources proches Le modèle géométrique traditionnel de l'onde "de contournement minimum" – où l'onde contralatérale suivrait une ligne de courant tangente à la tête avec une vitesse constante ($= c$) – est en réalité inexact: il ne s'agit que d'une tendance

26. Il serait néanmoins intéressant de vérifier la loi de l'ITD (KEMAR) *aussi* en fonction du site.

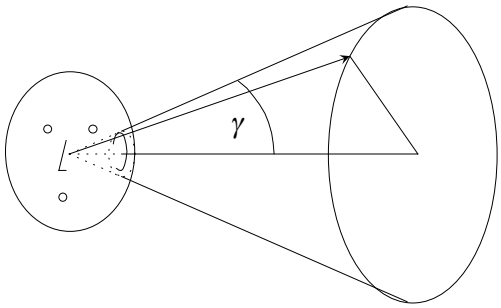

FIG. 1.10 – *Cône d'ambiguïté associé à une incidence quelconque (flèche): toute onde dont l'incidence appartient à ce même cône produit un ITD et un ILD semblable. La seule donnée de l'ITD ou de l'ILD ne permet donc pas une détection univoque de la direction d'incidence.*

asymptotique haute-fréquence! La trajectoire de contournement représentée Figure 1.5 n'a valeur que de "pseudo-ligne de courant", constituable par morceaux au regard de la description Figure 1.7: un morceau correspond à la propagation sur la surface cachée de la tête où la vitesse est à peu près égale à c en hautes-fréquences; pour trouver un deuxième morceau (demi-droite tangente à la sphère) il faut aller trouver un équivalent parallèle à distance de plusieurs longueurs d'onde de la tête (à l'écart des perturbations) en "glissant" sur un front d'onde pour se raccorder à une ligne de courant à peu près régulière qui remonte vers la source. C'est par ailleurs uniquement dans ce domaine haute-fréquence que les modèles géométriques "idéaux" pour les sources proches présentés dans ([Bla83], p.76) sont applicables!

Conclusion: utilisation des différences interaurales, effet de latéralisation

ITD, ILD, latéralisation "statique" et cône d'ambiguïté. Les lois d'ITD et d'ILD en fonction de l'incidence d'une source montrent de toute évidence qu'ils sont des *indicateurs de la position latérale d'une source*. Dans un contexte d'expériences d'écoute très général, il est bon d'ajouter que *l'ITD est un indice plus prégnant que l'ILD*. Des expériences d'écoute au casque l'ont montré (recensées dans [Bla83] ou encore [Mar96]), et on peut le comprendre intuitivement: l'ITD est en effet plus robuste que l'ILD (exemple d'une oreille modérément bouchée ou masquée), pas trop équivoque, car monotone, et relativement uniforme (le long des fréquences) par rapport à l'ILD. Cependant, *l'effet de latéralisation* – que nous qualifierons ici de "statique" – ne correspond qu'à une information de localisation très équivoque si la tête reste fixe: étant donné la symétrie – même approximative – de la tête par rapport à l'axe interaural, les différences interaurales ne permettent pas de différencier les incidences appartenant à un même cône, ayant lui-même pour axe de symétrie l'axe interaural $(O,\vec{y})$. C'est le *cône d'ambiguïté*, ou *cône de confusion* [Beg94] [Bla83] (Figure 1.10). Il sera illustré plus loin (en 1.3.3) comment les rotations de la tête permettent de résoudre cette ambiguïté.

Domaines d'action des mécanismes. Comme nous l'avons illustré, les mécanismes de détection de l'ITD d'après le déphasage interaural et d'après le retard de l'enveloppe d'amplitude se partagent respectivement, comme domaines d'action, les basses et les hautes fréquences. Plus précisément, il est communément admis ([Bla83], p.164) qu'une zone de chevauchement existe, située approximativement entre 1000 et 2000 Hz, selon le type de signal et son incidence. Pour l'ILD, qui devient significatif à partir de 2 ou 3 kHz, c'est une mesure globale – sur toute la bande de fréquence – qui est requise.

Effet global de latéralisation: pondération des indices ITD et ILD. A partir de l'estimation de l'ITD et de l'ILD, il est théoriquement possible de remonter à l'information (ambiguë) de position latérale. Mais dans les cas qui seront traités dans ce document, la "cohérence" de ces indices n'est en général pas garantie: l'ITD estimé en haute-fréquence peut par exemple indiquer une moindre latéralisation que l'ITD basse-fréquence. On peut imaginer une pondération des effets de l'ITD par bande critique en tenant compte de la répartition énergétique du signal, mais dans un cadre général, une telle modélisation reste discutable. Le problème de décision peut également se poser entre l'ITD et l'ILD: il serait alors possible, comme le suggère la théorie de Franssen [Fra64], de retenir comme effet global une pondération linéaire de la latéralisation due à chaque indice. Nous choisirons quant à nous de n'accorder à l'effet de l'ILD qu'une appréciation qualitative.

Variabilité fréquentielle de l'ITD: conséquence sur l'estimation. Si les lois "rigoureuses" que nous avons illustrées rejoignent partiellement les lois simplifiées traditionnelles, il est surtout important de *retenir qu'elles varient (en amplitude et en forme) en fonction de la fréquence*, l'ITD maximal diminuant quand la fréquence augmente. Le rapport maximal atteint entre les modèles basse- et haute- fréquence vaut $\max\left(\mathrm{ITD}^{LF}/\mathrm{ITD}^{HF}\right) = \frac{3}{1+\pi/2} \simeq 1{,}167$ dans le cas de la sphère, et encore plus, semble-t-il, avec un modèle plus réaliste (KEMAR). La conséquence sur le problème de l'estimation de l'ITD[27] est évidente – quoique souvent insoupçonnée –: elle justifie une estimation par bande, sinon continue en fréquence, plutôt que globale, y compris pour le cas de référence que constitue l'onde plane (section 1.4). Cette nécessité se fera encore plus cruciale lors de l'estimation de la variance de l'ITD (méthode originale pour mesurer l'"acuité de latéralisation"), basée sur l'étalement des enveloppes temporelles.

Loi d'ITD: intérêt de la forme sinusoïdale. Malgré la variation fréquentielle de l'amplitude de la loi d'ITD, l'intérêt de la forme sinusoïdale pour le domaine basse- et moyenne-fréquence reste entier: l'ITD peut être déduit à un facteur près par simple projection du vecteur incidence sur l'axe interaural. Cette propriété linéaire est avantageusement exploitée pour prédire l'effet de localisation (en terme de hauteur en plus de l'azimut, par exemple) selon les rotations de la tête (en 1.3.3 et en 1.5).

1.3.3 Compléments pour la détection du premier front d'onde: rotations et indices spectraux

Résolution des ambiguïtés par rotations de la tête: interprétation de la "latéralisation dynamique"

Les deux indices de localisation les plus couramment employés – l'ITD et l'ILD – sont donc responsables d'un *effet de latéralisation caractérisé par un angle* $\gamma = \arccos(\vec{u}.\vec{y})$ *que nous nommerons "***pseudo-angle***"*, qui, si la tête reste fixe, ne permet pas de discerner la provenance $\vec{u}$ d'un son parmi un ensemble de directions plausibles ($\{\vec{u}, \vec{u}.\vec{y} = \cos\gamma = \text{constante}\}$) définissant le *cône d'ambiguïté* (Figure 1.10). Nous montrons maintenant que des mouvements de la tête permettent de lever complètement l'ambiguïté [DRP99].

Supposons par exemple une rotation "*yaw*" de la tête (autour de l'axe $(O,\vec{z})$) d'angle azimutal γ, et une onde d'incidence $\vec{u}$ et de coordonnées absolues (initiales) (θ,δ). En utilisant les lois d'ITD basse-fréquence (1.45) et leur propriété linéaire (projection), l'ITD produit par rotation panoramique de la tête décrirait, pour une fréquence f donnée, une courbe $\mathrm{ITD}(\psi) = D(f)/c\,\sin(\theta-\psi)\cos\delta$, proportionnelle à la loi (1.45) telle qu'elle est présentée Figures 1.6 et 1.9, c'est-à-dire restreinte aux incidences horizontales ($\delta = 0$). Le rapport d'amplitude entre ces deux courbes vaut $\cos\delta$ et fournit donc l'information de site au signe près (indétermination haut-bas). Il suffit en fait d'une légère rotation pour avoir accès à cette information de site et à la connaissance complète de l'azimut θ [Ber75]. Ce dernier peut en effet être déduit – et notons-le,

27. De plus, cela remet en question certains modèles du processus psychoacoustique de localisation, basés sur une intercorrélation globale.

indépendamment du paramètre R/c – en considérant conjointement les valeurs de l'ITD et de sa dérivée par rapport à ψ [DRP99]:

$$\left\{ \begin{array}{ll} \mathrm{ITD}(\psi) & = \frac{D(f)}{c} \sin(\theta - \psi) \cos\delta \\ \frac{\partial \mathrm{ITD}(\psi)}{\partial \psi} & = -\frac{D(f)}{c} \cos(\theta - \psi) \cos\delta \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \sin(\theta - \psi) \\ \cos(\theta - \psi) \end{array} \right. \tag{1.48}$$

L'azimut ainsi détecté, la valeur absolue du site $|\delta|$ peut alors être déduite en fonction du paramètre R/c. Notons qu'une rotation de la tête s'accompagne des plus fortes variations d'ITD lorsque l'incidence est parallèle au plan de rotation de l'axe interaural (ou des oreilles). Pour désigner ce mécanisme de détection basé sur les variations de l'ITD lors des rotations de la tête, nous parlerons de "*latéralisation dynamique*" (ou "active"), par opposition à latéralisation "statique" (ou "passive"). Des développements similaires, quoiqu'un peu plus complexes, sont présentés dans [DRP99] et s'appliquent au modèle "haute-fréquence" (1.47).

Plus généralement, en usant de tous les degrés de liberté de rotation de la tête (de l'axe interaural), on comprend ([DRP99] [Beg94] par exemple) qu'il est objectivement possible de *détecter à la fois la direction d'incidence du front d'onde et sa vitesse de propagation* (ou bien, plus globalement, le paramètre R/c du modèle de l'ITD): l'auditeur peut trouver le plan de rotation de son axe interaural qui maximise les variations d'ITD (latéralisation dynamique). On peut supposer que cet "exercice", bien qu'inconscient, participe à un *processus d'apprentissage* de la localisation auditive (avec ou même sans vision associée), *au cours duquel la valeur du paramètre R/c est "étalonnée"*. Une fois cet étalonnage effectué et d'après la démonstration qui vient d'être énoncée, la seule rotation de la tête autour de son axe vertical doit permettre de cerner à la fois l'azimut et le site (en valeur absolue), dans l'hypothèse d'une vitesse de propagation ordinaire c. Comme commenté dans [DRP99] et plus loin (en 1.5.2), le cas d'*un front d'onde apparent de vitesse locale différente de c*– en général créé artificiellement *dans le cadre d'une restitution sur haut-parleurs – est susceptible de produire une image sonore "contre nature"* de caractère soit flou et diffus (ou avec une sensation de hauteur exagérée), soit instable et fuyant, *dans la mesure où l'auditeur est libre de tourner sa tête* même légèrement.

Indices spectraux et HRTF

A l'échelle des fréquences plus élevées (au-delà de 7 ou 8 kHz environ), les perturbations créées par la tête deviennent plus complexes. En particulier, l'onde sonore subit des réflexions dans le pavillon de l'oreille avant de parcourir le canal auditif, ce qui altère le spectre du signal perçu (effets de peignes: creux et pics) (Figure 1.11). Le signal sonore porté par l'onde subit donc une opération complexe de filtrage avant d'être reçu au niveau des tympans. Les fonctions de transfert binaurales associées par paire à chaque direction d'incidence sont appelées les HRTF (*Head-Related Transfer Functions*[28]).

Du fait de la complexité et de l'asymétrie de l'oreille, les *indices spectraux* sont caractéristiques du chemin parcouru, donc de la direction d'incidence (site (Cf.Thèse) et azimut) de l'onde. La figure 1.11 illustre la variété des spectres des HRTF (KEMAR) pour quelques incidences appartenant au même cône d'ambiguïté, là où un modèle de tête purement sphérique n'offre aucune possibilité de discrimination. Ce sont ces indices spectraux – en plus d'éventuelles rotations de la tête – qui laissent aux individus sourds d'une oreille une certaine aptitude à la localisation directionnelle: on parle d'ailleurs également d'*indices monoraux* (d'après R. Glasgal).

Note: les épaules et le torse participent à la perturbation du champ (diffraction et réflexions), et leur effet est normalement englobé dans la mesure des HRTF (sauf si elle est réalisée sur une tête artificielle sans

28. Les HRTF sont généralement exprimées dans le domaine fréquentiel, mais voient aussi leur nom couramment donné aux réponses impulsionnelles (HRIR: *Head-Related Impulse Responses*) dont elles sont les transformées de Fourier, et qui sont généralement mesurées à l'entrée du canal auditif.

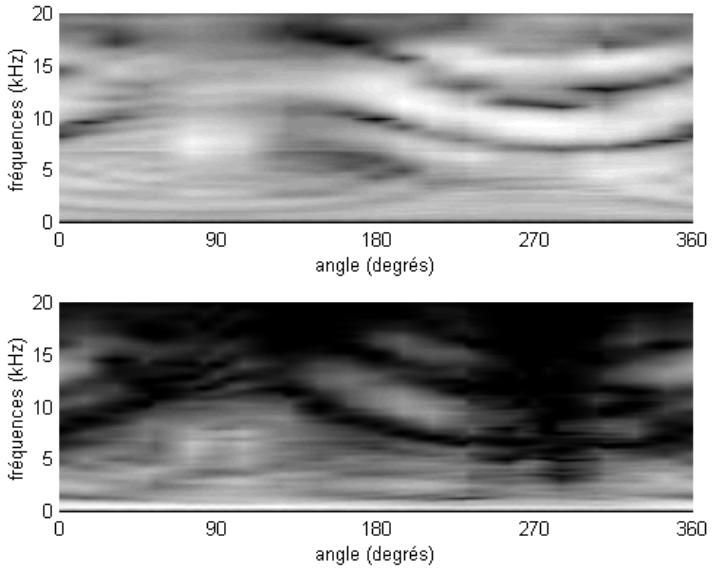

FIG. 1.11 – *Spectres d'énergie (gris clair pour les fortes énergies) des HRTF (KEMAR) gauche (en haut) et droite (en bas) pour une direction d'incidence parcourant approximativement le cône d'ambiguïté d'ouverture* 40° *par rapport à l'axe interaural (côté gauche). Parcours angulaire:* 0° *(incidence la plus de face),* 90° *(incidence la plus haute),* 180° *(incidence la plus arrière),* 270° *(incidence la plus basse).*

buste). Par contre, il faut signaler que la prise en compte de cette perturbation est figée une fois les mesures de HRTF réalisées, ce qui ne permet pas, à partir de ces seules mesures, de rendre compte des rotations de la tête avec exactitude (applications avec *Head-Tracking*, cf 2.5.1).

L'exploitation des indices spectraux pour la localisation est le résultat d'un apprentissage et d'une adaptation par l'individu qui dure depuis sa naissance. La localisation est d'ailleurs surtout efficace pour les sources familières, dont la connaissance du timbre permet de détecter et identifier l'altération du spectre. Les HRTF sont spécifiques à chaque individu, puisqu'elle dépendent directement de la morphologie (tête et pavillon de l'oreille), et il en est d'ailleurs de même pour le potentiel de discrimination directionnelle par les indices spectraux[29]. Ce qu'il est important de retenir lorsqu'on aborde la question de la restitution spatialisée, en particulier au casque, c'est que *la bonne restitution des indices spectraux* – si possible propres aux HRTF de l'individu – *est un des facteurs essentiels de la qualité de localisation et du naturel des images sonores*.

1.3.4 Effet des réflexions et de la réverbération, compléments psychoacoustiques

Loi du premier front d'onde ou effet d'antériorité

La localisation d'après le premier front d'onde, dont les mécanismes les plus connus viennent d'être décrits, suppose que celui-ci soit temporellement distinct des réflexions qui suivent. Des expériences basées sur la présentation artificielle de deux fronts d'ondes successifs ont dégagé des écarts temporels critiques de 0,630 ms et de 1 ms [Bla83], qui sont de l'ordre de grandeur de l'ITD maximal. Au-delà de 1 ms, le premier front d'onde qui détermine la direction perçue: c'est l'effet d'antériorité, encore connu sous le nom d'effet de Haas quand il s'agit d'expériences sur la parole. En-deçà de 630 μs, c'est une direction intermédiaire qui est perçue. Ces résultats doivent être cependant nuancés en fonction de la nature du signal émis, du rapport d'énergie entre les fronts d'onde successifs, de leur direction et de leur nombre. Quoiqu'il en soit, un écart de 1 ms correspondant à une différence de marche de 34 cm, l'arrivée de l'onde directe est en général bien distincte de celle des réflexions précoces.

Largeur apparente de source

Lorsque la source émet un signal aux faibles variations dynamiques, les réflexions précoces interagissent avec l'onde directe et produisent en quelque sorte une dispersion des informations directionnelles autour de la direction principale. Cela se traduit notamment par un effet de décorrélation interaurale. L'impression subjective qui en résulte est celle d'une extension apparente de la source en largeur.

Effet de distance

L'atténuation en $1/d$ de l'onde directe venant d'une source est à elle seule une donnée peu significative ou peu robuste pour l'estimation auditive de sa distance: même quand il est possible de se faire une idée de l'intensité émise d'après le timbre – dans le cas où les deux sont liés par le modèle de production du son, par exemple la voix – l'estimation de l'atténuation reste incertaine, faute de référence absolue du niveau sonore. Quant à l'absorption atmosphérique, qui altère le spectre du signal émis en atténuant particulièrement le domaine haute-fréquence [Beg94], elle ne devient significative qu'à partir de relativement grandes distances,

29. Néanmoins, le sujet humain est capable, au prix d'un ré-apprentissage, de s'adapter à d'autres HRTF, par l'intermédiaire d'une synthèse binaurale (section 2.5.1). Cette idée suscite un vif intérêt pour un certain nombre d'applications (aide auditive aux opérateurs, navigateurs aériens...), et pousse la recherche d'individus "super-localisateurs" [Beg94] ou de HRTF "supranormales" [SCDH88].

et ne constitue donc également qu'un indice peu fiable. Des informations beaucoup plus pertinentes, exploitables même sans que le timbre du signal émis soit connu ou familier, proviennent des réflexions et de la réverbération associées à l'onde directe. Pour décrire les choses de façon simplifiée: plus la source est proche, plus l'onde directe est perçue de façon"forte" et précoce par rapport aux réflexions et à la réverbération, et inversement.

Une tentative de *formalisation de l'information objective de distance* est proposée dans [Ger92f] sous le nom d'*hypothèse de Craven* (du nom de Peter Craven). Dans un modèle simplifié, l'amortissement des réflexions n'est supposé dû qu'à la décroissance en $1/d$ de l'onde sphérique réfléchie, de sorte que le rapport d'amplitude entre les contributions (onde directe et réflexions) et leur écart temporel, permettent de déduire leur retard absolu. Dans un deuxième temps, l'atténuation atmosphérique (exponentielle) est introduite dans le modèle. Gerzon exploite ce modèle pour la définition d'un *pan-pot* de distance. Il fait également de cette hypothèse un *critère de préservation de l'effet de distance par les systèmes de prise de son/restitution*: ce critère se base sur l'aptitude à restituer de façon conforme les antécédents physiques supposés de l'effet de distance, à savoir, selon Gerzon, sur la *préservation de l'énergie des réflexions venant de toutes directions*.

Le modèle de Gerzon devrait bien-sûr être corrigé en tenant compte de l'absorption des parois, du caractère spéculaire ou diffus des réflexions, mais aussi de la directivité et de l'orientation de la source: l'onde directe venant d'un chanteur est par exemple moins importante (surtout dans un domaine haute-fréquence) si celui si tourne le dos à l'auditeur. Ces facteurs introduisent autant de paramètres qui rendent difficile une estimation purement objective. Et c'est un fait: l'impression subjective de distance dépend également du son émis (voix chuchotée, voix criée, etc...).

Impression spatiale et enveloppement

Les sensations d'espace et d'enveloppement sont des qualités très recherchées pour l'écoute musicale dans une salle de concert, mais aussi dans un contexte de restitution stéréophonique et dans les applications de réalité virtuelle. Les nombreuses études qui se sont portées sur ces questions [Bar70][Gri96b] [Gri92] ont permis d'affiner progressivement la compréhension des causes physiques de ce qu'on appelle de façon plus générique "l'impression spatiale".

Il en ressort que ce sont les *réflexions latérales* qui sont principalement responsables de l'impression spatiale [Bar70]. En se manifestant à travers les différences interaurales, ces composantes latérales de l'effet de salle jouent en effet un rôle beaucoup plus informatif sur le plan de la perception, que les réflexions d'incidences proches du plan médian qui sont quant à elles surtout responsables d'un effet de coloration. Différents modèles ont été proposés pour tenter de corréler le degré et la qualité subjective de l'impression spatiale à une caractérisation physique de l'effet de salle mesuré au niveau de l'auditeur. Des outils mathématiques variés ont été introduits pour cela: la fonction d'intercorrélation appliquée aux réponses captés par deux microphones omnidirectionnels [PP88] ou bien appliquée à la réponse impulsionnelle binaurale (IACC); la fraction latérale d'énergie (LF) basée sur les réponses impulsionnelles de pression et de vélocité latérale (Barron); la mesure des fluctuations interaurales de phase (ITD) et d'intensité (ILD) [Gri92], etc... L'observation des seules réponses impulsionnelles ne suffit cependant pas à définir les qualités de l'impression spatiale. Griesinger distingue par exemple trois types d'impression spatiale qui dépendent des caractéristiques dynamiques du signal émis en plus de la répartition temporelle et énergétique des réflexions [Gri96b]:

- L'impression spatiale arrière (*Background Spatial Impression*) qui apparaît distinctement de la source en réponse à des signaux de type impulsif, impression dont l'intensité dépend de l'intensité de l'énergie réfléchie (après 55 ms).
- L'impression spatiale continue (*Continuous Spatial Impression*), en présence de sources continues ou d'extinction lente.

– L'impression spatiale précoce (*Early Spatial Impression*), essentiellement frontale et typique des petites salles. Elle n'est ni spacieuse ni enveloppante, et dépend du rapport d'énergie entre le son direct et l'énergie réfléchie (entre 2 et 50 ms). Pour des sources aux attaques peu marquées (legato), cette impression peut être interprétée comme l'effet d'une largeur apparente de source (*Apparent Source Width*).

Sans rapporter plus en détails les travaux sur l'impression spatiale, les interprétations qui s'en dégagent permettent de définir des critères pour juger l'*aptitude des systèmes de restitution à créer des effets d'espace ou d'enveloppement*. Pour doter la restitution de ces qualités, un système doit être capable d'assurer une décorrélation interaurale suffisante, ou encore de *reproduire des valeurs maximales d'ITD et d'ILD qui correspondent à l'effet de réflexions latérales, même si la scène sonore principale est frontale*.

Inhibition de l'effet de coloration

La sommation des réflexions et du son direct se traduit par un effet de coloration du signal de pression mesuré en un point: cette coloration peut être caractérisée physiquement en observant la transformée de Fourier de la réponse impulsionnelle mesurée. Cela se traduit également comme un effet auditif de coloration lorsque *ce même signal* est présenté aux deux oreilles, où à l'une d'entre elles[30]. En situation réelle cependant, si la diversité des incidences de réflexions par rapport à l'axe interaural est suffisante, *l'effet subjectif de coloration est inhibé par la décorrélation des signaux binauraux*. C'est pourquoi les réflexions d'incidences proches du plan médian de l'auditeur (réflexions par le sol ou le plafond), qui ne contribuent pas à l'impression spatiale, participent à l'effet subjectif de coloration. Au contraire, les composantes latérales des réflexions ont la fonction inverse.

Ces observations ont des implications non-négligeables dans un contexte de reproduction sonore spatialisée, notamment sur haut-parleurs: lorsque que la décorrélation des signaux reconstitués aux oreilles est réduction[31] par rapport aux signaux stéréophoniques issus d'une prise de son, les impressions spatiales qui auraient été éprouvées par l'auditeur dans la scène originale sont également dégradées **et** partiellement converties en un effet subjectif de coloration, selon le degré de dégradation.

Analyses perceptives approfondies

En complément des effets "classiques" des réflexions et de la réverbération, il est bon de souligner l'existence d'interprétations perceptives plus subtiles, même si elles ne s'imposent pas forcément à la conscience. En plus de la caractérisation en termes d'attributs subjectifs ("chaleur", "brillance", "présence", etc...) [JKM+93][32], l'expérience d'écoute peut par exemple conduire à l'identification de l'environnement acoustique, et même des matériaux qui constituent le revêtement des parois. L'analyse perceptive des données auditives peut encore se manifester de façon plus subtile bien qu'en général inconsciente – et sans doute moins nette que chez certains animaux, comme les chauve-souris! –: par la sensation, voire la détection, de la proximité de parois ou d'objets. C'est une aptitude qui a été observée chez certains aveugles [Beg94]. Dans les applications de réalité virtuelle, il est possible qu'une bonne modélisation des phénomènes de réflexion et de diffraction améliore la sensation de présence des objets et des parois représentés dans la scène virtuelle.

30. ... comme on peut en faire l'expérience, dans la vie courante, en se bouchant simplement une oreille (d'après Jean-Dominique Polack).

31. Comme il sera développé dans le chapitre suivant, cette réduction de la décorrélation est généralement le fait du *cross-talk*: le mélange (ou diaphonie) des signaux stéréophoniques pendant leur transmission entre les haut-parleurs et les oreilles.

32. Voir aussi Beranek, Plenge *et al*, Schroeder *et al*...

1.4 Méthodes mathématiques pour l'estimation des indices de localisation

1.4.1 Intérêts du problème

Le problème de l'estimation des indices de localisation – et de l'effet de localisation lui-même – peut se poser dans le cadre de préoccupations différentes:

- Tenter de comprendre et reproduire les *mécanismes psychoacoustiques de localisation*, *à partir de flux sonores* mesurés au niveau des oreilles. Cela exige d'intégrer des modèles complexes (rapportés par exemple dans [Bla83], [Dud], [BCDS89a]) reproduisant les différentes étapes du processus auditif: filtrage par l'oreille (externe et moyenne), transduction "mécano-nerveuse" (oreille interne, cochlée), traitement interaural... Ce type de modélisation doit permettre de réaliser la séparation des flux sonores (effet "cocktail-party") et leur localisation, et se contente difficilement, en tous cas, des méthodes présentées plus bas.
- Répondre au besoin technique d'un modèle efficace de représentation des HRTF, notamment pour une implémentation économique de la synthèse binaurale (voir plus loin en 2.5.1). Il s'agit en particulier, *à partir de réponses impulsionnelles binaurales*, d'estimer séparément les retards (ITD) et la structure des filtres à phase minimale [LJ97].
- Evaluer objectivement les performances relatives – en termes de localisation et de qualité d'image sonore – de systèmes de restitution spatiale. Les effets supposés de localisation sont en général estimés indépendamment du contenu sonore, c'est-à-dire *à partir des réponses impulsionnelles binaurales synthétiques*, associées à la restitution d'une image virtuelle à l'aide de plusieurs haut-parleurs, les HRTF originales (source unique) servant alors de référence. C'est ce dernier aspect qui motive principalement cette section, dont les méthodes sont appliquées au chapitre 3 (section 4.1) et plus partiellement dans [DRP98].

Les méthodes présentées ici se concentrent sur l'estimation de l'ILD et surtout de l'ITD, indice plus prégnant, mais aussi plus problématique à estimer. L'estimation de l'effet de localisation est donc restreinte ici à l'effet de *latéralisation*. Nous faisons remarquer par ailleurs que les expressions qui sont données sous forme intégrale – destinée aux signaux à support continu – sont évidemment directement transposables au domaine des signaux à support discret (typiquement les signaux numériques), en les remplaçant simplement par des sommes.

1.4.2 Méthodes classiques ou existantes

Rappel de méthodes d'après flux sonores

La méthode la plus classique pour mesurer le retard interaural associé à la perception d'une source sonore, repose sur l'inter-corrélation[33] des signaux binauraux $s_l(t)$ et $s_r(t)$, ici dans une version normalisée [Bla83]:

$$\mathrm{IACC}_{lr}(\tau) = \frac{\int_{-\infty}^{\infty} s_l(t-\tau) s_r(\tau)\, dt}{\sqrt{\int_{-\infty}^{\infty} s_l^2(t)\, dt \int_{-\infty}^{\infty} s_r^2(t)\, dt}} \tag{1.49}$$

La fonction $|\mathrm{IACC}_{lr}(\tau)|$ est majorée par 1. Le retard interaural se définit comme le lieu de son maximum:

$$\mathrm{ITD}_{IACC} = \underset{\tau}{\mathrm{Argmax}}\left(|\mathrm{IACC}_{lr}|\right) \tag{1.50}$$

33. IACC: *Inter-Aural Cross Correlation*.

Notons que le changement de signe de l'un des deux signaux s_l et s_r n'a pas d'incidence sur l'estimation, ce qui n'est pas pertinent sur le plan perceptif, en tous cas dans un domaine basse-fréquence ($f < 2kHz$). De plus, le processus modélisé n'est ni causal, ni adaptatif (ou plutôt, il ne tient pas compte d'un effet d'oubli). Un modèle traduisant mieux ces aspect est reporté dans [Bla83]. Il est en général plus juste de réaliser une inter-corrélation à court-terme, et d'opérer un traitement par bande de fréquence critique (échelle des *Barks* ou des *Erbs*) [Bla83] [BCDS89b] [ZF81] [Dur98], ou au moins introduire une pondération fréquentielle des signaux sonores qui reflète la courbe d'audition.

Méthodes d'après réponses impulsionnelles

Le travail d'après réponse impulsionnelle traduit la présence d'une seule source [image] sonore, réelle ou virtuelle (plusieurs sources corrélées). On doit garder à l'esprit qu'en réalité, la perception dépend étroitement du contenu (structure temporelle, fréquentielle, sans parler de la connotation sémantique, prise en compte dans l'étape précoce d'identification (Cf 1.1.2). Il faudra donc relativiser les mesures, selon que la source émet un signal stationnaire, modulé, comprenant des transitoires, restreint à une bande de fréquence, etc...

Commençons par l'estimation peu problématique de l'ILD et de l'ITD basse-fréquence (retard de phase).

ILD (*Interaural Level Difference*)

Pour chaque incidence, l'ILD peut être estimé pour chaque fréquence d'après la différence des spectres gauche et droit en dB:

$$\mathrm{ILD}(f) = 10\log_{10}\frac{|H_l(f)|^2}{|H_r(f)|^2} \tag{1.51}$$

Cette mesure n'est pas fréquemment employée dans la littérature. On préfère souvent une estimation globale, pleine bande:

$$\mathrm{ILD} = 10\log_{10}\frac{\int |H_l(f)|^2\,df}{\int |H_r(f)|^2\,df} = 10\log_{10}\frac{\int (h_l(t))^2\,dt}{\int (h_r(t))^2\,dt} \tag{1.52}$$

D'après nos conventions, l'ILD est positif pour une onde plane d'incidence latérale gauche.

ITD basse-fréquence (retard de phase)

L'ITD basse-fréquence est quant à lui défini en fonction de la fréquence comme *différence des retards de phase* entre les réponses droite et gauche. Il est lui aussi positif pour une onde plane d'incidence latérale gauche. En écrivant les réponses fréquentielles binaurales ainsi: $H_l(f) = |H_l(f)|e^{j\varphi_l(f)}$ et $H_r(f) = |H_r(f)|e^{j\varphi_r(f)}$, il s'écrit:

$$\mathrm{ITD}_{phase}(f) = \frac{\varphi_l(f) - \varphi_r(f)}{2\pi f} \tag{1.53}$$

ITD haute-fréquence: retard de groupe

L'ITD n'est détectable en haute-fréquence qu'en tant que retard d'enveloppe, d'après 1.3.2. Un autre outil du traitement du signal peut être utile, sous certaines conditions, pour son estimation d'après le signaux binauraux. Il s'agit du calcul du *retard de groupe* (par exemple [Bla83]):

$$\mathrm{ITD}_{group}(f) = \frac{1}{2\pi}\left(\frac{\partial\varphi_l}{\partial f} - \frac{\partial\varphi_r}{\partial f}\right) = \frac{1}{2\pi}\left(\frac{\partial \arg(H_l/H_r)}{\partial f}\right) \tag{1.54}$$

C'est encore une mesure qui dépend de la fréquence. Il faut être conscient qu'elle ne caractérise vraiment le retard d'enveloppe que pour des signaux à bande étroite, pour lesquels l'enveloppe varie assez lentement par rapport à la fréquence centrale. Calculée sur des réponses comme les HRTF, on constate qu'elle présente de fortes variations en fonction de la fréquence, ce qui la rend très difficilement exploitable sous cette forme [DRP98] (Cf Annexe B). La possibilité d'en déduire une estimation plus globale est discutée plus loin.

ITD global ou haute-fréquence: maximum de corrélation interaurale

La méthode basée sur la corrélation interaurale énoncée plus haut (équations 1.49 et 1.50) peut également être appliquée aux réponses impulsionnelles binaurales h_l et h_r – en nommant $\mathrm{IACC}_h(\tau)$ la fonction correspondante –, assurant du même coup les conditions de causalité et de limitation temporelle (puisqu'il s'agit de réponses causales et à support fini). Cette méthode, appliquée en pleine bande ou seulement sur les réponses préalablement filtrées, n'est cependant que partiellement satisfaisante: d'une part, elle donne peu d'indication sur l'acuité de localisation (à la rigueur à travers le maximum de la fonction d'intercorrélation); d'autre part, elle est moyennement fiable, dans le sens où lorsque plusieurs pics sont présents sur la courbe de $\mathrm{IACC}_h(\tau)$, elle ne prend en compte que le plus élevé et n'accorde aucune importance aux autres, même s'ils sont du même ordre de grandeur (Figure?). Cela peut créer des discontinuités d'estimation lors du déplacement d'une image sonore, qui ne correspondent d'ailleurs absolument pas à une réalité perceptive!

Remarque: Malgré l'étalement de la loi d'ITD entre les basses et les hautes fréquences, c'est une tendance "loi haute-fréquence" que marque nettement l'estimation globale par cette méthode (appliquée sur les HRTF de KEMAR). Cela peut s'expliquer par le fait que dans le domaine fréquentiel, l'importance de l'intercorrélation est traduite suivant une échelle linéaire, et que de ce fait, le domaine haute-fréquence s'en trouve favorisé.

Autres méthodes

Une étude comparative de différentes méthodes d'estimation de l'ITD a été offerte par Véronique Larcher [LJ97], où est évoquée, outre celles que nous avons ou allons citer, une méthode de détection par seuil. Citons encore Jot et al [JLP99], qui proposent une estimation d'après les segments quasi-linéaires d'excès de phase. Elle est intéressante dans la mesure où elle permet de discriminer l'effet de localisation par bande de fréquence. Cependant, dans [JLP99], la méthode est exploitée en ne tenant compte que du segment le plus long sur une bande haute-fréquence ou basse-fréquence, afin de donner une estimation unique de l'ITD pour chaque bande. De ce fait et de la même manière que pour l'estimation d'après le maximum de la corrélation interaurale, les résultats ne sont que partiellement représentatifs de l'ITD globalement perçu.

ITD haute-fréquence: intercorrélation des enveloppes d'amplitude

Il est reconnu par ailleurs que l'oreille n'est pas sensible à l'information de phase dans le domaine hautes-fréquences (au-delà de 5 kHz). Pour des signaux revêtant un caractère impulsionnel, transitoire ou modulé en amplitude, par exemple, cela justifie donc de baser la détection des époques d'arrivée de groupes – soit de l'ITD – sur l'enveloppe d'amplitude des signaux binauraux plutôt que sur les signaux eux-mêmes. Techniquement, le calcul de l'enveloppe est réalisé en utilisant la transformée de Hilbert, qui permet d'obtenir à partir d'un signal réel $s(t)$, une version $\sigma(t)$ en quadrature (déphasée de 90°), pour former un signal analytique $\Psi(t) = s(t) + j\sigma(t)$. On en déduit son enveloppe temporelle d'amplitude $a(t) = |\Psi(t)| = |s(t) + j\sigma(t)|$. L'application de la méthode aux réponses binaurales $h_l(t)$ et $h_r(t)$ donne les signaux d'enveloppe respectifs $a_l(t)$ et $a_r(t)$. Pour une estimation séparée des indices de localisation basses- et hautes-fréquences, chaque réponse $h_l(t)$ et $h_r(t)$ pourra subir préalablement un filtrage passe-haut. Cela est d'autant plus recommandé

que la loi de l'ITD en haute fréquence fournit des valeurs plus petites qu'en basse fréquence, comme nous l'avons montré en 1.3.2. Le calcul de la corrélation interaurale normalisée est alors transposé comme suit:

$$\mathrm{IACC}_a(\tau) = \frac{\int_{-\infty}^{\infty} a_l(t-\tau)a_r(\tau)\,dt}{\sqrt{\int_{-\infty}^{\infty} a_l^2(t)\,dt \int_{-\infty}^{\infty} a_r^2(t)\,dt}} \tag{1.55}$$

Là encore, l'inégalité $\chi_{lr}(\tau) \leq 1$ est toujours vérifiée. L'estimation traditionnelle de l'ITD à partir de l'intercorrélation se base sur l'observation du maximum de la fonction $\chi_{lr}(\tau)$:

$$\mathrm{ITD}_{iacc^a} = \mathrm{argmax}_\tau \chi(\tau) \tag{1.56}$$

Bien que cette méthode semble fournir dans des cas défavorables des résultats parfois plus judicieux que la précédente, elle fait l'objet de remarque similaires quant à sa fiabilité et l'absence d'indication sur l'acuité de localisation. Elle peut produire une estimation très discontinue – qui reflète peu l'impression subjective – de l'effet de localisation des images sonores lors d'une restitution sur haut-parleurs.

1.4.3 Méthodes originales avec intervalle de confiance ou indice d'acuité

ITD global ou haute-fréquence: époques moyennes des enveloppes d'énergie (théorie énergétique)

Outre les méthodes basées sur les maxima des fonctions d'intercorrélation, il existe une autre approche d'estimation globale du retard, peu exploitée semble-t-il d'après la littérature, mais pourtant séduisante. Se référant à la *théorie énergétique* de Mertens [Mer62] [Mer65], Condamines [Con78] rapporte une méthode d'"*évaluation de la différence des époques moyennes de groupe des signaux binauraux élémentaires, lesdits signaux correspondant à la décomposition en impulsions gaussiennes*[34] *du signal analytique initial*". Le signal analytique $\Psi(t)$ associé à une impulsion de Gauss d'époque moyenne de groupe t_0 et de fréquence moyenne f_0 s'écrit:

$$\Psi(t) = e^{-\mu^2(t-t_0)^2} e^{2\pi j f_0 t} \tag{1.57}$$

Rappelons que le propre d'une impulsion de Gauss est de traduire un compromis optimal entre définition temporelle et définition fréquentielle, l'une étant inversement proportionnelle à l'autre et fixée par le paramètre μ. Les cas extrêmes sont la sinusoïde pure ($\mu = 0$) et l'impulsion de Dirac $\delta(t)$ (si $\mu = \infty$). Dans le cadre de ses expérimentations axées sur la restitution stéréophonique, c'est en réalité à l'émission par les haut-parleurs que Mertens applique ce choix d'impulsions de Gauss, pour en quantifier les conséquences sur les indices de localisation. Nous allons nous même, dans un premier temps, appliquer cette analyse aux conséquences de l'émission d'une impulsion infiniment brève ($\mu = \infty$), c'est-à-dire effectuer l'estimation des époques moyennes des enveloppes sur les réponses binaurales "brutes" (pleine-bande). En réutilisant l'enveloppe d'amplitude $a(t)$ introduite plus haut, on peut montrer que l'époque moyenne d'arrivée de chaque réponse impulsionnelle s'écrit:

$$\tau = \frac{\int_{-\infty}^{\infty} a^2(t)t\,dt}{\int_{-\infty}^{\infty} a^2(t)\,dt} \tag{1.58}$$

Pour appliquer cette méthode en toute généralité, il serait plus respectueux des propriétés psychoacoustiques de choisir un intervalle d'intégration temporelle borné à une durée de l'ordre de 0,630 μs, limite de séparation de deux impulsions au-delà de laquelle l'effet de sommation n'est plus valide et l'on peut percevoir

34. Ou décomposition en ondelettes (de Gabor).

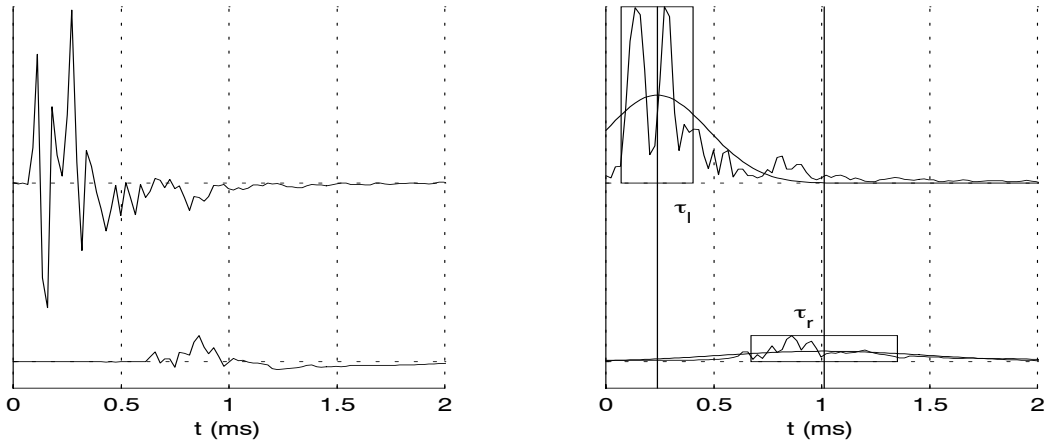

FIG. 1.12 – *Détection de l'ITD d'après la différence des époques moyennes des enveloppes d'amplitude. A gauche, les réponses binaurales temporelles h_l et h_r associées à l'azimut* 60°. *A droite, leurs enveloppes d'amplitude a_l et a_r, et leur modélisation statistique d'ordre 2: les époques moyennes τ_l et τ_r sont marquées par des lignes verticales dont l'écart définit l'ITD; les variances associées définissent les demi-largeurs des cadres rectangulaires; la caractérisation de chaque enveloppe étant restreinte aux moments d'ordre 1 (moyenne) et 2 (variance), un modèle d'enveloppe équivalent est donné par une gaussienne (forme de cloche).*

des événements auditifs distincts [Bla83]. Cela dit, les réponses temporelles que nous aurons à traiter étant en général relativement courtes, et, le cas échéant, souvent composées d'impulsions suffisamment denses, une telle correction sera inutile. L'équation (1.58) appliquée aux réponses impulsionnelles gauche et droite, fournit des valeurs τ_l et τ_r qui permettent d'extraire l'ITD:

$$\mathrm{ITD} = \tau_l - \tau_r \tag{1.59}$$

Dans le but d'affiner l'estimation et parce que l'amplitude de la loi d'ITD varie entre les basses et les hautes fréquences, nous recommandons plus loin d'effectuer au préalable un filtrage passe-haut des réponses binaurales (Figure 1.13). Cette méthode d'estimation est illustrée Figure 1.12.

Information sur l'acuité de l'estimation: utilisation de la variance

Aucune des méthodes "traditionnelles" ne semble accompagner les estimations fournies d'un indice clair sur le crédit à accorder à leur précision[35] ("*indice de confiance*"), alors même que ces estimations peuvent présenter des variations importantes d'une méthode à l'autre et/ou pour de faibles changements de conditions. Ces mesures se révèlent d'un caractère d'autant plus arbitraire que les discontinuités observées dans les estimations ne sont nullement perçues comme telles la plupart du temps. Il est pourtant dangereux, par exemple lors de l'évaluation objective des qualités d'image sonore (souvent en terme de localisation) produite par un système de restitution (cf 4.1), de ne se fier qu'à une mesure catégorique de l'ITD sans savoir si elle est vraiment pertinente. Pour pallier ce manque, nous proposons de compléter l'estimation des époques moyennes, de nature semble-t-il relativement stable et robuste, par celle de leur variance (écart-type), qui donne une

35. ... sinon le maximum de l'IACC normalisée, mais celui-ci ne permet pas de définir un intervalle de confiance.

indication de l'étalement temporel. A partir de la variance temporelle associée à chaque enveloppe:

$$\sigma_\tau = \sqrt{\frac{\int_{-\infty}^{\infty} a^2(t)(t-\tau)^2\,dt}{\int_{-\infty}^{\infty} a^2(t)\,dt}}, \tag{1.60}$$

nous déduisons une variance associée à l'estimation de l'ITD:

$$\sigma_{ITD} = \sqrt{\sigma_{\tau_l}^2 + \sigma_{\tau_r}^2} \tag{1.61}$$

σ_{ITD} donne en quelque sorte une marge d'incertitude sur l'estimation de l'ITD (Figure 1.13). L'équation (1.61) signifie bien que cet indice traduit une double incertitude, schématisée Figure 1.12 pour chaque oreille par l'étalement d'une gaussienne de forme $e^{-\mu^2(t-\tau)^2}$, où $\mu = 1/(2\sigma_\tau)$.

Variabilité fréquentielle: sélection de régions haute-fréquence

On peut constater Figure 1.13 que la variance est relativement importante, même mesurée pour des paires de HRTF (source unique). Au moins une raison semble s'imposer: il a été mis en évidence en 1.3.2 que la loi d'ITD est de plus grande amplitude pour les basses fréquences que pour les hautes fréquences. A cause de cette variabilité fréquentielle de l'ITD, l'information de la variance à partir des réponses pleine-bande ne permet pas distinguer très nettement le cas des réponses de références (HRTF originales, source unique) du cas des réponses synthétiques, plus complexes (voir 4.1).

Il semble donc nécessaire de dissocier au moins les parties haute- et basse-fréquence des réponses binaurales à traiter, en ignorant carrément la partie basse-fréquence qui ne s'inscrit pas dans le mécanisme de détection par enveloppe. La figure 1.13 montre que l'estimation de l'ITD par cette méthode est progressivement affinée lorsque les réponses binaurales subissent préalablement un filtrage passe-haut avec une fréquence de coupure de plus en plus élevée: l'ITD colle de mieux en mieux à la loi haute-fréquence, et la variance associée diminue par rapport au traitement sans pré-filtrage. Cette dernière finit cependant par stagner à une valeur non négligeable. Deux causes peuvent être avancées: même en considérant une région haute-fréquence, chaque réponse a un étalement temporel naturel qui résulte de réflexions multiples et d'effets complexes de diffraction au niveau l'oreille, particulièrement pour les petites longueurs d'onde[36]; par ailleurs, un filtrage tend à augmenter l'étalement temporel d'autant plus qu'il est plus sélectif.

Discussion: pour un affinement des méthodes

Il aurait pu paraître astucieux de vouloir appliquer cette opération sur des bandes critiques pour recueillir un ITD ou un pseudo-angle[37] spécifique à chaque bande, puis établir la moyenne[38] et la variance de ces estimations pour caractériser précisément le degré et l'acuité de latéralisation. Malheureusement, le principe d'incertitude temps-fréquence veut qu'un filtrage à bande étroite, même d'un signal impulsionnel, provoque l'étalement de sa réponse temporelle, rendant impuissante la détection d'enveloppe, et même la méthode par intercorrélation, qu'elle soit appliquée aux signaux filtrées[39] comme à leur enveloppe.

36. C'est sans-doute dans ce type de phénomène qu'il faut chercher l'explication de la singularité observée aux alentours de l'azimut 110° (cas du contournement le plus long pour la réponse contralatérale).

37. On définit un pseudo-angle comme l'angle d'ouverture, par rapport à l'axe interaural, du cône d'ambiguïté correspondant à l'ITD pour la fréquence considérée.

38. ... éventuellement pondérée par l'énergie contenue dans chaque bande.

39. En appliquant cette méthode à l'aide des filtres de [Sla], il a semblé que leur sélectivité fréquentielle était parfois trop sévère, en dehors d'un domaine basse-fréquence, pour une estimation univoque des retards lorsqu'ils étaient importants (problèmes de sauts

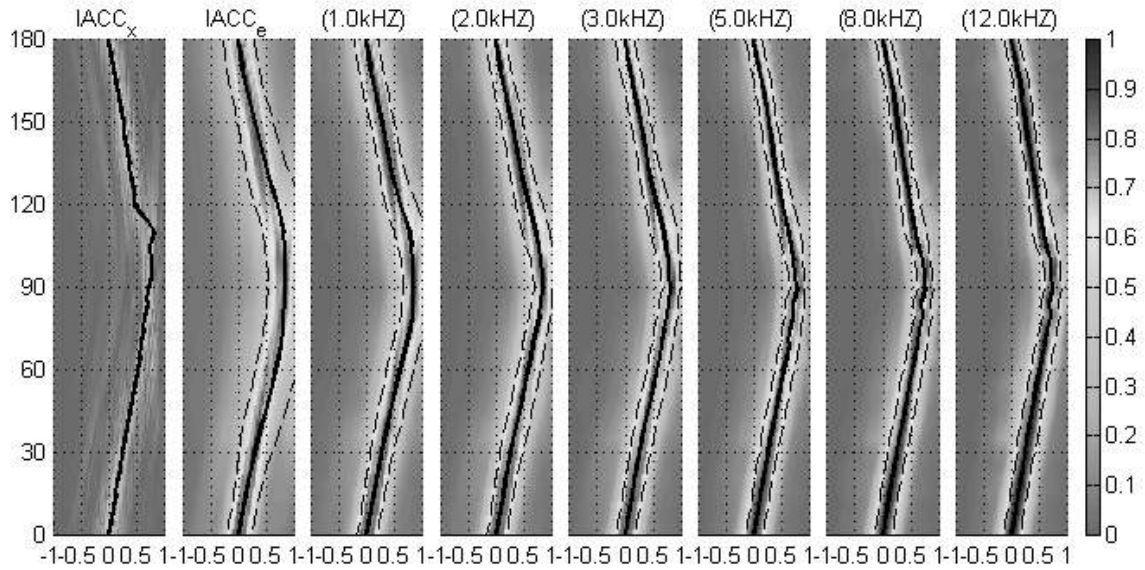

FIG. 1.13 – *Exemples d'estimation de l'ITD (en ms et en abscisses) et de l'éventuelle variance associée, pour un parcours semi-panoramique des HRTF du KEMAR (angles en degrés, en ordonnées). De gauche à droite: détection par maximum (courbe pleine) de la fonction de corrélation interaurale (niveau de gris); puis, sur fond d'intercorrélation des enveloppes d'amplitudes, estimation de l'ITD comme différence des époques moyennes des enveloppes d'énergie (courbe pleine), la variance associée définissant autour de la courbe de l'ITD un intervalle de confiance (bande délimitée par les courbes tiretées); les autres cas représentent le même procédé, mais précédé d'un filtrage passe-haut (filtre du second ordre, ici), de fréquences de coupure respectives 1 kHz, 2 kHz, 3 kHz, 5 kHz, 8 kHz et 12 kHz.*

D'un point de vue pratique, le filtrage préalable par bande critique semble, du fait de l'étroitesse des bandes, imposer des conditions justement trop critiques aux méthodes d'estimation que nous cherchons à leur appliquer. Il aurait été par ailleurs illusoire de croire qu'on traduirait plus fidèlement des propriétés psychoacoustiques par le simple biais d'un filtrage par bande critique, sans élaboration de modèles plus subtils des traitements en aval, d'autant plus que nous travaillons ici à partir de réponses impulsionnelles. Plutôt que des mécanismes psychoacoustiques[40], les méthodes mathématiques que nous développons doivent avoir comme objectif de *traduire au mieux le potentiel de discernement de l'information directionnelle* offerte par les réponses gauche et droite aux oreilles.

La *première exigence* qui s'impose, c'est que l'estimation objective fournie ne soit pas en contradiction flagrante avec l'effet subjectif de localisation, et qu'un indice de confiance associé à l'estimation définisse d'une part un intervalle d'indétermination qui couvre les valeurs issues d'estimations subjectives, et traduise d'autre part l'acuité directionnelle subjective. Parmi les méthodes présentées plus haut, celle utilisant les époques moyennes d'enveloppes et les variances associées paraît remplir le mieux ces conditions, quoique l'indication d'acuité reste grossière et très floue, même dans le seul cas d'une source unique (HRTF originales). L'*objectif suivant* est donc de pouvoir traduire l'effet subjectif supposé (*degré et acuité de latéralisation*) avec plus de pertinence et de discernement. Le problème, nous l'avons vu, semble résider dans l'étalement temporel des réponses binaurales, ce qui fournit une variance importante à l'estimation de l'ITD même dans les cas d'identité gauche-droite (par exemple une source frontale: $\theta = 0$). Par ailleurs l'anonymat du contenu fréquentiel de chaque réponse dû à la seule considération de leurs enveloppes calculées sur une large bande constitue un biais, donc un danger pour l'interprétation.

Intégration pondérée du retard de groupe

Pour contourner ce problème, l'idéal serait de faire précéder le traitement par la réduction des réponses par un facteur commun, soit encore par leur déconvolution par une réponses commune[41]. Suivant des idées du même goût, *une solution relativement simple se dégage* de l'association de plusieurs approches parmi celles évoquées plus haut. En effet et de façon assez naturelle, l'estimation des retards de groupes en fonction de la fréquence (1.54) *élimine le "facteur commun" des réponses*. On peut en déduire une estimation globale par calcul d'une *moyenne avec pondération énergétique*:

$$\mathrm{ITD} = \frac{\displaystyle\int_0^\infty |H_l H_r| \frac{1}{2\pi} \frac{\partial \arg(H_l/H_r)}{\partial f}\, df}{\displaystyle\int_0^\infty |H_l H_r|\, df} = \frac{\displaystyle\int_0^\infty |H_l H_r^*| \frac{1}{2\pi} \frac{\partial \arg(H_l H_r^*)}{\partial f}\, df}{\displaystyle\int_0^\infty |H_l H_r^*|\, df} \tag{1.62}$$

On obtient la même valeur en remplaçant H_l et H_r par les transformées de Hilbert Φ_l et Φ_r de h_l et h_r dans le domaine fréquentiel. Il est intéressant de noter la similitude de cette expression avec une définition

de phase), se traduisant par l'apparition (épisodique) de grosses erreurs d'estimation pour les incidences latérales. Il semble pourtant que ce soit cette méthode qui a été adoptée par Wightman et al (cités par [LJ97]). Peut-être employaient-ils des filtres un peu moins sévères?

40. D'ailleurs nos méthodes ne peuvent pour le moment prétendre traduire les imperfections de l'ouïe directionnelle, comme on pourrait le faire dans le domaine de la compression audio-numérique [Dur98], puisqu'elles ne sont déjà pas capables de la même acuité avec la même robustesse!

41. Notons que ce traitement peut faire disparaître des réponses initiales des propriétés symptomatiques qui ne sont pas neutres pour la perception auditive: il s'agit en particulier de la coloration qui leur est associée, et qui peut détériorer le naturel de l'image sonore.

équivalente à (1.58) que donne Mertens [Mer62] pour l'estimation des époques moyennes de groupe:

$$\tau = -\frac{1}{2\pi}\frac{\int_0^\infty \Phi\Phi^* \frac{\partial \arg \Phi}{\partial f}\, df}{\int_0^\infty \Phi\Phi^*\, df} \quad , \tag{1.63}$$

où $\Phi(f)$ est la transformée de Fourier du signal analytique $\Psi(t)$, en utilisant les notations exposées plus haut.

Nous avons donc détourné, en quelque sorte, l'expression (1.63) pour obtenir une *estimation globale de l'ITD à partir d'informations élémentaires relatives* (retard interaural de groupe pour chaque fréquence), *beaucoup plus efficace que la différence de deux estimations globales indépendantes* (1.59). La pondération par le terme $|H_l H_r|$ permet de minimiser l'effet de la plupart des sauts de phase de chaque réponse $\arg H_l(f)$ ou $\arg H_r(f)$, qui coïncident souvent avec un "passage à vide" dans le spectre d'énergie associé $|H_l(f)|$ ou $|H_r(f)|$. Pour une meilleure acuité, il est recommandé de réserver ce traitement au domaine haute-fréquence (au-delà de 1 ou 2 kHz). Mais il n'est pas forcément avantageux de l'effectuer par sous-bandes, car choisir des bandes plus étroites rend en fait l'estimation plus inégalitaire, même sur des réponses "simples" comme les HRTF! On peut par contre corriger l'intégration sur l'axe des fréquences en adoptant l'échelle psychoacoustique des Erbs (en utilisant par exemple la fonction `ERBSpace` de [Sla]), ou plus simplement une échelle logarithmique (largeur de bande proportionnelle à la fréquence), si les composantes basse-fréquence (<1 kHz) ont été éliminées.

Il s'avère très avantageux d'étendre cette approche au calcul de la variance de l'estimation:

$$\sigma^2_{\mathrm{ITD}} = \frac{\int_0^\infty (|H_l H_r|)\left(\frac{1}{2\pi}\frac{\partial \arg(H_l/H_r)}{\partial f} - \mathrm{ITD}\right)^2 df}{\int_0^\infty (|H_l H_r|)\, df} \quad , \tag{1.64}$$

où la valeur ITD est calculée d'après (1.62).

En pratique, les réponses fréquentielles H_l et H_r sont obtenues par FFT (*Fast Fourier Transform*: Transformée de Fourier Rapide) des réponses à support discret h_l et h_r. Le calcul du temps de groupe en fonction de la fréquence (1.54) peut fournir des valeurs ponctuellement très grandes (très au-delà des valeurs typiques d'ITD) qu'il est nécessaire d'ignorer pour éviter des artefacts d'estimation indésirables (déplacement de la moyenne, variance excessive). Le faible poids énergétique $|H_l H_r|$ qui leur est naturellement associé la plupart du temps, rend normalement leur contribution négligeable. Par sécurité, un seuil peut être fixé (par exemple 1,5 ou 2 ms) pour attribuer aux valeurs qui le dépassent une pondération énergétique nulle. La figure 1.14 illustre le gain d'efficacité très appréciable que cette méthode confère aux estimations, par rapport à la précédente.

Application du modèle gaussien aux estimations classiques

Le reproche principal qui peut être fait aux méthodes basées sur le maximum de l'IACC, qu'il s'agisse de l'intercorrélation des réponses binaurales ou de leurs enveloppes d'amplitude, c'est qu'elles ne retiennent aucune information des pics (maxima) secondaires, que ceux-ci soient proches ou non – en amplitude comme en écart temporel – du pic retenu, alors qu'ils sont susceptibles de peser dans la "décision" ou la détection de l'ITD. Un moyen d'extraire des informations plus globales de l'IACC, consiste à lui appliquer une modélisation gaussienne, c'est-à-dire calculer le retard moyen pondéré par l'IACC ou son carré, ainsi que la variance

associée:

$$\mathrm{ITD} = \frac{\int \tau |\mathrm{IACC}(\tau)|^2 \, d\tau}{|\mathrm{IACC}(\tau)|^2 \, d\tau} \tag{1.65}$$

$$\sigma_{\mathrm{ITD}} = \sqrt{\frac{\int (\tau - \mathrm{ITD})^2 |\mathrm{IACC}(\tau)|^2 \, d\tau}{|\mathrm{IACC}(\tau)|^2 \, d\tau}} \tag{1.66}$$

De la même manière qu'avec la méthode basée les époques moyennes des enveloppes, cela produit une estimation beaucoup plus continue que par extraction du maximum lorsqu'on a affaire au déplacement panoramique d'une source – réelle ou virtuelle. Cela reflète d'ailleurs mieux, en général, l'impression subjective de l'auditeur. Là encore, la valeur de σ_{ITD} (1.66) est censée traduire l'acuité de détection de l'ITD, ou encore une marge d'incertitude autour de l'ITD retenu (Figure 1.14).

Comparaisons et conclusions

Nous venons de proposer un ensemble de méthodes pour l'estimation objective d'un ITD global ou haute-fréquence. La diversité des outils employés suscite un certain nombre de questions:

- Quelle est la validité psychoacoustique de ces mesures? Les différents critères d'extraction de l'ITD restent en effet de nature mathématique et arbitraire.
- Faut-il préférer les estimations les plus continues (1.59) (1.65), ou celles qui souligne le mieux la complexité des structures temporelles et leur variation (1.62) (1.64)?
- Enfin, le rapport entre les marges d'incertitude (variances) d'une méthode à l'autre, reflète-t-il l'acuité relative d'estimation de chaque méthode?

Le fait que ce sont des méthodes mathématiques qui sont présentées ici, limite leur prétention à décrire le potentiel de discrimination directionnelle offert par les réponses binaurales, plutôt qu'à traduire fidèlement les mécanismes psychoacoustiques de localisation. On peut estimer de ce point de vue que la nouvelle méthode (1.62) nommée IRGD sur la figure 1.14 remplit de façon très satisfaisante le cahier des charges établi plus haut. Sur un exemple apparemment simple, elle a montré son aptitude à souligner une singularité qui apparaît peu, voire est complètement gommée avec les autres méthodes, même si certaines d'entre elles semblent plus séduisantes parce qu'elles produisent des courbes plus lisses et régulières. Elle est aussi la seule à s'accompagner d'une variance d'estimation nulle lorsque les réponses binaurale gauche et droite sont identiques (source frontale par exemple), ce qui semble optimal comme traitement objectif de l'information directionnelle. Cependant, l'incertitude subjective sur l'acuité de latéralisation existe bel et bien lors d'une écoute réelle, même pour des sources frontales: c'est ce que Blauert désigne sous le nom de tache de localisation (*localization blur*) [Bla83]). De ce point de vue, l'intérêt porté aux autres méthodes (GMD et GMIACC sur la figure 1.14) mérite d'être préservé. Il est bon d'ajouter que dans le cas d'une source unique, la "dispersion" ou l'incertitude dans l'estimation d'un ITD haute-fréquence global n'est pas une tare pour le système perceptif: elle n'est que le "pendant temporel" des indices spectraux, qui aident bel et bien la localisation, mais sont ici ignorés. Notre incertitude de mesure est donc "plus large" que la tache de localisation.

Ces méthodes sont appliquées en 4.1 pour l'évaluation objective de la restitution ambisonique, elle-même confrontée aux résultats d'expériences d'écoute informelles.

1.5 Prédiction objective de la localisation d'après caractérisation acoustique

Nous avons décrit en 1.3 les mécanismes connus de la perception auditive spatiale et les interprétations en termes de localisation, de qualité d'image ou d'impression spatiale qu'on peut leur associer, en mettant

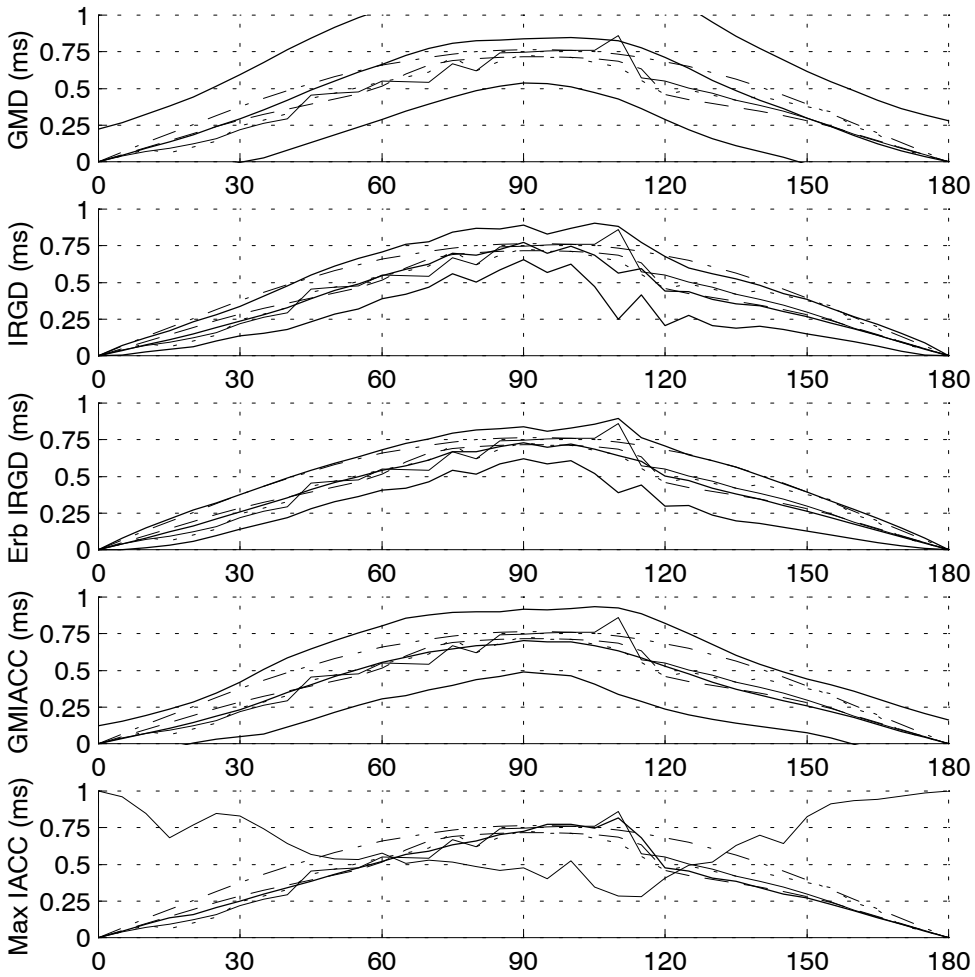

FIG. 1.14 – *Estimations de l'ITD (traits gras) et d'un intervalle de confiance (pour les quatre premières, délimité par deux traits gras) d'après différentes méthodes, appliquées aux HRTF du KEMAR. De haut en bas: estimation (ITD et écart-type) d'après différence de modèles gaussiens (GMD:* Gaussian Model Difference*); intégration du retard de groupe relatif sur l'échelle linéaire des fréquences avec pondération énergétique et calcul de la variance associée (IRGD:* Integrated Relative Group Delay*); méthode analogue, mais en utilisant l'échelle des Erbs (Erb IRGD); ITD et écart-type d'après la modélisation gaussienne de la fonction d'intercorrélation (GMIACC:* Gaussian Modeled InterAural Cross-Correlation*); Détection de l'ITD d'après le maximum de l'IACC (IACC normalisée superposée à la courbe de l'ITD). Pour référence, sont superposées à chaque estimation les estimations de l'ITD comme retard de phase pour les fréquences 500 Hz (-.), 2 kHz (- -), 8 kHz (–), 16 kHz (...).*

en avant l'hypothèse d'un apprentissage nourri d'expériences auditives "naturelles", c'est-à-dire liées à des conditions de production naturelle du champ sonore: une ou plusieurs sources distinctes et interaction avec l'environnement. Pour rappel, on considère en particulier que les mécanismes de localisation d'après le premier front d'onde se basent sur la référence d'une onde plane (avec une approximation justifiée). En gardant en vue la possibilité d'une application directe au domaine de la reproduction stéréophonique[42], nous tentons à présent d'appliquer ces mécanismes à une situation d'écoute dans un champ sonore quelconque, notre ambition étant de prédire l'effet subjectif – en terme de création d'image sonore – d'après la caractérisation objective du champ et selon les conditions de mobilité de la tête.

En complément de la caractérisation locale de la propagation par le vecteur vélocité $\vec{V}$ (introduit en 1.2.2), le souci d'une caractérisation globale (en 1.5.1) va mettre en évidence la réalité physique que revêt une autre grandeur: le vecteur énergie $\vec{E}$. A l'origine, ces vecteurs vélocité et énergie ont été définis par Gerzon dans un contexte de restitution stéréophonique[42]. On regrette cependant que leur identification sur le plan acoustique et leur valeur de prédiction quant aux indices et effets de localisation n'aient été que très partiellement explicitées dans la littérature existante – voire nullement en ce qui concerne le vecteur $\vec{E}$. C'est cette carence que nous tentons ici de pallier.

Le travail de démonstration auquel nous nous livrons procède en trois phases:

1. *Identification* des vecteurs $\vec{V}$ et $\vec{E}$ comme grandeurs caractéristiques de la propagation acoustique. Si celle du vecteur vélocité $\vec{V}$ (en 1.2.2) est assez naturelle, celle du vecteur énergie $\vec{E}$ (en 1.5.1) est plus équivoque et ne traduit des propriétés acoustiques que par l'intermédiaire de considérations statistiques. Il est en particulier nécessaire de bien expliciter les domaines – spatial, fréquentiel, temporel – d'application de la caractérisation.
2. *Etablissement de relations objectives* entre ces caractérisations à l'endroit où l'auditeur prend place et les indices de localisation (ITD et ILD), en fonction de l'orientation de la tête.
3. *Interprétation* des indices de localisation et de leur variation (selon la mobilité de la tête) en termes de localisation et de qualité de l'image sonore (précision, flou, "évanescence" ou instabilité), d'après les mécanismes de localisation associés à une onde plane. Autrement dit, il s'agit de la *prédiction d'un effet subjectif plausible*.

1.5.1 Vecteurs vélocité et énergie: conditions de définition et identifications

Définition des vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ d'après Gerzon

Gerzon place sa définition des vecteurs vélocité et énergie dans le contexte très particulier d'une restitution sur haut-parleurs ne faisant pas intervenir de décalage temporel entre les signaux émis, et les haut-parleurs étant placés de façon concentrique (Figure 1.15). Considérant la création d'*une* source virtuelle associée à un signal S, chaque haut-parleur i, placé dans la direction $\vec{u}_i$ vu du centre du dispositif, émet un signal S_i lié à S par un gain G_i, éventuellement complexe: $S_i = G_i.S$. Pour simplifier, les ondes émises sont supposées planes sur la zone d'écoute centrale, et chaque signal S_i est traduit au centre directement en terme de pression $p_i(\vec{r} = 0) = S_i$, contribution à la pression totale $p = \sum p_i$. En d'autres termes, on a affaire à une *convergence synchrone d'ondes portant des signaux très fortement corrélés*.

Adoptant le centre du dispositif comme point d'écoute ou de mesure privilégié, Gerzon définit le vecteur vélocité par la valeur unique:

$$\vec{V} = \frac{\sum G_i \vec{u}_i}{\sum G_i}, \tag{1.67}$$

42. "Stéréophonique" au sens large (cf introduction du chapitre suivant).

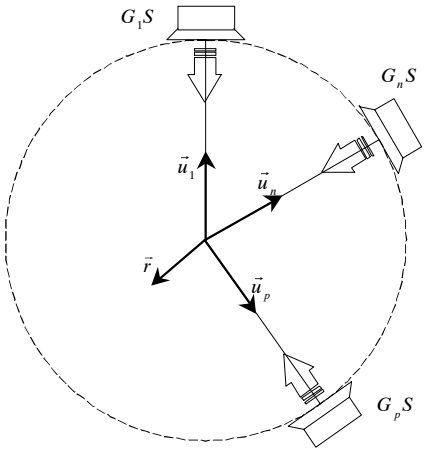

FIG. 1.15 – *Description générale d'un dispositif de restitution.*

et le vecteur énergie par:

$$\vec{E} = \frac{\sum |G_i|^2 \vec{u}_i}{\sum |G_i|^2} = r_E \, \vec{u}_E \tag{1.68}$$

D'après Gerzon, mais sans plus d'explication de sa part, ce vecteur est censé donner la direction apparente de la source sonore au regard des mécanismes de localisation haute-fréquence.

Si $\vec{V}$ s'apparente à la caractérisation de la propagation de phase telle que nous l'avons décrite en 1.2.2, le vecteur énergie $\vec{E}$ ne représente pas directement une grandeur physique, bien qu'homogène à une intensité acoustique (comme souligné dans [Nic99]). En adoptant une approche probabiliste, on comprend que le vecteur $\vec{E}$ indique la *direction moyenne* $\vec{u}_E$ *de provenance de l'énergie*, et par sa norme $r_E = |\vec{E}| \leq 1$, le *taux de concentration de l'énergie* dans cette direction [DRP98] [Nic99], le cas idéal ($r_E = 1$) étant celui d'une onde plane unique. Cherchant à dépasser cette interprétation intuitive du vecteur énergie, nous l'identifions plus loin comme une grandeur synthétique, caractéristique de la propagation acoustique globale, en explicitant quelques unes des propriétés statistiques qu'il reflète. Nous donnons ensuite des interprétations quantitatives et qualitatives des manifestations de cette grandeur[43] sur les indices de localisation, ainsi que sur l'effet subjectif de localisation attendu.

Il faut remarquer par ailleurs que les définitions (1.67) et (1.68) sont totalement indépendantes du contenu sonore S associé à la source virtuelle. Alors qu'on cherche à établir des connections entre la caractérisation du champ et les indices de localisation induits, cela incite à observer précisément la structure du champ acoustique, et ses caractéristiques invariantes pour une source virtuelle donnée mais indépendamment de son contenu sonore. Cette question est replacée au paragraphe suivant dans un contexte acoustique plus général.

Observation structurée d'un champ acoustique associé à une source sonore subjective unique

Le critère général adopté pour l'évaluation d'un système de restitution stéréophonique repose sur sa capacité à reproduire l'*effet* d'un front d'onde. L'étude présentée dans les sections suivantes sur la prédiction de l'effet et de la qualité de localisation d'après les vecteurs vélocité et énergie, repose sur ce même critère. Dans la mesure où nous y manions des grandeurs synthétiques ($\vec{V}$ et $\vec{E}$), dont il est sommairement prétendu qu'elles définissent la qualité d'image sonore, il a semblé bon, avant même de préciser leur domaine de

43. ... c'est-à-dire des implications des phénomènes acoustiques qu'il caractérise, sur la perception spatiale.

définition et/ou d'identification comme caractéristiques acoustiques, de proposer une démarche formelle applicable dans un cadre général. Il s'agit de structurer l'observation du champ acoustique en vue de dégager et de caractériser des événements acoustiques élémentaires facilement interprétables sur le plan perceptif, et finalement caractériser l'image sonore dans son ensemble.

On s'intéresse au bout du compte à la caractérisation d'une image sonore unique, ou bien de chaque image sonore séparément lorsqu'il y en a plusieurs. Il est donc important d'isoler la *propriété générique* qui caractérise *un champ acoustique propre à induire* – lorsque l'auditeur s'y plonge – *une seule image sonore*, quelle que soit sa qualité spatiale subjective, fût-elle intériorisée[44]. Précisons d'emblée que la notion d'image ou de source sonore subjective implique une dimension temporelle: la durée d'observation des phénomènes doit être suffisante pour en percevoir les invariants directionnels. Cela n'en exclue pas, cependant, une analyse sur le plan fréquentiel.

La définition objective que nous proposons pour une telle classe de champ se résume à la condition suivante: qu'il existe un signal $s(t)$ et, en tout point $\vec{r}$ de la région considérée (champ libre), une fonction de transfert $h_{\vec{r}}(t)$ telle que le champ de pression mesuré en $\vec{r}$ puisse s'écrire comme le produit de convolution $p(\vec{r},t) = h_{\vec{r}} * s(t)$. Il est clair que $h_{\vec{r}}(t)$, considérée comme fonction spatiale et temporelle, vérifie l'équation des ondes en champ libre (1.11): $(\triangle - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})h_{\vec{r}}(t) = 0$. On considère que cette fonction est causale: il s'agit d'une réponse impulsionnelle.

Il faut noter que si dans un contexte de production acoustique "naturelle", la notion de source subjective est associée à l'émission par une source acoustique unique, cette unicité ne s'impose absolument pas à la définition donnée ci-dessus. Le contre-exemple est bien-sûr donné par les systèmes de restitution sur haut-parleurs, où l'on a affaire à plusieurs sources acoustiques placées en des lieux différents émettant des signaux fortement, sinon parfaitement corrélés, typiquement issus (pour une source virtuelle) de la transformation linéaire d'un unique signal. On peut alors objecter que dans un contexte acoustique naturel, le principe de réflexion reproduit cet effet de multiplicité de sources corrélées (sources-miroirs, cf 1.2.4). La multiplicité artificielle des sources stéréophoniques se différencie fondamentalement de la multiplicité naturelle des sources-miroirs, par la simultanéité ou la quasi-simultanéité de l'arrivée des fronts d'onde sur le lieu d'écoute, qui est la marque d'un phénomène acoustique typiquement artificiel[45].

Démarche "analogique": l'onde plane comme événement acoustique élémentaire de référence. L'objet de notre étude (cette section) est d'aboutir, à partir de la description des événements acoustiques, à leur interprétation – au moins partielle – en terme de perception spatiale subjective. La démarche que nous choisissons pour cela s'appuie sur les mécanismes de perception spatiale établis au gré d'expériences "naturelles" et "ordinaires" (Cf Introduction). Appliquée à une situation quelconque, elle nécessite d'organiser l'observation des phénomènes acoustiques d'après la structure typique des événements associés à une expérience de référence: le cas d'une source physique unique. Dès lors, s'intéressant – pour les qualités de localisation – à la partie précoce de cette structure (onde directe et réflexions précoces, cf 1.3), l'effet d'une onde plane naturelle émerge comme événement élémentaire, souvent dissociable temporellement des autres événements.

Observation localisée dans le temps et dans l'espace. Pour mener à bien une telle analyse en restant dans un cadre très général, il est donc nécessaire de procéder à une observation localisée dans le temps (typiquement à un intervalle d'intégration de l'ordre de 1 ms) et dans l'espace (emplacement d'écoute). Pour

44. L'unicité d'une image sonore est délicate à définir sans équivoque: dire qu'elle est associée à une seule source subjective primaire peut être dangereux (sons de voix et de piano diffusés par le même haut-parleur); parler d'une matière sonore spatialement indissociable ne doit pas interdire l'extension subjective de l'image dans l'espace (sa largeur apparente).

45. Avec une production acoustique naturelle, cette simultanéité n'apparaît que dans des cas très particuliers: conditions de symétrie des parois réfléchissantes par rapport à l'axe source-auditeur, et non-prédominance du son direct ou faible différence de marche avec les réflexions, comme dans un couloir étroit réverbérant, par exemple.

une observation plus pertinente de la propagation de phase à travers le vecteur vélocité, la définition de ce dernier devrait alors reposer sur la transformée de Fourier du champ de pression p et de son gradient en considérant un support temporel borné (fenêtre temporelle glissante), sans quoi sa variabilité temporelle ne peut pas être traduite. Nous commenterons plus loin comment se traduit son observation statistique sur une région spatiale limitée.

Restriction à la réponse impulsionnelle spatiale $h(\vec{r},t)$: qualité potentielle de l'image. Un objectif majeur étant de décrire les qualités générales d'image sonore qui peuvent être induites par un mode de production sonore particulier (naturel ou bien artificiel: stéréophonique), c'est à la réponse impulsionnelle spatiale $h(\vec{r},t)$ que nous préconisons d'appliquer l'analyse. Mais l'observation de la réponse $h(\vec{r},t)$ ne permet pas, si elle est localisée dans le temps, de mettre en évidence les interactions entre l'onde directe et les réflexions qui peuvent avoir lieu dans le cas d'un signal $s(t)$ de type continu ou stationnaire. La caractérisation acoustique (par $\vec{V}$ et $\vec{E}$) associée à $h(\vec{r},t)$ doit donc faire en général l'objet d'une *interprétation différenciée*, suivant notamment les paramètres dynamique du signal (attaque, durée, extinction, ou stationnarité: voir 1.3.4). Dans le cas particulier d'une comparaison entre une production acoustique "naturelle" et son imitation par un système stéréophonique – tentant d'en reproduire l'effet des fronts d'onde successifs –, nous nous focaliserons sur le potentiel de restitution de chaque front d'onde indépendamment des autres, en reléguant la question subséquente de leurs interactions à l'analyse de la situation de référence (interprétation de l'effet de salle).

Analyse statistique de la propagation, identification(s) du vecteur énergie

La démarche sus-énoncée suggère, *dans un cas très général*, d'effectuer l'analyse fréquentielle de la réponse spatio-temporelle $h(\vec{r},t)$ avec pondération par une fenêtre temporelle $\kappa(t-\tau)$ à support borné et centrée sur τ:

$$\hat{h}_\tau(\vec{r},f) = \int \kappa(t-\tau)\, h(\vec{r},t)\, e^{-2\pi j f t}\, dt, \tag{1.69}$$

qui donne lieu à une description espace-temps-fréquence du vecteur vélocité:

$$\vec{V}_{\hat{h}_\tau}(\vec{r},f) = -\frac{j}{k}\frac{\vec{\nabla}\hat{h}_\tau(\vec{r},f)}{\hat{h}_\tau(\vec{r},f)} \tag{1.70}$$

Il faut bien noter que l'observation sur une durée finie induit une différence de $\vec{V}_{\hat{h}_\tau}$ avec la description $\vec{V}_{\hat{p}_\tau}$ du champ $p = h * s$, qui traduit certes une réalité acoustique plus concrète, mais qui dépend du signal s. Dans la suite, nous posons $\vec{V} = \vec{V}_{\hat{h}_\tau}$ et $h = \hat{h}_\tau$ pour simplifier l'écriture, avec en tête que dans le cas de la reproduction de l'effet d'*un* front d'onde par un système stéréophonique, la réponse h est de toutes façons "de courte durée", considérée sur une zone d'écoute privilégiée centrale.

Ceci étant précisé, nous nous plaçons maintenant à nouveau dans le cas particulier d'une convergence synchrone d'ondes planes, tel que la reproduction stéréophonique de l'effet d'un front d'onde. Dans les lignes suivantes, nous montrons quelques aspects à travers lesquels le vecteur énergie $\vec{E}$ (1.68) traduit le comportement statistique de la propagation, s'agissant, dans le cas présent, de la propagation de phase telle qu'elle est caractérisée par le vecteur vélocité $\vec{V}$.

Nous choisissons pour illustration des domaines d'estimation statistique susceptibles d'avoir rapport, de près ou de loin, avec la captation des flux sonores directionnels par la tête (et les oreilles!), ou avec une situation globale d'écoute dans le champ. Par analogie très naïve avec la forme sphérique de la tête et en considérant que les oreilles peuvent s'y déplacer à la surface par rotation de la tête, on peut s'intéresser à la propagation moyenne à la surface d'une sphère imaginaire. Nous définissons pour cela la moyenne pondérée

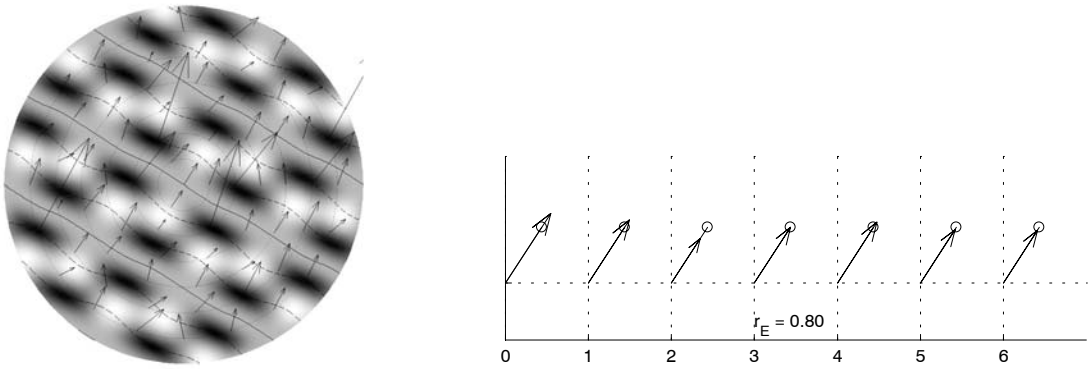

FIG. 1.16 – *Champ d'interférence monochromatique (à gauche): le champ d'énergie est représenté en niveau de gris (clair = plus forte énergie). Les courbes d'équi-phase (front d'onde) sont représentées en traits continus (phase 0 modulo* 2π*) et en tirets (phases* $2\pi/3$ *et* $4\pi/3$ *modulo* 2π*). Les vecteurs vélocité sont estimés sur des cercles concentriques, et leur partie réelle, perpendiculaire aux fronts d'onde, est représentée par des flèches (le plus petit cercle sert de cercle unitaire à leur échelle). On constate que les valeurs excessives du module* r_V *de* $\vec{V}$ *sont associées à des zones de faible énergie: elles pèsent donc peu dans la moyenne* $\vec{\mathcal{V}}_S(f,r)$*. A droite, les moyennes pondérées* $\vec{\mathcal{V}}_S(f,r)$ *associées aux cercles de*

$\vec{\mathcal{V}}_S$ du vecteur vélocité sur une sphère de rayon r et pour chaque fréquence f:

$$\vec{\mathcal{V}}_S(f,r) = \frac{\oint |h(r\vec{u}_{d\Omega},f)|^2 \vec{V}(r\vec{u}_{d\Omega},f)\, d\Omega}{\oint |h(r\vec{u}_{d\Omega},f)|^2\, d\Omega}, \tag{1.71}$$

où $\vec{u}_{d\Omega}$ est le vecteur unitaire pointant dans l'angle solide élémentaire $d\Omega$. Cette définition pourrait s'apparenter à une description du flux acoustique global traversant la sphère imaginaire. Dans le cas d'ondes concourantes de gains fréquentiellement uniformes, la quantité $\vec{\mathcal{V}}_S(f,r)$ ne dépend que de $kr = 2\pi fr/c$. Par extension, la moyenne peut être définie sur le domaine spatial $\mathcal{D}$ contenu dans cette sphère et centré en $\vec{r} = 0$:

$$\vec{\mathcal{V}}_{\mathcal{D}}(f,r) = \frac{\int_0^r \oint_{S(\rho)} |h(\rho\vec{u}_{d\Omega},f)|^2 \vec{V}(\rho\vec{u}_{d\Omega},f)\, d\Omega}{\int_0^r \oint_{S(\rho)} |h(\rho\vec{u}_{d\Omega},f)|^2\, d\Omega}, \tag{1.72}$$

Il est facile de vérifier que $\vec{\mathcal{V}}_S(f,r)$ tend vers $\vec{V}(\vec{r}=0,f)$ quand kr tend vers 0 ($f = kc/2\pi$). On pourrait aussi montrer que $\vec{\mathcal{V}}_S(kr)$ tend rapidement vers $\vec{E}$ lorsque kr augmente (au-delà d'une certaine valeur), comme l'illustre la figure 1.16. Ces deux propriétés sont également vraies pour la moyenne $\vec{\mathcal{V}}_{\mathcal{D}}(f,r)$, ce qui suggère, en étendant le rayon r aux dimensions d'une zone d'écoute élargie, que $\vec{E}$ décrit la propagation moyenne dans ce domaine.

On peut encore chercher à caractériser la propagation globale sur toute une bande de fréquence, mesurée en un point $\vec{r}$ qu'on pourra assimiler à une position d'écoute, en introduisant la moyenne $\vec{\mathcal{V}}_f$ des valeurs que prend le vecteur vélocité suivant l'axe des fréquence:

$$\vec{\mathcal{V}}_f(\vec{r}) = \frac{\int |h(\vec{r},f)|^2 \vec{V}(\vec{r},f)\, df}{\int |h(\vec{r},f)|^2\, df}, \tag{1.73}$$

Si l'on se place maintenant dans le cas – typiquement "stéréophonique" – de la convergence synchrone au point $\vec{r}=0$, d'ondes planes d'incidences $\vec{u}_i$ et de gains relatifs G_i (éventuellement complexes), comme décrit au début de cette section. En posant $G_i=|G_i|e^{j\varphi_i}$, on définit le vecteur des phases $\underline{\varphi}(\vec{r},f)=[\varphi'_1(\vec{r},f)\ \dots\ \varphi'_N(\vec{r},f)]^t$, où $\varphi'_i(\vec{r},f)=\varphi_i+2\pi f\,\vec{u}_i.\vec{r}/c$ la phase de l'onde i au point $\vec{r}$ et à la fréquence f. Le vecteur vélocité, considéré en un point $\vec{r}$ et pour une fréquence f, peut également s'exprimer en fonction du vecteur des phases $\underline{\varphi}(\vec{r},f)$:

$$\vec{V}(\vec{r},f)=\vec{V}(\underline{\varphi})\frac{\sum|G_i|e^{j\varphi'_i}\vec{u}_i}{\sum|G_i|e^{j\varphi'_i}},\qquad\text{avec}\quad\varphi'_i=\varphi_i+k\,\vec{u}_i.\vec{r}\quad\text{et}\quad k=\frac{2\pi f}{c},\tag{1.74}$$

alors que la réponse h s'écrit, dans le domaine fréquentiel:

$$h(\vec{r},f)=h(\underline{\varphi})=\sum|G_i|e^{j\varphi'_i}\tag{1.75}$$

En s'éloignant du centre dans une direction qui n'est pas une direction de symétrie[46], c'est-à-dire de préférence telle que les projections $\vec{u}_i.\vec{r}$ soient différentes deux à deux, on peut considérer en première approximation que le vecteur des phases $\underline{\varphi}$, à une position $\vec{r}$ et comme fonction de la fréquence f, se comporte comme une variable aléatoire à valeurs uniformément distribuées dans $[0\ 2\pi]^N$. Notons $E(\underline{\varphi})=|h(\underline{\varphi})|^2$ la pondération énergétique dans la définition (1.73):

$$\vec{\mathcal{V}}_f(\vec{r})=\langle\vec{V}(\underline{\varphi})\rangle_E=\frac{\displaystyle\int_{[0\,2\pi]^N}E(\underline{\varphi})\vec{V}(\underline{\varphi})\,d^N\underline{\varphi}}{\displaystyle\int_{[0\,2\pi]^N}E(\underline{\varphi})\,d^N\underline{\varphi}}\tag{1.76}$$

Par développement du numérateur et du dénominateur:

$$\begin{cases}E(\underline{\varphi}) &= \displaystyle\sum_{i=1}^{N}G_i^2+2\sum_{1\le p<q\le N}|G_p|\,|G_q|\cos(\varphi'_p-\varphi'_q)\\ E(\underline{\varphi})\vec{V}(\underline{\varphi}) &= \displaystyle\sum_{i=1}^{N}|G_i|^2\vec{u}_i+\sum_{1\le p\ne q\le N}|G_p|\,|G_q|e^{j(\varphi'_p-\varphi'_q)}(\vec{u}_p+\vec{u}_q)\end{cases}\tag{1.77}$$

et du fait que les termes dépendant des phases sont nuls en moyenne, d'après l'hypothèse d'un $\underline{\varphi}$ uniformément distribué sur $[0\ 2\pi]^N$, il apparaît clairement que:

$$\vec{\mathcal{V}}_f(\vec{r})=\langle\vec{V}\rangle_E=\frac{\displaystyle\sum_{i=1}^{N}|G_i|^2\vec{u}_i}{\displaystyle\sum_{i=1}^{N}|G_i|^2}=\vec{E}\tag{1.78}$$

Le vecteur énergie caractérise donc la propagation moyenne, considérée sur une large bande de fréquence et pour une position excentrée quelconque $\vec{r}$. Cette propriété sera directement exploitée en 1.5.4 pour la prédiction de la localisation en position d'écoute excentrée. On remarque de plus que le terme de *phasiness* (partie imaginaire de $\vec{V}$), c'est-à-dire le gradient d'énergie, est nul en moyenne.

Les développements précédents ont mis en évidence que le comportement statistique de la propagation traduit par $\vec{E}$ pouvait être observé suivant une dimension spatiale et/ou fréquentielle. On peut conjecturer

46. Dans le cas de deux ondes planes interférentes, la relation de phase entre les ondes est invariante par déplacement parallèle au plan de symétrie des deux incidences.

qu'on obtiendrait la même caractérisation synthétique en partant d'une définition temporelle de la propagation telle qu'on a pu en introduire en 1.2.2 (vélocité d'énergie $\vec{U}_E$, $\langle \vec{U}_E \rangle$ ou $\vec{U}$). Rappelons enfin qu'un autre aspect est mis en jeu implicitement dans toutes ces considérations statistiques: la pondération énergétique par le terme $|h|^2$ est liée à la réponse impulsionnelle indépendamment du signal s, alors qu'en réalité, les caractéristiques de la propagation moyenne sont susceptibles de varier suivant le contenu fréquentiel du signal au cours du temps! Les estimations présentées plus haut – dont le vecteur énergie – sont donc applicables au regard d'un comportement statistique homogène du signal s.

1.5.2 Effet de localisation basse-fréquence pour une propagation locale $\vec{V}(f)$ homogène

On se place dans le cas où la propagation locale – au centre de la tête, mais en l'absence de l'auditeur – est uniforme au moins sur une bande basse-fréquence: $\vec{V}(f) = \text{constante}, \forall f < f_h$. On suppose cette caractérisation est valide sur une zone qui engloberait la tête, jusqu'à la fréquence f_h. On verra dans les applications évoquées aux chapitres suivants, que l'extension spatiale de cette zone de validité est en général proportionnelle à la longueur d'onde. Cette question sera discutée au cas par cas, en fonction de la géométrie des ondes en interférence, puis en fonction de l'indice r_E.

Il a été mis en évidence, en 1.2.2, que le vecteur vélocité reflète la propagation de phase par sa partie réelle $\Re(\vec{V})$, et le gradient spatial du champ d'énergie par sa partie imaginaire $\Im(\vec{V})$. Nous traitons tout d'abord le cas particulier d'un vecteur vélocité réel.

Cas où $\vec{V}(f)$ est réel et uniforme sur la bande basse-fréquence

On notera donc dans le cas présent: $\vec{V}(f) = r_V \, \vec{u}_V$. D'après les développements exposés en 1.2.2, le champ acoustique a localement l'apparence d'un front d'onde d'incidence $\vec{u}_V$ et de vitesse de propagation $c_V = c/r_V$, où c est la vitesse du son. Reprenons maintenant la loi d'ITD (1.45), valable pour les basses fréquences. En substituant la vitesse apparente c_V à la vitesse c, on peut *prédire l'ITD* perçu par l'auditeur (en fonction de l'orientation $\vec{y}$ de son axe interaural) s'il se place au lieu considéré:

$$\text{ITD}^{\vec{V}}(\vec{u}_V) = \frac{D(f)}{c_V} \vec{u}_V.\vec{y} = \frac{D(f)}{c} r_V \, \vec{u}_V.\vec{y} = \frac{D(f)}{c} \vec{V}.\vec{y}, \tag{1.79}$$

soit encore, comparée à la loi de référence (cas d'une onde plane naturelle) [DRP98]:

$$\text{ITD}^{\vec{V}}(\vec{u}_V) = r_V \, \text{ITD}^{LF}(\vec{u}_V) \tag{1.80}$$

Cette équation suggère, lorsque c'est possible, la localisation dans un cône d'ambiguïté, d'angle d'ouverture γ par rapport à l'axe interaural tel que:

$$\cos\gamma = \vec{V}.\vec{y} \tag{1.81}$$

Remarque: l'affirmation de l'équation (1.79) par simple extrapolation de la loi (1.45) pourrait presque paraître abusive, l'extrapolation étant basée sur la seule vitesse de propagation apparente c_V. La loi équivalente (1.80) est pourtant très bien vérifiée par les simulations ([DRP98] (Annexe B) et autres illustrations présentes dans ce document), comme l'illustre la figure 1.17. Le caractère intuitif de (1.79) se double par ailleurs d'une justification théorique tout à fait rigoureuse, à travers l'expression du champ diffracté à la surface d'une sphère (Annexe A.3), dont le développement au premier ordre met en jeu le rapport des composantes harmoniques sphériques d'ordres 0 et 1, lequel n'est autre (en substance) que le vecteur vélocité. Lors de mesures ou de simulations des indices binauraux cependant, la dérive de l'ITD mesuré par rapport à la prédiction (1.80) est dans certains cas observée dès le début de la bande basse-fréquence, alors que dans

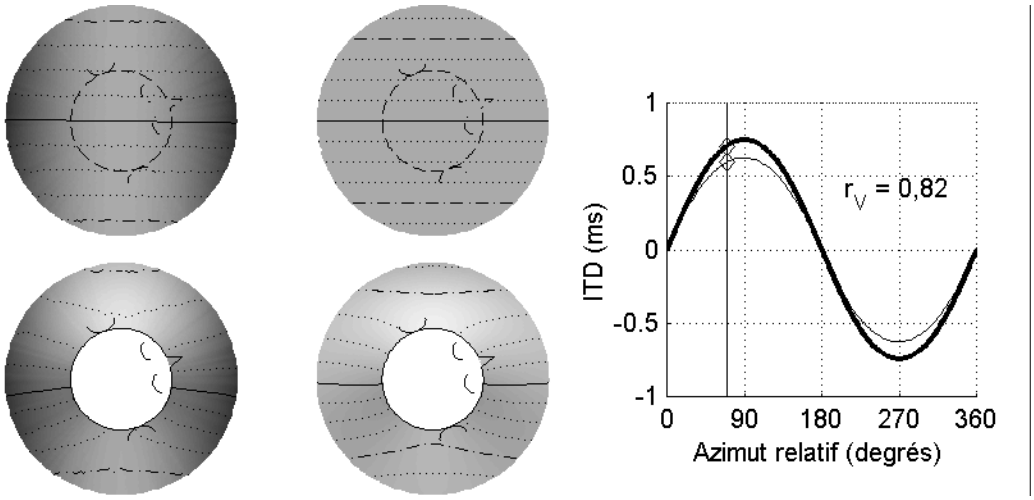

FIG. 1.17 – *Effets comparés d'un front d'onde local (à gauche) de vitesse apparente $c_V = c/r_V > c$ ($r_V = 0{,}82$), et d'une onde plane de même direction de propagation et de vitesse naturelle c (à droite). En haut: visualisation en champ libre (tête en transparence). En bas: visualisation du champ diffracté. Champ monochromatique de fréquence 600 Hz. Le rapport r_V des longueurs d'onde apparentes entre les deux situations se reporte directement sur le rapport des ITD perçus selon l'orientation de la tête: la loi de l'ITD "naturel" étant tracée en gras, la loi issue de la situation de gauche (front d'onde plus rapide) lui est proportionnelle dans un rapport $r_V < 1$. Pour les besoins de l'illustration, le front d'onde de gauche a été généré par interférence de deux ondes planes (d'angles $\pm 35^\circ$).*

d'autres cas la prédiction est excellente avec la même constance sur toute la bande. Comme nous l'illustrerons à plusieurs occasions au cours de ce document, il semble que la portée et la qualité de cette prédiction soit liées à la similarité du vecteur vélocité avec le vecteur énergie $\vec{E}$, ainsi qu'à une valeur r_V (ou plutôt de $r_E = |\vec{E}|$) proche de 1.

La loi de prédiction (1.80) ou (1.79) étant établie, nous pouvons en tenter une interprétation telle que pourrait la faire le système perceptif, par confrontation aux mécanismes de localisation basés sur une onde plane naturelle, et en fonction des rotations de la tête. *jusqu'à la prédiction* – partielle, compte-tenu de la restriction aux mécanismes basse-fréquence – *d'un effet perceptif plausible*, en termes de localisation et de qualité d'image sonore. Nous aurons à distinguer trois cas: $r_V < 1$, $r_V > 1$, et enfin $r_V = 1$ [DRP99].

Le cas $r_V < 1$ ($c_V > 1$) est celui que l'on rencontrera le plus souvent dans un contexte de restitution stéréophonique (sur deux haut-parleurs ou plus):

- **Tête fixe (latéralisation statique)**. L'ITD perçu est inférieur (pour chaque fréquence) à l'ITD maximal donné par (1.45) et définit un cône d'ambiguïté plausible. Mais la "latéralité perçue" est inférieure au cas d'une onde plane naturelle de même incidence (Figure 1.18).
- **Rotation *yaw***. Elle permet une bonne détection de l'azimut, mais y associe un *effet de hauteur artificiel ou exagéré* (Figure 1.18) du fait d'une moindre latéralisation dynamique (variations de l'ITD (1.48)): pour une incidence de site δ, le site apparent δ' est tel que $\cos\delta' = r_V \cos\delta$. Cet effet de hauteur, typique de nombreuses techniques stéréophoniques, est reporté dans la littérature ([Bau61], par exemple), en plus d'avoir été constaté durant cette thèse au cours d'écoutes informelles[47].

47. Selon d'autres observations, la hauteur peut être également fixée par la position des *tweeters*, mais c'est lié surtout aux phénomènes haute-fréquences, alors que nous parlons ici de phénomènes BF.

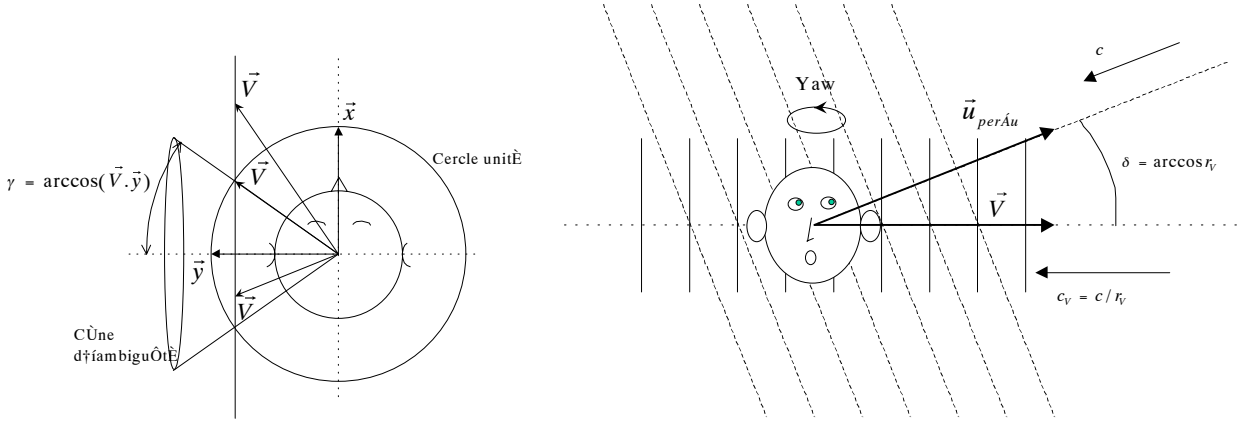

FIG. 1.18 – **A gauche***: Lorsque la tête est fixe, des vecteurs vélocité* $\vec{V}$ *de projections identiques sur l'axe interaural* $\vec{y}$ *produisent un effet de localisation traduit par le même cône d'ambiguïté de pseudo-angle (ou angle d'ouverture)* $\gamma = \arccos(\vec{V}.\vec{y})$*. La direction du vecteur vélocité n'est comprise dans ce cône que lorsque* $|\vec{V}| = r_V = 1$*.* **A droite***: Effet de hauteur artificiel par rotation* yaw *de la tête (rotation autour de l'axe vertical), et en présence d'un front d'onde local horizontal de caractéristique* $r_V < 1$*, soit de vitesse apparente* $c_V = c/r_V > c$*. Pour chaque fréquence, la longueur d'onde apparente est plus courte que une onde plane naturelle de vitesse c. Le retard de phase (ITD basse-fréquence) perçu aux oreilles et ses variations par rotation* yaw *(les oreilles restant dans le plan horizontal) correspondent à l'effet d'une onde plane naturelle d'incidence oblique par rapport au plan horizontal, et dont le front d'onde local – considéré dans le plan horizontal – serait la "coupe horizontale". De façon duale, la direction perçue est donnée par un vecteur unitaire* $\vec{u}_{perçu}$ *dont le vecteur vélocité* $\vec{V}$ *est la projection orthogonale sur le plan horizontal. Notons qu'en principe, le site* $\delta = \arccos r_V$ *est perçu avec une indétermination haut-bas.*

– **Mouvements libres**. Poursuivant la discussion entamée en 1.3.3, on peut émettre l'hypothèse que pour lever toute ambiguïté (par exemple haut-bas) et détecter la direction du front d'onde (*i.e.* la direction "de la source", lors d'expériences naturelles) avec la meilleure précision possible, le sujet "cherche" l'axe ou le plan de rotation qui lui offre la meilleure latéralisation dynamique, c'est-à-dire les plus grandes variations d'ITD par rotation autour de cet axe ou dans ce plan (qui doit contenir l'incidence de l'onde). Or la latéralisation dynamique maximale est ici inférieure – dans un rapport r_V – à celle qui correspond à une image "naturelle", c'est-à-dire induite par une onde plane naturelle. Même si cela reste du domaine de l'inconscient, on peut supposer que cette quête insatisfaite d'une variation dynamique de l'ITD "optimale", se traduise par l'*impression subjective d'une localisation peu précise, soit d'une image floue*. Nous verrons plus loin que cette interprétation est transposable aux phénomènes et mécanismes de localisation en haute-fréquence. Et c'est un fait: le caractère "flou" fait partie des attributs les plus souvent mentionnés pour l'image sonore lors d'écoutes stéréophoniques.

Plus rare, le cas $r_V > 1$ ($c_V < 1$) – sinon ses symptômes lorsque l'usage de la caractérisation par le vecteur $\vec{V}$ n'est pas appropriée – fait partie des artefacts possibles de la reproduction sur haut-parleurs (voir 2.5.2). Considérons une incidence $\vec{u}_V$ d'azimut et site absolus θ et δ. L'interprétation qui peut y être associée est la suivante:

– **Tête fixe**. Tant que $|\vec{V}.\vec{u}_{lr}| \leq 1$, c'est-à-dire tant l'angle d'incidence par rapport à l'axe interaural est supérieur à $\arccos(1/r_V)$, l'ITD définit un cône d'ambiguïté plausible pour une image naturelle (Figure 1.18). Dans le cas contraire, l'ITD maximal possible pour une source réelle est excédé, avec le risque que l'impression subjective d'une image sonore naturelle soit détruite.

– **Rotation *yaw***. Si l'incidence a une inclinaison verticale suffisante (telle que $\cos\delta < 1/\kappa_V$), on peut s'attendre à l'impression d'une image sonore ayant une moindre hauteur $\delta = \pm\arccos(r_V\cos\delta)$. Sinon, les variations de l'ITD ne sont pas naturelles (latéralisation dynamique excessive), l'image tend à se déplacer dans le sens opposé à la rotation. Si une cohérence suffisante des indices de localisation est maintenue, la consistance de l'image peut survivre, mais un effort de "ré-étalonnage" de la latéralisation dynamique est imposé à l'auditeur[48].
– **Mouvements libres**. Le caractère non-naturel voire fuyant de l'image (d'autant plus que $\kappa_V > 1$) tel qu'il vient d'être commenté, a de fortes chances d'apparaître.

Enfin le cas $r_V = 1$ permet, sur la base des mécanismes basse-fréquence, une détection de la direction de l'onde en parfaite conformité avec le cas d'une image naturelle. Il est donc recommandé que cette condition soit vérifiée au point d'écoute, ou au moins approchée, dans un contexte de restitution sur haut-parleurs.

Cas où $\vec{V}(f)$ est imaginaire et uniforme sur la bande basse-fréquence

Le fait que le vecteur vélocité comporte une partie imaginaire $\Im(\vec{V}(f)) = \vec{\Phi}(f) = jr_\Phi\,\vec{u}_\Phi$ indique, rappelons-le (section 1.2.2), un gradient d'énergie du champ. Nous avions proposé, dans [DRP99], une loi approximative de prédiction de l'ILD d'après $\vec{\Phi}$. Cette loi se base sur l'extrapolation linéaire du gradient d'énergie en une différence d'énergie entre les emplacements des oreilles, et en l'absence de la tête:

$$\Delta\text{ILD} \simeq -\frac{20}{\ln 10} 2kR\vec{\Phi}.\vec{u}_{lr}, \tag{1.82}$$

où $\vec{u}_{lr}$ est le vecteur unitaire qui oriente l'axe interaural. Cette loi grossière ne tient pas compte de la diffraction par la tête. En la confrontant à des simulations exactes du phénomène de diffraction dans un domaine basse-fréquence, il semble qu'il faille corriger la valeur prédite par un facteur $3/2$ (Figure 1.19), le même que pour la correction de l'ITD (Figure 1.5).

Il est physiquement impossible, au vu de l'équation des ondes, que le gradient d'énergie du champ soit constant suivant sa propre direction $\vec{\Phi}$. Il est donc évident que l'extrapolation linéaire atteint vite une limite de validité, au-delà de laquelle le rapport d'énergie peut même s'inverser. Rapportée à l'échelle de la tête, elle n'est valable que pour les larges longueurs d'onde, c'est-à-dire les basses fréquences.

L'ILD produit de ce domaine basse-fréquence lorsque $\vec{\Phi}$ a une direction latérale, revêt un caractère artificiel par comparaison à l'effet d'une onde plane. On pourrait à la rigueur interpréter cet ILD comme la manifestation d'une source proche, à la différence près que l'ILD haute-fréquence n'est pas garanti être de la même teneur.

L'effet associé à cette partie imaginaire $\vec{\Phi}$ du vecteur vélocité est appelé "*phasiness*" dans le contexte de la restitution stéréophonique, du fait qu'il apparaît typiquement dans le cas d'une interférence d'ondes en quadrature, ou présentant plus généralement une différence de phase au point considéré. Il est souvent considéré comme gênant et à éviter.

1.5.3 Prédiction haute-fréquence d'après le vecteur énergie $\vec{E}$

Nous avons montré (en 1.5.2) le caractère prédictif d'un vecteur vélocité uniforme au lieu d'écoute, quant à l'effet de localisation dans un domaine basse-fréquence. Le vecteur énergie est quant à lui présenté par Gerzon comme un critère de localisation haute-fréquence, mais une justification précise semble faire défaut. Nous tentons ici d'en présenter une, avec l'espoir de dériver une loi du même type que (1.80) ou (1.81), pour

48. Ce type d'effet et l'effort (de réapprentissage) consécutif peut se rencontrer de façon analogue dans le domaine visuel: avec le port de lunettes, ou le passage des lunettes aux verres de contact, ou encore lors de la manipulation d'une loupe...

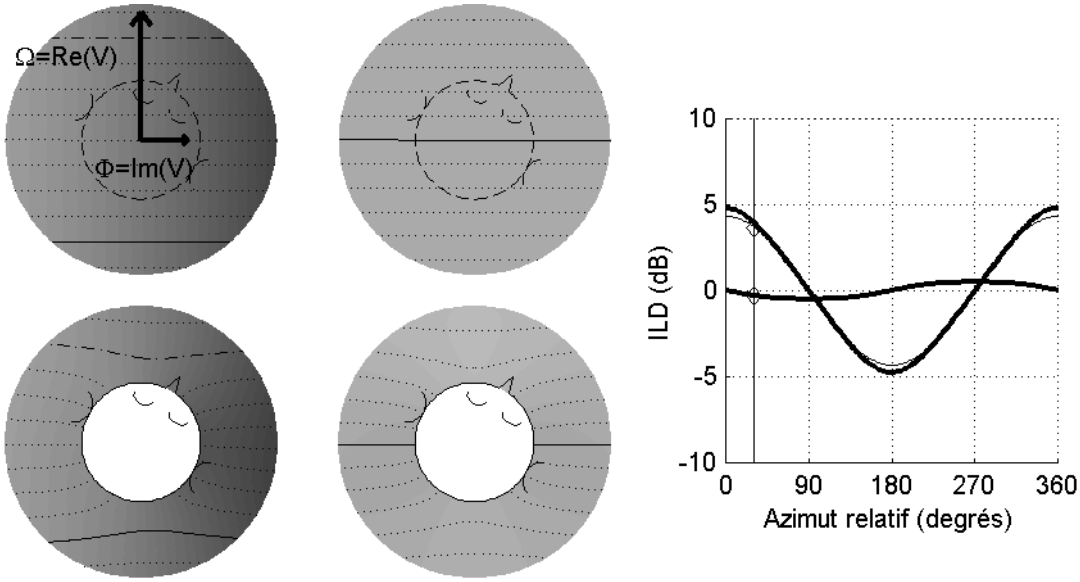

FIG. 1.19 – *Altération de l'ILD basse-fréquence en présence d'un gradient d'énergie (terme de* phasiness $\vec{\Phi} = \Re(\vec{V})$*). Courbes en gras: ILD dû à une onde plane (loi la moins ample) et ILD ajouté par la présence du gradient d'énergie. Cette dernière courbe coïncide quasiment avec la prédiction (1.82) corrigée par un facteur* $3/2$ *(courbe fine). Cas d'un champ monochromatique (300 Hz).*

la prédiction des indices de localisation haute-fréquence et de leurs effets d'après le vecteur énergie. En nous focalisant tout d'abord sur les aspects temporels (signaux impulsifs ou transitoires), nous allons nous inspirer de la "théorie énergétique" de Mertens [Mer62] [Mer65] que nous avons évoquée en 1.4.3, en l'adaptant ici avec quelques hypothèses simplificatrices. Nous commenterons ensuite le cas de signaux plus continus, en nous intéressant à la relation entre le vecteur énergie et l'ILD reconstitué.

Prédiction de l'ITD haute-fréquence: cas de convergence synchrone

Le mécanisme le plus prégnant pour la localisation à partir des informations haute-fréquences se base, rappelons-le, sur la détection du retard d'enveloppe, et s'applique à des signaux de type impulsif, transitoire, ou bien encore modulant. Dans la démonstration qui suit, l'estimation de l'ITD haute-fréquence repose plus précisément sur l'époque moyenne des enveloppes d'énergie des réponses gauche et droite (Cf 1.4), en admettant qu'elle constitue une mesure pertinente sur le plan perceptif.

Dans un premier temps et afin de faire apparaître clairement le vecteur énergie, la tête est *supposée acoustiquement transparente*. On ignore donc les phénomènes de masquage, et les temps d'arrivée d'une onde plane d'incidence $\vec{u}$ au niveau des oreilles gauche et droite par rapport au centre sont donnés par:

$$\tau_l = -\frac{R}{c}\vec{u}.\vec{y}, \qquad \tau_r = \frac{R}{c}\vec{u}.\vec{y}, \tag{1.83}$$

où $\vec{y}$ est le vecteur unitaire dirigeant l'axe interaural de la droite vers la gauche, et R est la distance de chaque oreille au centre de la tête (Figure 1.5). Nous nous plaçons dans le cas particulier d'ondes convergeant de façon synchrone au point d'écoute, tel que cela a été décrit au début de 1.5.1. Cela revient au fait de pouvoir

décomposer le champ de pression sous la forme[49]:

$$p(\vec{r},t) = \int_{\mathbb{U}_3} a(\vec{u}) s(t) e^{jk\vec{u}.\vec{r}} d\vec{u}, \tag{1.84}$$

Dans un domaine haute-fréquence sur lequel les phases des signaux élémentaires sont statistiquement incohérentes au niveau des oreilles, on peut considérer en première approximation que ces signaux s'ajoutent en énergie, c'est-à-dire que les enveloppes d'énergie $e_l(t)$ et $e_r(t)$ des signaux $l(t)$ et $r(t)$ mesurés au niveau des oreilles dérivent de l'enveloppe d'énergie $e_s(t)$ du signal $s(t)$ de la façon suivante:

$$e_l(t) = \int_{\mathbb{U}_3} a^2(\vec{u}) e_s\left(t + \frac{R}{c}\vec{u}.\vec{y}\right) d\vec{u} \qquad e_r(t) = \int_{\mathbb{U}_3} a^2(\vec{u}) e_s\left(t - \frac{R}{c}\vec{u}.\vec{y}\right) d\vec{u} \tag{1.85}$$

Pour les besoins de la démonstration, il est commode de raisonner avec une réponse impulsionnelle du type "impulsion de Dirac" $s(t) = \delta(t)$. Toujours avec l'hypothèse d'une tête acoustiquement transparente, il vient alors rapidement que:

$$\tau_l = \frac{\int t e_l(t)\, dt}{\int e_l(t)\, dt} = -\frac{R}{c}\vec{E}.\vec{y} \qquad \text{et} \qquad \tau_r = \frac{R}{c}\vec{E}.\vec{y}, \tag{1.86}$$

soit:

$$\text{ITD} = \tau_r - \tau_l = \frac{2R}{c}\vec{E}.\vec{y} \tag{1.87}$$

Il est intéressant de compléter la mesure par l'estimation de la variance ou de l'écart-type σ_{TD} (équations 1.60 et 1.61). En notant $E_y = \vec{E}.\vec{y}$, il vient:

$$\sigma_{\text{ITD}} = \sqrt{2}\frac{R}{c}\sigma_{E_y}, \qquad \text{avec:} \quad \sigma_{E_y} = \sqrt{\frac{\int_{\mathbb{U}_3} (\vec{u}.\vec{y} - E_y)^2\, a^2(\vec{u})\, d^2\vec{u}}{\int_{\mathbb{U}_3} a^2(\vec{u})\, d^2\vec{u}}} \tag{1.88}$$

La quantité σ_{E_y} peut être décrite comme un *indice de dispersion latérale* (suivant la direction $\vec{y}$) *des incidences*. On pourrait être tenté de la compléter par des composantes σ_{E_x} et σ_{E_z} pour former un "vecteur variance", afin de déduire à volonté l'indice $\sigma_{E_{y'}}$ associé à un nouvel axe $\vec{y}'$ par simple projection de ce vecteur sur cet axe. Malheureusement un tel vecteur ne fait pas sens du fait de la définition [quadratique] de ses composantes, et sa projection ne produirait pas l'indice $\sigma_{E_{y'}}$ escompté.

Cette démonstration, bien qu'elle repose sur un certain nombre d'hypothèses simplificatrices, a au moins l'avantage de dégager une loi de prédiction explicite de l'ITD haute-fréquence à partir du vecteur énergie. Avant de pousser plus loin l'interprétation de l'effet de localisation attendu, il reste à déterminer de quelle manière la loi (1.87) est modifiée par la prise en compte d'un modèle plus réaliste.

Précisons tout d'abord que l'hypothèse de sommation en énergie des signaux ou de leurs enveloppes n'est pas en général pas rigoureusement vérifiée: elle l'est dans le cas d'impulsions courtes qui ne se chevauchent pas temporellement à l'arrivée aux oreilles; dans l'autre cas, cela supposerait que tous les signaux soient en quadrature deux à deux, ce qui ne semble pas raisonnable. Dans le cas d'une tête transparente, cela peut avoir pour effet de déplacer légèrement l'estimation de l'ITD (Figure 1.20). On considère cependant que cette hypothèse reste "globalement" applicable.

49. La démonstration est évidemment transposable au cas d'un nombre fini d'ondes planes.

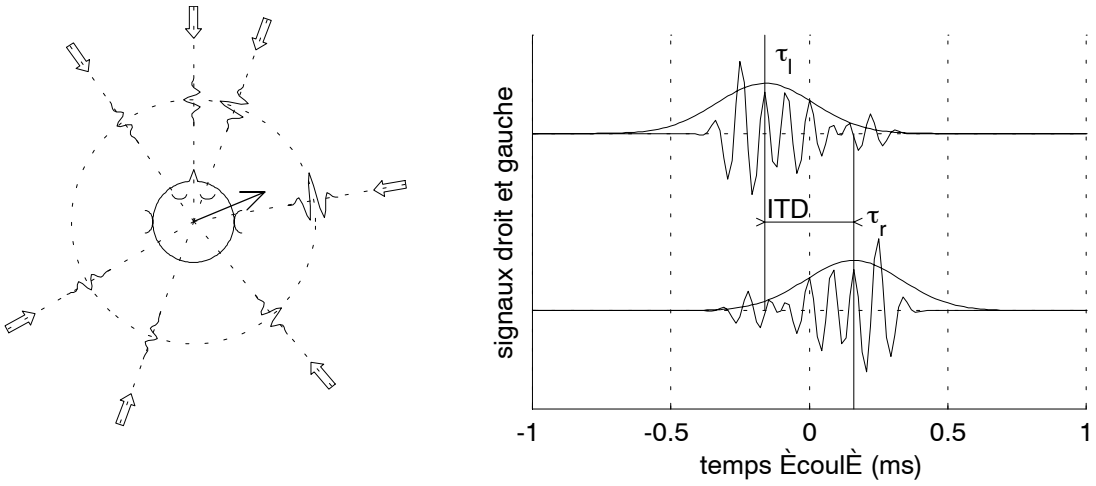

FIG. 1.20 – *Effet de localisation par détection d'enveloppe en présence de fronts d'onde impulsifs (impulsion gaussienne) convergeant de façon synchrone vers le point d'écoute. Le vecteur énergie, moyenne des différentes incidences pondérée par les énergies, est indiqué par une flèche grasse, le cercle pointillé étant utilisé comme référence pour la norme 1. Les signaux résultant au niveau des oreilles dans l'hypothèse d'une tête acoustiquement transparente, sont présentés avec la modélisation gaussienne de leur enveloppe d'amplitude (forme de cloche). L'écart des époques moyennes (centres des gaussiennes) définit l'ITD.*

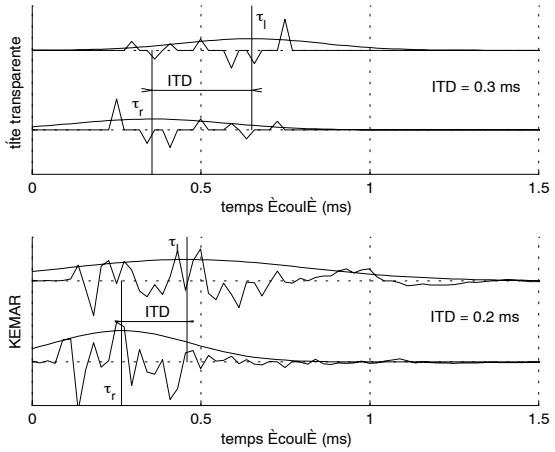

FIG. 1.21 – *Signaux gauche et droit reçu aux oreilles d'une tête acoustiquement transparente et du KEMAR, dans la situation de la figure 1.20, où l'on a remplacé les impulsions de Gauss par des impulsions de Dirac. Les modélisations gaussiennes des enveloppes sont assez étalées. L'ITD mesuré sur la tête transparente est conforme à la prédiction (1.87). Malgré le contournement de la tête, l'ITD mesuré sur KEMAR est notablement inférieur à la prédiction, car la pondération énergétique des contributions élémentaires est biaisée par l'effet de masquage par la tête.*

En considérant maintenant le modèle d'une tête "solide" non-transparente acoustiquement, *deux corrections* sont à introduire du fait de la diffraction, à savoir, grossièrement: une correction temporelle qui reflète le contournement, et une correction énergétique qui reflète l'effet de masquage, qui était totalement négligé. Plutôt que d'expliciter un nouveau modèle des enveloppes d'amplitude reconstruites, il est avantageux de raisonner sur les pseudo-angles ($\arccos(\vec{u}_i.\vec{y})$) ou les cônes d'ambiguïté associés aux différentes contributions. Plus précisément, nous allons montrer dans quel sens le cône d'ambiguïté suggéré par la valeur E_y (pseudo-angle $\arccos E_y$) doit être réévalué si la valeur de σ_{E_y} n'est pas nulle. Cette dernière condition signifie une dispersion des incidences de part et d'autres du cône en question. A cause de l'effet de masquage, les ondes d'incidences proches du plan médian (frontales, par exemple) participent plus équitablement aux deux enveloppes temporelles gauche et droite que les ondes latérales, dont les participations contralatérales sont en particulier atténuées, et laissent donc une empreinte plus forte sur les époques moyennes. L'ITD est donc d'autant plus atténué que la dispersion σ_{E_y} est grande. C'est ce qu'illustre la figure 1.21. On peut se faire un idée approximative du cône d'ambiguïté "apparent": son pseudo-angle γ est sans-doute tel que $\max(E_y - \sigma_{E_y}, 0) < \cos\gamma < E_y$ si $E_y > 0$. Pour une estimation plus précise, la seule donnée de la variance σ_{E_y} ne suffit pas. Le cas $\sigma_{E_y} = 0$ signifie quant à lui que toutes les incidences sont placées dans un même cône d'ambiguïté (pseudo-angle $\arccos E_y$), qui correspond alors normalement à l'effet subjectif de latéralisation.

Une prédiction plus précise que (1.87) peut être obtenue en incorporant dans l'estimation une pondération énergétique $\mathcal{W}$, qui traduit le masquage global – c'est-à-dire pleine-bande – par la tête en fonction de l'incidence. A partir des HRTF du KEMAR [GM94], on donne pour l'oreille gauche l'expression approchée $\mathcal{W}(\eta) = (1+\eta)^2 + \iota$, où $\eta = \cos\gamma_i = \vec{u}.\vec{y}$ et $\iota = 0{,}1$. Pour l'oreille droite, il faut changer η en $-\eta$. Reste à intégrer cette pondération dans le calcul du pseudo-angle moyen $\bar{\gamma}$, ou plutôt du "pseudo-cosinus" moyen $\bar{\eta} = \cos\bar{\gamma}$. Dans le cas d'une distribution discrète de fronts d'onde (directions $\vec{u}_i$ et amplitudes G_i), on obtient la formule de prédiction empirique:

$$\bar{\eta} = \frac{1}{2}\left(\frac{\sum G_i^2 \vec{u}_i.\vec{y}\,\mathcal{W}(\vec{u}_i.\vec{y})}{\sum G_i^2 \mathcal{W}(\vec{u}_i.\vec{y})} + \frac{\sum G_i^2 \vec{u}_i.\vec{y}\,\mathcal{W}(-\vec{u}_i.\vec{y})}{\sum G_i^2 \mathcal{W}(-\vec{u}_i.\vec{y})}\right) \qquad \Rightarrow \bar{\gamma} = \arccos\bar{\eta}, \tag{1.89}$$

où $\bar{\gamma}$ caractérise le cône d'ambiguïté représentatif de l'effet de latéralisation. On pourra vérifier, en 4.1, que cette formule fournit des courbes d'allure très semblable aux estimations de l'ITD par les méthodes décrites en 1.4.3. Une mesure de l'incertitude associée à cette estimation pourrait être calculée suivant le même modèle de pondération énergétique, et donner une indication sur la tache de localisation associée à l'image sonore.

Interprétation selon les mouvements de la tête: on peut directement transposer l'interprétation des mécanismes basse-fréquence en relation avec le vecteur vélocité (1.5.2, cas $\kappa \leq 1$), au domaine haute-fréquence et au vecteur énergie. Cependant, étant donné que la latéralisation prédite par $\vec{E}.\vec{y}$ (angle $\arccos(\vec{E}.\vec{y})$ par rapport à l'axe interaural) est souvent surestimée, il faut revoir à la baisse les effets de latéralisation statique et dynamique à attendre. Autrement dit, pour un $r_E < 1$ et en ne considérant que les mécanismes haute-fréquence, il faut s'attendre à une image encore plus floue (ou avec un effet de hauteur plus exagéré) que ce qui pourrait ressortir des mécanismes basse-fréquence avec un κ de même valeur.

Cas d'une convergence non-synchrone des fronts d'ondes

Lorsqu'il y a des différences de temps non-négligeables dans l'arrivée des fronts d'onde au point d'écoute, ou encore lorsque l'auditeur se trouve hors du point de convergence synchrone, le raisonnement qui vient d'être présenté n'est plus valide. La réception des impulsions (ou des attaques) par les oreilles risque en effet d'être trop étalée dans le temps pour que l'effet de sommation puisse être appliqué. Si le premier front d'onde

est d'énergie suffisante par rapport aux suivants, c'est *a priori* la loi du premier front d'onde (ou *effet d'antériorité*, 1.3.4) qui fait autorité; c'est lui qui détermine la direction apparente, et les fronts d'onde suivants peuvent être interprétés comme des réflexions, voire comme des échos si l'écart temporel est suffisamment grand. S'il est d'énergie plus faible, un phénomène d'*inhibition* peut avoir lieu au profit d'un front d'onde suivant (*primary sound inhibition* [Bla83]), selon les rapports d'énergie et les écarts temporels. Aucune loi explicite ne semble avoir été dégagée qui permette une prédiction mathématique générale de la direction perçue dans ces conditions, mais on peut se référer avec intérêt aux quelques résultats d'expériences présentés à ce sujet dans [Bla83] (p.225 et suivantes). Par exemple, le rapport d'énergie entre deux fronts successifs (signal de parole), nécessaire pour inhiber l'effet du premier, passe de 15 dB (rapport d'*amplitude* 5,6) pour un écart de 10 ms (différence de marche de 3,4 m) à 30 dB (rapport 31,6) pour 100 ms (34 m). Bien qu'il faille sans-doute relativiser ces données en fonction du nombre de fronts d'onde (souvent plus de deux), de leurs directions, et de la nature du signal (netteté de l'amorce des fronts d'onde, décroissance...), il sera bon d'avoir ces ordres de grandeur en tête lorsqu'on s'intéressera à des conditions d'écoute excentrée par rapport à un dispositif de restitution[50].

Etant donné que la fonction de distribution des incidences élémentaires n'est pas contenue de façon explicite dans la grandeur synthétique $\vec{E}$, il est d'autant plus difficile de s'essayer à une prédiction explicite de la direction perçue à partir de ce seul vecteur énergie, même connaissant la position de l'auditeur par rapport au point de convergence synchrone et à la direction $\vec{u}_E$. On peut quand même se livrer à quelques commentaires: il est évident que la direction perçue est d'autant plus stable (par éloignement du point de convergence) que l'indice r_E est proche de 1, puisque la situation s'apparente alors au cas d'une source unique; dans le cas d'une valeur r_E inférieure à 1, on peut conjecturer que le cas le plus favorable est tel que la fonction de distribution de l'énergie $|a(\vec{u})|^2$ est décroissante à mesure que l'incidence $\vec{u}$ s'éloigne de la direction $\vec{u}_E$, et que cette fonction est dense, au moins dans la direction $\vec{u}_E$. Cette propriété se manifeste partiellement à travers l'indice de dispersion σ_{E_y}, où l'on aurait choisi $\vec{y} = \vec{u}_E$. Ces aspects sont observés avec attention lors du choix d'une stratégie de décodage ambisonique dédiée à des auditoires étendus (voir en 3.1.3 et 4.2.2).

Vecteur énergie et réduction du masquage interaural (ILD)

Nous avons commenté plus haut les effets du masquage par la tête sur la latéralisation par détection des enveloppes temporelles d'énergie: le poids relatif d'une onde latérale, même prédominante, est diminué par la présence d'incidences proches du plan médian, et encore plus par les ondes venant du côté opposé. En s'affranchissant des aspects temporels, et en s'intéressant en particulier aux cas de flux sonores continus, il apparaît clairement que la latéralisation due à l'ILD souffre de la même manière de la dispersion latérale des incidences. En effet, en admettant que la sommation des contributions au niveau des oreilles se fait statistiquement hors-phase, c'est-à-dire en énergie, les effets de masquage propres aux différentes incidences s'atténuent mutuellement, de façon analogue aux effets d'ombre autour d'un objet éclairé suivant différentes incidences, dans le domaine visuel (Figure 1.22).

Nous donnons maintenant quelques commentaires pour justifier la valeur prédictive du vecteur énergie quant à l'effet de latéralisation et au regard de ces aspects. En première approximation: le *taux de concentration d'énergie* r_E indique donc à quel point l'effet de masquage – ou l'ILD – est atténué ou bien préservé, par rapport au masquage (ILD) dû à une onde unique d'incidence $\vec{u}_E$. Pour être plus précis et plus juste: la dispersion latérale σ_{E_y} indique la réduction de latéralisation par rapport au cône d'ambiguïté de pseudo-angle $\arccos(\vec{E}.\vec{y})$, $\vec{y}$ orientant l'axe interaural. *Ce comportement s'apparente complètement à celui décrit plus*

50. En plus de cela, il pourra être nécessaire de remettre en cause l'hypothèse d'ondes planes sur la zone d'écoute, et corriger leurs énergies relatives et directions en fonction du rapprochement ou de l'éloignement des sources 3.1.3, 4.2.2

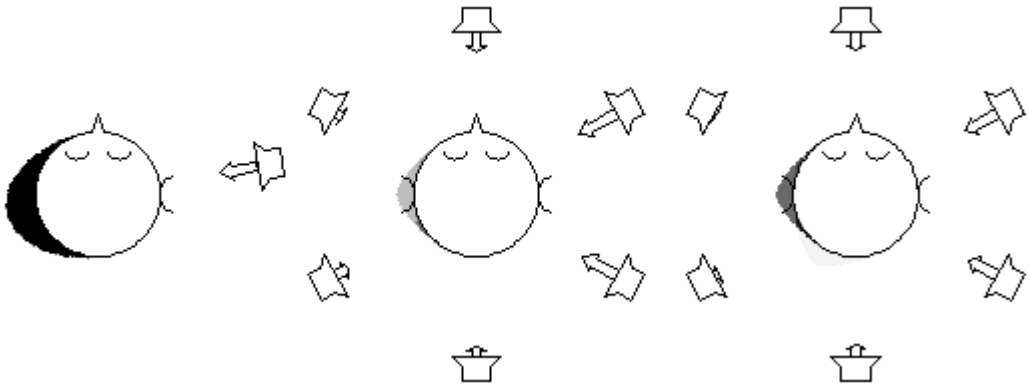

FIG. 1.22 – *Illustration schématique de l'effet de masquage par la tête pour les fréquences élevées, en présence d'une ou de multiples ondes transportant le même signal (à un gain près représenté par la longueur des flèches). L'effet de masquage est symbolisé par des zones ombrées, par analogie aux phénomènes lumineux. Son atténuation (au centre et à droite), donc celle de l'ILD, est liée à la dispersion angulaire de l'énergie des contributions, que traduit le module r_E du vecteur énergie $\vec{E}$. Le masquage est renforcé (à droite par rapport au centre) en "concentrant" l'énergie dans la même direction meilleur r_E. [MODIFIER LA FIGURE!!!] Au centre et à droite: restitution ambisonique d'ordre 1, décodages basique et "max r_E (Cf 2.4).*

haut, basé sur la détection de l'ITD haute-fréquence. Une légère nuance mérite d'être rappelée cependant: au contraire de l'ITD, l'ILD associé à une incidence unique n'est pas une fonction tout à fait monotone de l'incidence quand celle-ci se déplace d'un côté vers l'autre, ce qui fait de lui une information moins robuste et plus qualitative.

Ces commentaires doivent pouvoir s'appliquer à des positions d'écoute à l'écart du point de convergence synchrone si l'on a affaire à des flux sonores plutôt stationnaires.

1.5.4 Prédiction basse-fréquence dans le cas d'une propagation de phase non-cohérente

L'effet de localisation basses fréquences repose, d'après la théorie de Makita [Mak62] et comme nous l'avons développé plus haut, sur le processus de détection de front d'onde en terme de direction et vitesse apparentes de propagation. Nous avons montré en détail, en 1.5.2 et dans [DRP99], que cette détection, qui utilise l'information de déphasage interaural et de ses variations par rotation de la tête, est directement prédite par le vecteur vélocité $\vec{V}$ (ou plus précisément sa partie réelle), mesuré à l'endroit de l'auditeur et défini pour chaque fréquence. Alors que cette démonstration s'appuyait sur une caractérisation uniforme de la propagation locale sur toute la bande basse-fréquence considérée, donc sur une valeur unique du vecteur vélocité, on aimerait maintenant exhiber un prédicteur synthétique qui puisse s'appliquer au cas d'une distribution $\vec{V}(f)$ non-uniforme, typiquement celle observée en une position excentrée $\vec{r}$ lors d'une restitution sur haut-parleurs. Pour cela, nous mettons à profit l'interprétation de la distribution $\vec{V}(f)$ au point $\vec{r}$ comme la fonction d'une variable aléatoire – le vecteur de phase $\underline{\varphi}$ –, ainsi que nous l'avons introduite en 1.5.1.

Prédiction en écoute excentrée par le vecteur énergie $\vec{E}$

La moyenne pondérée du vecteur vélocité $\mathcal{V}_f = \langle \vec{V} \rangle_E$ définie par (1.73) et (1.76) se présente comme un bon candidat à la prédiction: à chaque "événement" $\vec{V}$ est associé un ITD (retard de phase, équation 1.79),

et sa pondération énergétique semble justifiée sur le plan perceptif. En considérant que la variable aléatoire $\underline{\varphi}$ a bien la propriété d'être à valeurs uniformément distribuées, l'équation (1.78) met à jour l'identité du prédicteur recherché: le *vecteur énergie* $\vec{E}$. Ce dernier peut donc d'après cette approche, être interprété comme *critère de localisation basse-fréquence en position d'écoute excentrée*, c'est-à-dire hors de la zone de convergence des ondes planes, lorsque la cohérence des phases n'est plus assurée. Sous réserve de validité du modèle proposé, il est censé refléter l'effet de localisation basse fréquence caractéristique d'un vecteur vélocité $\vec{V}$ qui prendrait sa valeur sur toute la bande des fréquences, à quelques nuances près toutefois: il est en effet raisonnable de penser que la qualité de l'image soit dégradée du fait de la non-cohérence des indices de localisation basse-fréquence.

Prédiction d'après le maximum de probabilité

Si ce premier modèle de prédiction a le mérite de faire apparaître explicitement le vecteur énergie, d'autres hypothèses peuvent être proposées. En conjecturant que l'effet de localisation est associé à l'événement (front d'onde) le plus probable, l'analyse statistique présentée dans [DRP98] (voir annexe) se base sur l'estimation du maximum de probabilité de $\Re(\vec{V})$ d'après tirage aléatoire du vecteur $\underline{\varphi}$. Les restitutions caractérisées par les indices r_E les plus proches de 1 y sont mis en avantage: meilleure valeur du $\Re(\vec{V})$ le plus probable (bonne direction et norme plus proche de 1), et moindre dispersion autour de cette valeur. Toutefois, cette approche n'a pas permis de mettre en évidence une relation explicite entre cette prédiction et le vecteur énergie $\vec{E}$. Par ailleurs, il n'y est pas tenu compte du poids énergétique $E(\underline{\varphi})$ de l'événement associé au vecteur de phases $\underline{\varphi}$.

Post-citation

En conclusion de cette étude, une intéressante citation de Mertens semble tomber à point, qui suggère ([Mer65], p.155) un *rapprochement entre la théorie* du front d'onde *de Makita et sa propre théorie* (théorie des époques moyennes de groupe), "*qui ferait intervenir la distribution de la vélocité de groupe du front d'onde dans une région de l'espace au moins aussi large que la tête*".

Chapitre 2

Principes de restitution spatialisée et représentations associées

2.1 Introduction

Les techniques de restitution sonore spatialisée qui prennent place dans le contexte très multi-forme des applications multimédias forment un large éventail: elles vont des procédés stéréophoniques traditionnels issus de prises de son ou de simples *pan-pots*, aux techniques plus élaborées de synthèse binaurale et de restitution transaurale, en passant par les procédés *surround*, *Ambisonics*, etc... Dédiées à une diffusion sur des dispositifs variés (haut-parleurs ou casque), elles ont pour objet la création d'illusion d'images sonores localisées (positionnement 2D ou 3D) ou plus globalement la reproduction d'une scène ou d'un espace sonore, qu'il soit issu d'une prise de son naturelle ou bien composé artificiellement.

A cette notion de *reproduction* est jointe celle de *représentation intermédiaire du champ sonore* restitué, autour de laquelle gravitent plusieurs enjeux d'importance. Cette représentation contient l'information spatiale du champ sonore sous la forme d'un certain nombre de canaux audio (ou signaux) qui définissent le format de transmission (ou de stockage), et est associée à des procédés spécifiques d'*encodage* (prise de son ou synthèse du champ) et de *décodage* (produisant les signaux à diffuser sur haut-parleurs ou au casque)[1]. La première préoccupation est que la restitution qui en découle offre des qualités d'illusion satisfaisantes, autant en terme d'image sonore individuelle que de propriétés spatiales plus globales. La concision de la représentation (nombre de canaux) est évidemment appréciée pour la transmission (débit ou volume de stockage limité). Dans le cadre d'une composition et création d'une scène sonore complexe au niveau même de l'utilisateur, l'usage d'une représentation intermédiaire compacte (avec un coût de décodage fixe) peut permettre de réduire le coût de traitement global. Enfin, on porte un intérêt particulier aux modes de représentation qui offrent les caractéristiques suivantes: la portabilité (vers différents dispositifs de restitution), l'universalité (par rapport à l'auditeur), la prédictibilité et le contrôle des propriétés spatiales (pour la composition et la manipulation du champ sonore).

Objectifs

Les techniques étudiées dans ce chapitre sont donc abordées à la fois en tant que procédés de création d'illusion sonore, et pour le mode de représentation du champ acoustique qui leur est associé le cas échéant.

1. Les formats les plus courants (stéréo et multi-canal) ne nécessitent pas de décodage: les signaux diffusés sont les signaux transmis.

Dans une optique d'application immédiate à la composition de scène virtuelle et de la navigation 3D, ce chapitre répond aux trois préoccupations suivantes:

- Recenser et commenter les principales approches de restitution envisageables, et en décrire les principes, d'autant qu'un certain nombre d'entre elles sont mises en oeuvre (Chapitre 5) à des fins de comparaison et d'expérimentation. On s'intéresse également à l'usage combiné de ces techniques pour une adaptation à un dispositif de restitution restreint (sur paire de haut-parleurs ou au casque).
- Prendre en compte les différents formats de représentation de champ acoustique (dont les plus courants: stéréo et multi-canal) susceptibles de participer à la composition d'une scène virtuelle.
- Trouver un mode de représentation intermédiaire satisfaisant (*Ambisonics*) qui permette à la fois: le mélange et la manipulation de sources de nature différentes (sources monophoniques et champs complexes pré-composés) tout en contrôlant et préservant leurs propriétés spatiales, une restitution adaptable à des dispositifs de géométrie diverses (en fonction des ressources de l'utilisateur), un coût de traitement raisonnable, un coût de transmission réduit.

Plus fondamentalement, ce chapitre a pour objectif d'expliciter le lien entre mode de représentation et propriétés de restitution (notamment sur haut-parleurs), qui couvrent à la fois la qualité de restitution, dont nous détaillons les aspects plus loin, les contraintes d'écoute et la robustesse aux mouvements de l'auditeur. Des outils objectifs de caractérisation sont pour cela introduits et interprétés. Les potentiels et limites des diverses approches apparaissent ainsi à travers les compromis qu'elles réalisent entre ces différentes propriétés. Cette étude se conclue sur une synthèse générale où l'approche ambisonique se voit mise en valeur de par ses atouts pratiques et ses propriétés de restitution associées.

Démarche adoptée

Chacune des techniques abordées dans ce chapitre est d'abord introduite d'après l'approche ou les théories sur lesquelles elle se fonde à l'origine. En dépit de la grande diversité des approches, nous tenterons de leur appliquer une démarche d'analyse commune, afin d'apporter des éléments d'interprétation complémentaires sur la qualité et la robustesse de la restitution, particulièrement lorsqu'il s'agit d'une reproduction sur haut-parleurs. Poursuivant une intention présente dans [DRP99], la restitution de chaque image sonore individuelle est jugée au regard des propriétés suivantes:

- La qualité de l'image localisée (en position statique) d'après la consistance des indices de localisation;
- Le comportement (naturel ou pas) des indices de localisation et la robustesse de l'image sonore par rotation de la tête;
- La robustesse par translation de la tête, ou inversement, la dégradation de la qualité spatiale et du contenu sonore (coloration) en position d'écoute non-idéale.

La connaissance des qualités de restitution de ces événements élémentaires en fonction de leur direction, débouche sur une appréciation des qualités spatiales de la reproduction de l'espace sonore complexe dans son ensemble, dont la préservation des impressions spatiales et le potentiel d'enveloppement.

Ces analyses s'appuient, dans la mesure du possible, sur une observation structurée des phénomènes acoustiques synthétisés à l'échelle de la tête, et mettent à profit les outils de caractérisation introduits au chapitre précédent: les vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$. Ces grandeurs servent à caractériser les artefacts de l'illusion sonore, la dégradation des qualités spatiales du champ complexe, mais aussi la robustesse de l'illusion.

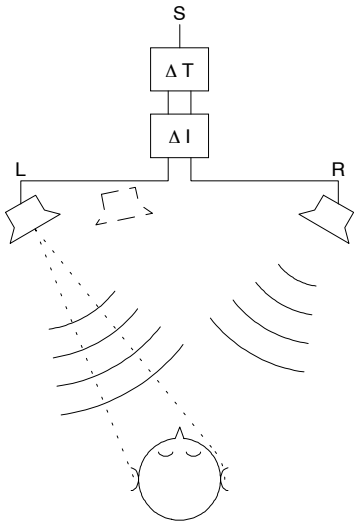

FIG. 2.1 – *Principe de création d'une image sonore entre deux haut-parleurs à partir d'un seul signal sonore S, avec contrôle de sa position par introduction d'une différence d'intensité* ΔI *et/ou d'un retard* ΔT *entre les canaux stéréophoniques L et R. L'angle habituellement recommandé entre les haut-parleurs est de* 60°*, ceux-ci formant un triangle équilatéral avec l'auditeur.*

2.2 Stéréophonie traditionnelle sur deux haut-parleurs

2.2.1 Définition et techniques microphoniques associées

Ce que nous appelons ici stéréophonie "traditionnelle" ou "conventionnelle"[2] désigne les techniques de restitution sur deux haut-parleurs, sans-doute encore les plus répandues, à l'origine liées aux procédés de prise de son utilisant un couple de microphones. De façon analogue aux expériences d'écoute au casque où un effet de latéralisation découle directement des différences d'intensité "ΔI" et/ou de temps "ΔT", le *principe de stéréophonie sur deux haut-parleurs* repose ici sur le constat suivant: il est possible de créer l'*impression subjective d'une image sonore* située quelque part entre les deux haut-parleurs[3] en alimentant ceux-ci à partir d'un même signal, la position de l'image virtuelle pouvant être contrôlée en jouant sur le ΔI et/ou sur le ΔT introduits entre les haut-parleurs. Dans la figure 2.1, l'angle de 60° recommandé entre les haut-parleurs est un compromis entre la création d'images frontales suffisamment stables et consistantes et une largeur de scène sonore satisfaisante, compromis qui sera expliqué dans les sections suivantes.

Les différences d'intensité (ou plus généralement d'amplitude) et/ou de temps entre les deux *canaux stéréophoniques* peuvent être **produites naturellement** lors de l'enregistrement d'une source sonore à l'aide d'un couple de microphones, en fonction de sa position relative et selon le couple utilisé. La figure 2.2 regroupe les principaux couples utilisés: trois couples de microphones coïncidents (M-S, *Stereosonic*, XY) ne produisant qu'une différence d'intensité entre les canaux pour chaque source enregistrée ($\Delta T = 0$), et deux couples à microphones non-coïncidents ou espacés (ORTF et AB), produisant un ΔT pour les sources latérales, ainsi qu'un ΔI pour le couple ORTF uniquement.

Les couples *Stereosonic* et M-S[4] sont associés à la stéréophonie sur deux haut-parleurs telle que Blumlein

2. Tout en réservant au terme "stéréophonie", considéré seul, la définition donnée en 2.1, c'est à la classe présentée ici que nous associerons les notions de "prise de son stéréophonique" et "canaux stéréophoniques".

3. Il s'agit donc d'une *image virtuelle*, puisqu'"elle n'a pas de support tangible" (*dixit* Rozenn Nicol [Nic99]).

4. *Mitte-Seite* en Allemand, ou bien *Mid-Side* en Anglais, et non "mono-stéréo" comme on pourrait le croire!

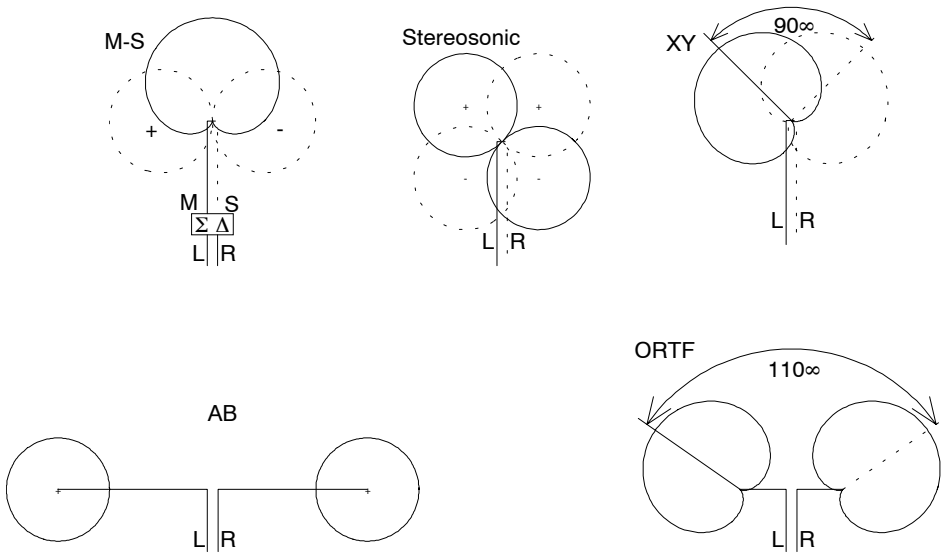

FIG. 2.2 – *Couples stéréophoniques les plus usités: M-S (*Mid-Side*) composé d'un microphone cardioïde (M) et d'un bidirectif (S);* Stereosonic*, deux capsules bidirectives montées à* $\pm 45^\circ$*; XY, deux capsules cardioïdes formant un angle* $80^\circ < \phi < 130^\circ$*; AB, deux capsules omnidirectives relativement espacées (de l'ordre de 1 m); ORTF (du nom de l'institution où cette technique a été développée), deux capsules cardioïdes généralement espacées de 17 cm et montées à* 110°.

l'a conçue à partir des années 30 ([Blu33] par exemple). Remarquons au passage que le couple *Stereosonic* perçoit autant les sons arrière que les sons frontaux, ce qui a comme regrettable conséquence de "convertir" l'effet de salle (réflexions arrière) en une désagréable coloration, outre un "repliement frontal" des contributions arrière. Pour cette raison et en dépit des très bonnes qualités d'image sonore qu'on lui reconnaît[5], les conditions d'utilisation de cette technique sont relativement sévères et restrictives. Ce problème n'est pas présent avec les techniques M-S et XY. La technique M-S, où les signaux gauche et droit (L et R) sont produits par somme et différence d'une captation frontale et symétrique (M), et d'une captation latérale et anti-symétrique (S) des sons, laisse de surcroît la *possibilité de régler la largeur apparente* de la scène sonore. Il suffit pour cela d'un simple contrôle du gain de la composante S, alors qu'avec le couple XY, il serait nécessaire de modifier l'angle entre les capsules pour modifier la largeur de la scène restituée.

Les couples non-coïncidents AB et ORTF semblent être quant à eux l'apanage d'une "école française" de la prise de son. Comme il est commenté plus loin, l'introduction d'une différence de temps entre les canaux stéréophoniques ne se soumet pas aussi bien aux modélisations théoriques de la création d'image sonore, que les approches purement "ΔI". Le couple ORTF, avec son espacement de 17 cm, suggère certes des similarités avec la façon dont les oreilles perçoivent les signaux, mais dans le cas présent, les signaux stéréophoniques ne sont pas conduits individuellement et séparément jusqu'aux oreilles. Les techniques de ce type, qui relèvent donc plutôt d'une approche empirique, sont en général appréciées pour les qualités d'"ambiance" et de "profondeur" qui sont associées à leur restitution.

Enfin, les signaux stéréophoniques peuvent être **produits artificiellement** à partir de signaux monopho-

5. C'est par exemple le seul des procédés présentés qui préserve l'énergie des sources sonores indépendamment de leur direction: $\cos^2(\theta - \pi/4) + \cos^2(\theta + \pi/4) = 1$!

niques, selon un procédé appelé potentiomètre panoramique (*pan-pot*), schématisé Figure 2.1. Ce procédé est désormais indissociable du travail de mixage pour les productions de studio, le pan-pot d'intensité (ΔI) étant de loin le plus représenté, d'autant qu'il était plus facile à réaliser avec des techniques analogiques. Les signaux qui font l'objet de ce pan-pot sont généralement issus d'une prise de son de proximité ("microphone d'appoint"), afin d'éviter sa coloration par l'effet de salle, particulièrement perceptible en prise de son monophonique. Il est bon de signaler une distinction fondamentale avec une prise de son utilisant un couple, qui montre les limites d'une production stéréophonique par simple pan-pot. Le couple de microphones capte naturellement, en plus du son direct de la source, l'effet de salle et les réflexions en respectant – au moins partiellement – leur distribution spatiale, ce qui se traduit par une décorrélation naturelle des signaux stéréophoniques, elle-même responsable de la largeur apparente de la source lors de la restitution. Partant d'un signal monophonique, un simple pan-pot n'est pas capable d'une telle décorrélation, et donne lieu à une image "nue", sans largeur et *sans perspective stéréophonique*! Pour pallier ce défaut, quelques procédés ont émergé, dans les premiers temps de la production par pan-pot, qui consistent à produire deux signaux décorrélés à partir d'un seul signal monophonique. Ces techniques dites de "*pseudo-stéréophonie*" [Bla83] n'apportent qu'une maigre satisfaction, à côté des procédés de synthèse numérique d'effet de salle qui ont depuis pris leur place dans la production de studio.

Mécanismes et théories sous-jacentes

Malgré des conséquences similaires sur le déplacement latéral des images sonores, il faut bien souligner que les *mécanismes* impliqués dans la reconstitution de l'image sonore sont *très différents*, selon que les signaux stéréophoniques sont présentés *au casque ou sur haut-parleurs*. Alors que les ΔT et ΔI se reportent directement sur les indices de localisation (ITD et ILD) lors d'une présentation au casque, chaque oreille perçoit, lors d'une restitution sur haut-parleurs, un mélange des deux signaux, l'un (contribution contralatérale) subissant un retard et une atténuation par rapport à l'autre (contribution ipsilatérale). Ce mélange – ou *diaphonie* – des signaux stéréophoniques à cause de ces chemins croisés est désigné sous le terme anglais de *cross-talk*. Par contraste avec la présentation au casque, donc, on montre par exemple qu'à cause du *cross-talk*, *une différence* d'intensité (ou plutôt *d'amplitude*) *se reporte au niveau des oreilles comme un retard* dans un domaine basse-fréquence (paragraphe suivant). Au contraire des expériences au casque, la restitution sur haut-parleurs offre par ailleurs une extériorisation naturelle des images sonores, distribuées sur une scène frontale. En revanche, celle-ci se trouve confinée entre les haut-parleurs, à cause du *cross-talk*.

2.2.2 Théories basse-fréquence et interprétation acoustique

La modélisation de l'impact des paramètres ΔI et ΔT sur l'effet de localisation, à travers l'ITD particulièrement, est relativement accessible au moins dans un domaine basse fréquence où les longueurs d'onde sont assez grandes par rapport à la taille de la tête. Plusieurs théoriciens s'y sont essayés [Bau61] [Ber75] [Mak62] (et Mertens, PLUS LOIN?) qui, à force de développements mathématiques parfois ardus, ou usant de formalismes plus légers, ont finalement produit des lois de pan-pot au moins compatibles quand elles n'étaient pas identiques.

Sans reporter les détails de ces calculs, nous en présentons ici les résultats les plus connus et utilisés. En notant G_L et G_R les gains associés au haut-parleurs gauche et droit, et ϕ_F l'angle séparant chaque haut-parleur de l'axe médian (Figure 2.1), l'angle θ_S décrivant la position apparente de la source est donné suivant l'une

des deux lois suivantes, selon la liberté de mouvement accordée la tête:

$$\begin{cases} \dfrac{\sin\theta_S}{\sin\phi_F} = \dfrac{G_L - G_R}{G_L + G_R} \\ \dfrac{G_R}{G_L} = \dfrac{\sin\phi_F - \sin\theta_S}{\sin\phi_F + \sin\theta_S} \end{cases} \quad \text{Loi des sinus, pour une tête fixe dirigée suivant l'axe médian} \tag{2.1}$$

$$\begin{cases} \dfrac{\tan\theta_S}{\tan\phi_F} = \dfrac{G_L - G_R}{G_L + G_R} \\ \dfrac{G_R}{G_L} = \dfrac{\tan\phi_F - \tan\theta_S}{\tan\phi_F + \tan\theta_S} \end{cases} \quad \text{Loi des tangentes, pour une tête libre ou dirigée vers la source virtuelle} \tag{2.2}$$

A l'aide d'un habile formalisme (*pressure phasors* [Bau61]), Bauer a également montré que l'introduction d'une pure différence de phase $(\varphi_1 - \varphi_2)$ – fréquentiellement uniforme – entre les deux canaux (gains complexes de même norme), se reporte comme une différence d'intensité au niveau des oreilles (ILD), alors qu'un rapport réel d'amplitudes est converti en ITD pour donner la loi (2.1). Notons qu'en pratique, les lois (2.1) et (2.2) restent très voisines et coïncident lorsque $\theta_S = 0$ ou $\theta_S = \pm\phi_F$.

Il faut signaler que l'émergence de ces lois a nécessité la simplification draconienne des calculs et des paramètres pris en compte. Cela s'est notamment traduit par la restriction aux procédés de pan-pot "sans ΔT", ainsi que l'usage de développements limités au premier ordre seulement, confinant leur portée à un domaine basse-fréquence. En outre, l'hypothèse simplificatrice d'une tête acoustiquement transparente a été adoptée d'emblée par les différents auteurs.

Interprétation acoustique

Malgré la multiplicité des interprétations ou des formalismes, tous ces calculs traduisent la même réalité. *Nous préférons retenir une interprétation acoustique* telle que Makita [Mak62] a pu introduire en identifiant la propagation du front d'onde synthétique créé par combinaison des contributions des haut-parleurs. Précisons à ce sujet que c'est sa direction d'incidence (2.2) qui fournit selon lui la direction apparente de la source, à condition, devrait-on ajouter, que la tête soit mobile. C'est sur la base de ces résultats que Gerzon définit et généralise le vecteur vélocité $\vec{V}$ au cas de multiples haut-parleurs (équation 1.28 ou 2.21), ce dont il est question en 2.3 et 2.4.

Les lois (2.1) et (2.2) peuvent être retrouvées rapidement à partir de la description synthétique du champ donnée par le vecteur vélocité $\vec{V}$, que nous avons par ailleurs redéfini dans un contexte général en 1.2.2 et dont nous avons montré et interprété les implications sur la localisation et les qualités de l'image sonore en 1.5.2, et cela *sans avoir recours à l'hypothèse de tête acoustiquement transparente*. Dans le cas présent, le vecteur vélocité a pour expression:

$$\vec{V} = \frac{G_L\vec{u}_L + G_R\vec{u}_R}{G_L + G_R} = \cos\phi_F\vec{u}_x + \frac{G_L - G_R}{G_L + G_R}\sin\phi_F\vec{u}_y \tag{2.3}$$

Rappelons (section 1.2.2) qu'à une fréquence donnée, le vecteur $\vec{V}$ – ou plutôt sa partie réelle $\Re(\vec{V}) = r_V\,\vec{u}_V$ dans un contexte plus général – indique la direction d'incidence $\vec{u}_V$ et la vitesse apparente $c_V = c/r_V$ du front d'onde synthétique localement reproduit. Sous la condition que ce front d'onde (caractérisé en champ libre) soit au moins de la dimension de la tête (Figure 2.4), l'ITD basse-fréquence produit correspond à l'effet d'une onde plane naturelle de même direction, mais à un facteur r_V près (1.80). Ce raisonnement doit être cependant nuancé: la prédiction de l'ITD n'est exacte que lorsque le front d'onde synthétique est très large par rapport à la tête; à mesure que le rapport d'échelle devient moins important (Figure 2.4 à droite), l'ITD

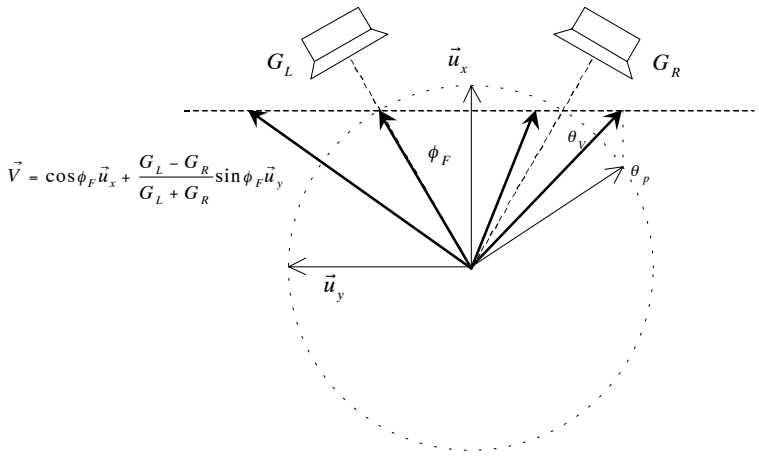

FIG. 2.3 – *Vecteurs vélocité $\vec{V}$ possibles en stéréophonie à deux haut-parleurs. Leur composante suivant l'axe médian $\vec{u}_x$ est fixe et ne dépend que de l'angle ϕ_F. Le module r_V est inférieur à 1 quand $\vec{V}$ pointe entre les deux haut-parleurs et supérieur à 1 "à l'extérieur" des haut-parleurs, le cercle unité étant représenté en pointillé. Par effet de rotation de la tête et en considérant seulement les mécanismes de localisation basse-fréquence, les images sonores n'ont donc pas la même qualité selon leur direction: effet de hauteur ou image floue ($r_V < 1$) pour les positions entre les haut-parleurs, et "sur-latéralisation" à l'extérieur.*

réellement mesuré diminue progressivement par rapport à celui prédit d'après $\vec{V}$. Quoiqu'il en soit, si $r_V < 1$, l'image est donc ramenée vers le plan médian de l'auditeur, par rapport à la direction $\vec{u}_V$, et un effet artificiel de hauteur ($\delta = \arccos r_V$) est produit par rotation de la tête. Cet effet de hauteur est par ailleurs évoqué et expliqué par Bauer à l'aide d'un autre formalisme [Bau61].

Si les gains G_L et G_R sont fréquentiellement uniformes, le front d'onde synthétique a les mêmes caractéristiques de propagation sur toute la bande de fréquence considérée, y assurant la cohérence des indices de localisation. Cette propriété est donc assurée par les pan-pots d'amplitude et les prises de son avec couple coïncident. En revanche, l'introduction d'une différence de temps ΔT (prise de son ORTF par exemple) se traduit par un rapport de gains G_L/G_R complexe dont la phase varie avec la fréquence:

$$\frac{G_L}{G_R}(f) = \left|\frac{G_L}{G_R}\right| e^{j2\pi f \Delta T} \tag{2.4}$$

Pour les très basses fréquences ($f \ll 1/\Delta T$), ce pan-pot peut être assimilé à un pan-pot d'amplitude. Au-delà, le vecteur vélocité devient complexe, sa partie réelle varie en fonction de la fréquence, dégradant la cohérence de la localisation basse-fréquence, tandis que sa partie imaginaire indique l'apparition d'un gradient d'énergie du champ, qui se reporte sur la création d'un ILD artificiel (cf 1.5.2).

L'équation (2.3) et la figure 2.3 montrent que le champ des vecteurs vélocité $\vec{V}$ possibles par interférence de deux ondes planes est limité: son module r_V et sa direction $\vec{u}_V$ ne peuvent pas être contrôlés indépendamment, le vecteur $\vec{V}$ décrivant une droite affine. En écrivant θ_V l'angle de $\vec{u}_V$ dans le repère $(\vec{u}_x, \vec{u}_y)$, c'est lui qui donne l'azimut perçu par effet de rotation de la tête, et est conforme à la loi des tangentes (2.2), alors qu'avec l'hypothèse d'une tête fixe et d'une image dans le plan horizontal, c'est la projection du vecteur vélocité sur le cercle unité et parallèlement à l'axe médian $(0, \vec{u}_x)$ qui donne l'azimut θ_p perçu, conformément à la loi des sinus 2.1).

Les lois (2.1), (2.2) ou encore l'équation (2.3) suggèrent qu'en imposant des gains G_L et G_R de signes opposés, il est possible de *créer des images "à l'extérieur des haut-parleurs"* ($|\theta_S| > \phi_F$), ce dont semble

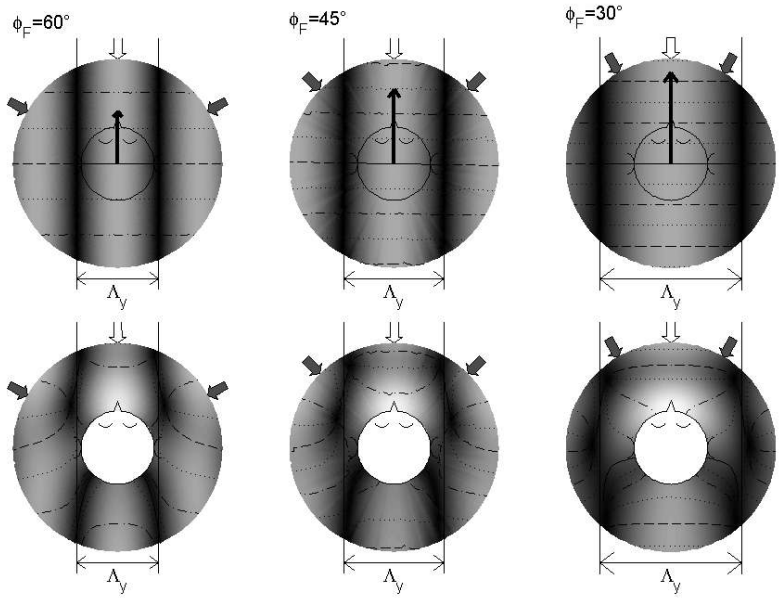

FIG. 2.4 – *Figure d'interférence entre deux ondes planes, avec création ici d'un front d'onde local de direction apparente comprise entre celle des ondes élémentaires. Le niveau de gris représente le champ d'énergie (nul dans les zones sombres), et les lignes en tirets ou pointillés représentent les courbes iso-phase (fronts d'onde). Pour une même fréquence* $f = 1000$ *Hz, la période spatiale* Λ_y *de la figure d'interférence – donc la largeur du front d'onde synthétique – est plus petite lorsque l'angle* $2\phi_F$ *entre les directions des ondes contributives est plus grand, mais est indépendante de la direction du front d'onde synthétique:* $\Lambda_y = \frac{\lambda}{2\sin\phi_F}$*. De gauche à droite: interférence de deux ondes planes d'incidences* $\pm\phi_F = \pm 60°$*,* $\pm\phi_F = \pm 45°$*, et* $\pm\phi_F = \pm 30°$*. En haut: propagation en champ libre; vecteur vélocité représenté par une flèche dirigée vers le haut, le rayon du disque étant considéré comme unitaire pour le module* r_V*. En bas: effet de diffraction en présence de la tête. Dans le cas de gauche (*$\phi_F = 60°$*), la largeur du front d'onde synthétique, rapportée à l'échelle de la tête, est trop critique pour produire un effet de localisation qui corresponde au front d'onde synthétique central. La fréquence* f *est supérieure à la fréquence dite "de* cross-talk*". Pour* $\phi_F = 45°$*, les oreilles se trouvent dans un "trou d'énergie". Pour* $\phi_F = 30°$*, un effet de localisation convenable peut être espéré, approximativement prédit par la valeur du vecteur vélocité.*

pourtant privée la stéréophonie traditionnelle. Malheureusement, cette possibilité est mise en défaut au regard des indices de localisation haute-fréquence que ce type de pan-pot est capable de générer, comme nous l'expliquons dans la section 2.2.3 suivante. Il faut ajouter, considérant à nouveau les seuls mécanismes de localisation basse-fréquence en présence d'une onde synthétique telle que $r_V > 1$ – cas des incidences $|\theta_V| > \phi_F$ –, que les variations de l'ITD par rotation de la tête ne correspondent à rien de "naturel" (variations exagérées), alors qu'elle peuvent correspondre à un effet de hauteur lorsque $r_V < 1$ (se reporter en 1.5.2 pour plus de détails).

Puisqu'il faut s'attendre à la reproduction d'une scène sonore confinée entre les deux haut-parleurs, le seul moyen d'élargir la scène restituée consiste à choisir un angle ϕ_F plus grand. Mais ce choix n'est pas gratuit, car il se traduit par une dégradation des images frontales (Figure 2.4): d'une part, la diminution de l'indice $r_V = \cos\phi_F$ pour une image purement frontale lui donne une qualité plus floue; et d'autre part, le front d'onde localement synthétisé atteint une largeur critique – telle que les oreilles se trouvent dans un "trou d'énergie" – à une fréquence plus basse. Cette fréquence est appelée *fréquence de cross-talk* et peut être définie en première approximation par:

$$\Lambda_y(f_{XT}) = D \qquad \Rightarrow \qquad f_{XT} = \frac{c}{2D\sin\phi_F}, \tag{2.5}$$

où D est la diamètre de la tête. Précisons que la limite de validité du système stéréophonique, pour satisfaire les mécanismes de localisation basse-fréquence, est inférieure à cette fréquence de cross-talk. Le choix de l'angle ϕ_F s'arrête traditionnellement à 30°. Avec un choix $\phi_F = 45°$ ou supérieur, la restitution risque de souffrir de l'effet du "*trou au milieu*", dû au contraste entre la pauvreté de l'image frontale et la précision des images à proximité des haut-parleurs. Ce phénomène s'explique non seulement par l'observation des phénomènes basse-fréquence (Figure 2.4), mais aussi par l'étude des mécanismes haute-fréquence, développée dans les pages suivantes.

2.2.3 Mécanismes et artefacts haute-fréquence: manifestations du *cross-talk*

Alors que les lois de pan-pot d'amplitude se prévalent d'une reconstruction cohérente de l'information de localisation (déphasage interaural) sur toute une bande basse-fréquence, il en est autrement sur le domaine haute-fréquence complémentaire. En effet, le rapport d'échelle entre le front d'onde reproduit localement et la tête (Figure 2.4), met hors de portée une recomposition des indices de localisation au niveau des oreilles avec la même précision ou la même cohérence. En dépit d'une reconstruction imprécise, les mécanismes de localisation propres au domaine haute-fréquence peuvent être pourtant satisfaits (au moins partiellement) pour donner lieu à la création d'une image sonore subjective. Les paragraphes suivants en montrent quelques aspects, selon la stratégie de prise de son envisagée (ΔI ou ΔT). Puis les artefacts de localisation dus au *cross-talk* et propres à ce domaine haute-fréquence sont discutés.

Stéréo "ΔI" ($\Delta T = 0$): mécanismes et modélisation

L'auditeur étant placé et orienté suivant l'axe médian des haut-parleurs, l'observation des réponses temporelles aux oreilles révèle la présence de deux impulsions successives – impacts ipsi- et contra-latéraux – espacées d'un temps τ_{ic} fixe et identique pour les deux oreilles. Seule l'énergie relative des impulsions change en fonction de la direction de la source virtuelle, puisque seuls des gains G_L et G_R sont mis en jeu. Si l'intervalle τ_{ic} est assez faible, ce qui suppose que l'angle ϕ_F soit raisonnable, une fusion des deux événements peut s'opérer. La détection des instants d'arrivée à chaque oreille est alors déplacée vers la première ou la seconde contribution, selon leur poids énergétique relatif (Figure 2.5). Ce déplacement s'effectue symétriquement et dans un sens opposé d'une oreille à l'autre. Par un raisonnement purement qualitatif, on comprend aisément

que l'ITD détecté est borné par ceux que peuvent produire les haut-parleurs comme sources seules, et que l'image sonore créée ne peut être que confinée entre les deux haut-parleurs.

Pour quantifier l'ITD supposé détecté, il est commode d'utiliser le modèle de détection des époques moyennes des enveloppes d'énergie, suggéré par Mertens [Mer62] [Mer65], et dont nous avons adapté le principe en 1.4.3. Nous avons également montré un lien de prédiction entre l'ITD détecté ou la direction perçue, et le *vecteur énergie* $\vec{E}$, critère synthétique introduit par Gerzon dans un cadre plus général de restitution sur haut-parleurs. Dans le cas présent, le vecteur énergie s'écrit comme pondération des incidences $\vec{u}_L$ et $\vec{u}_R$ des haut-parleurs par les énergies (ou puissances) associées G_L^2 et G_R^2:

$$\vec{E} = \frac{G_L^2 \vec{u}_L + G_R^2 \vec{u}_R}{G_L^2 + G_R^2} = \cos\phi_F \ \vec{u}_x + \frac{G_L^2 - G_R^2}{G_L^2 + G_R^2} \ \sin\phi_F \ \vec{u}_y, \tag{2.6}$$

en réutilisant les notations de la figure 2.3. D'après Gerzon, ce vecteur indique la direction perçue au regard des mécanismes de localisation haute-fréquence et avec l'aide des rotations de la tête [Ger74](?), avec une précision indiquée par son module $r_E = |\vec{E}|$, la meilleure valeur $r_E = 1$ étant atteinte dans le cas d'une source unique. De la même façon que nous l'avons fait avec le vecteur vélocité, il semble utile d'apporter à ce propos quelques nuances et précisions. En supposant valide la modélisation d'après les époques moyennes, on peut admettre que $\vec{u}_E = \vec{E}/rE$ prédit assez bien la *direction perçue en lui faisant face*. Ensuite, une valeur $r_E < 1$ indique une réduction de l'effet de latéralisation, en général plus conséquente que celle associée en basse-fréquence à une valeur $r_V = r_E$. Cela se traduit par un déplacement de l'image sonore vers le plan médian, ou encore un effet de hauteur, et plus généralement par une image de caractère flou, selon les libertés de rotation de la tête.

Pour une loi de pan-pot ou des gains identiques en haute- et basse-fréquence, les vecteurs vélocité et énergie n'ont en général pas la même direction ni le même module, exceptés dans les directions $\pm\phi_F$ des haut-parleurs ($r_V = r_E = 1$) et dans la direction médiane ($r_V = r_E = \cos\phi_F$). Lorsque $\vec{V}$ pointe à l'extérieur du secteur angulaire $[\phi_F \ -\phi_F]$ des haut-parleurs (S_L et S_R en opposition de phase), $\vec{E}$ y reste confiné et revient vers le centre. On note que la "qualité haute-fréquence" d'une image centrale, indiquée par $r_E = \cos\phi_F$, pâtit d'un élargissement de l'angle $2\phi_F$ de la même manière que sa "qualité basse-fréquence" r_V: c'est un argument de plus pour limiter de choix de ϕ_F à une valeur consensuelle de 30°.

Modélisation avec ΔT et ΔI

Le ΔT et le ΔI introduits avec le couple ORTF pourraient fournir des indices de localisation très satisfaisants s'ils se reportaient directement sur l'ITD et l'ILD, c'est-à-dire si le masquage contralatéral était suffisant. A défaut d'un masquage complet, chaque réponse ipsi-latérale est là encore légèrement "polluée" par la contribution contralatérale. Mais du fait de la non-simultanéité des fronts d'onde à leur arrivée aux oreilles ($\Delta T \neq 0$ dans le cas d'une source latérale), les mécanismes psychoacoustiques mis en jeu pour l'extraction des informations de localisation ne sont plus les mêmes que dans le cas d'un pan-pot purement ΔI (émission synchrone par les haut-parleurs). Cette différence tient au fait que l'oreille exposée au haut-parleur le plus précoce reçoit les contributions ipsilatérale et contralatérale avec un écart temporel trop important pour qu'il y ait fusion des deux événements successifs [Bla83] (si ΔT est suffisamment grand), du moins pour les signaux de type impulsif ou transitoire (Figure 2.5). Dans ce cas, la détection du temps d'arrivée à cette oreille se fait uniquement d'après le premier événement, d'autant qu'il est d'amplitude plus importante, alors qu'à l'autre oreille, la détection se base sur la fusion des contributions ipsi- et contra-latérales qui sont plus rapprochées, à la fois en temps et en énergie. Ainsi, le ΔT entre les signaux émis se reporte assez bien – quoique partiellement – sur la différence de temps détectée, tandis que le résidu contralatéral tardif perçu par l'oreille précocement exposée peut être interprété comme l'effet d'une réflexion.

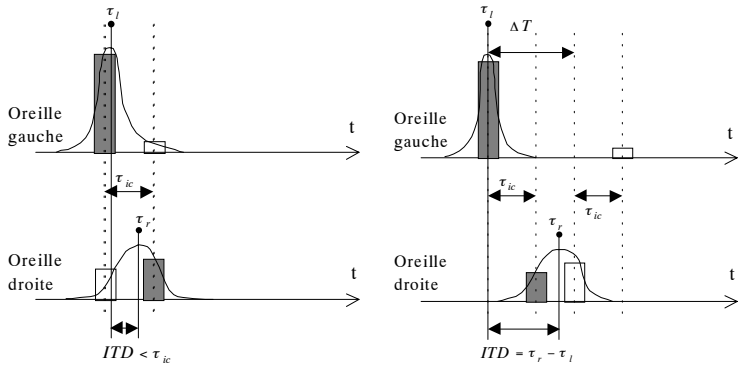

FIG. 2.5 – *Détection des instants d'arrivée à chaque oreille. A gauche: cas de deux incidences synchrones (stéréo ΔI), avec hypothèse de fusion des événements. Les haut-parleurs étant placés en $\pm\phi_F = \pm 30°$, les contributions impulsives ipsi- et contralatérales se succèdent avec un écart de $\tau_{ic} \simeq 1{,}3\frac{D}{c}\sin\phi_F$, où D est le diamètre de la tête. Ici, $\tau_{ic} \simeq 0{,}25$ ms est suffisamment petit pour qu'il y ait fusion psychoacoustique des événements. A droite: cas de deux ondes impulsives non-simultanées (stéréo ΔT), avec de surcroît une différence d'amplitude (couple ORTF). A l'oreille gauche précocement exposée, l'écart temporel entre les contributions ipsi- et contra-latérales est important et la détection du temps d'arrivée τ_l se fait sur la première, par ailleurs prédominante en énergie. A l'oreille droite, les deux contributions sont plus rapprochées et la détection de l'instant d'arrivée τ_r se base sur leur fusion. L'ITD résultant $\tau_r - \tau_l$ a une valeur intermédiaire entre le ΔT et l'ITD τ_{ic} associé à l'angle de haut-parleurs.*

D'après ce mécanisme et en se basant sur les transitoires d'attaque, *il paraît donc possible de créer des effets d'incidence perçue hors du secteur angulaire confiné entre les haut-parleurs*, ce qui semblait interdit aux procédés purement ΔI. En contrepartie, le comportement des images s'avère *moins robuste et moins contrôlable* dans le cas de signaux plus continus, et surtout ceux à bande-étroite. En effet, l'introduction d'une différence de temps ΔT – donc d'une différence de phase dépendant linéairement de la fréquence – implique en régime harmonique une excursion assez importante des retards de phase et de groupe perçus aux oreilles en fonction de la fréquence. Dans un contexte de restitution musicale, cela se traduit par exemple par la fluctuation de la position apparente d'un violon en fonction de sa mélodie.

Le couple AB, dont l'espacement entre les microphones est notablement plus grand que celui du couple ORTF, mérite également de brefs commentaires. Avec ce couple, la prise de son d'une *source très latérale* introduit en effet une différence de temps ΔT bien supérieure à l'ITD maximal. A la restitution, c'est alors la loi du premier front d'onde qui détermine la direction perçue, l'effet du deuxième front étant interprétable comme une réflexion: *l'image est localisée sur un haut-parleur*, et non "à l'extérieur" comme pouvait en être capable le couple ORTF. Lorsqu'au contraire le couple AB capte un ensemble de *sources confinées dans un secteur angulaire frontal relativement réduit*, les différences de temps ΔT sont plus importantes qu'avec un couple ORTF, tout en restant raisonnables par rapport à l'ITD maximal: le mécanisme de détection présenté Figure 2.5 est alors applicable pour les signaux de type impulsif ou transitoire, à la différence près qu'aucun ΔI n'est introduit pour une source lointaine[6]. En générant des différences ΔT significatives entre des sources de faible écart angulaire, *le couple AB fait donc l'effet d'une "loupe sonore"*.

6. Les directivités des microphones existants n'étant de toutes façons pas suffisamment sélectives pour fournir des variations d'amplitude significatives pour de faibles écarts angulaires, ce sont des microphones omnidirectifs qui sont utilisés.

Qualité des images

Bien que des mécanismes psychoacoustiques aient pu être mis en évidence pour expliquer la création subjective d'images sonores localisées ailleurs que sur les haut-parleurs, cela ne leur confère pas la qualité qui peut être associée à une source sonore naturelle. Concernant la *qualité individuelle d'une image fantôme*, les tares de la reproduction stéréophonique revêtent deux aspects. D'une part, *le cross-talk dégrade la cohérence des différences interaurales*, et surtout de l'ITD. Cette cohérence s'observe suivant l'axe des fréquences, mais elle est également éprouvée lors de rotations de la tête, même légères. On peut voir en r_E une indication approximative du degré de cohérence, et de la consistance de l'image produite. D'autre part, aucune des stratégies employées n'est apte à restituer les *indices spectraux* associés à une image naturelle, les indices spectraux reconstitués portant l'empreinte des deux haut-parleurs.

Limites à la décorrélation, conséquences sur la largeur et la coloration

Le *cross-talk* impose au niveau de l'auditeur une limite implacable à la décorrélation interaurale, celle-là même que fournissent les réflexions précoces et la réverbération en situation d'écoute "directe", et qui est responsable de la largeur apparente de source et de l'impression spatiale. Il tend donc à convertir partiellement ces qualités subjectives naturelles en un effet de coloration, ainsi que l'on peut en faire l'expérience à l'issue d'une prise de son/restitution monophonique ou encore en se bouchant une oreille. C'est d'ailleurs probablement pour cette raison que le meilleur placement du couple de microphones, pour l'enregistrement d'un concert par exemple, ne correspond généralement pas à la place qu'un spectateur pourrait juger idéale pour l'écoute directe du concert: en choisissant un placement plus proche des sources, on augmente l'importance relative du son direct et on élargit la distribution latérale des réflexions précoces, ce qui a pour effet de limiter leur conversion en effet de coloration lors de la restitution sur haut-parleurs.

C'est l'identification du cross-talk comme l'artefact nuisible majeur de la stéréophonie traditionnelle, qui motive également Ralph Glasgal à proposer une solution assez radicale à travers son système *Ambiophonics*: éliminer ce cross-talk par une barrière acoustique, physique [Gla95].

Restriction à deux haut-parleurs: risques de rupture de l'illusion sonore et problème de stabilité

A défaut de produire des images fantômes ayant les qualités d'image naturelle, l'une des conditions pour préserver l'illusion sonore est de leur assurer une qualité relativement homogène sur l'ensemble de la scène sonore ou lorsqu'elles se déplacent, afin que le système perceptif puisse s'y accommoder. Dans le cas présent d'une reproduction sur deux haut-parleurs où l'on ne peut contrôler indépendamment la direction ($\vec{u}_V$, $\vec{u}_E$) et la qualité (r_V, r_E) des images fantômes, il est donc nécessaire de faire le choix d'un angle ϕ_F raisonnable pour éviter un trop fort contraste de qualité entre les images parfaitement frontales et celles à proximité des haut-parleurs.

Une condition corollaire à la préservation de l'illusion est d'éviter l'émergence des haut-parleurs comme sources individuelles. Si celle-ci est inévitable pour une image latérale issue d'un pan-pot artificiel, elle peut être moins sensible dans le cas d'une restitution à partir d'une prise de son réelle... et bien faite! Mais la dissociation des haut-parleurs comme sources individuelles, même lorsqu'ils participent tous deux à la fois, peut avoir lieu de façon plus insidieuse quoique moins flagrante, par légers mouvements de la tête[7]: la présence des deux sources simultanées est en effet trahie par des changements de la coloration perçue. A l'extrême, l'émergence d'un des deux haut-parleurs a bien-sûr lieu de façon beaucoup plus évidente lorsque l'auditeur s'écarte trop de l'axe médian des haut-parleurs: le haut-parleur le plus proche est en effet perçu

7. C'est du moins la thèse de Günther Theile.

plus tôt (effet d'antériorité) et avec plus d'intensité. Le problème de stabilité et de dissociation possible des sources réelles est là encore attribuable à la restitution sur deux haut-parleurs: un nombre plus important de contributions simultanées permettrait en effet d'éviter une commutation binaire entre la perception d'un haut-parleur ou de l'autre par déplacement latéral par rapport à l'axe médian, et favoriserait plus généralement le gommage de leur individualité.

2.2.4 Bilan

Notions émergentes

Tout en démontrant les limitations de la stéréophonie traditionnelle sur deux haut-parleurs, l'étude qui vient d'être présentée a donné l'occasion d'introduire un certain nombre de notions, d'outils et de préoccupations qui resteront présents dans les sections suivantes pour l'étude de systèmes plus élaborés. En particulier, l'approche basée sur l'observation et la caractérisation des phénomènes acoustiques synthétisés pour l'interprétation de la restitution sonore, pourra être exploitée directement ou bien extrapolée suivant les stratégies considérées.

Un premier enseignement est la mise en évidence de **deux domaines fréquentiels qui offrent des conditions distinctes et inégales aux mécanismes/[à la réalisation] d'illusions sonores**, considérant un dispositif et une position d'écoute donnée:

- *Un domaine basse-fréquence*, pour lequel il est *possible de reconstruire, au moins à l'échelle de la tête, un front d'onde synthétique* avec des caractéristiques de propagation (direction et vitesse) fréquentiellement uniformes, et différentes – au moins en direction – des ondes contributives élémentaires venant des haut-parleurs: il s'agit alors d'un *événement acoustique synthétique perceptivement indissociable*, propre à fournir des indices de localisation cohérents sur tout le domaine basse-fréquence considéré, y compris lors d'une rotation de la tête, même si leur variation au cours de cette rotation peut ne pas correspondre à l'effet d'une onde plane naturelle.
- *Au-delà, un domaine haute-fréquence*, où se manifestent les *artefacts du cross-talk*, i.e. du "mélange" des contributions concourantes: la reconstruction des informations auditives de localisation ne peut en général [cas stéréo] pas y être assurée avec la même cohérence qu'une image sonore naturelle due à une onde plane. Lorsque les conditions ne sont pas trop critiques, le processus de création d'une image sonore localisée (image fantôme) repose sur des mécanismes psychoacoustiques de fusion (ΔI) et/ou d'inhibition/sélection (ΔT) des différentes contributions. Au contraire du domaine basse-fréquence, *la distinction des différentes sources réelles contributives devient objectivement possible* par mouvements de la tête (dont rotations), et ce, d'autant plus qu'elles sont distinctes dans l'espace. Même si, en vertu d'une certaine tolérance [ou imperfection] psychoacoustique, le système perceptif n'est pas capable dans tous les cas de faire émerger à la conscience une telle distinction, *l'image sonore subjective reconstituée*, prise individuellement, ne possède pas les qualités d'une image naturelle, selon le degré de cohérence statique et/ou dynamique des indices de localisation produits. Considérant plus globalement la restitution associée à la prise de son dans un champ complexe – avec réflexions et effet de salle, par exemple – les artefacts du *cross-talk* se manifestent également par *la transformation et la dégradation des qualités spatiales ordinairement attendues des réflexions et du champ réverbéré*: *la décorrélation* des signaux stéréophoniques – due aux réflexions latérales captées par le couple microphonique, dans le cas présent – est en effet *dégradée* par le mélange des signaux à leur réception aux oreilles, et par ce fait, l'effet des réflexions est partiellement *"transformé" en effet subjectif de coloration*[8] (section 1.3.4).

8. C'est surtout vrai pour les techniques "ΔI", pour lesquelles la décorrélation des signaux stéréophoniques est à la base moins

La notion de cross-talk est couramment assimilée à ses artefacts ou manifestations “anti-naturelles" en haute-fréquence. La distinction entre les deux domaines fréquentiels conduit à la définition d'une *fréquence de cross-talk* $f_{XT} \approx \frac{c}{2D \sin \phi_F}$ qui les délimite. Bien que l'observation du cross-talk se base ici sur le cas de la stéréophonie conventionnelle, la description de ses artefacts haute-fréquence est tout à fait transposable, comme nous le verrons, à bien d'autres systèmes. Même avec les systèmes qui ont pour vocation de contrôler la reconstruction des signaux au niveau des oreilles (section 2.5), la séparation fondamentale entre les deux domaines restera présente à travers des questions de stabilité.

Deuxièmement, cette étude a mis en évidence l'**utilité des vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$** comme outils synthétiques pour caractériser la restitution et son potentiel. Les propriétés attachées à ces vecteurs ont été démontrées dans un cadre très général au chapitre précédent: d'abord définis comme caractérisant la propagation acoustique – à une échelle locale pour $\vec{V}$, et globale pour $\vec{E}$ –, il a été montré qu'ils permettaient de prédire, sous certaines conditions, les indices et les effets de localisation. Leur définition étant ici étroitement liée à la géométrie du dispositif, ils permettent des déductions très rapides sur l'effet et la qualité de localisation à attendre de la restitution. Leur usage est tout particulièrement dédié aux procédés ΔI, qui assurent leur uniformité en fonction de la fréquence. On peut également voir en le vecteur énergie $\vec{E}$ un indice indirect du degré de cross-talk, dont une manifestation, considérant un champ complexe dans son ensemble, est le manque de décorrélation interaurale et l'exagération de la coloration perçue.

Les **implications de la géométrie du dispositif** sont particulièrement sensibles dans le cas d'une restitution sur deux haut-parleurs: il en résulte *des contraintes et des compromis délicats*.

- De l'angle $2\phi_F$ séparant les haut-parleurs dépend directement la *périodicité spatiale Λ_y de la figure d'interférence* des deux ondes planes, qui, rapporté au diamètre de la tête, définit la fréquence de cross-talk.
- A la géométrie du dispositif est associée une *classe d'événements acoustiques reproductibles*, caractérisés par exemple par les vecteurs vélocité et énergie. La restriction à deux haut-parleurs – cas auquel nous avons affaire ici – apparaît comme très limitant: on ne peut pas contrôler à la fois la direction et le module de $\vec{V}$ (propagation locale), et il en est de même pour $\vec{E}$. Cela se traduit directement par un manque d'homogénéité de la qualité et de la stabilité des images sonores en fonction de leur direction apparente.

Deux stratégies de base ont été décrites pour l'instant, liées, sur le plan de la prise de son, à la coïncidence ou la non-coïncidence des microphones utilisés, ces propriétés se traduisant sur le plan de la restitution par la **convergence synchrone ou bien non-synchrone des ondes**. Les deux sont défendables dans un *contexte de prise de son artistique*: elles ont en effet leurs qualités et leurs propriétés propres, et leur choix pratique relève bien souvent d'une question d'école et d'esthétique. Cela dit, l'approche ΔI apparaît nettement avantageuse comme stratégie de restitution basse-fréquence, de par la cohérence des informations de localisation (retard de phase) offerte en toutes conditions. Rappelons que cette cohérence est directement liée à l'uniformité du vecteur vélocité sur toute la bande basse-fréquence. D'un autre côté, la restitution associée aux prises de son par microphones non-coïncident (ΔT) possède certaines qualités “artistiques" appréciables – souvent décrites par en termes de “largeur", “air", “espace" –, auxquelles l'étude des mécanismes haute-fréquence de recréation d'image sonore a pu donné quelques explications. Mais sans-doute faut-il souligner comme le fait Lipshitz [Lip86], que ces qualités sont peut-être plus le fait des artefacts de l'approche ΔT que d'une reproduction fidèle du champ original enregistré, et qu'en contrepartie cette approche est caractérisée par un comportement peu prédictible et peu contrôlable des images sonores, dont l'effet de localisation peut être très variable selon la nature du signal. Par contraste, si l'approche synchrone n'offre aucun moyen de créer des images sonores hors du secteur angulaire des haut-parleurs, elle propose un rendu direction-

---

importante.

nel moins fantaisiste et plus robuste en assurant une meilleure cohérence globale des indices de localisation haute-fréquence, avec l'avantage de disposer d'un outil de prédiction: le vecteur énergie. Ce sont autant de raisons qui feront porter un intérêt privilégié à l'approche ambisonique (2.4).

Enfin, la comparaison des techniques microphoniques permet de **mettre en relation les qualités globales de la restitution et la "représentation" de la scène sonore issue de la prise de son, c'est-à-dire de l'encodage acoustique du champ original**. Dans le cas de la stéréophonie sur deux canaux, il semble que le choix d'un couple de microphone fasse à chaque fois l'objet d'un dilemme: le couple *StereoSonic* de Blumlein a par exemple la qualité de préserver l'équilibre énergétique de l'ensemble des sources sonores, et le défaut de capter les sources arrière, et de les superposer à la restitution de la scène frontale. Aucune de ces techniques ne permet par ailleurs d'assurer à la restitution la cohérence entre la prédiction de la localisation basse-fréquence et la prédiction haute-fréquence, c'est-à-dire la colinéarité des vecteurs vélocité et énergie. [...] Comme nous le verrons plus loin, ces dilemmes sont résolus de façon naturelle avec l'approche ambisonique, qui peut d'ailleurs être vue comme l'extension logique de la prise de son *StereoSonic* [cite Malham?].

Conclusions sur la stéréophonie conventionnelle

La stéréophonie conventionnelle, dont les principaux procédés viennent d'être présentés, reste probablement le mode de diffusion et de reproduction d'une scène sonore le plus répandu, notamment en matière de production musicale. Celle-ci bénéficie en effet de techniques de prise de son naturelle et de mixage éprouvées. Le mode de diffusion sur deux canaux laisse quant à lui une très forte empreinte et conditionne en partie le développement de procédés plus évolués (*surround*, *Ambisonics*, binaural, transaural, etc...) par des problèmes de compatibilité.

2.3 Stéréophonie panoramique et périphonique

Les procédés de création d'image sonore (dits "procédés ΔI et ΔT") mis en jeu en stéréophonie conventionnelle ont l'avantage de pouvoir dériver de prises de son naturelles ou de pan-pots de très faible complexité. Mais utilisés avec seulement deux haut-parleurs frontaux, ils ont l'inconvénient de restreindre la reproduction sonore à une scène frontale confinée et de limiter le potentiel d'impression spatiale et d'immersion. Compte-tenu des mécanismes stéréophoniques mis en jeu, ces limites ne peuvent être dépassées qu'en enrichissant le dispositif de haut-parleurs, ce qui donne lieu à la stéréophonie panoramique ou *surround* (dispositif horizontal entourant l'auditeur) ou bien encore périphonique (restitution 3D avec haut-parleurs en hauteur). Dans cette section, l'extension des techniques stéréophoniques est abordée suivant différents aspects:

- Les configurations de haut-parleurs recommandées pour la restitution panoramique (en 2.3.1).
- L'extension des techniques de pan-pot 2D et 3D (en 2.3.2).
- Les techniques de prise de son multi-canal (en 2.3.3).
- Les systèmes de codage/décodage *surround* (en 2.3.4) utilisant deux canaux de transmission compatibles avec le standard de diffusion stéréophonique.

L'approche ambisonique, qui y est évoquée à plusieurs reprises, se voit réservée la section 2.4 entièrement, en raison de ses spécificités.

2.3.1 Extension et stabilisation de la scène sonore: pourquoi et comment

Tentative et échec de la quadriphonie

Depuis l'établissement d'un mode de restitution sur deux haut-parleurs comme standard stéréophonique, l'expérience quadriphonique dans les années 60, a été la première tentative concrète et à prétentions commerciales pour la reproduction d'un espace sonore entourant l'auditeur de toutes parts, reproduction jusque là confinée à une scène frontale relativement étroite. Comme son nom l'indique, la restitution quadriphonique repose sur l'utilisation de quatre haut-parleurs, disposés en carré. Ce dispositif intuitivement suffisant pour une restitution panoramique complète se révèle en réalité très ingrat, si l'on considère que l'angle de 90° entre deux haut-parleurs adjacents est déjà critique pour une restitution conventionnelle sur deux haut-parleurs. L'artefact du "trou au milieu", émergeant par contraste entre une image très précise lorsqu'elle est située sur un haut-parleur et une qualité médiocre entre les haut-parleurs, conduit en effet inéluctablement à la rupture de l'illusion. La plupart des stratégies de restitution quadriphonique n'ayant pas offert de solution adaptée à ce problème, l'expérience quadriphonique s'est soldée par un échec. Il existait pourtant une approche qui offrait les conditions d'une restitution homogène, donc apte à préserver l'illusion sonore: l'approche ambisonique (abordée section 2.4), qui malgré ses qualités et la possibilité d'une restitution sur des dispositifs de haut-parleurs plus favorables, a probablement souffert des retombées de cet échec. Quoiqu'il en soit, la configuration carrée reste assez défavorable à une production d'images sonores à la fois précises et stables en fonction des déplacements de l'auditeur.

Recommandations sur le dispositif pour la restitution multi-canal

C'est en grande partie un contexte de télévision haute-définition (TVHD) qui a dicté ses exigences et présidé aux recommandations sur le choix des dispositifs de restitution panoramique, dite encore restitution *surround* ou multi-canal. La culture d'une scène frontale privilégiée, mais aussi la présence de l'image, ont fait porter la préoccupation dominante sur la précision et la stabilité des images sonores frontales, impliquant naturellement une densification des haut-parleurs frontaux. Celle-ci doit pouvoir minimiser la distorsion directionnelle entre l'image et le son, non seulement pour un auditeur situé au centre du dispositif, mais aussi pour un auditoire plus élargi. La distorsion audio-visuelle jugée tolérable est $V = 10\%$. Un consensus s'est établi [CCI92] sur la base de trois haut-parleurs frontaux ($\phi_F = 30°$) et seulement deux haut-parleurs arrière ou latéraux ($60° \leq \phi_B \leq 90°$), la distribution des événements sonores latéraux et arrière requérant moins de précision directionnelle (Figure 2.6). Plus que de participer à la création d'images fantômes latérales, la vocation majeure des haut-parleurs latéraux et arrière est de restituer les sons d'ambiance et de favoriser l'enveloppement et l'impression spatiale. A ces recommandations sur la géométrie du dispositif, dite "$3-2$", est désormais associé un standard de diffusion multi-canal[9] sur cinq canaux séparés: le format 5.1 (cinq canaux discrets plus un canal basse-fréquence LFE[10]), en vigueur en radio et télévision numériques, de même que pour le DVD (*Digital Versatile Disc*) et le SACD (*Super-Audio Compact Disc*), successeurs du CD.

Dans un contexte d'application plus général – création électro-acoustique ou renforcement sonore pour large assemblée par exemple – des dispositifs plus fournis sont également envisageables, permettant à la fois d'assurer une restitution relativement robuste des images frontales et d'envisager la création d'images-fantômes latérales consistantes. Par souci d'homogénéité de la restitution, des configurations en polygones réguliers peuvent être recommandées (Figure 2.6).

9. "Multi-canal", ou bien "multi-canaux", ou encore moins couramment "multi-sons", chacun de ces termes étant invariable d'un point de vue grammatical!

10. *Low Frequency Effect*: il s'agit d'effets basse-fréquence non-directifs.

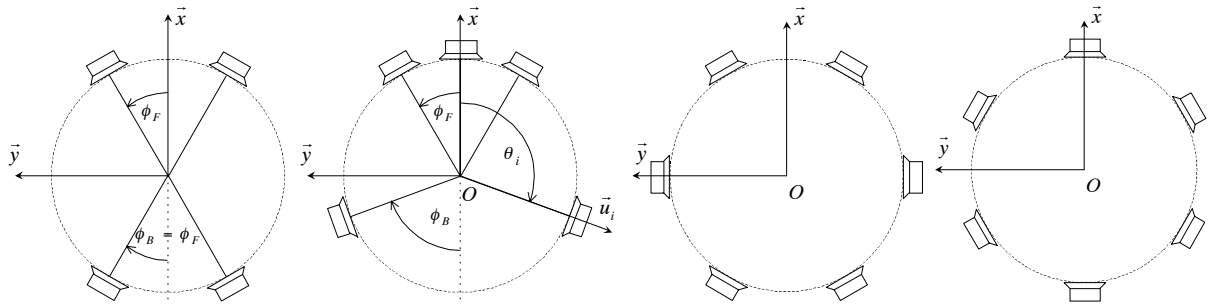

FIG. 2.6 – *Configurations typiques de haut-parleurs pour une restitution panoramique (de gauche à droite): rectangulaire (quadriphonique si $\phi_F = 45°$); configuration $3-2$ (typiquement, $\phi_F = 30°$ et $60° \leq \phi_B \leq 90°$); configurations hexagonales régulières.*

2.3.2 Techniques de pan-pot

Procédés de pan-pot panoramique (2D) et périphonique (3D)

Nous présentons d'abord des procédés de pan-pot prenant modèle sur les pan-pots d'amplitude (2.2) ou d'intensité (2.6) définis pour la stéréophonie sur deux haut-parleurs. Ils sont dotés d'outils objectifs de prédiction de la localisation relativement fiables: les vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$. Les pan-pots avec ΔT, dont les lois dérivent de données essentiellement empiriques, ne sont pas exploités ici parce que l'effet de localisation qui leur est associé est moins prédictible et généralement moins robuste, notamment pour les sources latérales et lors des rotations de la tête. Les procédés ΔT seront par contre à nouveau présents dans un contexte de prise de son panoramique, en 2.3.3.

Le VBAP (*Vector Based Amplitude Panning*) [Pul97] se présente comme une formulation vectorielle de la loi des tangentes (2.2), formalisme qui permet de définir une méthode générique de pan-pot pour des dispositifs panoramiques (2D) et même périphoniques (3D). Le VBAP repose donc implicitement sur la théorie de Makita [Mak62], selon laquelle la direction perçue – si la tête est mobile – est celle du vecteur vélocité $\vec{V}$, c'est-à-dire celle du front d'onde synthétisé au centre du dispositif. Plaçons-nous d'abord dans le cadre d'une restitution horizontale et supposons une source virtuelle à restituer dans la direction $\vec{u}_S$ (angle θ_S) à l'aide de deux haut-parleurs de directions $\vec{u}_1$ et $\vec{u}_2$ (vecteurs unitaires, ou angles θ_1 et θ_2). Le pan-pot proposé par Pulkki consiste à attribuer aux haut-parleurs des gains respectifs g_1 et g_2 tels que:

$$\vec{u}_S = g_1 \vec{u}_1 + g_2 \vec{u}_2 \tag{2.7}$$

On vérifie immédiatement que la direction du vecteur unitaire $\vec{u}_S$ est celle du vecteur vélocité $\vec{V} = \frac{g_1 \vec{u}_1 + g_2 \vec{u}_2}{g_1 + g_2}$. L'équation (2.7) s'écrit sous forme matricielle:

$$\mathbf{u_S}^t = \mathbf{g}.\mathbf{U_{12}} \qquad \text{avec} \quad \mathbf{g} = [g_1\, g_2]^t \qquad \text{et} \quad \mathbf{U_{12}} = [\mathbf{u_1}\, \mathbf{u_2}]^t = [\vec{u}_1\, \vec{u}_2]^t = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \end{bmatrix} \tag{2.8}$$

Les gains g_1 et g_2 sont alors obtenus par inversion du système:

$$\mathbf{g} = \mathbf{u_S}^t\, \mathbf{U_{12}}^{-1}, \tag{2.9}$$

la matrice $\mathbf{U_{12}}$ étant inversible dès que $\vec{u}_1$ et $\vec{u}_2$ ne sont pas colinéaires. Il s'agit en quelque sorte d'une projection du vecteur $\vec{u}_S$ sur la base vectorielle $(\vec{u}_1, \vec{u}_2)$. Pulkki propose en outre de corriger ces gains pour

une préservation de l'énergie totale restituée:

$$\tilde{\mathbf{g}} = \frac{\mathbf{g}}{|\mathbf{g}|} = \frac{\mathbf{g}}{\sqrt{\mathbf{g}.\mathbf{g}^t}}, \tag{2.10}$$

où $\tilde{\mathbf{g}} = [\tilde{g}_1\ \tilde{g}_2]$, de sorte que $\tilde{g}_1^2 + \tilde{g}_2^2 = 1$. Notons au passage une légère contradiction entre ce souci de préservation de l'énergie, et l'hypothèse de sommation des ondes en amplitude, sous-jacente à la loi des tangentes (au moins dans un domaine basse-fréquence). Lorsque le dispositif comprend plus de deux haut-parleurs, le pan-pot fait intervenir la paire de haut-parleurs adjacents à la source virtuelle[11].

Le principe s'étend aisément au cas de configurations 3D de haut-parleurs. Il convient alors de choisir comme base trois haut-parleurs adjacents (de directions $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$) formant un triangle dans lequel pointe la direction $\vec{u}_S$ de la source virtuelle, vu de l'auditeur. Les gains g_1, g_2, g_3 attribués aux haut-parleurs doivent alors vérifier, en notant $\mathbf{g} = [g_1\ g_2\ g_3]$ et $\mathbf{U_{123}} = [\mathbf{u_1}\ \mathbf{u_2}\ \mathbf{u_3}]^t$:

$$\mathbf{u_S}^t = \mathbf{g}\,\mathbf{U_{123}} \qquad \Rightarrow \mathbf{g} = \mathbf{u_S}^t\,\mathbf{U_{123}}^{-1} \tag{2.11}$$

La conservation de l'énergie nécessite là encore d'appliquer la correction (2.10).

Il a été souligné en 2.2 que la validité de la loi des tangentes – c'est-à-dire la prédiction par le vecteur vélocité – est restreinte à un domaine basse-fréquence au-delà duquel c'est le vecteur énergie $\vec{E}$ qui se présente comme le meilleur prédicteur de la localisation, conformément aux théories de Gerzon [Ger92b] étayées par notre étude en 1.5. Aussi a-t-il été proposé [PBJ98] d'étendre le principe du VBAP en VBIP (*Vector Based Intensity Panning*), où il s'agit de définir des gains g_i tels que le vecteur énergie $\vec{E} = \frac{\sum g_i^2\,\vec{u}_i}{\sum g_i^2}$ ait la direction souhaitée $\vec{u}_S$. Le calcul est une variante très simple du VBAP: il suffit de remplacer les g_i issus de (2.9) ou (2.11) par les valeurs $\sqrt{g_i}$ avant d'appliquer la normalisation (2.10). Il est suggéré [PBJ98] que les deux méthodes VBAP et VBIP soient appliquées complémentairement sur des domaines respectivement basse- et haute-fréquence au moyen d'un filtrage par sous-bandes, la fréquence de transition suggérée étant de 700 kHz. Cette méthode est désignée sous le nom de VBP (*Vector Based Panning*).

Pour Pulkki, le VBAP offre en toute direction une "précision maximale" à l'image sonore compte-tenu des haut-parleurs présents. En particulier, seul un haut-parleur est alimenté lorsque la source virtuelle est placée dans sa direction. En réalité, la propriété de "précision maximale" est inexacte: pour les images placées entre plusieurs haut-parleurs, l'indice $r_V < 1$ indique une "qualité basse-fréquence" sous-optimale en position d'écoute centrée, contrairement à ce que pourra offrir une restitution ambisonique (section 2.4) en faisant participer l'ensemble des haut-parleurs; le VBIP est plus recommandable en haute-fréquence, voire en pleine bande si l'on envisage des positions d'écoute excentrées. Mais surtout, ce souci de "précision maximale" a son revers de médaille, à savoir une qualité d'image sonore (indices r_V et r_E) non-homogène en fonction de sa position, qui peut s'avérer gênante particulièrement dans le cas de sources en mouvements [PBJ98]. Les haut-parleurs émergent individuellement au passage des sources virtuelles ("*ping-pang-pung effect*") et les contrastes de précision – d'autant plus forts que les haut-parleurs sont écartés – risquent de briser l'illusion d'image fantôme. A cela peut s'ajouter des variations de la coloration perçue. D'un autre côté, l'utilisation de haut-parleurs confinés dans un secteur angulaire limité offre une relative stabilité de la restitution lors d'une écoute en position excentrée. Enfin, le VBAP (ou le VBIP ou le VBP) présente l'avantage pratique d'être facilement applicable à toutes sortes de configurations de haut-parleurs, y compris non-régulières. L'ensemble des directions couvertes est défini par l'union des secteurs angulaires (ou angles solides pour une configuration 3D) formés par les paires ou triplets de haut-parleurs.

11. C'est ainsi que nous avons implémenté le VBAP 2D (pour configurations horizontales) parmi d'autres techniques de spatialisation, comme nous le présentons au chapitre 5

Lois de pan-pot plus élaborées

Les stratégies qui viennent d'être présentées utilisent un nombre minimal de haut-parleurs simultanés – deux en 2D et trois en 3D – pour contrôler la direction apparente de la source, supposée prédite par le vecteur vélocité $\vec{V}$ (en basse-fréquence *et* position centrée) pour le VBAP ou bien par le vecteur énergie $\vec{E}$ (en haute-fréquence *ou* position excentrée) pour le VBIP. En conséquence, la *qualité* de la restitution – r_V ou r_E – fluctue en fonction de la direction apparente. De surcroît, le fait que $\vec{V}$ et $\vec{E}$ ne sont en général pas colinéaires signifie que la localisation basse-fréquence est peu stable par déplacement de l'auditeur [Ger92d].

En présence de trois ou quatre haut-parleurs frontaux, Gerzon [Ger92d] préconise donc des lois de pan-pot susceptibles de les exploiter simultanément. L'une des approches évoquées consiste à satisfaire des caractéristiques naturelles du front d'onde localement reconstruit. A l'aide de trois haut-parleurs, il est en effet possible de produire un front d'onde de direction apparente quelconque ($\vec{u}_V = \vec{u}_S$) et de vitesse apparente naturelle ($r_V = 1$) afin de satisfaire les mécanismes de localisation basse-fréquence, ce qui se résume dans la condition sur le vecteur vélocité: $\vec{V} = \vec{u}_V = \vec{u}_S$. Malheureusement, la loi de pan-pot qui vérifie cette condition produit d'importantes distorsions de la direction du vecteur énergie $\vec{E}$ par rapport à la direction attendue $\vec{u}_S = \vec{u}_V$, qui se traduisent par une contradiction entre les effets de localisation haute-fréquence et basse-fréquence, ainsi qu'une instabilité basse-fréquence par déplacement de l'auditeur. Les images sonores tendent à se condenser exagérément autour des haut-parleurs. Aussi Gerzon préfère-t-il réaliser un compromis, et définit une loi de pan-pot qui garantit des vecteurs vélocité et énergie de même direction $\vec{u}_V = \vec{u}_E = \vec{u}_S$, mais ne satisfait plus exactement la condition $r_V = 1$. Cette loi assure: une meilleure consistance des indices de localisation sur toute la bande de fréquence (et moins de fatigue d'après Gerzon); une meilleure homogénéité de l'image sonore en fonction de sa direction (évolution continue de sa qualité, moins de risque d'émergence des haut-parleurs); une meilleure stabilité aux déplacements de l'auditeur. Par ailleurs, l'énergie totale investie varie peu en fonction de la direction de la source virtuelle.

Les lois de pan-pot de Gerzon ont été proposées pour trois ou quatre haut-parleurs frontaux fixes. Il pourrait être envisagé d'en appliquer le principe à une configuration panoramique plus fournie, en sélectionnant pour une source virtuelle donnée les trois ou quatre haut-parleurs les plus voisins. Le schéma du *pair-wise pan-pot* précédemment évoqué (VBAP ou VBIP) serait ainsi étendu en "pan-pot par triplet ou quadruplet de haut-parleurs", avec l'avantage d'éviter l'effet d'individualisation – ou matérialisation – sporadique des haut-parleurs, et de favoriser l'homogénéité de la restitution.

Une autre stratégie consiste à utiliser l'ensemble des haut-parleurs présents pour optimiser la reconstruction d'une onde plane au centre du dispositif (*sweet-spot*). Poletti [Pol96a] définit ainsi des lois de pan-pot *"optimales" au sens d'une reconstruction locale*, introduites sous le nom de "fonctions *asinc*" (*angular sinus cardinal*) lorsque la configuration de haut-parleurs est régulière. Cette approche a la particularité de ne faire participer qu'un seul haut-parleur pour une source virtuelle placée dans sa direction, et tous les haut-parleurs dans les autres cas. L'effet de localisation basse-fréquence en position centrée est bien assuré, et jusqu'à une fréquence qui croît avec la densité angulaire des haut-parleurs. Mais cela s'accompagne de fluctuations de qualité du vecteur énergie $\vec{E}$ – distorsion directionnelle et variations de r_E en fonction de l'azimut – qui dénotent un effet de localisation haute-fréquence peu satisfaisant (moins performant qu'avec le VBIP ou le VBAP), des problèmes de stabilité (sensibilité du *sweet-spot*) et un défaut d'homogénéité de la restitution. Cette approche aura l'occasion d'être abordée plus en détail au chapitre 3, en tant que cas très particulier de restitution ambisonique d'ordre supérieur à 1 [Pol96a] [DRP98]. Précisons dès à présent que ce type de restitution, que nous qualifierons de "minimale" ou "super-minimale", ne respecte pas la "philosophie ambisonique" qui est intimement attachée à la question d'homogénéité de la restitution, telle qu'elle se manifeste à travers les vecteurs vélocité et énergie.

Enfin, nous verrons que l'approche ambisonique abordée plus loin, en 2.4 pour l'ordre 1 et au chapitre 3

pour les ordres supérieurs, peut se présenter comme une stratégie de pan-pot à part entière qui offre un compromis optimal entre qualité et homogénéité des images sonores pour une position centrée. En contrepartie, elle est soumise à un problème de *sweet-spot* souvent plus critique, et la définition du pan-pot ambisonique pour les configurations non-régulières de haut-parleurs se montre plus problématique qu'avec le VBAP [Pul97].

2.3.3 Techniques de prise de son

Comme pour la stéréophonie conventionnelle sur deux haut-parleurs, on peut distinguer deux grandes formes de prise de son panoramique ou multi-canal: l'une basée sur l'usage de microphones coïncidents (approche ΔI) et l'autre utilisant des microphones espacés (ΔI avec ΔT), le développement actuel de ces techniques s'orientant plus spécifiquement vers une restitution de type 3-2 (Figure 2.6). Elles ont pour objectif de restituer *l'ensemble du champ sonore original* autour de l'auditeur, ce qui signifie[12]: pouvoir reproduire l'effet de présence localisée des objets sonores autour de l'auditeur; mais aussi restaurer les qualités spatiales du champ – notamment les composantes latérales de l'effet de salle (réflexions et réverbération) – pour un effet d'enveloppement et des impressions spatiales qui enrichissent le plaisir de l'expérience auditive. Etant donné la culture d'une scène sonore frontale, les limitations du dispositif 3-2, ainsi que celles de la prise de son, les haut-parleurs latéraux-arrière ont surtout pour fonction de produire les impressions spatiales et diffuser des sons d'ambiance, plutôt que d'assurer la création d'images fantômes latérales stables.

ΔT *versus* ΔI

Les prises de son de type ΔT sont généralement telles qu'à chaque canal (dédié à un haut-parleur) est associé un microphone. Un schéma très courant consiste en un triplet de microphones frontaux – sorte de couple ORTF augmenté d'un microphone central – destiné aux haut-parleurs frontaux et un couple arrière – de type AB ou ORTF – placé en retrait et souvent dirigé vers l'arrière (Figure 2.7). Ce retrait du couple arrière introduit un ΔT entre haut-parleurs avant et arrière qui permet de faire jouer l'effet d'antériorité pour la séparation des demi-espaces avant et arrière[13]. Le choix des distances entre les microphones ainsi que de leur directivité et orientation relève d'*approches essentiellement empiriques*, dont on retient quelques principes:

- Considérant que chaque paire de microphones adjacents couvre un certain secteur angulaire en fonction de l'écart et l'angle relatif des microphones, la couverture angulaire du panorama sonore doit être complète (360°), en évitant les recouvrements et trous entre les angles de couverture des différentes paires[14].
- Les distances et angles des microphones peuvent être optimisés pour satisfaire des lois de pan-pot $\Delta I + \Delta T$, définies au préalable par des expériences d'écoute, pour chaque paire de haut-parleurs séparément[15]. Comme en réalité les haut-parleurs participent tous simultanément et non paire par paire, une seconde phase d'optimisation du système – dans son ensemble, cette fois – est réalisée à partir de tests d'écoute (correction des distorsions angulaires).
- Le retrait D_{FB} du couple arrière doit être moins important s'il doit participer à la création d'images fantômes latérales, que s'il s'agit de fournir des sons d'ambiances et une diffusion spatiale décorrélée de la scène frontale.

12. Etant donné les limites matérielles (nombre restreint de haut-parleurs), ce n'est pas une reproduction fidèle – au sens physique – du champ acoustique original qui est attendue.

13. Le ΔT est parfois introduit comme un retard électronique fixe, qui ne joue qu'en faveur de la stabilité des sources frontales.

14. Mike Williams, Workshop sur la *Comparaison de systèmes de prise de son multi-canal*, AES Paris 2000. Voir aussi les preprint 4997 et 5157.

15. Nicolas Jacques et Arnaud Mora, même Workshop.

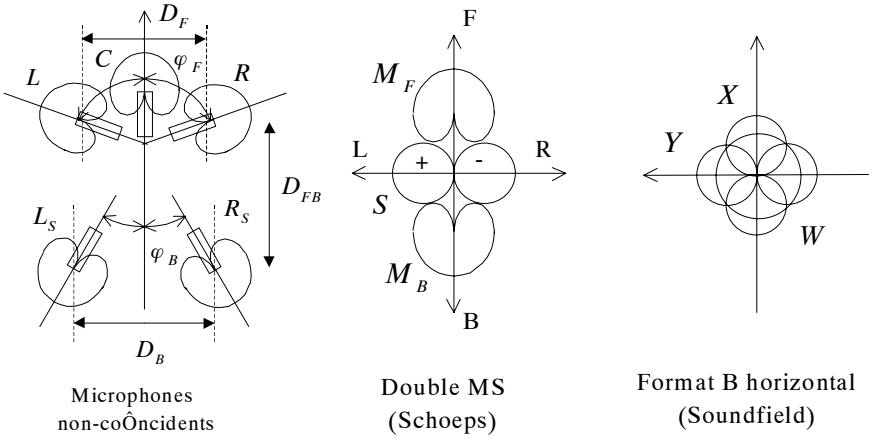

FIG. 2.7 – *A gauche: schéma typique de prise de son multi-canal utilisant des microphones non-coïncidents. Au centre et à droite: exemples de prise de son avec trois microphones coïncidents ou quasi-coïncidents, dont la combinaison permet en théorie de synthétiser des directivités quelconques de type* $a + b\cos(\theta - \theta_0)$.

– L'espacement D_B du couple arrière est essentiel pour parvenir à une bonne décorrélation des signaux L_S et R_S, condition d'une impression spatiale satisfaisante.

Les approches ΔI avec microphones coïncidents présentent quelques différences notables. Il n'est déjà pas possible, en pratique, de faire coïncider rigoureusement cinq microphones. Ceux-ci capteraient de toutes façons une information spatiale redondante, étant donnée leur directivité limitée à l'ordre 1 (en $a + b\cos(\theta - \theta_0)$). Trois microphones suffisent donc en théorie à l'encodage sur les cinq canaux, par combinaison de leurs signaux de sortie. Le système double-MS (Figure 2.7) s'inspire ainsi du couple MS évoqué en 2.2 pour la stéréophonie sur deux haut-parleurs, pour produire deux ou trois canaux frontaux ($C = M_F$, $L = M_F + \alpha S$, $R = M_F - \alpha S$) avec contrôle possible de la largeur de scène, et deux canaux arrière ($L_S = M_B + \beta S$, $R_S = M_B - \beta S$). Le microphone *Soundfield*, dont le principe est détaillé en 2.4.2, délivre quant à lui les composantes omnidirectionnelle W et bidirectionnelles X et Y (en $\cos\theta$ et en $\sin\theta$) du champ (Figure 2.7), avec l'avantage de résoudre les problèmes de coïncidence que peut connaître par exemple le double-MS, ou bien l'assemblage de capsules omni- et bi-directionnelles. Bien qu'il permette de synthétiser des microphones équivalents de directivité quelconque – et pourquoi pas le double-MS –, son utilisation est surtout associée à la technique de reproduction ambisonique, à laquelle nous consacrons la section suivante 2.4.

Un point fort des techniques ΔI est d'assurer un effet de localisation assez robuste – au moins pour les images frontales – et prédictible[16] par les vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$, alors qu'avec les prises de son avec ΔT, l'effet de localisation est plus souvent approximatif et peu stable en ce qui concerne les images latérales, notamment lors de rotations de la tête. Le point faible des prises de son avec microphones coïncidents est lié à la relativement faible séparation des canaux entre eux, due aux directivités limitées des microphones existants: c'est surtout le manque de *séparation latérale* du champ réverbéré diffusé par les haut-parleurs latéraux-arrière qui est regrettable, parce qu'il appauvrit l'impression spatiale et l'enveloppement. Contraintes aux mêmes limites de directivité des microphones, les prises de son avec microphones espacés (donc avec ΔT) permettent de rétablir une décorrélation satisfaisante de la partie latérale du champ réverbéré entre les canaux gauches et droits (frontaux et surtout arrière). C'est la raison pour laquelle les approches "ΔT" semblent plus prisées que les approches ΔI dans un contexte de prise de son, au contraire

16. Nous verrons qu'*en ce sens*, l'approche ambisonique se présente comme une technique optimale, de par son aptitude à respecter les propriétés directionnelles originales de la scène sonore.

des techniques de pan-pot homologues. Mais quelle que soit l'approche de prise de son, les images sonores – surtout latérales – sont généralement peu robustes aux déplacements de l'auditeur hors du centre du dispositif (*sweet-spot*), du fait que chaque source est captée par tous les microphones donc diffusée sur tous les haut-parleurs, problème qui ne se rencontre pas avec les pan-pots par paire – ou triplet – de haut-parleurs évoqués plus haut.

Conclusion et perspectives

La résolution des problèmes ou compromis énoncés passe par un certain nombre de mesures, de nature plus ou moins prospective.

Le dispositif 3-2 rendant difficile, sinon utopique, la quête d'une restitution panoramique (sur 360°) stable, il doit pouvoir être envisagé d'enrichir et de rééquilibrer la configuration de haut-parleurs. Une fois réalisées, les prises de sons ΔT sont dédiées à un dispositif particulier et ne se prêtent guère à des procédés de mixage (vers plus de voies), susceptibles de dégrader la qualité sonore (coloration par filtrage en peigne) et spatiale. L'approche ambisonique (section suivante) se montre au contraire tout à fait disposée à une restitution "à géométrie variable", tout en ne nécessitant la transmission que de trois canaux (W, X, Y) contre cinq (C, L, R, L_S, R_S).

Des progrès notables semblent promis aux moyens de prise de son dans un futur proche, avec l'avènement attendu de microphones de directivités d'ordres supérieurs[17]. De telles directivités permettraient une meilleure sélectivité angulaire des haut-parleurs participant à la création d'images fantômes, avec pour conséquences: une moindre sensibilité au *sweet-spot*; une meilleure séparation latérale du champ réverbéré, donc des impressions spatiales mieux préservées; dès lors, il serait moins nécessaire d'avoir recours aux procédés ΔT, ce qui signifierait une plus grande stabilité et prédictibilité des images sonores, ainsi qu'une plus grande portabilité de l'enregistrement vers d'autres dispositifs. Nous verrons que cette progression est intimement liée aux développements de l'approche Ambisonique aux ordres supérieurs, qui font l'objet de la partie II de ce document.

2.3.4 Compatibilité stéréo deux canaux: *surround matrix systems*

Le standard de transmission et stockage stéréophonique sur deux canaux a été, et est encore un élément fortement présent dans le domaine de la production et de la restitution *surround*, qu'il s'agisse de production cinématographique ou musicale. Depuis les années 70, des procédés de codage et de décodage ont été affinés, avec pour ambition de véhiculer sur seulement deux canaux l'information "*surround*" en plus de la scène frontale habituelle. Dans le cas présent d'un champ sonore panoramique, il semble évident que l'information spatiale, originellement encodée sur quatre ou cinq canaux distincts, se trouve dégradée par le mélange sur les deux canaux intermédiaires. La description succincte des principes de codage/décodage qui suit, montre de quelle manière et dans quelles limites une différenciation des informations spatiales originales est possible. Elle en indique les conséquences sur les contraintes de production, et signale la structure très particulière de l'espace sonore qui est imposée.

Codeurs et décodeurs *surround*: description succincte

Les systèmes de codage/décodage *surround* via deux canaux (dits *surround matrix systems*) consistent très basiquement en des opérations de matriçage. Les premiers systèmes (systèmes 4-2-4, dont le *Dolby Surround*) ne considèrent que quatre canaux originaux: C (centre), L (gauche), R (droit) et un seul canal

17. Directivités de type $g_0 + g_1 \cos\theta + g_2 \cos(2\theta)$ par exemple.

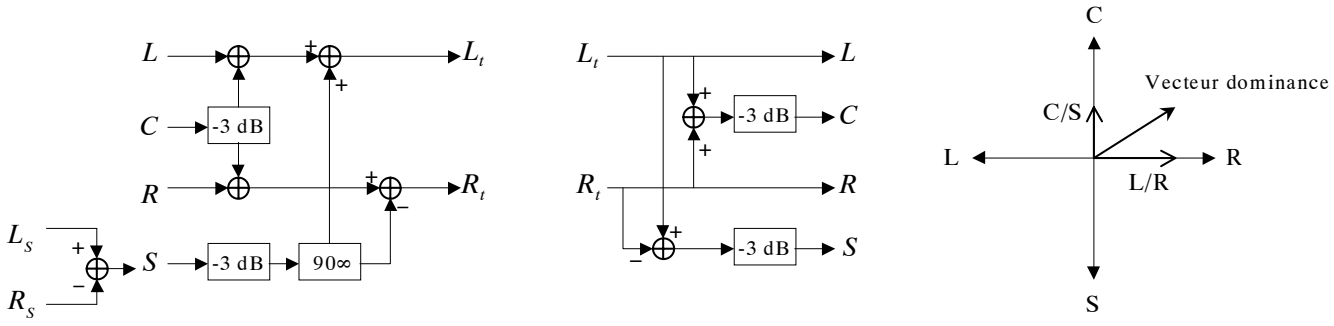

FIG. 2.8 – *Codage et décodage passif 4-2-4 (Dolby Surround simplifié), avec variante d'encodage pour deux canaux surround L_S et R_S. Détection de la direction prédominante à un instant donné (à droite): le "vecteur dominance" a pour coordonnées les rapports d'énergie L/R et C/S en dB.*

surround S généralement diffusé sur les haut-parleurs arrière en opposition de phase. En notant L_t et R_t les canaux stéréo transmis, le codage *Dolby Surround* (Figure 2.8) s'écrit de manière simplifiée:

$$\begin{cases} L_t &= L + \frac{1}{\sqrt{2}}C + j\frac{1}{\sqrt{2}}S \\ R_t &= R + \frac{1}{\sqrt{2}}C - j\frac{1}{\sqrt{2}}S \end{cases} \tag{2.12}$$

Le facteur j désigne un déphasage[18] de 90°: la mise en quadrature les signaux qui seraient présents à la fois dans S et dans L ou R avec une simple différence d'amplitude et/ou de signe, permet d'éviter de les annuler mutuellement et en préserve l'énergie totale. Par extension, l'encodage passif des cinq canaux C, L, R, L_S et R_S pour *Dolby ProLogic* utilise ce même schéma en posant $S = L_S - R_S$ (Figure 2.8). Les décodeurs de la première génération consistent en un simple *matriçage passif* (David Haffler, Dynaco):

$$\begin{pmatrix} L_o \\ R_o \\ C_o \\ S_o \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} L_t \\ R_t \end{pmatrix} \tag{2.13}$$

Si ce système purement passif est capable d'assurer une séparation complète (pas de mélange d'information) entre les canaux "diamétralement opposés" (entre C et S, ou L et R), il n'offre qu'une piètre séparation de 3 dB (effet de diaphonie) entre les canaux adjacents (entre L ou R et C ou S), qui se traduit par une dégradation de la qualité spatiale de la restitution en termes de localisation et d'impression spatiale. Il faut noter que ce schéma de codage/décodage (Equations 2.12 et 2.13 ou Figure 2.8) est associé de façon implicite à une organisation topographique très particulière de l'espace sonore: séparation entre une scène frontale (objets sonores précis) et un plan sonore *surround* (ambiance et fonction d'enveloppement), et séparation latérale pour la localisation frontale et l'enveloppement arrière. Les deux axes avant-arrière (somme-différence) et gauche-droite ne sont pas interchangeables et le premier n'est pas inversible, notamment à cause du déphasage de 90°.

Pour rétablir une séparation convenable entre les canaux et renforcer les effets directionnels à tout moment, il apparaît nécessaire de modifier les coefficients de la matrice en fonction du contenu des signaux L_t et R_t, ou plus précisément en fonction de l'information directionnelle suggérée par les signaux issus du décodage passif (2.13). Il s'agit donc d'un **décodage actif** mettant en jeu une matrice adaptative (*steering logic*),

18. En réalité, le codage *Dolby Surround* traite préalablement le signal *surround S* (filtrage passe-bande 100 Hz-7 kHz et réduction de bruit *Dolby B*).

principe que l'on retrouve avec quelques variantes dans les systèmes actuels (Dolby ProLogic [Dre93], Circle Surround [Wal96], Lexicon Logic7 [Gri96a]). Le renforcement directionnel se base sur la détection d'une direction prédominante à tout instant, définie d'après les différences de niveaux sonores (en dB) entre L_o et R_o, et entre C_o et S_o (Figure 2.8). Dans le cas d'une détection frontale par exemple, le signal C_o est retranché aux signaux L_o et R_o afin qu'il soit diffusé uniquement par le haut-parleur central ("procédé d'annulation" [Dre93]).

Le danger associé au décodage actif est la fluctuation intempestive des qualités sonores et spatiales: la dégradation et l'instabilité des "arrière-plans sonores" (dispersion, déplacement, voire renversement des sources jugées secondaires) au profit d'une source jugée prépondérante à un instant donné, les variations du niveau sonore ("effets de pompe")... autant de manifestations qui se révèlent rapidement gênantes et fatigantes en dépit d'un effet de focalisation espéré sur l'événement prépondérant. Pour définir une action plus pertinente, les décodeurs doivent tenir compte de paramètres dynamiques supplémentaires – degré de prédominance, niveau sonore et rapidité de leur variation – en fonction desquels différents modes d'action sont déclenchés [Dre93]. Cela ne suffit pas en général à assurer une restitution fiable et cohérente par rapport au matériel *surround* original. Il semble capital, particulièrement avec le *Dolby ProLogic*, que le travail de production (mixage) soit réalisé en fonction de ce qui ressort de la chaîne de codage-décodage, et que les pan-pot mis en jeu soient de type ΔI (pas de ΔT)[19]. Le principe de *codage actif* des systèmes 5-2-5, comme *Circle Surround* et ceux de *Lexicon*, permet de mieux prévenir contre les aberrations directionnelles et en même temps d'améliorer la séparation latérale, en particulier entre les canaux L_S et R_S.

Dans un contexte de diffusion musicale – aux exigences esthétiques plus contraignantes que pour le cinéma – Griesinger [Gri97] énonce quelques critères essentiels pour un système *surround* performant: la préservation de l'équilibre énergétique des sources (niveaux sonores relatifs); le respect de l'effet de localisation original, objectif pas toujours réalisable; une diffusion spatiale maximale du champ sonore d'arrière-plan (dont le champ réverbéré), qui exige une bonne décorrélation latérale des canaux. En vérifiant ces critères, un décodage *surround* peut aussi bien s'appliquer à un enregistrement stéréo conventionnel pour en améliorer de façon très appréciable la qualité de restitution en termes d'impression spatiale et d'enveloppement, tout en respectant les propriétés de perspective de la scène frontale.

Conclusion

Dans un contexte de diffusion cinématographique ou musicale, les systèmes de codage/décodage évoqués apportent des solutions appropriées à la contrainte d'une représentation intermédiaire de l'espace sonore sur deux canaux. Il est toutefois recommandé que le travail de production (mixage) soit réalisé en fonction du résultat de la chaîne d'encodage/décodage. Mais un tel mode de représentation s'accompagne de limitations qui rendent son exploitation incertaine dans un contexte plus interactif de navigation 3D. L'information directionnelle est en effet peu explicite et son traitement devient peu fiable dès que plusieurs événements sonores surviennent simultanément. Ce mode de représentation – et de codage/décodage associés – n'est donc pas adapté à la transmission de scènes sonores de composition arbitraire, dès lors qu'elles n'obéissent pas à la structure "*scène frontale*/*surround*" de type cinéma ou diffusion musicale, et que l'étape d'encodage ne peut être supervisée.

19. Au cours de démonstrations réalisées au CCETT, le codage-décodage ProLogic appliqué à un enregistrement multi-canal (prise de son musicale de Radio-France) a donné des résultats catastrophiques.

2.4 *Ambisonics*

Les analyses de la restitution qui ont ponctué les sections précédentes ainsi que l'étude plus générale présentée en 1.5, font ressortir un enseignement majeur: la reproduction d'images sonores naturelles au regard des mécanismes de localisation basse-fréquence – comprenant les rotations de la tête – passe par la synthèse d'un front d'onde local ayant les caractéristiques de propagation d'une onde plane, quelle que soit la direction apparente voulue. L'approche ambisonique, apparue dans les années 70 sous l'impulsion de Michael Gerzon, a été développée autour de cette spécification minimale de la propagation, à savoir la donnée du champ de pression et de son gradient en un point ("point de vue" de l'auditeur). Cette représentation locale du champ acoustique, qui peut être vue comme la restriction au premier ordre d'une décomposition du champ en harmoniques sphériques, définit le format ambisonique original (B-format): une composante omnidirectionnelle W et trois composantes bidirectionnelles X, Y, Z. La prise en compte de composantes d'ordres supérieurs, qui permet d'accroître la résolution spatiale de la représentation, laisse envisager la définition de systèmes ambisoniques d'ordres supérieurs [BV95] [DRP98].

Ce sont d'abord les systèmes ambisoniques traditionnels (d'ordre 1) qui sont traités dans cette section. La description de l'étape d'encodage, de la prise de son et des formats dérivés est suivi d'un rappel des principes de décodage d'après Gerzon. Une analyse de la restitution est alors complétée par une interprétation approfondie des critères objectifs utilisés pour la caractérisation de chaque image sonore (vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$), mais aussi pour une caractérisation plus globale des qualités spatiales restituées et des effets du *cross-talk* (indice $r_E = |\vec{E}|$). Cette étude dessine ainsi les préoccupations et précise les outils qui présideront à la généralisation des systèmes ambisoniques aux ordres supérieurs (Partie II).

2.4.1 Un encodage directionnel: définition et avantages

Une spécificité majeure de l'approche ambisonique, par rapport aux techniques précédemment évoquées, repose sur le fait que les canaux transmis contiennent de façon explicite l'*information directionnelle* des images sonores dans la scène sonore à reproduire, *indépendamment du dispositif de haut-parleurs employé.*

Les canaux ambisoniques[20] résultent d'un encodage directionnel du champ acoustique en un point[21], qui consiste à mesurer, en plus de la pression p (composante omnidirectionnelle W), les composantes de la vélocité particulaire $\vec{v}$ – ou encore du gradient de pression – suivant trois directions orthogonales $\vec{x}$, $\vec{y}$, $\vec{z}$ (composantes bidirectionnelles X, Y, Z). Mathématiquement, l'encodage directionnel d'une source sonore S dans la direction du vecteur unitaire $\vec{u}$ – ou plutôt une onde plane d'incidence $\vec{u}$ portant un signal S – se traduit par les équations suivantes [Ger73][22]:

$$\left\{\begin{array}{lll} W & = S & \\ X & = \sqrt{2}\,\vec{u}.\vec{x}\,S & = \sqrt{2}\,\cos\theta\cos\delta\,S \\ Y & = \sqrt{2}\,\vec{u}.\vec{y}\,S & = \sqrt{2}\,\sin\theta\cos\delta\,S \\ Z & = \sqrt{2}\,\vec{u}.\vec{z}\,S & = \sqrt{2}\,\sin\delta\,S \end{array}\right. , \tag{2.14}$$

où le vecteur incidence $\vec{u}$ est décrit par (θ,δ) en coordonnées sphériques (Figure 1.4). Dans le cas d'un champ complexe (plusieurs ondes planes), les canaux encodés résultent de la somme des différentes contributions. Les quatre canaux W, X, Y et Z constituent le *B-format* (ou format B). Il est fréquent, dans le cadre d'une

20. Nous parlerons encore de "représentation ambisonique".

21. Point de "vue" qui sera proposé à l'auditeur au moment de la restitution.

22. Nous avons ajouté l'expression de l'encodage sous forme de projection vectorielle, parce qu'elle nous semble commode et synthétique, bien qu'elle ne soit pas très fréquemment utilisée dans la littérature.

restitution purement horizontale, de ne retenir que les trois premiers canaux W, X et Y, et d'ignorer la composante verticale Z. Il est alors d'usage de conserver le nom de format B pour cette version restreinte (2D). Le facteur de normalisation $\sqrt{2}$ a été introduit, à l'origine, pour assurer des canaux (W, X, Y) de puissances moyennes équivalentes dans le cas de sources sonores *horizontales* réparties uniformément dans toutes les directions, ou encore d'un champ diffus "horizontal". Mentionnons une légère variante de ces conventions d'encodage: celle de Malham [Mal92] qui propose une correction de l'ensemble des composantes par un facteur $1/\sqrt{2}$.

Outre l'avantage de fournir, grâce à la composante verticale Z, une description de la scène sonore en *trois dimensions*, pouvant donner lieu à une restitution "*périphonique*", ce format a la propriété de proposer une *représentation homogène* des événements sonores, c'est-à-dire qui ne privilégie pas une direction plus qu'une autre. Enfin, cette représentation se prête très naturellement à des manipulations du champ, et en particulier des rotations (cf 3.1.5).

2.4.2 Compléments techniques: prise de son et formats dérivés

Prise de son ou encodage acoustique

Dans le principe, réaliser une prise de son ambisonique revient à mesurer à la fois et en un même point le signal de pression et les composantes de son gradient (ou encore de la vitesse ou vélocité particulaire) suivant trois axes orthogonaux $\vec{x}$, $\vec{y}$ et $\vec{z}$. Il existe évidemment des microphones adéquats, pour accéder individuellement à ces différentes mesures: microphone omnidirectionnel (capteur de pression), microphones bidirectionnels (capteurs de vélocité), dits "figure en 8" à cause de leur diagramme de directivité. En admettant qu'ils soient de qualité acceptable (une bonne approximation des directivités idéales n'est pas toujours acquise), il est en revanche problématique de faire coïncider ces quatre capteurs en un même point, et d'éviter d'autre part une perturbation mutuelle de la mesure du champ du fait de la non-transparence acoustique des instruments de mesure. En pratique, il semble que cette approche soit réduite à la prise de son 2D (horizontale), qui ne nécessite que trois capteurs, et assez couramment appliquée dans un contexte de production musicale par exemple (*Nimbus Records*...? pour la production du format UHJ). Quelques exemples d'assemblage ont été cités sur la liste de discussion *sursound* [Sur].

Une autre méthode a été développée par Craven et Gerzon [CG77] pour accéder à une mesure homogène du champ ambisonique 3D, donc des quatre composantes W, X, Y, Z du B-format. De façon analogue à une approximation d'un gradient de pression par la juxtaposition de deux capteurs de pression, mais avec plus de robustesse et de subtilité(!), elle repose sur les mesures de quatre capsules cardioïdes peu espacées, placées aux centres des faces d'un tétraèdre régulier, et dirigées vers l'extérieur (Figure 2.9). La réalisation d'un tel appareil est connue sous le nom de **microphone *SoundField*** [Sou]. Les quatre signaux mesurés – nommés *LF*, *LB*, *RB* et *RF* – sont dits constituer le *format A*. En adoptant l'hypothèse de quasi-coïncidence des capsules – hypothèse acceptable à l'échelle des grandes longueurs d'ondes –, il est assez facile de recomposer les quatre directivités associées à W, X, Y, et Z par combinaisons linéaires[23] des directivités cardioïdes – indépendantes de par leurs orientations. En effet, on n'a affaire qu'à des fonctions de directivités d'ordre 1 ou 0 (fonctions affines en quelque sorte), de la forme générale: $\mathcal{G}(\vec{u}) = \alpha + \beta\,\vec{u}.\vec{u}_{mic} = \alpha + \beta\cos\theta$, pour une incidence $\vec{u}$ formant un angle θ avec l'orientation $\vec{u}_{mic}$ du microphone. On déduit donc les composantes du

23. Réciproquement, on a déjà souligné que l'on pouvait imiter n'importe quelle forme de prise de son (d'ordre 1) par combinaison des composantes ambisoniques W, X, Y, Z.

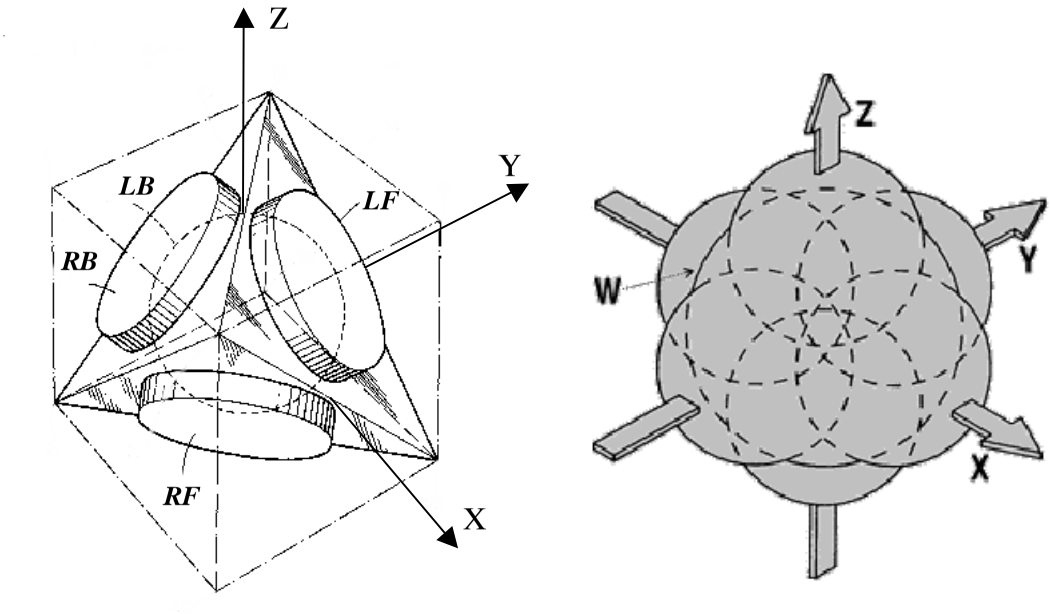

FIG. 2.9 – *A gauche: Schéma du microphone* Soundfield *(d'après [CG77]): quatre capsules cardioïdes placées sur les faces d'un tétraèdre régulier, et captant des signaux LF (pour* Left-Front*), RF (*Right-Front*), LB (*Left-Back*), RB (*Right-Back*), qui constituent le format A. A droite: Composantes directionnelles du B-format: W, X, Y, Z.*

format B de celles du *format A* d'après les relations:

$$\begin{cases} W &= LF+LB+RF+RB \\ X &= LF-LB+RF-RB \\ Y &= LF+LB-RF-RB \\ Z &= LF-LB-RF+RB \end{cases}, \tag{2.15}$$

à un facteur près entre W et X, Y, Z, selon les conventions d'encodage utilisée.

Considérant des fréquences croissantes (longueurs d'onde décroissantes), un écart de phase croissant apparaît dans l'arrivée d'une onde aux différentes capsules, du fait de leur non-coïncidence. Au-delà d'une certaine fréquence, les directivités de W, X, Y, Z ne peuvent pas être rigoureusement reconstituées. Basant leurs arguments sur [Str71], les inventeurs proposent une égalisation pour établir un meilleur rapport entre les composantes d'ordre 0 et 1 à mesure que la fréquence augmente. Les filtres proposés dans le brevet sont susceptibles d'être ou d'avoir été affinés [CG77]. Leur définition originale est la suivante, en notant $\mathcal{F}_W$ celui qui s'applique à W et $\mathcal{F}_X$ celui qui s'applique à X, Y et Z:

$$\begin{cases} \mathcal{F}_W &= \dfrac{1+j\omega\tau-\frac{1}{3}\omega^2\tau^2}{1+\frac{1}{3}j\omega\tau} \\ \mathcal{F}_X &= \sqrt{6}\dfrac{1+\frac{1}{3}j\omega\tau-\frac{1}{3}\omega^2\tau^2}{1+\frac{1}{3}j\omega\tau} \end{cases} \tag{2.16}$$

où $\tau = R/c$. Le facteur $\sqrt{6}$ corrige les relations 2.15 pour des raisons de conformité avec les conventions d'encodage de Gerzon (2.14).

Il semble que les prétentions générales du brevet couvrent l'application du principe – utilisation de capsules de type (hypo-, hyper-)cardioïde placées sur les faces d'un polyèdre régulier – à la synthèse de directivités d'ordres supérieurs à 1, donnant lieu à une représentation du champ – basée sur une décomposition en harmoniques sphériques – d'ordre supérieur au B-format. Tout en se fondant sur des considérations sur les intégrales multiples [Str71], aucun développement explicite n'y est donné. La conception d'un microphone ambisonique d'ordre 2 fait actuellement l'objet d'une thèse[24]. Nous avons nous-même dégagé et formalisé une approche générique pour la prise de son ambisonique aux ordres supérieurs (section 3.4), dans le cadre de l'extension des techniques ambisoniques aux ordres supérieurs (partie II).

Formats dérivés pour la diffusion compatible stéréo ou multi-canal

Au contraire du *A-format* "primitif", peu parlant et difficilement exploitable directement, le *B-format* et les équations d'encodage (2.14) associées trouvent une interprétation directe en terme d'information directionnelle ou spatiale. Mais ce sont des signaux de connotation et de fonctionnalité encore différentes qui sont habituellement transmis ou stockés, conformément aux standards stéréophonique (2 canaux) et désormais multi-canal (5.1). L'adaptation nécessaire des productions ambisoniques aux modes standard de diffusion (radio, disques, CD, et maintenant DVD, télévision et radio numériques) a suscité la création de formats dérivés, dont les plus anciens sont rassemblés sous la désignation *C-formats* [Ger85], et plus connus pour dériver du matriçage UHJ:

$$\begin{cases} \Sigma &= 0.9397W + 0.1856X \\ \Delta &= j(-0.3420W + 0.5099X) + 0.6555Y \\ T &= j(-0.1432W + 0.6512X) - 0.7071Y \\ Q &= 0.9772Z \end{cases}, \tag{2.17}$$

où le facteur imaginaire pur j correspond à un déphasage de $\pi/2$ (signal en quadrature, par transformation de Hilbert). L'opération de matriçage inverse, au besoin appliquée à l'autre bout de la chaîne de diffusion, s'écrit alors:

$$\begin{cases} W &= 0.982\Sigma + 0.197j(0.828\Delta + 0.768T) \\ X &= 0.419\Sigma - j(0.828\Delta + 0.768T) \\ Y &= 0.187j\Sigma + (0.796\Delta - 0.676T) \\ Z &= 1.023Q \end{cases} \tag{2.18}$$

Par analogie avec la technique M-S, les signaux gauche et droit retenus pour la restitution stéréo deux canaux sont $L = (\Sigma + \Delta)/2$ et $R = (\Sigma - \Delta)/2$, le choix des canaux Σ et Δ relevant d'un mode de diffusion compatible stéréo-mono. Les canaux L et R ainsi définis forment le *format BHJ*[25], le seul à avoir survécu commercialement, d'ailleurs sous le nom de *format UHJ* (apanage de la maison de disque *Nimbus Record*). A la façon des systèmes *surround* précédemment évoqués (Cf 2.3.4), ce format n'est pas seulement destiné à une diffusion directe sur deux haut-parleurs (mode dit "super-stéréo"), mais peut être décodé pour une restitution *surround* acceptable (*décodeur UHJ*), même en l'absence du canal T (2.18). Les formats plus complets sont: le *format THJ*, incorporant la troisième composante T pour une restitution horizontale complète; le *format PHJ*, qui grâce à l'ultime et quatrième canal Q, permet une restitution "périphonique" (3D) complète. La tentative d'exploitation de ces formats et matriçage dans un contexte de diffusion radiophonique (au Royaume-Uni) s'est soldée par un échec.

Dans un contexte de diffusion pour la Télévision Haute-Définition (TVHD), Gerzon a également proposé [Ger92c] une extension du format B horizontal ("*Enhanced B-format*": *BF-format* et *BEF-format*), en l'augmentant de canaux E et F que l'on pourrait qualifier de "sélectivement redondants", destinés à renforcer la

24. Thèse de Philip Cotterell, University of Reading, UK.

25. Pour simplifier, nous confondons la dénomination du matriçage et celle du format auquel il donne naissance.

stabilité des images frontales et la séparation avant/arrière des événements sonores. Cela s'applique donc à une production de studio et non à un prise de son ambisonique.

Malgré quelques pertes familiales, la saga des "*X-formats*" se poursuit aujourd'hui avec la proposition du *G-Format* [Ele98], destinée à favoriser la diffusion de la production ambisonique au sein de la production multi-canal dans sa globalité, assujettie au standard 5.1 pour support DVD. Il s'agit de produire les canaux discrets par un pré-décodage du format B, voire du format BEF, pour la configuration $3/2$ standard, en utilisant de préférence la nouvelle génération de décodeurs ambisoniques dédiée à ce type de configuration (*Vienna decoder* [Ger92a]). Cela a le double avantage de livrer l'*excellence du "son" ambisonique* sans la nécessité d'un décodeur ambisonique au niveau de l'utilisateur, et de se prêter à un décodage-inverse (ou ré-encodage) pour une restitution sur un autre dispositif si l'on dispose du système adéquat. Il est suggéré, pour pouvoir accéder à une restitution périphonique (3D), d'utiliser le canal supplémentaire LFE (*Low Frequency Effect*) pour transmettre des informations de hauteur (composante verticale). Les deux canaux supplémentaires LPCM du DVD (pour restitution restreinte à deux haut-parleurs) peuvent être dédiés à la transmission du format UHJ pour un restitution "super-stéréo".

2.4.3 Le décodage "psychoacoustique" selon Gerzon

Afin de restituer sur haut-parleurs la scène sonore encodée, il est nécessaire de procéder à un *décodage* du format B, qui prend la forme d'un *matriçage*, c'est-à-dire d'une combinaison linéaire des signaux W, X, Y (et Z) dont résultent les signaux qui alimenteront les haut-parleurs. Disposant des informations directionnelles du champ sonore mesurées en un point, on pourrait naïvement espérer offrir à l'auditeur, au moment de la restitution, les mêmes informations auditives spatiales (la même "perception") que s'il avait pris place dans le champ original encodé au lieu où a été effectuée la mesure. Malheureusement, la tête (de l'auditeur, avec ses oreilles) est un "instrument de mesure" volumineux qui perturbe fortement le champ sonore, et qui ne peut pas se suffire, du moins à l'échelle des petites longueurs d'onde (hautes fréquences), de la reproduction *locale* de caractéristiques *locales* du champ (que traduit le format B). A défaut d'une reproduction fidèle de l'expérience d'écoute et en restant pour l'instant dans des considérations très générales, le décodage doit avoir pour but d'*exploiter le mieux possible les informations directionnelles* contenues dans le format B afin de *produire* lors de la restitution *le meilleur effet subjectif de localisation* pour chaque source sonore originale. C'est en ce sens que la *théorie du décodage ambisonique* développée par Gerzon, que nous présentons maintenant, mérite le qualificatif de "*psychoacoustique*".

Formalisme

Posons tout d'abord le formalisme du décodage qui doit conduire à son optimisation "psychoacoustique", en considérant le cas d'une source unique (signal S) de direction $\vec{u}(\theta,\delta)$, encodée suivant les équations (2.14). Le dispositif de restitution est constitué de N haut-parleurs placés sur un cercle (pour une restitution horizontale) ou sur une sphère (pour une restitution 3D "périphonique"), dans les directions $\vec{u_i}(\theta_i,\delta_i)$, vus du centre du dispositif. Le décodage est pour l'instant considéré comme une simple opération de matriçage. En notant $\mathbf{D}$ la matrice de décodage à définir, et $\mathbf{S} = [S_1\ S_2\ \ldots\ S_N]^t$, les signaux S_i délivrés par les haut-parleurs sont tels que:

$$\mathbf{S} = \mathbf{D}.\begin{pmatrix} W \\ X \\ Y \\ Z \end{pmatrix} \tag{2.19}$$

Dans le cas présent d'une source encodée unique, le vecteur des composantes ambisoniques s'écrit comme le produit $[W,X,Y,Z]^t = \mathbf{B}_{\vec{u}}.S$, où $\mathbf{B}_{\vec{u}} = [1, \sqrt{2}\vec{u}.\vec{x}, \sqrt{2}\vec{u}.\vec{y}, \sqrt{2}\vec{u}.\vec{z}]^t$ est le *vecteur coefficient d'encodage* (au format B) associé à la direction d'incidence $\vec{u}$. On condensant les opérations d'encodage et de décodage, on écrit que les signaux S_i sont proportionnels au signal d'origine S: $S_i = G_i S$, de sorte que l'opération de décodage (2.19) se traduise sous la forme:

$$\mathbf{G} = \mathbf{D}.\mathbf{B}_{\vec{u}}, \tag{2.20}$$

où l'on a introduit le vecteur gains $\mathbf{G} = [G_1\ G_2\ \ldots\ G_N]^t$.

Vecteurs vélocité et énergie

L'optimisation du décodage telle que Gerzon la conçoit est *dédiée à un auditeur placé au centre du dispositif*. S'attachant à caractériser, puis optimiser la restitution pour cette place privilégiée, Gerzon introduit deux quantités mathématiques censées traduire l'effet de localisation dans des domaines basse- et haute fréquence respectivement. Il s'agit du *vecteur vélocité* $\vec{V}$:

$$\vec{V} = \frac{\sum_{i=1}^{N} G_i \vec{u}_i}{\sum_{i=1}^{N} G_i} = r_V\, \vec{u}_V, \tag{2.21}$$

et du *vecteur énergie* $\vec{E}$:

$$\vec{E} = \frac{\sum_{i=1}^{N} G_i^2 \vec{u}_i}{\sum_{i=1}^{N} G_i^2} = r_E\, \vec{u}_E \tag{2.22}$$

Parce que des justifications complètes semblaient faire défaut dans la littérature, une interprétation approfondie de ces grandeurs – qui décrivent avant tout des propriétés de propagation – en tant que prédicteurs de l'effet de localisation a été proposée dans la section 1.5. Pour simplifier, *et si l'auditeur est au centre du dispositif*, $\vec{V}$ et $\vec{E}$ sont associés respectivement aux effets de localisation en basse- et haute-fréquence.

On retient en particulier que plus r_V est faible et inférieur à 1, plus l'image (d'après le contenu basse-fréquence) est ramenée vers le plan médian (moindre latéralité) si la tête est fixe, ou est localisée avec un effet de hauteur exagérée par rotation *yaw* de la tête, ou est localisée de façon "floue" dans la direction $\vec{u}_V$ lorsque tous les mouvements de tête sont permis.

De même que pour le vecteur vélocité, le vecteur énergie peut prédire la direction apparente de la source virtuelle (d'après le contenu haute-fréquence) à l'aide des rotations de la tête, mais avec une d'autant moins bonne précision (image plus floue) que $r_E = |\vec{E}|$ est faible (moindre ITD et moindre ILD), le cas idéal $r_E = 1$ étant celui d'une onde plane unique. L'étude 1.5.3 laisse pressentir qu'à valeur r_E et r_V égales et inférieures à 1, la dégradation de l'image en terme de précision est plus importante au regard des mécanismes haute-fréquence qu'au regard des basses fréquences. Nous précisons plus tard le domaine d'application de cet "indice" dans le cadre du décodage ambisonique.

Critères "psychoacoustiques" et règles de décodage

Le bien fondé des vecteurs vélocité et énergie étant établi, Gerzon préconise très logiquement, pour un décodage optimal, que l'opération de matriçage (2.20) soit *différenciée selon les deux domaines fréquentiels* d'application de ces vecteurs. Ainsi, la matrice de décodage basse-fréquence $\mathbf{D}^{LF}$ doit être telle que pour tout vecteur d'encodage $\mathbf{B}_{\vec{u}}$, les gains G_i résultant de l'opération (2.20) "produisent" un vecteur vélocité $\vec{V}$ de même direction $\vec{u}_V$ que celle de la source encodée $\vec{u}$, et de module r_V le plus proche de 1, bref, idéalement: $\vec{V} = \vec{u}$. De façon analogue, la matrice de décodage haute-fréquence $\mathbf{D}^{HF}$ doit être telle que le vecteur énergie produit soit de direction $\vec{u}_E = \vec{u}$ et de module r_E le plus proche de 1. Une autre contrainte que Gerzon impose

comme règle supplémentaire de décodage pour les deux domaines fréquentiels, est la colinéarité du vecteur énergie $\vec{E}$ avec le vecteur vélocité $\vec{V}$: $\vec{u}_E = \vec{u}_V$. Cette condition s'interprète, dans le domaine basse-fréquence, comme un souci de "robustesse", ou plus précisément comme l'assurance d'une expansion radiale maximale du front d'onde local synthétisé. Nous aurons à plusieurs reprises au cours de ce document, l'occasion d'observer et de commenter ce lien entre l'expansion spatiale d'un front d'onde local et la similitude de la caractérisation locale de propagation $\vec{V}$ avec la caractérisation globale ou statistique de propagation $\vec{E}$.

Un autre aspect spécifié dans la définition du codage, est la *conservation globale* de l'énergie: l'énergie restituée $\sum G_i^2 S^2$ doit être égale à celle de la source encodée S^2, ce qui implique: $\sum G_i^2 = 1$, à supposer les ondes s'additionnent *globalement* "en quadrature" (avec des relations de phase aléatoires). Pour être plus juste, le gain d'énergie $\sum G_i^2$ n'est observable qu'en moyenne et non pas pour chaque fréquence en toute position: moyenne fréquentielle ou bien temporelle pour une position suffisamment excentrée, voire moyenne statistique sur l'ensemble des positions d'écoute. On peut noter une légère contradiction entre ce critère, censé être appliqué "pleine-bande", et le fait qu'on s'appuie, dans le domaine basse-fréquence, sur une sommation des ondes en amplitude pour recréer au centre un front d'onde locale satisfaisant.

Résolution du problème de décodage

La résolution des règles de décodage s'avère relativement facile dans le cas de **configurations de haut-parleurs vérifiant certaines propriétés géométriques de régularité**[26] (configurations carrées, rectangulaires, cubiques, parallélépipédiques, en polygones réguliers). Gerzon en donne les solutions pour quelques structures typiques [Ger92b]. Plutôt que se référer à chacune des ses démonstrations, nous rappelons ici une approche générique de résolution dans ces cas simples, mise en évidence dans [DRP98] (Annexe B) pour des systèmes 2D d'ordres 1 ou supérieurs, et développée de façon plus approfondie et plus générale encore dans la partie II de ce document.

Il est assez immédiat de constater que pour reproduire un front d'onde local de mêmes caractéristiques de propagation $\vec{V}$ que le front d'onde original, il suffit de préserver le rapport entre les composantes directionnelles X, Y, Z ("vitesse particulaire") et la composante W (pression). Cela revient, à un gain global g_{LF} près, à reconstruire l'ensemble de ces composantes au centre du dispositif. En notant $\mathbf{B}_{\{\vec{u}_i\}} = [\mathbf{B}_{\vec{u}_1} \ \ldots \ \mathbf{B}_{\vec{u}_N}]$ la matrice de "réencodage" associée aux directions des haut-parleurs, la reconstruction du champ ambisonique par les différentes contributions venant des haut-parleurs s'écrit:

$$\mathbf{B}_{\vec{u}} = \mathbf{B}_{\{\vec{u}_i\}}.\mathbf{G} \tag{2.23}$$

Une manière simple de déduire les gains G_i consiste alors à inverser le système (2.23), en mettant en jeu comme matrice de décodage la pseudo-inverse (au sens des moindres carrés) de $\mathbf{B}_{\{\vec{u}_i\}}$, qui a la propriété – parmi les autres solutions possibles – de minimiser globalement l'énergie restituée:

$$\mathbf{G} = \mathbf{B}_{\{\vec{u}_i\}}{}^t(\mathbf{B}_{\{\vec{u}_i\}}.\mathbf{B}_{\{\vec{u}_i\}}{}^t)^{-1}.\mathbf{B}_{\vec{u}} \quad \Rightarrow \quad \mathbf{D}^{LF} = g_{LF}\mathbf{D}_{pinv} \quad \text{avec} \quad \mathbf{D}_{pinv} = \mathbf{B}_{\{\vec{u}_i\}}{}^t(\mathbf{B}_{\{\vec{u}_i\}}.\mathbf{B}_{\{\vec{u}_i\}}{}^t)^{-1}, \tag{2.24}$$

où l'on suppose que le nombre N de haut-parleurs est supérieur au nombre de composantes ambisoniques. Dans les cas *géométriquement simples* qui nous intéressent ici, le produit $(\mathbf{B}_{\{\vec{u}_i\}}.\mathbf{B}_{\{\vec{u}_i\}}{}^t)$ est diagonal, ce qui rend facile la résolution du problème. De surcroît, on montre [DRP98] que ce produit vaut $N\mathbf{I}$ ($\mathbf{I}$ étant la

26. La propriété de régularité qui rend le décodage si commode sera discutée en terme d'échantillonnage, en 3.2.3, dans le cadre des systèmes ambisoniques aux ordres supérieurs 2D et 3D.

matrice identité) pour les configurations horizontales régulières (*polygones réguliers*), de sorte que:

$$\mathbf{D}_{pinv} = \frac{1}{N}\mathbf{B}_{\{\vec{u}_i\}}{}^t \quad \Rightarrow \quad \mathbf{D}^{LF} = \frac{g_{LF}}{N}\mathbf{B}_{\{\vec{u}_i\}}{}^t = \frac{g_{LF}}{N}\begin{bmatrix} 1 & \sqrt{2}\vec{u}_1.\vec{x} & \sqrt{2}\vec{u}_1.\vec{y} \\ 1 & \sqrt{2}\vec{u}_2.\vec{x} & \sqrt{2}\vec{u}_2.\vec{y} \\ \vdots & \vdots & \vdots \\ 1 & \sqrt{2}\vec{u}_N.\vec{x} & \sqrt{2}\vec{u}_N.\vec{y} \end{bmatrix} \tag{2.25}$$

Le décodage revient alors en substance à une opération de *projection* des composantes d'encodage $\mathbf{B}_{\vec{u}}$ sur les composantes de réencodage $\mathbf{B}_{\{\vec{u}_i\}}$. Les tenants mathématiques fondamentaux qui se cachent derrière cette propriété seront abordées dans le détail et dans un contexte plus général en 3.2.3. Pour obtenir une expression de la même forme dans le cas d'une restitution 3D avec des configurations polyédrales régulières, il serait nécessaire de remplacer le *facteur de normalisation* 2D $\sqrt{2}$ par un facteur de normalisation 3D $\sqrt{3}$ dans les conventions d'encodage (2.14).

Une seconde propriété accompagne la matrice de décodage $\mathbf{D}^{LF}$ (2.24) dans le cas de configurations (régulières et "semi-régulières") telles que le produit $\mathbf{B}_{\{\vec{u}_i\}}.\mathbf{B}_{\{\vec{u}_i\}}{}^t$ est diagonal [DRP98]: le vecteur énergie produit est automatiquement colinéaire à $\vec{u}$ et $\vec{V}$. De plus, il le reste si l'on modifie le gain relatif g_1/g_0 des composantes X, Y (et Z) par rapport à la composante W, en amont du décodage (2.20). On peut donc optimiser le module r_E en jouant sur le gain g_1/g_0, tout en vérifiant la contrainte $\vec{u}_E = \vec{u}_V = \vec{u}$. La matrice de décodage haute-fréquence prend alors la forme:

$$\mathbf{D}^{HF} = \mathbf{D}_{pinv}.\mathrm{Diag}([g_0\ g_1\ g_1\ (g_1)]^t) \tag{2.26}$$

Le rapport optimal g_1/g_0 qui maximise r_E, vaut $1/\sqrt{2}$ pour une restitution horizontale (2D), et $1/\sqrt{3}$ pour une restitution périphonique 3D. Notons au passage, dans ce cas optimisé, l'égalité $r_E = r_V = g_1/g_0$. Bien qu'à notre connaissance cela ne soit pas mentionné dans la littérature, le fait que $r_E^{3D} < r_E^{2D}$ suggère donc que la restitution ambisonique (à l'ordre 1) des sources – placées dans le plan horizontal – est moins précise avec un dispositif périphonique (3D) de haut-parleurs qu'avec un dispositif panoramique (horizontal)! Cette observation sera commentée dans un cadre plus général en partie II.

Le choix des gains absolus g_{LF} et g_0 relève du critère de préservation en énergie. Là encore, le lecteur est invité à consulter [DRP98] (Annexe B) et les développements de la partie II.

Le cas des **configurations non-régulières**, comme celles recommandées pour la diffusion multi-canal (Configuration $3/2$, Figure 2.6), est autrement plus délicat à résoudre. La vérification des règles et l'optimisation du décodage exigent l'intervention de méthodes d'optimisation non-linéaire [Ger92a] [Tre97], et il n'existe pas à ce jour de solution générique adaptée à toutes les configurations. D'ailleurs, plusieurs matrices de décodage peuvent satisfaire les conditions haute-fréquence. En ajoutant une contrainte supplémentaire (fixer le gain d'énergie dans une direction donnée), Gerzon explicite des solutions pour quelques configurations dans [Ger92a]. Des travaux menés au CCETT [Tre97] ont donné lieu à des solutions identiques pour le décodage basse-fréquence (avec $r_V = 1$), et différentes, bien qu'aussi convenables, pour le décodage haute-fréquence. Dans ce cas de figure, la qualité de restitution "haute-fréquence" (r_E) et l'énergie totale restituée varient selon la direction de la source virtuelle encodée: la précision r_E est meilleure là où les haut-parleurs sont plus rapprochés (typiquement les haut-parleurs frontaux), et l'énergie suit la tendance inverse (Figure 3.17, chapitre suivant).

Fréquence de transition

La fréquence de transition entre l'application de la matrice basse-fréquence et de la matrice haute-fréquence doit être définie selon les domaines de validité des critères $\vec{V}$ et $\vec{E}$. La prédiction d'après $\vec{E}$ s'applique dans le domaine haute-fréquence qui commence à la fréquence au-delà de laquelle le front d'onde

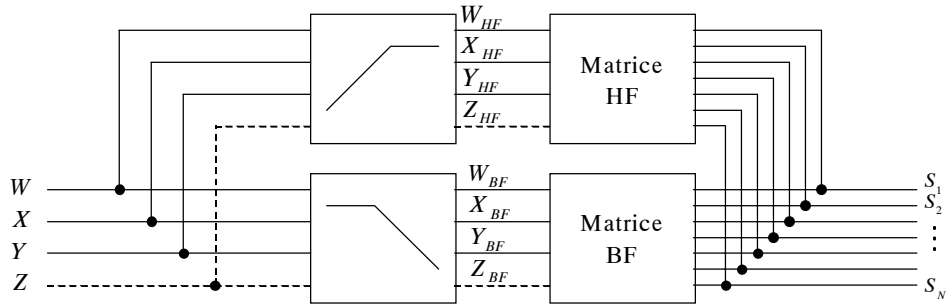

FIG. 2.10 – *Structure générale d'un décodeur ambisonique en deux sous-bandes, adapté aux configurations non-régulières.*

synthétique local est "plus petit" que la tête, c'est-à-dire là où s'arrête la validité de la prédiction par $\vec{V}$. Dans le cas des systèmes ambisoniques du premier ordre comme ici, cette fréquence de transition est de l'ordre de 700 Hz. Il paraît important à ce sujet, de *dissiper une confusion* assez fréquente qui semble provenir du principe de transition fréquentielle – propre à la restitution et de surcroît à une position d'écoute centrée – entre un mode de localisation (retard de phase, basse-fréquence) qui peut être complètement prédit par $\vec{V}$, et un mode "haute-fréquence" auquel $\vec{E}$ s'applique en fait comme "moins mauvais prédicteur" de la localisation. Cette confusion peut s'exprimer de plusieurs façons: confusion entre ce schéma propre à la restitution ambisonique[27] et les principes ou mécanismes de la localisation dans un cadre général; application abusive, pour l'effet de localisation en position d'écoute excentrée, de la prédiction d'après des vecteurs $\vec{V}$ et $\vec{E}$ définis pour le centre!

Structure du décodeur, *Shelf-Filtering*

Le décodeur dans son ensemble doit assurer un matriçage différencié selon les domaines basse- et haute-fréquence. Dans un cas très général, il faut donc d'abord dissocier en deux sous-bandes chacun des signaux W, X, Y (et Z) avant de leur appliquer respectivement les matriçages haute- et basse-fréquence, puis additionner les deux groupes de signaux S_i d'alimentation des haut-parleurs (Figure 2.10).

Les propriétés du décodage pour les configurations régulières et semi-régulières permettent une implémentation beaucoup plus efficace. Comme il a été expliqué, le décodage ne nécessite qu'une unique opération de matriçage (2.25) (2.26), qu'il suffit de faire précéder par une correction du gain de chaque composante, dépendant de la fréquence. Cette correction met en jeu des *shelf-filters*, c'est-à-dire des filtres dont la courbe fréquentielle d'amplitude *s'étage en deux paliers* qui correspondent aux gains basse- et haute-fréquence (Figure 2.11). Il est important que les relations de phase soit préservées entre le filtre qui s'applique à W et les filtres – identiques – qui s'appliquent à X, Y, Z.

2.4.4 Analyse critique de la restitution

Observations à partir d'une configuration rectangulaire

Pour faire écho à l'analyse entamée en 2.2, il est intéressant d'entamer l'étude de la restitution ambisonique en commentant le cas d'une géométrie simple et juste suffisante pour une restitution 2D: une configuration rectangulaire ou carrée, géométrie qui a d'ailleurs été le premier support d'application de la technique

27. Il peut être appliqué à d'autres types de restitution, y compris ambisonique aux ordres supérieurs, mais la fréquence de transition peut-être remise en cause.

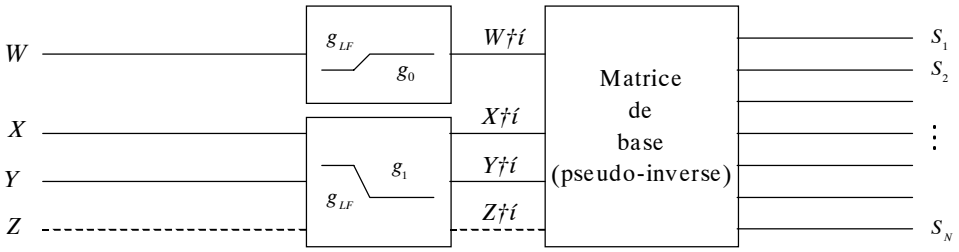

FIG. 2.11 – *Structure d'un décodeur ambisonique en deux sous-bandes spécifiques aux configurations régulières (polygones ou polyèdres réguliers) ou semi-régulières (rectangle, parallélépipède rectangle, ...).*

ambisonique. Deux questions se posent à nouveau: «Quelle est maintenant la classe des événements acoustiques – et des effets de localisation – reproductibles?», et: «Quelles sont les implications de la géométrie, et les compromis liés au nombre de haut-parleurs?».

Contrairement à ce que permet l'exploitation simultanée de seulement deux haut-parleurs, les caractéristiques de la propagation locale (vecteur vélocité $\vec{V}$) peuvent être ici contrôlées "à volonté", par simple modification des gains relatifs de W, X et Y. S'intéressant en particulier à la restitution des conditions d'une image sonore naturelle **dans un domaine basse-fréquence**, il est possible de synthétiser un front d'onde de vitesse de propagation naturelle ($r_V = 1$) dans n'importe quelle direction: c'est ce que réalise le décodage "basique" basse-fréquence pour une source encodée comme une onde plane (2.14).

Il faut remarquer que pour parvenir à ce contrôle, tous les haut-parleurs participent et *des ondes arrière sont mises en jeu*! Contrairement au cas de deux ondes d'incidences symétriques par rapport à l'axe frontal $(O, \vec{x})$ (Figure 2.4), l'extension spatiale de la figure d'interférence créée par les quatre ondes concourantes est limitée suivant $(O, \vec{x})$. Plus précisément, la géométrie rectangulaire lui impose une périodicité $\Lambda_x = \frac{\lambda}{2\sin(\pi/2-\phi_F)} = \frac{\lambda}{2\cos\phi_F}$ suivant $\vec{x}$ en plus de la périodicité $\Lambda_y = \frac{\lambda}{2\sin\phi_F}$ suivant $\vec{y}$ (Figure 2.12). Une première conclusion s'impose. En contrôlant à la fois la direction et la vitesse du front d'onde synthétique, c'est-à-dire en assurant un *comportement naturel des indices de localisation basse-fréquence (ITD) par rotation de la tête*, on prive l'auditeur d'un autre degré de liberté de mouvement: *la stabilité* de l'effet de localisation basse-fréquence *par translation de la tête se trouve limitée* suivant l'axe $(O, \vec{x})$, alors qu'elle ne l'était théoriquement pas avec une restitution sur deux haut-parleurs. Notons qu'il reste en théorie un degré de liberté de translation: le déplacement suivant l'axe vertical, peu exploité en pratique. La stabilité suivant cet axe vertical serait d'ailleurs elle-même limitée dans le cas d'une restitution périphonique.

On constate que même en ne s'intéressant qu'aux aspects basse-fréquence de la restitution, le choix de l'angle ϕ_F des haut-parleurs par rapport à l'axe médian se pose déjà comme un compromis: pour les images frontales, la stabilité par déplacement latéral de la tête peut être améliorée en diminuant l'angle ϕ_F, de même que le domaine de reconstruction basse-fréquence se voit élargi en fréquence si la tête reste orientée suivant l'axe $\vec{x}$; mais dans le même temps, la stabilité par translation avant-arrière se trouve limitée, et la reconstruction basse-fréquence des images latérales devient plus difficile à l'échelle de la tête[28].

Ce compromis lié aux dimensions Λ_x et Λ_y se retrouve avec au moins la même importance, avec des stratégies se focalisant sur le contrôle de la reconstruction du champ au niveau des oreilles de l'auditeur, comme il est développé en 2.5.3.

Cela étant, c'est surtout au regard des **aspects hautes-fréquences** que la question de l'angle ϕ_F se pose de façon critique. Certes, la correction haute-fréquence du décodage apporte une amélioration capitale à la qualité de localisation: l'indice $r_E \approx 0{,}707$ (contre $r_E \approx 0{,}667$ pour le décodage basique) montre une "re-

28. En dépit de l'augmentation de la fréquence de *cross-talk*!

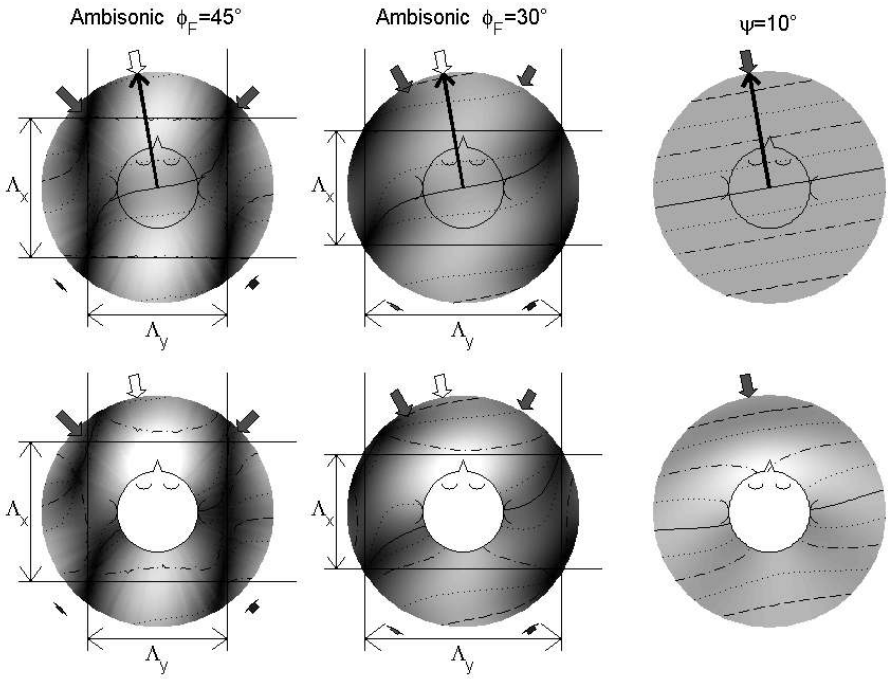

FIG. 2.12 – *Interférence de quatre onde plane concourantes, lors d'une restitution ambisonique sur un dispositif rectangulaire de haut-parleurs ($\phi_F = 45°$ et $\phi_F = 30°$). Comparaison avec l'onde plane de référence de même direction apparente $\psi = 10°$. Champ monochromatique de fréquence 800 Hz. Champ libre en haut, la flèche grasse centrée désignant un vecteur vélocité unitaire; champ diffracté par la présence de la tête en bas. En haut à droite, les courbes de phase aux "extrémités" du disque (en tiret) représentent un déphasage de π ou $-\pi$ par rapport à la ligne centrale continue de phase nulle. Les flèches extérieures pleines indiquent les directions des ondes planes élémentaires contributives, leurs longueurs étant proportionnelles à leurs amplitudes. Chaque motif d'interférence s'étend sur une zone rectangulaire dont les proportions $\Lambda_x/\Lambda_y = \sin\phi_F/\cos\phi_F$ sont les proportions inverses du rectangle décrit par les haut-parleurs. La configuration $\phi_F = 30°$ est ainsi plus favorable aux incidences frontales. A 800 Hz, la configuration carrée $\phi_F = 45°$ devient critique: les oreilles entrent dans le creux d'énergie.*

concentration" de l'énergie dans la direction de la source virtuelle, ce qui se traduit par le rétablissement d'un ITD et d'un ILD plus significatifs pour les images latérales [DRP98] (voir aussi en 4.1). Mais le potentiel de latéralisation reste très limité: pour une configuration carrée, l'effet le plus latéral correspond au cône d'ambiguïté dont font partie les deux haut-parleurs latéraux, soit un angle d'ouverture de 45° seulement!

La stabilisation des images frontales par réduction de l'angle ϕ_F se fait aux dépends de la latéralisation maximale, et inversement. Par ailleurs, le choix d'une configuration non-régulière (rectangulaire) déséquilibre la distribution d'énergie, l'"effort" de restitution étant en quelque sorte plus important pour les directions où les sources réelles sont plus rares, ce qui est illustré Figure 3.15 au chapitre suivant.

S'arrêtant sur la configuration carrée ($\phi_F = 45°$), qui assure une qualité égale de la restitution dans toutes les directions, la qualité de restitution des images frontales peut donc paraître bien pauvre, comparée à une restitution stéréophonique conventionnelle ($\phi_F = 30°$): abstraction faite des aspects basse-fréquence, l'écart de 90° entre les haut-parleurs est habituellement jugé critique et suffisant pour donner lieu à l'artefact du *trou au milieu* ("*The hole in the middle*".). Mais à la différence d'une restitution sur deux haut-parleurs qui auraient cet écart angulaire, ou encore des procédés quadriphoniques qui n'ont pas survécu, les conditions de l'illusion auditive ne sont pas rompues: d'une part, le naturel de l'image sonore au regard des mécanismes basse-fréquence est assuré pour toutes les directions ($r_V = 1$); d'autre part, la qualité relativement peu précise attachée aux indices de localisation haute-fréquence est garantie être la même pour toutes les directions (r_E constante), permettant au système perceptif de s'en accommoder. En dehors d'Ambisonic, certains procédés quadriphoniques ou plus généralement multi-canal [CS72] [Pol96b] vérifient la première condition, mais pas la seconde, des images précises pouvant émerger sporadiquement dans les directions des haut-parleurs.

L'avantage d'une multiplication possible des haut-parleurs

L'une des spécificités de l'approche ambisonique est de ne pas être assujettie à une configuration de haut-parleurs. Le dispositif carré, qui constitue la configuration régulière minimale pour une restitution adéquate, est particulièrement sujet aux problèmes d'instabilité: lorsque l'auditeur s'écarte du centre (*sweet-spot*), les images voisines du haut-parleur dont il se rapproche tendent à s'y rabattre. Ce haut-parleur est en effet perçu isolément si les temps d'arrivée des différentes contributions, ainsi que leurs énergies relatives, sont trop distincts, ce qui advient facilement lorsque les haut-parleurs sont très écartés (Figure 2.13). En augmentant le nombre de haut-parleurs pour la restitution, leur écart angulaire diminue et la densité temporelle des contributions perçues en favorise la fusion perceptive. Il est attendu que cette fusion résiste à une position d'autant plus excentrée que le nombre de haut-parleurs est important.

Notons qu'en principe, l'augmentation du nombre de haut-parleur n'améliore pas la précision ou la résolution spatiale des images sonores perçues en position centrée, puisque l'indice r_E reste le même.

Conditions d'écoute critiques: une adaptation du décodage encore possible

La qualité d'homogénéité propre à la restitution ambisonique tient en partie au fait que plusieurs, voire tous les haut-parleurs participent à la création d'une même image sonore. On constate même, de la part des haut-parleurs opposés à la direction de la source virtuelle, une participation non négligeable bien que réduite par le décodage haute-fréquence. Considérons maintenant une position d'écoute assez excentrée, se rapprochant des haut-parleurs opposés à la source virtuelle. La contribution de ces derniers risque alors de devenir prédominante dans l'effet perçu de localisation, et ce pour deux raisons: d'une part, le changement des distances relatives entre l'auditeur et les différents haut-parleurs a modifié le poids énergétique relatif de chaque contribution, ce qui se traduit par une distorsion du vecteur énergie $\vec{E}$ perçu; d'autre part, les différentes contributions n'étant plus perçues de façon synchrone, c'est le haut-parleur le plus proche (perçu

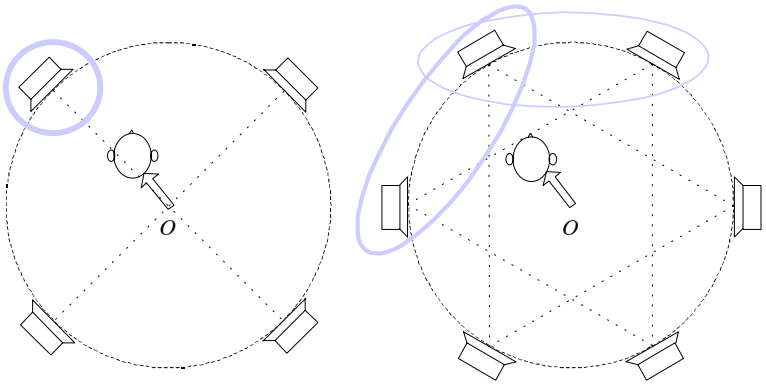

FIG. 2.13 – *L'avantage de multiplier les haut-parleurs pour la préservation de l'illusion sonore, tout en gardant la même résolution spatiale (*Ambisonics *ordre 1). En s'écartant du centre du dispositif carré (sans rester sur un axe médian), l'auditeur se rapproche d'un seul haut-parleur et s'éloigne des autres: le haut-parleur le plus proche risque d'émerger comme source individuelle (instabilité par effet de bascule). Dans le cas d'une configuration hexagonale, l'auditeur se rapproche toujours d'au moins deux haut-parleurs à la fois tant qu'il reste dans l'hexagone intérieur (délimité par les segments en tirets), et l'éloignement des haut-parleurs adjacents est moins critique: cela favorise la fusion perceptive des sources réelles et la préservation de l'illusion.*

le plus précocement) qui risque de définir la direction apparente par effet d'antériorité (Figure 2.14). Lorsque c'est une scène complexe qui est restituée, celle-ci risque donc de souffrir d'un repliement vers les haut-parleurs les plus proches, laissant des trous dans le panorama sonore (pour les auditeurs très excentrés). Quant aux sources sonores en mouvement, censées décrire un cercle autour de l'auditeur par exemple, leur trajectoire est susceptible d'être perçue de façon discontinue, subissant des rebroussements anormaux.

Pour éviter ces anomalies gênantes lorsque l'auditoire s'étend à proximité des haut-parleurs, voire à l'extérieur de leur périmètre, Malham [Mal92] a proposé une troisième forme de décodage, s'ajoutant aux décodages *basique* et max r_E. Elle a la propriété d'annuler la contribution d'un haut-parleur lorsque la source virtuelle est située dans la direction opposée, la participation des haut-parleurs étant croissante à mesure que leur direction se rapproche de la source virtuelle (Figure 2.14). Tous les haut-parleurs délivrent alors des signaux en phase, *i.e.* de même signe, ce qui vaut à ce décodage le nom de baptême de "décodage *in-phase*".

S'appliquant aux configurations régulières de haut-parleurs, ce décodage *in-phase* de Malham se présente comme une autre façon de pondérer les composantes d'ordre 1 (X, Y, Z) par rapport à la composante W avant l'application de la matrice de décodage basique. Il se caractérise par un rapport $g_1/g_0 = 1/2$ contre un rapport $g_1/g_0 = 1/\sqrt{2}$ pour le décodage *max* r_E (2.26). Ce rapport $1/2$ est valable pour une restitution périphonique (3D) aussi bien que pour une restitution horizontale (2D), contrairement au décodage *max* r_E. Il faut bien être conscient que le décodage *in-phase* est en revanche sous-optimal pour une position d'écoute centrée: l'indice r_V associé ne vaut que $1/2$ (restitutions 2D et 3D), et l'indice r_E ne vaut que $2/3 = 0{,}667$ (restitution 2D) ou 0,5 (restitution 3D) contre 0,707 (2D) ou 0,577 (3D) avec un décodage *max* r_E.

Une analyse plus approfondie et appliquée aux systèmes ambisoniques généralisés aux ordres supérieurs est donnée au chapitre 4.

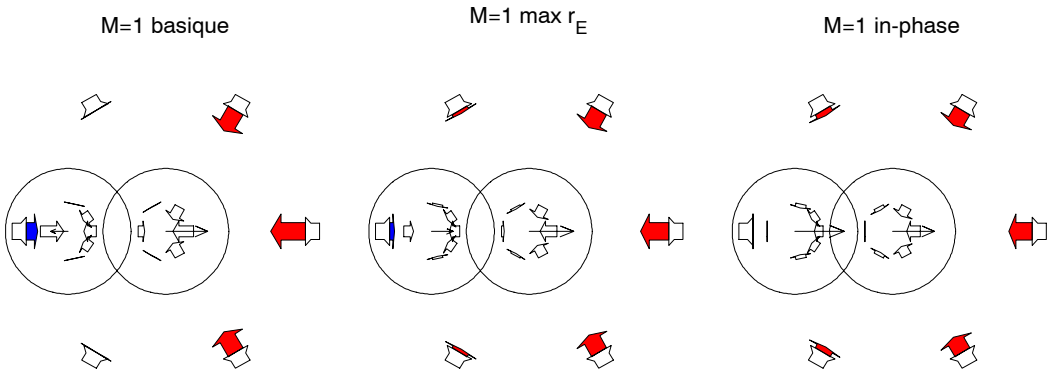

FIG. 2.14 – *Comparaison de trois décodages ambisoniques pour la restitution d'une source virtuelle (orientée vers la droite): décodages* basique, max r_E *et* in-phase *à l'ordre 1. Les gains des haut-parleurs sont indiqués par la longueur des flèches pleines. L'amplitude et la direction apparentes des différentes contributions en des positions centrée et excentrée sont indiquées par des flèches larges et creuses, et le vecteur énergie résultant par des flèches fines. Chaque cercle est considéré de rayon unité pour le vecteur énergie. En position centré, l'arrivée des ondes est synchrone, la direction et la longueur du vecteur énergie indiquent la direction perçue et la "qualité" de l'image. En position excentrée (vers la gauche), le poids perceptif du haut-parleur gauche – opposé à la source virtuelle – croît: avec le décodage basique, son amplitude apparente devient prédominante et le vecteur énergie est retourné vers la gauche; avec le décodage* max r_E*, elle concurrence celle du haut-parleur droit et l'image risque de se rabattre sur le haut-parleur gauche par effet d'antériorité, même si le vecteur énergie reste orienté vers la droite; avec le décodage* in-phase*, l'amplitude du haut-parleur gauche s'annule, l'effet d'antériorité est évité et le vecteur énergie est satisfaisant.*

Interprétations du *cross-talk* et caractérisation par r_E

La restitution ambisonique sur quatre ou plus de haut-parleurs n'échappe pas au problème du *cross-talk*, dont les manifestations négatives en haute-fréquence ont été commentées dans le cas de la restitution sur deux haut-parleurs en 2.2. Là encore, il se caractérise par le fait que les contributions des différentes sources réelles par la création d'une image fantôme sont objectivement dissociables[29] au-delà d'une certaine fréquence – la fréquence de cross-talk –, lorsque le front d'onde synthétisé localement n'est plus assez grand par rapport à la tête. C'est surtout d'un point de vue théorique que cette dissociation est possible (traitement objectif de l'information sonore), elle n'est en général pas effective sur le plan perceptif, particulièrement avec une restitution ambisonique qui fait collaborer simultanément plusieurs haut-parleurs, voire tous: la fusion psychoacoustique (ou effet de sommation) des événements parvenant aux oreilles est alors favorisée par une plus grande densité temporelle. Il n'en reste pas moins que le *cross-talk* entache quelque peu la qualité de restitution spatiale.

Considérant individuellement la reproduction d'un front d'onde original, le cross-talk est responsable de l'élargissement de sa tache de localisation (*localization blur* [Bla83]): il offre une moindre cohérence des indices de localisation haute-fréquence et diminue l'effet de latéralisation statique et dynamique (lors de rotations de la tête). L'indice r_E (associé au décodage haute-fréquence) décrit le taux de concentration – ou inversement, d'étalement – des participations autour de la direction $\vec{u}_E$. Il traduit en ce sens le *degré de cross-talk*, et l'on peut voir en lui un bon indice de la dégradation de la précision de l'image sonore (si $r_E < 1$) par rapport à une image naturelle ($r_E = 1$).

L'interprétation de l'indice r_E peut maintenant être étendue à la reproduction d'une scène sonore complexe, afin de caractériser la préservation ou la dégradation des qualités spatiales – impressions spatiales et effet d'enveloppement. Ces qualités dépendent notamment de la composante latérale des réflexions et de la réverbération associées à chaque source, qui se manifeste à travers la décorrélation interaurale et les fluctuations temporelles d'ITD et d'ILD (cf 1.3.4). Le cross-talk étant responsable d'une diminution générale de l'effet de latéralisation et d'un "étalement" de chaque contribution élémentaire, les qualités spatiales originales se trouvent dégradées à la reproduction: la direction apparente des réflexions latérales est ramenée vers le plan médian de l'auditeur et *la décorrélation interaurale est réduite*, comparée à l'expérience d'écoute naturelle dans la scène originale. Dans la mesure où la direction globale de propagation de chaque événement élémentaire (onde directe ou réflexion) est préservée ($\vec{u}_E = \vec{u}$), *r_E est donc propre à indiquer le degré de dégradation des qualités spatiales dans leur ensemble*[30]. Cependant, pour un auditeur placé au centre du dispositif qui bénéficie d'une bonne reproduction basse-fréquence ($\aleph_V = 1$), les effets d'enveloppement, de largeur apparente de source, et autres impressions spatiales, ne sont que partiellement dégradés puisqu'ils reposent sur des phénomènes (réflexions) qui interviennent en grande partie dans un domaine basse-fréquence. Il reste malgré tout que la réduction de la décorrélation interaurale tend à "convertir" l'effet spatial des réflexions en un effet de coloration perçue (1.3.4). Cette dernière remarque, qui fait écho à une discussion entamée en 2.2, est étayée par l'expérience des preneurs de "son ambisonique" [Sur] qui constatent en général un effet de coloration exagéré à l'issu de la prise de son/restitution ambisonique, et recommandent de réduire la présence des réflexions arrière lors de la prise de son.

2.4.5 Vers une extension aux ordres supérieurs

L'approche ambisonique est dotée d'un cadre mathématique qui permet d'envisager de manière rationnelle une extension de son format de représentation et de ses qualités de restitution vers des degrés supérieurs

29. ... d'après la cohérence des indices de localisation et de leur comportement par rotation de la tête.

30. On ne pourrait donc pas dire que cet indice traduit le même "degré" de cross-talk pour une restitution stéréophonique sur deux haut-parleurs, étant donné la distorsion systématique des directions apparentes $\vec{u}_E$ vers l'axe médian.

de résolution spatiale. La représentation ambisonique résulte en effet d'une décomposition du champ en harmoniques sphériques, dont le B-format ne représente que la restriction à l'ordre 1. En s'intéressant tout d'abord à une représentation et une restitution restreintes au plan horizontal, ainsi que l'a exposé Bamford [BV95], la décomposition d'une onde plane d'incidence ψ et portant un signal S s'écrit en tout point $\vec{r}(r,\theta)$ du champ:

$$p(\vec{r}) = SJ_0(kr) + S\left(\sum_{m=1}^{\infty} 2j^m J_m(kr)[\cos(m\psi)\cos(m\theta) + \sin(m\psi)\sin(\theta)]\right) \tag{2.27}$$

Ce développement du champ en série de Fourier-Bessel fait apparaître des fonctions radiales – les fonctions de Bessel $J_m(kr)$ – et des fonctions angulaires, dites harmoniques cylindriques[31], dans lesquelles on reconnaît les fonctions d'encodage ambisonique d'ordre 1:

$$\begin{array}{lcl} 1 & \rightarrow & W \\ \sqrt{2}\cos\psi & \rightarrow & X \\ \sqrt{2}\sin\psi & \rightarrow & Y \end{array} \tag{2.28}$$

Les harmoniques cylindriques suivantes définissent alors des fonctions d'encodage d'ordres supérieurs:

$$\begin{array}{lcll} \sqrt{2}\cos(2\psi) & \rightarrow & U & \\ \sqrt{2}\sin(2\psi) & \rightarrow & V & \text{ordre 2} \\ \ldots & & & \\ \sqrt{2}\cos(m\psi) & \rightarrow & \ldots & \\ \sqrt{2}\sin(m\psi) & \rightarrow & \ldots & \text{ordre } m \end{array} \tag{2.29}$$

Le décodage d'un système d'ordre supérieur (ordre 2 par exemple) peut être défini à la manière du décodage basique d'ordre 1 (2.24), avec comme objectif de "reconstruire" les composantes encodées (W,X,Y,U,V) par combinaison des ondes planes émises par les haut-parleurs (principe de "réencodage", équation 2.23), requérant pour cela plus de haut-parleurs. L'expansion (2.27) suggère que le champ encodé est alors reconstruit sur un voisinage plus large – pour une longueur d'onde donnée – autour du point central.

C'est autour de cette idée de reconstruction physique étendue du champ acoustique, que l'investigation des systèmes ambisoniques d'ordres supérieurs a d'abord été orientée [BV95] [Pol96a]. La conséquence d'une telle extension est l'élargissement du domaine basse-fréquence de reconstruction des informations binaurales lorsque la tête est centrée [DRP98] (Annexe B, plus étude complémentaire en 4.1), ou encore l'élargissement de la zone d'écoute si l'on s'attache à la reconstruction dans un domaine basse-fréquence fixé. On peut ainsi voir en la restitution ambisonique un cas particulier de reconstruction holographique du champ ("*Holophonie*") [NE98] [NE99] [Nic99].

Ce type d'analyse qui permet de quantifier l'extension du domaine de reconstruction, ne donne malheureusement pas les moyens de caractériser la restitution au-delà de ces limites et de démontrer l'amélioration qu'y apportent les ordres supérieurs. Les propriétés de la restitution qu'il est possible de caractériser objectivement ne se réduisent pourtant pas à l'expression de ces limites de reconstruction. On a pu en effet constater que l'indice r_E associé à un décodage caractérise à la fois:

- la "précision" de chaque image sonore au regard des mécanismes haute-fréquence (résolution spatiale),
- la préservation – ou la dégradation – des qualités spatiales plus globales comme l'enveloppement et autres impressions spatiales,

31. ... que l'on peut considérer d'une certaine manière comme un sous-ensemble des harmoniques sphériques, comme expliqué en 3.1.2.

– la robustesse des images au déplacement de l'auditeur hors du centre (moindre sensibilité au *sweet-spot*).

A mesure que la résolution spatiale de la représentation ambisonique est affinée par l'introduction de composantes d'ordres supérieurs, l'indice r_E associé à la restitution est amélioré[32] [DRP98], ce qui traduit le fait que le système tire meilleur profit de la densité angulaire des haut-parleurs. C'est ainsi une amélioration de l'ensemble des propriétés de restitution sus-énoncées qui est attendue avec le développement des systèmes d'ordres supérieurs.

Le développement de systèmes d'ordres supérieurs ne constitue pas une rupture avec les principes des systèmes traditionnels d'ordre 1. Les qualités propres à *Ambisonics*, comme l'homogénéité de la restitution doivent pouvoir être préservées, et l'optimisation du décodage (*max* r_E et *in-phase*) en fonction du domaine de fréquences et des conditions d'écoute est généralisable aux ordres supérieurs. C'est ce qui est réalisé de façon partielle dans [DRP98] (Annexe B), et plus complète au chapitre 3.

2.5 Contrôle de la reconstruction au niveau des oreilles

En contrepartie de leur relative simplicité, les techniques de restitution sur haut-parleurs qui ont été présentées jusqu'ici sont caractérisées par un artefact commun: le *cross-talk*, ou encore l'incapacité des systèmes à contrôler les informations sonores reconstituées aux oreilles de l'auditeur, notamment dans un domaine haute-fréquence. Il en résulte une limitation de la qualité et de la précision de chaque image sonore individuelle, et une limitation des qualités spatiales associées au champ sonore dans son ensemble, comme la largeur de la scène et l'enveloppement.

Les techniques présentées maintenant ont pour vocation de contrôler la reconstruction des informations sonores spatiales directement au niveau des oreilles. Il s'agit en premier lieu de ce que l'on désigne couramment par techniques binaurales (restitution au casque, en 2.5.1) et transaurales (restitution sur deux haut-parleurs, en 2.5.2), qui donnent lieu à une reconstruction des informations directionnelles *statiques* (tête fixe). Le principe de restitution transaurale connaît également, à travers quelques stratégies récentes de restitution sur quatre haut-parleurs (cf 2.5.3), une extension similaire à celle qu'a connu la stéréophonie conventionnelle avec la stéréophonie panoramique et ambisonique. Ces procédés tendent à reconstituer les informations directionnelles non-seulement statiques mais aussi *dynamiques* (variations par légère rotation de la tête).

Au-delà d'une simple description des principes mis en jeu, nous cherchons à nouveau dans cette section à caractériser les propriétés de restitution sur haut-parleurs, en nous attachant notamment au comportement des images fantômes par rotation de la tête, à la robustesse par translation de la tête, et à l'effort de reconstruction investi. Cette étude s'appuie sur une caractérisation des phénomènes acoustiques synthétisés à l'échelle de la tête, ainsi que sur l'observation des aspects énergétiques. A cette occasion, des outils de caractérisation comme les vecteurs vélocité et énergie sont à nouveau utilisés, et au besoin réinterprétés.

2.5.1 Écoute au casque: techniques binaurales et variantes

Les techniques binaurales ont connu un développement considérable depuis deux décennies. Souvent présentées comme *le* moyen d'offrir enfin à l'auditeur des qualités d'image sonore exceptionnelles concurrençant l'expérience auditive naturelle, elles ont acquis une place de choix dans les applications de réalité virtuelle. Parce qu'elles ont pour objectif de contrôler finement les informations sonores perçues par les oreilles, leur mode de restitution privilégié est l'écoute au casque – dite "présentation binaurale". La présentation des signaux binauraux par l'intermédiaire de deux haut-parleurs est possible par l'intervention d'un

32. r_E tend vers 1 lorsque l'ordre du système tend vers l'infini.

filtrage transaural, qui a pour but d'annuler les chemins croisés (cross-talk) entre les haut-parleurs et les oreilles. Ces techniques transaurales ou assimilées sont abordées en section 2.5.2 pour traiter des questions spécifiques à la reproduction sur haut-parleurs, comme celle de la robustesse aux mouvements de l'auditeur. En se limitant à la restitution au casque, l'étude des techniques binaurales – et assimilées – se concentre ici sur des questions liées à la représentation et la synthèse du champ acoustique binaural, comme l'efficacité et la qualité de la représentation, le coût de calcul, la portabilité, etc...

Principe

Le principe d'une restitution binaurale est finalement assez simple: il consiste à présenter directement aux oreilles de l'auditeur, à l'aide d'un casque (ou écouteurs), les informations sonores (signaux binauraux) qu'il aurait perçues s'il avait pris place dans l'environnement naturel. Ces informations incluent de façon naturelle l'effet des diffractions et réflexions de chaque événement acoustique par le corps, la tête et les oreilles de l'auditeur. Celui-ci dispose donc d'un ensemble riche et cohérent d'indices de localisation qui lui sont familiers: il s'agit des différences interaurales (ITD et ILD) qui donnent lieu à une détection latérale des événements sonores, mais aussi des indices spectraux qui complètent la localisation directionnelle dans le plan médian et résolvent notamment les indéterminations avant-arrière[33] et sur la position verticale. La restitution binaurale est donc susceptible de donner lieu à une reconstitution subjective du paysage sonore en trois dimensions, s'affranchissant des limitations propres aux techniques traditionnelles de restitution sur haut-parleurs.

Prise de son et synthèse binaurale

La réalisation la plus directe du principe binaural découle de la prise de son au moyen de capteurs logés dans le canal auditif de chaque oreille d'un individu, les signaux obtenus étant restitués ultérieurement aux oreilles du même individu (Figure 2.15). En réalité, les signaux originaux subissent plusieurs transformations avant d'être à nouveau restitués au sein des canaux auditifs: leur enveloppe spectrale en particulier est modifiée par les fonctions de transfert électro-acoustiques des transducteurs (microphones et écouteurs), et par le filtrage par le pavillon lorsque les signaux ne sont pas diffusés directement dans le canal auditif. Il est donc nécessaire de procéder à une égalisation des signaux binauraux en conséquence.

Les procédés de synthèse binaurale offrent quant à eux la possibilité de produire le champ acoustique binaural en fonction d'une composition arbitraire de l'environnement sonore virtuel. Ils se basent sur l'utilisation des réponses impulsionnelles binaurales associées à la position d'une source par rapport à l'auditeur dans l'espace virtuel. Il peut s'agir dans un premier temps d'une mesure effectuée dans une salle: la convolution d'un signal monophonique – de préférence issu d'une prise de son anéchoïque – avec les réponses binaurales mesurées permet alors de reproduire l'effet d'une diffusion de ce signal dans cette même salle. Ce procédé offre un unique degré de liberté: le choix du signal. L'effet de la salle et les positions de l'auditeur et de la source sont en revanche figés, ainsi que les propriétés de directivité de la source.

Pour des applications de type "réalité virtuelle" où l'on souhaite un contrôle complet de la composition de la scène sonore et de la situation de l'auditeur par rapport aux objets sonores, la démarche adoptée consiste à simuler l'effet de chaque événement acoustique élémentaire – onde directe et réflexions associées à chaque source – en le modélisant comme une onde plane[34]: il s'agit d'une opération de filtrage du signal par la paire de HRTF (*Head-Related Transfer Functions*, 1.3.3) associée à la direction de l'onde[35] (Figure 2.16). Pour

33. Du moins, en théorie.

34. Pour les sources proches, l'onde directe devrait être modélisée comme une onde sphérique. Mais l'effet de champ proche semble rarement pris en considération en synthèse binaurale.

35. En général, un traitement plus grossier est réservé à la réverbération tardive, voire aux réflexions précoces.

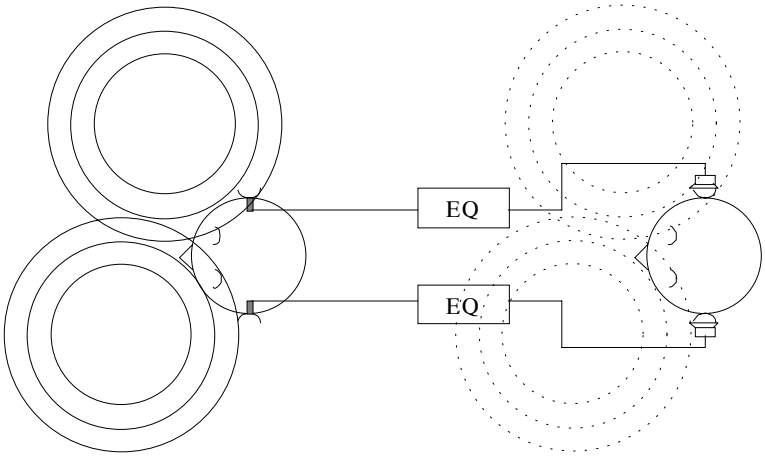

FIG. 2.15 – *Principe de restitution binaurale à partir d'un enregistrement.*

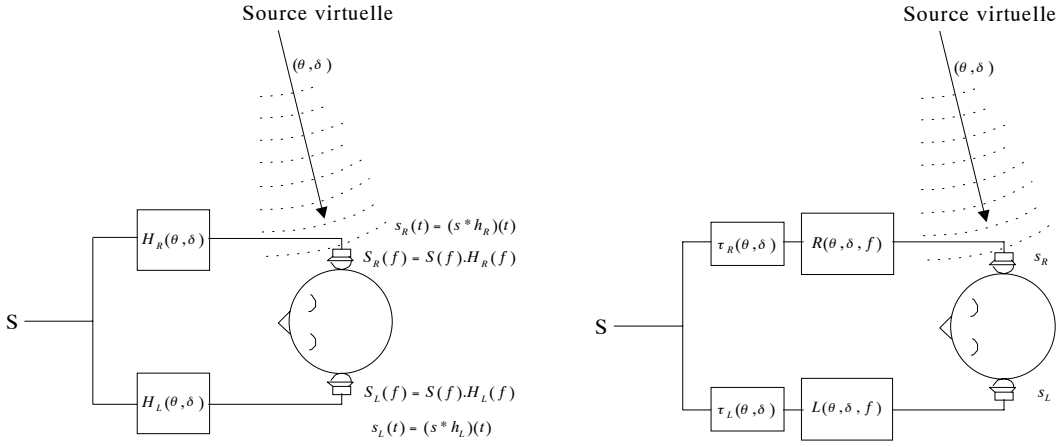

FIG. 2.16 – *Synthèse binaurale: simulation de l'effet d'un événement acoustique élémentaire (onde supposée plane), par filtrage utilisant une paire de HRTF. A gauche: les fonctions de transfert binaurales sont définies d'après les réponses impulsionnelles (HRIR) $h_L(t)$ et $h_R(t)$, l'opération de filtrage est équivalente à un produit de convolution. A droite, une implémentation plus efficace: chaque HRTF est modélisée par la combinaison d'un retard pur (τ_L ou τ_R) et d'un filtre à phase minimale ($L(f)$ ou $R(f)$), le filtrage qui suit le retard étant de type récursif (RII).*

pouvoir simuler une scène de composition arbitraire, il faudrait donc disposer d'un jeu de HRTF couvrant l'ensemble des directions possibles. Ces HRTF sont mesurées en chambre sourde, sous forme de réponses impulsionnelles (HRIR: *Head-Related Impulse Responses*[36]). De même que pour un enregistrement binaural, il est nécessaire de compenser les réponses des transducteurs par une phase d'égalisation des réponses mesurées. Un jeu assez complet de HRTF a été mesuré par Martin et Gardner [GM94] sur une tête artificielle (mannequin KEMAR) et est très largement utilisé par la communauté scientifique. Disposant des réponses impulsionnelles, la *réalisation la plus directe* du filtrage binaural consiste en une *opération de convolution* du signal monophonique avec chaque HRTF (HRIR) de la paire (Figure 2.16 gauche). Le coût de calcul d'une convolution dans le domaine temporel étant rapidement rédhibitoire, il est généralement fait appel à des algorithmes de convolution rapide qui consiste à effectuer la transformation duale dans le domaine fréquentiel: il s'agit pour simplifier d'un produit terme à terme des transformées de Fourier de la réponse impulsionnelle et du signal, le passage entre les domaines temporel et fréquentiel étant réalisé à l'aide de FFT (Transformée de Fourier Rapide). Ce type d'algorithme est abordé plus en détail en 5.2.1.

Problèmes et enjeux spécifiques

Pour des raisons pratiques, la mesure des HRTF est en général effectuée sur une tête artificielle (un mannequin), censée traduire une caractéristique moyenne. Etant donné la diversité morphologique des individus, il est malheureusement courant que les indices spectraux propres aux HRTF mesurées ne soient pas familiers à l'auditeur, et donnent lieu à un effet de localisation erroné. La manifestation la plus typique est le rejet des sources frontales en arrière ou en hauteur, c'est-à-dire hors du champ visuel. Cet effet de rejet est en général limité avec l'assistance d'une image visuelle qui puisse être identifiée comme la source sonore. La qualité de reproduction issue de la synthèse binaurale souffre fréquemment d'une autre tare: celle d'une perception intracrânienne des images sonores. Si ce problème est en partie attribuable à des indices spectraux non-familiers, il faut préciser qu'il est particulièrement encouragé par l'absence d'effet de salle[37] associé aux sources monophoniques spatialisées.

Le problème de l'individualité des HRTF et des indices spectraux ne constitue pas la seule limitation aux performances de la restitution binaurale. Dans des conditions d'écoute naturelles, les *mouvements de rotation de la tête* induisent des variations des indices de localisation – dont les différences interaurales – qui permettent de lever les indéterminations lorsque les indices spectraux ne sont pas suffisamment exploitables, ce qui peut advenir dans le cas de sources sonores peu familières. Pour satisfaire ces mécanismes de localisation dynamique, le système binaural doit pouvoir prendre en compte l'orientation de la tête – à l'aide d'un système de suivi de la tête (*Head-Tracking*) – et adapter la restitution des informations binaurales en conséquence, ce qui est réservé à une production par synthèse en temps réel. C'est sans-doute une condition – avec l'assistance visuelle – pour qu'une adaptation à des HRTF non-individuelles soit possible par effet d'apprentissage.

En tant que *mode de représentation* du champ sonore, l'approche binaurale telle qu'elle est présentée jusque là, souffre donc de deux principales carences. Une fois la production binaurale réalisée (prise de son ou synthèse), elle n'est appropriée qu'à un auditeur particulier[38], et ne permet pas une adaptation aux mouvements de la tête lors de la restitution.

En matière de synthèse binaurale dans un contexte de réalité virtuelle ou d'applications interactives, le problème des spécificités individuelles des HRTF est toujours présent, et nécessiterait pour être résolu que chaque utilisateur dispose d'une *banque de HRTF personnalisées*. Une autre question prédominante

36. Par souci de simplicité, nous utilisons la désignation "HRTF" même lorsqu'il s'agit des réponses impulsionnelles HRIR.

37. L'expérience d'écoute en chambre sourde n'est pas des plus naturelles!

38. ... ou à un groupe d'auditeurs ayant des caractéristiques semblables.

est bien-sûr le *coût de calcul*: le filtrage binaural doit pouvoir être effectué en temps-réel, et avec un délai suffisamment court pour assurer la transparence de l'interaction homme-machine, surtout si les positions relatives des sources virtuelles doivent être réactualisées en fonction des mouvements de la tête (applications avec *head-tracking*). Malgré l'explosion de la puissance de calcul disponible sur les ordinateurs courants, la minimisation des coûts de calcul reste un enjeu important dans la mesure où les ressources sont partagées avec d'autres traitements – audio ou vidéo. Bien que fonctionnelle en temps-réel grâce aux algorithmes de convolution rapide, l'implémentation du filtrage sous forme RIF (*Réponse Impulsionnelle Finie*) n'est pas des plus performantes.

Le paragraphe suivant évoque brièvement un ensemble de stratégies qui répondent progressivement aux différents enjeux évoqués.

Réponse aux enjeux: stratégies émergentes

L'idée de base pour l'optimisation du filtrage binaural revient à dissocier l'information temporelle (retard d'arrivée aux oreilles) des informations d'amplitude et indices spectraux. La méthode générique consiste à décomposer chaque réponse $H_L(f)$ en un filtre passe-tout $\tau_L(f)$, qui contient intrinsèquement l'information temporelle, et un filtre à phase minimale $L(f)$ qui contient toutes les informations spectrales, et dont l'énergie est condensée en début de réponse temporelle:

$$\begin{array}{ll} H_L(f) & = \tau_L(f).L(f) \\ H_R(f) & = \tau_R(f).R(f) \end{array} \tag{2.30}$$

Constatant que la partie passe-tout est bien souvent à phase approximativement linéaire, elle est généralement modélisée comme un retard pur – indépendant de la fréquence – donc peu coûteux. Les retards estimés τ_L et τ_R traduisent un ITD $(\tau_R - \tau_L)$ qui obéit approximativement à une loi "haute-fréquence"[39], proche de celle que décrit l'équation (1.47) pour un modèle de tête sphérique (Cf 1.3.2). Quant à la réponse à phase minimale, il est d'usage de la modéliser par un filtre récursif (RII) qui peut se révéler plus économique que le filtre convolutif (RIF) d'origine. De nombreuses méthodes du traitement du signal ont été définies et éprouvées pour une telle modélisation. Il en est fait un état de l'art dans [JLW95] ou [HK97], par exemple. Un avantage de l'implémentation RII est de faciliter le compromis entre qualité d'approximation des HRTF et coût de calcul (ordre des filtres). Pour la simulation d'une source virtuelle en mouvement, une interpolation des fonctions de transfert est nécessaire, ce qui a pour effet momentané de doubler approximativement le coût de calcul.

Si ces stratégies permettent de rendre le filtrage binaural plus économique, le traitement est toujours proportionnel au nombre de sources virtuelles. Une deuxième étape vers l'optimisation de la simulation d'une scène complexe comprenant de nombreuses sources virtuelles, consiste alors à *factoriser les opérations de filtrage*. Il s'agit de trouver une base de décomposition commune à l'ensemble des HRTF, et composée d'un nombre fini d'éléments $L_i(f)$, $i \leq N$. De cette sorte, une approximation de chaque HRTF $H_L(\theta,\delta,f)$ ou $H_R(\theta,\delta,f)$ est obtenue par une combinaison linéaire de ces *filtres de base* $L_i(f)$, dont les facteurs de

39. Il faut se rappeler (section 1.3.2) que l'ITD basse-fréquence (retard de phase) décrit une loi notablement plus ample que l'ITD haute-fréquence. Cela suggère que pour une incidence hors du plan médian, la différence entre ces deux ITD se retrouve sous forme de retard de phase résiduel en basse-fréquence au sein du couple de filtres à phase minimale associé, ce qui pourrait sembler étonnant étant donné la construction des filtres à phase minimale.

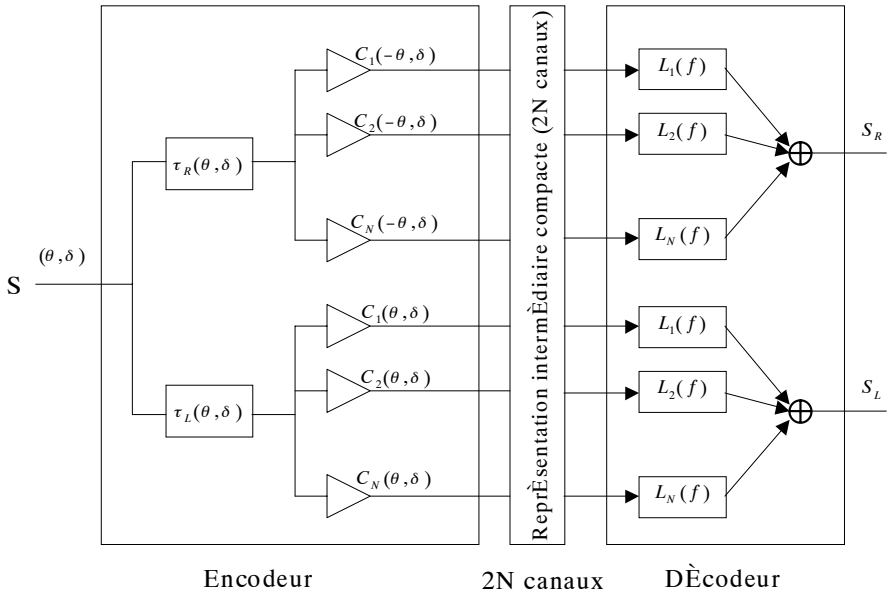

FIG. 2.17 – *Implémentation multi-canal d'une synthèse binaurale, basée sur une décomposition linéaire des HRTF à phase minimale (d'après [LJGW00]).*

pondération $C_i(\theta,\delta)$ sont appelés *fonctions spatiales* [JWL98] [LJGW00]:

$$\begin{aligned} H_L(\theta,\delta,f) &= \sum_{i=1}^{N} C_i(\theta,\delta).L_i(f) \\ H_R(\theta,\delta,f) &= \sum_{i=1}^{N} C_i(-\theta,\delta).L_i(f) \qquad \text{(pour des raisons de symétrie)} \end{aligned} \tag{2.31}$$

La décomposition est plus facile, c'est-à-dire requiert moins de filtres de base pour une qualité d'approximation équivalente, si elle est appliquée aux filtres à phase-minimale [JWL98]. C'est le choix préférentiellement retenu. La structure générique du système (Figure 2.17) se compose alors [LJGW00]:

- d'un *encodeur binaural multi-canal*, qui produit *une représentation intermédiaire compacte*, à savoir N canaux par oreille, où chaque signal monophonique est injecté avec un retard $\tau_L(\theta,\delta)$ ($\tau_R(\theta,\delta)$ pour l'oreille droite) et des gains d'amplitude $C_i(\theta,\delta)$ ($C_i(-\theta,\delta)$ pour l'oreille droite), en fonction de l'incidence (θ,δ) qui lui est associée;
- d'un *décodeur*, constitué de deux bancs de filtres de reconstruction parallèles $L_i(f)$, dont les sorties de chacun sont sommées pour fournir les signaux binauraux S_L et S_R.

L'étape d'encodage étant peu coûteuse, le traitement global a un coût quasi-indépendant du nombre de sources virtuelles. Le problème d'interpolation des filtres dans le cas de sources en mouvements est par ailleurs résolu de façon naturelle et transparente, au niveau de l'étape d'encodage.

La définition des fonctions spatiales et des filtres de base (ou filtres de reconstruction) fait l'objet de stratégies variées, très clairement exposées et comparées dans [LJGW00]. Les méthodes statistiques comme l'analyse en composantes principales (ACP) ou l'analyse en composantes indépendantes (ACI) réalisent une optimisation conjointe des fonctions $C_i(\theta,\delta)$ et filtres $L_i(f)$ en fonction du jeu de HRTF à décomposer. Les fonctions spatiales $C_i(\theta,\delta)$ – ainsi que l'encodage – sont alors sensibles aux différences inter-individuelles

(d'un jeu de HRTF à l'autre) et ne sont donc pas "universelles". Il en est de même pour une méthode récemment introduite par Gardner qui consiste à choisir d'abord les filtres de reconstruction parmi les HRTF d'origine ("*Subset Selection*"), puis en déduire les fonctions spatiales $G_i(\theta,\delta)$ optimales.

Une autre approche retient finalement notre attention, où ce sont les fonctions spatiales $G_i(\theta,\delta)$ qui sont d'abord fixées, indépendamment de la "tête" dont on modélise les HRTF. L'encodage peut alors être considéré comme "universel", à ceci près que la modélisation des retards τ_L et τ_R est susceptible de varier d'une tête à l'autre. Le choix particulier de fonctions harmoniques sphériques comme fonctions spatiales donne lieu au *binaural B-format* [JWL98], dont le nom évoque le format ambisonique[40] évoqué plus haut, en 2.4.1. Les filtres de reconstruction se déduisent alors par projection orthogonale des HRTF sur ces fonctions spatiales. La contrainte de fonctions spatiales prédéfinies rend la reconstruction des HRTF moins performante en haute-fréquence qu'avec les autres méthodes, à nombre égal de fonctions spatiales – ou de filtres de reconstruction. En compensation, cette approche offre quelques avantages pratiques très appréciables:

- un *encodage universel*, *i.e.* indépendant de l'individu (à la question près de la modélisation des retards τ_L et τ_R);
- la possibilité d'une *prise de son naturelle* (encodage acoustique) dans le cas de la restriction au harmoniques sphériques d'ordre 1, à l'aide de deux microphones Soundfield espacés approximativement d'un diamètre de tête (Figure 2.18);
- en plus des modes de diffusion binaural (au casque) et transaural (deux haut-parleurs, cf 2.5.2) ordinaires, la possibilité d'une *restitution améliorée sur un dispositif multi-canal* (ou au moins rectangulaire [JWL98]), le binaural B-format permettant notamment une discrimination entre les sources avant et arrière. Nous développons cet aspect en 2.5.3;
- enfin, la possibilité de n'exploiter que les canaux intermédiaires associés à une seule des oreilles (B-format simple), et d'en réaliser un *décodage ambisonique ordinaire* pour une restitution sur haut-parleurs (section 2.4).

Il faut noter que du schéma d'encodage acoustique tel qu'il est décrit par exemple Figure 2.18, découle une loi d'ITD $\tau_R(\theta,\delta) - \tau_L(\theta,\delta)$ en $\sin\theta\cos\delta$ qui n'est pas conforme à l'encodage théorique, pour lequel l'ITD suit approximativement la loi (1.47) [LJ97]. Deux attitudes sont possibles face à ce problème: tolérer une distorsion de l'ITD, et la minimiser globalement en choisissant un espacement D des microphones légèrement supérieur au diamètre de la tête [JWL98]; ou bien définir les filtres de reconstruction à partir d'une décomposition des HRTF originales dont on aurait ôtés les retards définis Figure 2.18, et non plus à partir des HRTF à phase minimale. Dans ce dernier cas, l'ITD est reconstitué de façon correcte au prix d'une reconstruction moins performante des indices spectraux.

Du fait de sa familiarité avec la représentation ambisonique, *il a semblé important* d'aborder succinctement l'approche *binaural B-format*. On peut en effet penser que le développement à venir de chacune de ces approches puisse bénéficier à l'autre, et que les deux puisse connaître ainsi d'un essor conjoint.

Utilisation détournée de la synthèse binaurale: haut-parleurs virtuels

Les techniques binaurales offrent le moyen de restituer au casque l'effet d'une restitution sur haut-parleurs d'un enregistrement stéréophonique, multi-canal ou ambisonique, lorsque l'utilisateur ne possède pas le dispositif de haut-parleurs requis ou lorsqu'il ne veut pas l'utiliser pour des questions de discrétion par rapport à son entourage. La *méthode dite "des haut-parleurs virtuels"* consiste à simuler par filtrage binaural l'effet de chaque haut-parleur considéré comme une source virtuelle disposée autour de l'auditeur. Appliqué

40. En restreignant les fonctions spatiales aux harmoniques sphériques d'ordre 0 et 1, le champ binaural intermédiaire issu de l'encodage est représenté par un double format B (un pour chaque oreille).

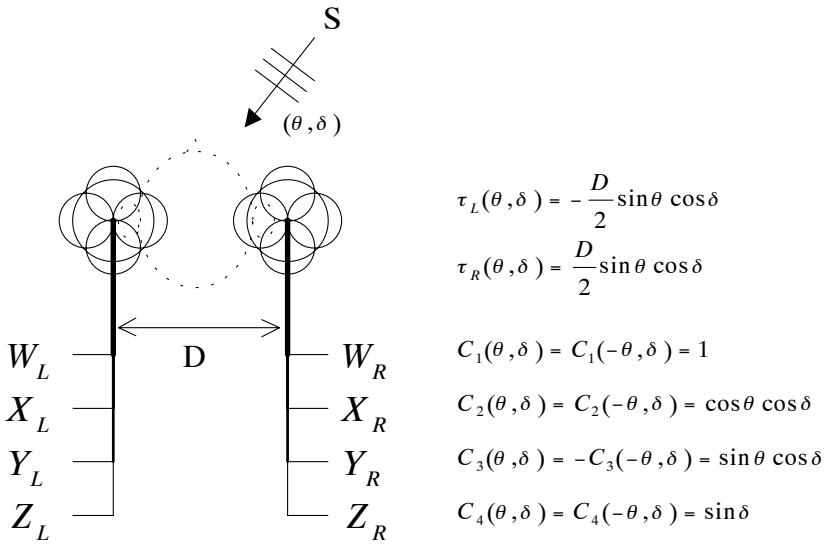

FIG. 2.18 – *Encodage acoustique au binaural B-format, à l'aide de deux microphones Soundfield espacés d'un peu plus d'un diamètre de tête et en l'absence de tête (prise de son en champ libre). Noter la légère différence des fonctions spatiales (absence du facteur $\sqrt{2}$) avec les conventions d'encodage décrites par (2.14).*

à la restitution *surround* ou multi-canal, ce procédé est connu sous le nom commercial de *Virtual Surround*. Dans ce type de contexte, un effet de salle – salle virtuelle de restitution – est en général associé aux haut-parleurs virtuels afin de favoriser l'extériorisation des images sonores. Par l'intermédiaire de techniques de type transaural décrites plus loin (annulation du cross-talk), cette approche s'étend naturellement à l'écoute sur deux haut-parleurs. Nous nous intéresserons à ces deux modalités de restitution en complément de la restitution ambisonique traditionnelle sur haut-parleurs.

Dans un *contexte expérimental*, la méthode des haut-parleurs virtuels peut également se montrer très utile pour l'évaluation subjective et la comparaison de techniques de restitution sur haut-parleurs, qu'il s'agisse de différentes formes de pan-pot ou de restitution ambisonique (Chapitres 3 et 5). Elle offre la possibilité de tester des configurations virtuelles requérant un nombre de haut-parleurs qui pourrait être prohibitif en pratique, et d'effectuer des commutations rapidement d'une configuration à une autre, sans contrainte matérielle. Elle permet enfin une comparaison des différents rendus avec l'effet d'une source unique produit par synthèse binaurale directe, qui peut alors servir de référence. Dans un contexte d'évaluation, il est important de réaliser une synthèse "exacte" du champ binaural recomposé. Lorsque nous appliquons cette méthode en 4.1 ou en 5, nous optons donc pour une implémentation par filtrage convolutif rapide utilisant les HRIR d'origine, et aucun effet de salle n'est associé aux haut-parleurs virtuels (salle de restitution virtuelle anéchoïque). Cela nous permet de contrôler précisément les conditions de reconstruction du champ acoustique au niveau des oreilles, mais dans le même temps, cette méthode ne peut se substituer à l'expérience d'une véritable écoute sur haut-parleurs, même en chambre anéchoïque: pour s'en rapprocher, il faudrait au moins tenir compte des rotations de la tête à l'aide d'un *head-tracker*, ce dont nous ne disposons pas.

Conclusions

Les techniques binaurales ont pour l'instant été abordées en tant qu'approche de la spatialisation dédiée à une restitution au casque, et pour les formes de représentation compacte du champ acoustique qu'elles proposent.

La reproduction binaurale est supposée offrir d'excellentes conditions d'illusion sonore, notamment grâce à des indices de localisation (ITD, ILD et indices spectraux) adaptés à l'auditeur, et surtout plus consistants (non-détériorés) qu'avec les techniques de reproduction présentées jusqu'ici. Mais il s'avère qu'en pratique, la restitution au casque est fréquemment entachée de problèmes d'intériorisation et de renversement directionnel des images sonores, du fait que l'enregistrement ou la synthèse binaurale est bien souvent dédié(e) à un autre auditeur, et qu'il manque en général une condition naturelle de localisation: l'adaptation des indices aux mouvements de la tête. On verra que la restitution sur deux haut-parleurs après traitement transaural ne résout que partiellement ces problèmes. Cependant, la restitution au casque – et par extension sur deux haut-parleurs – étant très abordable d'un point de vue matériel, elle peut être envisagée, *via* le principe des haut-parleurs virtuels, comme un appendice à d'autres techniques de spatialisation originellement dédiées à une restitution sur haut-parleurs. C'est une modalité que nous prendrons en compte pour la restitution ambisonique, dans un cadre applicatif comme dans un contexte d'évaluation.

2.5.2 Restitution sur deux haut-parleurs: *Transaural, Stereo-Dipole*, etc.

On recense plusieurs techniques qui ont la même vocation, à l'aide de deux haut-parleurs, de reconstruire au niveau des oreilles des signaux binauraux qui correspondraient à une expérience d'écoute naturelle (enregistrement binaural), ou bien encore des signaux quelconques. Ayant pour objet principal d'annuler les chemins croisés (*cross-talk*) entre les haut-parleurs et les oreilles (Figure 2.19), elles sont souvent désignées par le nom générique de "techniques transaurales", bien que le terme *Transaural* soit une marque déposée qui correspond à une technique particulière, les autres connues étant le *Stereo-Dipole* et la *Spectral-Stereo*. Outre un rappel succinct de la définition du filtrage transaural, l'objet des paragraphes suivants est de donner quelques éléments d'interprétation sur les questions de stabilité et de robustesse de la restitution sur haut-parleurs, notamment lorsque l'objectif est de produire une image sonore localisée (pan-pot transaural).

Définition

Le principe de l'annulation du *cross-talk* a connu ses premières tentatives pratiques avec Schroeder et Atal (procédé TRADIS) [SA63], mais c'est également à Cooper et Bauck [CB89] que l'on associe la définition formelle du procédé *transaural*. La formulation mathématique du problème et sa résolution sont ici succinctement rappelées[41].

Les fonctions de transfert de chaque oreille vers chaque haut-parleur sont notées H_{LL}, H_{LR}, H_{RR} et H_{RL} et exprimées dans le domaine fréquentiel (Figure 2.19, gauche). Le phénomène de *cross-talk* peut alors être assimilé à une opération de filtrage croisé qui transforme les signaux $\hat{X}_L$ et $\hat{X}_R$ émis par les haut-parleurs en signaux y_L et y_R reçus par les oreilles, opération qui s'écrit sous la forme mathématique suivante:

$$\mathbf{y} = \mathbf{H}\,\hat{\mathbf{x}}, \qquad \text{avec:} \quad \mathbf{y} = \begin{bmatrix} y_L \\ y_R \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}, \quad \hat{\mathbf{x}} = \begin{bmatrix} \hat{x}_L \\ \hat{x}_R \end{bmatrix} \tag{2.32}$$

Très souvent, les fonctions de transfert H_{AB} sont assimilées aux HRTF associées aux directions des haut-parleurs, ce qui suppose en particulier qu'on ne tient pas compte de l'effet de la salle sur la transformation

41. Les lignes suivantes sont très largement inspirées de [Gar95]

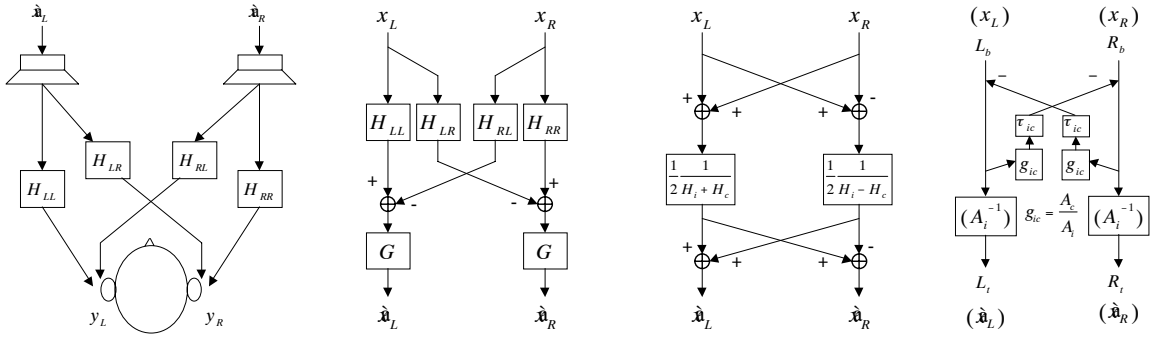

FIG. 2.19 – *Modélisation du phénomène de* cross-talk *par les fonctions de transfert des haut-parleurs vers les oreilles (à gauche), et schémas de filtres transauraux pour son inversion: au centre-gauche, une structure générique (où $G = 1/(H_{LL}H_{RR} - H_{LR}H_{RL})$); au centre-droit, une structure dite "*Shuffler*", plus économe, tirant profit des propriétés de symétrie ($H_i = H_{LL} = H_{RR}$ et $H_i = H_{LL} = H_{RR}$). Ces trois premiers schémas sont reproduits d'après [Gar95]. A droite enfin, inversion directe du cross-talk par une structure récursive (peu utilisée en pratique): chaque contribution contralatérale est annulée par la contribution ipsilatérale venant du haut-parleur opposé, moyennant une réinjection croisée, retardée (τ_{ic}), pondérée par $g_{ic} = \frac{A_c}{A_i}$ avec changement de signe.*

des signaux. Il est aussi envisageable de faire appel à des filtres basés sur un modèle simplifié de la diffraction par la tête – modèle de tête sphérique, par exemple – [RJ93] [WF95]. Disposant de signaux binauraux x_L et x_R qu'on aimerait restituer respectivement à chaque oreille, le problème à résoudre consiste à définir les signaux $\hat{x}_L$ et $\hat{x}_R$ à délivrer aux haut-parleurs[42] de sorte que la transformation (2.32) produise aux oreilles des signaux identiques y_L et y_R aux signaux x_L et x_R. Il s'agit donc d'inverser le système (2.32), ce qui ne pose pas de grande difficulté d'un point de vue formel:

$$\hat{\mathbf{x}} = \mathbf{H}^{-1}\,\mathbf{x}, \qquad \text{avec} \quad \frac{1}{H_{LL}H_{RR} - H_{LR}H_{RL}} \begin{bmatrix} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{bmatrix} \tag{2.33}$$

Cette opération de *filtrage transaural* est schématisée Figure 2.19 (centre gauche).

La formulation (2.33) est généralement simplifiée par des considérations de symétrie, le système étant défini en pratique pour une tête placée sur l'axe médian des haut-parleurs et orientée suivant cet axe. Il ne subsiste donc que deux fonctions de transfert distinctes $H_i = H_{LL} = H_{RR}$ et $H_c = H_{LR} = H_{RL}$. Ces propriétés de symétrie permettent une mise en oeuvre plus économique du filtrage transaural. Grâce à une structure dite "*shuffler*" [CB89], les opérations de filtrage ne nécessitent plus que deux filtres, qui s'appliquent à la somme et la différence des signaux x_L et x_R (Figure 2.19, droite). En termes mathématiques, la matrice inverse $\mathbf{H}^{-1}$ est diagonalisée par la matrice de mélange $\mathbf{D}$:

$$\mathbf{H}^{-1} = \frac{1}{H_i^2 - H_c^2} \begin{bmatrix} H_i & -H_c \\ -H_c & H_i \end{bmatrix} = \mathbf{D} \begin{bmatrix} \frac{1}{H_i + H_c} & 0 \\ 0 & \frac{1}{H_i - H_c} \end{bmatrix} \mathbf{D}^{-1}, \qquad \text{où} \quad \mathbf{D} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.34}$$

Lorsque les filtres transauraux sont calculés par application pure et simple des formules d'inversion dans le domaine fréquentiel et du fait de la complexité des HRTF, la causalité des réponses impulsionnelles qui

42. En toute rigueur, il est nécessaire de prendre en compte les fonctions de transfert électro-acoustiques des transducteurs. Par souci de simplicité, et parce que cela ne change en rien le principe décrit, on confond ici le signal électrique délivré au haut-parleur et le signal acoustique qu'il émet.

leur correspondent dans le domaine temporel n'est pas garantie [KNH97]. Pour cette raison, un paramètre de régularisation est en général introduit pour le calcul des filtres inverses (inversion au sens des moindres carrés).

En pratique, le filtrage transaural est surtout destiné à être utilisé en aval d'une synthèse binaurale ou d'un enregistrement binaural, bien qu'il puisse dans le principe être appliqué à des signaux x_L et x_R quelconques. Les performances du système sont d'ailleurs souvent jugées d'après sa capacité à éliminer purement et simplement le *cross-talk*, le test-type consistant à annuler complètement le signal reconstruit à l'une des deux oreilles ($y_R = 0$ par exemple). Toutefois, la seule préoccupation d'annuler le cross-talk sans porter attention à la nature de l'image sonore reproduite, ne permet pas une interprétation très pertinente des efforts globalement investis par le système – l'énergie totale mise en jeu – et des problèmes de stabilité de l'illusion auditive, qui dépendent en particulier de la dissimilitude spatiale entre la source virtuelle et les sources réelles (haut-parleurs).

Qualités générales

Les techniques transaurales apportent une amélioration très appréciable à la qualité de l'illusion sonore, comparées par exemple aux techniques stéréophoniques conventionnelles qui utilisent le même dispositif de haut-parleurs: la reconstitution des indices spectraux associés aux différentes sources virtuelles – encodées dans le matériel binaural, en aval – participent à leur différenciation et précise leur qualité de localisation dans l'espace sonore restitué; mais surtout, l'élimination du mélange (diaphonie) des informations binaurales *préserve la cohérence et la consistance des différences interaurales* sur une très large bande de fréquence, condition essentielle à une qualité d'image naturelle et à un effet d'immersion au sein de la scène sonore, laquelle n'est plus limitée au secteur angulaire des haut-parleurs. De plus, l'*extériorisation des images sonores* est naturellement acquise, du fait que les contributions sonores réelles sont effectivement extérieures et subissent le filtrage par les oreilles (pavillons) propres à l'auditeur, au contraire d'une restitution au casque. Pour ces dernières raisons – consistance préservée des différences interaurales et extériorisation naturelle –, le problème des spécificités individuelles des HRTF associées à la production du matériel binaural original, ne se posent plus de façon aussi critique que pour une restitution au casque, où le risque d'intériorisation est présent. En revanche, l'usage de seulement deux haut-parleurs frontaux pour un contrôle de la reconstruction en deux points – les oreilles –, impose à l'auditeur de fortes contraintes de position, et peut être responsable d'un renversement frontal des images arrière[43].

Propriétés "signal" du système - Vers le *Stereo-Dipole*

La définition purement formelle et mathématique qui vient d'être donnée du filtrage transaural ne permet pas à elle seule de commenter les propriétés du système sous des aspects "traitement du signal", et encore moins de se faire une idée sur la robustesse des images sonores aux déplacements de l'auditeur, sur le plan de la restitution.

L'inversion directe du cross-talk par la structure décrite Figure 2.19(droite) met en évidence la nature récursive du filtrage transaural, qui reste présente intrinsèquement [KNH97] même lorsque le filtre est implémenté sous forme RIF[44]. Ce schéma se base sur une modélisation simplifiée des différences entre les contributions ipsi- et contra-latérales[45]: une différence temporelle τ_{ic} et un rapport d'amplitude $g_{ic} = \frac{A_c}{A_i}$, en écrivant $H_i(f) = A_i(f)$ et $H_c(f) = A_c(f)\,e^{-j\tau_{ic} f}$. τ_{ic} et g_{ic} dépendent en principe de la fréquence, mais de

43. Tendance inverse de la restitution au casque, mais pour des raisons complètement différentes: la perception frontale vient ici probablement plutôt des variations de l'ITD par légère rotation de la tête, en dépit d'une bonne restitution des indices spectraux.

44. ... ce qui est généralement le cas

45. Dans une implémentation très simplifiée [Gar95] suggérée par Griesinger, g_{ic} est modélisé par un filtre passe-bas d'ordre 1.

moins en moins à mesure que les haut-parleurs sont proches l'un de l'autre (*i.e.* que ϕ_F diminue). Du fait que $g_{ic} = \frac{A_c}{A_c} < 1$ et que $\tau_{ic} > 0$, ce filtrage récursif est stable et causal (Atal et al [SA63]?). La structure récursive (Figure 2.19, droite) met en lumière le fait qu'un signal L_b ou R_b se voit réinjecté dans la même ligne avec un retard $2\tau_{ic}$ et avec une atténuation g_{ic}^2. Ainsi, chaque ligne gauche et droite prise individuellement se comporte comme un filtre en peigne, avec pour *fréquence de résonance* fondamentale $f_r = \frac{1}{\tau_{ic}}$, nommée *ringing frequency* dans [KNH97]. Cette résonance est particulièrement flagrante lorsque l'un des deux signaux d'entrée L_b ou R_b est nul (tâche d'annulation du cross-talk).

A la restitution, cet effet devient perceptible lorsque l'auditeur n'est pas correctement placé. Appliqué à des signaux impulsifs d'étalement temporel plus court que la période $2\tau_{ic}$ (ou d'énergie spectrale située au-dessus de f_r), le système a pour effet de produire un événement acoustique à répétition (dans le temps), donc également de le "dupliquer" dans l'espace, et l'on comprend que ce soit un facteur d'instabilité de la reconstruction au niveau des oreilles lorsque l'auditeur se déplace. Nous apportons dans la suite une interprétation plus complète de la stabilité de la reconstruction à partir de considérations spatiales sur le champ acoustique synthétisé, lesquelles permettront d'établir un lien étroit entre la fréquence dite "de cross-talk" (déjà introduite en 2.2) et la fréquence de résonance f_r.

Pour favoriser la stabilité du système et de la restitution, il paraît donc logique de chercher à élever cette fréquence f_r, donc diminuer τ_{ic} [KNH97]. Cela conduit directement à diminuer l'angle $2\phi_F$ entre les haut-parleurs, partant de la relation approximative $\tau_{ic} \approx D \sin\phi_F$, D désignant le diamètre de la tête. Ce sont ces arguments qui ont donné naissance à la technique *stereo-dipole*, qui doit son nom au fait qu'en rapprochant les deux haut-parleurs, ils tendent à se comporter comme la combinaison d'une source monopolaire et d'une source dipolaire, vu de l'auditeur. Le choix d'angle $\phi_F = 5°$ permet d'élever la fréquence de résonance – et de cross-talk – au-delà de 10 kHz. Par ailleurs, le rapprochement des haut-parleurs étant tel quel les réponses ipsi- et contra-latérales deviennent de plus en plus semblables, l'inversion du cross-talk devient peu dépendante des spécificités morphologiques de l'auditeur jusqu'à une fréquence plus élevée.

En contrepartie, le fait que τ_{ic} diminue se traduit par une amplification particulièrement importante des basses fréquences, ce qui doit être pris en compte au moment de l'implémentation pratique du système[46] [KNH97]. En simplifiant la modélisation du cross-talk (g_{ic} et τ_{ic} constants) et en posant $R_b = 0$ et $L_b = 1$ par exemple, les signaux issus du filtrage s'expriment dans le domaine fréquentiel sous la forme [KNH97]:

$$\begin{aligned} L_t &= \frac{1}{1 - g_{ic}^2 e^{-j2\omega\tau_{ic}}} \quad , \qquad \omega = 2\pi f \\ R_t &= -g_{ic} e^{-j\omega\tau_{ic}} L_t \end{aligned} \tag{2.35}$$

Considérant que g_{ic} est très proche de 1 au moins en basse-fréquence, on peut donner une mesure globale de l'énergie "mise en jeu":

$$E = L_t^2 + R_t^2 = \frac{1}{1 - \cos(2\tau_{ic}\omega)} \approx \frac{1}{1 - \cos(3D \sin\phi_F \omega / c)} \underset{\omega \to 0}{\rightarrow} \frac{2}{\left(\frac{3D}{c}\omega \sin\phi_F\right)^2} \tag{2.36}$$

Attention, il ne s'agit pas ici d'une énergie acoustique mesurable dans le champ synthétique! En effet, les haut-parleurs étant en pratique peu distants (typiquement quelques centimètres pour le stereo-dipole) par rapport à la longueur d'onde en basse-fréquence (par exemple 3,4 m pour 100 Hz), le déphasage entre les deux ondes émises varie peu dans l'espace. Puisque l'amplification basse-fréquence s'accompagne dans notre cas

C'est ce que nous adoptons en 5.2.1.

46. C'est ce qui a encouragé Bauck à proposer un système d'inversion de cross-talk dédoublé: une paire de haut-parleurs (*tweeters*) très rapprochés ($\phi_F = \pm 3°$) est dédiée au traitement haute-fréquence, alors que le traitement basse-fréquence est pris en charge par des haut-parleurs plus espacés ($\phi_F = \pm 20°$) [Bau97].

d'une quasi-opposition de phase entre les contributions, leur interférence reste d'énergie modérée dans tout l'espace. Si par contre les haut-parleurs étaient très lointains (modèle d'ondes planes), l'effet d'amplification basse-fréquence devrait être perceptible à quelque distance de l'axe médian, par effet de déphasage entre les deux ondes. La mesure (2.36) ne nous est donc utile que pour indiquer la puissance que les transducteurs et le système d'amplification en amont doivent supporter.

Ainsi, pour des angles ϕ_F de 30°, 10° et 5°, le gain d'énergie est d'environ 3 dB, 12 dB, 18 dB pour $f = 200$ Hz, de 9 dB, 18 dB, 24 dB pour $f = 100$ Hz et de 15 dB, 24 dB, 30 dB pour $f = 50$ Hz. Il ne cesse d'augmenter lorsque la fréquence diminue. Ce problème est beaucoup moins marqué lorsque la tâche du système est de produire l'effet d'une source virtuelle, auquel cas l'effet d'amplification est borné en basse-fréquence pour une direction de source virtuelle donnée.

Propriétés spatiales du champ reproduit - Interprétation géométrique

Pour l'observation des phénomènes acoustiques, nous reprenons ici le modèle d'ondes planes pour décrire les contributions des haut-parleurs, bien qu'en pratique ceux-ci puissent être relativement proches de l'auditeur. Limitant nos observations au voisinage de la tête, on peut considérer que *les ordres de grandeur* que nous tirons de ce modèle simplifié restent pertinents (Figure 2.20). D'après [DRP99] et ainsi qu'il a été rappelé en 2.2, on observe, dans le domaine fréquentiel, une figure d'interférence de *caractéristiques d'énergie et de propagation invariantes parallèlement à l'axe médian* $(O,\vec{x})$ *et présentant une périodicité spatiale suivant l'axe transversal* $(O,\vec{y})$ pour chaque fréquence f. La période spatiale $\Lambda_y(f) = \lambda/(2\sin\phi_F)$ se trouve être inversement proportionnelle à la fréquence $f = c/\lambda$ (Figure 2.20).

Cette description géométrique du champ d'interférence fournit un premier indice pour caractériser **la robustesse de l'image sonore par déplacement latéral de la tête**: la *largeur ou période spatiale* Λ_y, paramètre qui traduit partiellement la variabilité des caractéristiques du champ selon $(O,\vec{y})$, à une fréquence f donnée. On comprend intuitivement que plus Λ_y est grand – pour la fréquence f donnée –, plus l'auditeur peut déplacer sa tête latéralement sans constater de variation excessive des informations sonores reconstruites à ses oreilles, au moins jusqu'à une fréquence de tolérance donnée. Mais pour caractériser la robustesse de l'image sonore par déplacement latéral, il faut aussi tenir compte d'un deuxième paramètre de variabilité: *l'amplitude des variations* des caractéristiques du champ, qui peut être observée à travers le champ d'énergie [DRP99]. Cette amplitude est d'autant plus grande que les deux contributions sont d'énergie semblable. Dans un contexte de création d'une source virtuelle, cela se produit surtout lorsque la direction de la source virtuelle est éloignée de celles des sources réelles – les haut-parleurs –, comme il est commenté par la suite.

L'intérêt de l'approche *stereo-dipole* se manifeste ici de façon évidente: en diminuant l'angle ϕ_F jusqu'à 5°, la largeur Λ_y augmente considérablement, de sorte que la figure d'interférence englobe la tête jusqu'à une fréquence f_{XT} beaucoup plus élevée. En première approximation, la fréquence de cross-talk f_{XT} peut être définie avec l'hypothèse d'une tête acoustiquement transparente, de sorte que $\Lambda_y(f_{XT}) = \frac{c}{2f_{XT}\sin\phi_F} = D$, soit $f_{XT} = \frac{c}{2D\sin\phi_F}$. En tenant compte du contournement de la tête par effet de diffraction, on constate que f_{XT} s'identifie avec la fréquence de résonance f_r du filtre transaural: $f_{XT} = \frac{1}{2\tau_{ic}} = f_r$. Pour information, $f_{XT} \approx 11$ kHz pour $\phi_F = 5°$ contre $f_{XT} \approx 1{,}9$ kHz pour $\phi_F = 30°$. Ce qui était un caractéristique du signal définit ici une propriété spatiale de l'événement acoustique synthétisé: lorsque le contenu spectral est situé en dessous de la fréquence f_{XT}, le phénomène acoustique généré est en quelque sorte non-fragmenté à l'échelle de la tête.

Quand Λ_y est assez grand par rapport à D, c'est-à-dire pour les fréquences petites devant f_{XT}, les caractéristiques du champ synthétique varient peu à l'échelle de la tête. La prédiction des indices et de l'effet de localisation d'après la caractérisation locale par le vecteur vélocité $\vec{V}(\vec{r} = 0)$ doit donc être applicable, dans des limites qui sont définies plus loin. Lorsque l'objectif est de créer une image sonore dans une di-

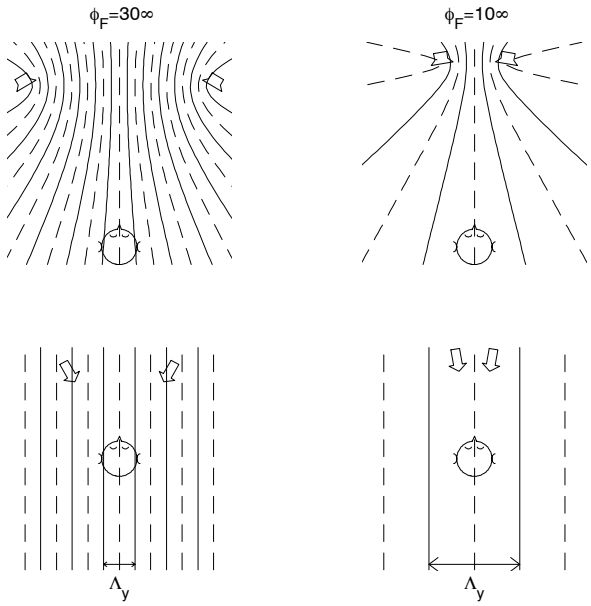

FIG. 2.20 – *Structure géométrique de l'interférence entre deux ondes monochromatiques de fréquence* $f = 2000$ *Hz et d'incidences* $\pm\phi_F$ *par rapport à l'auditeur, avec* $\phi_F = 30°$ *à gauche et* $\phi_F = 10°$ *à droite. En haut: ondes sphériques émises par deux haut-parleurs, éloignés de 1 m par rapport à l'auditeur. Les courbes sur lesquelles les ondes ont une différence de phase (ou encore une différence de marche) constante décrivent des hyperboles dont les haut-parleurs sont les foyers: en tirets, la différence de marche est multiple de la longueur d'onde* λ*; en traits continus, elle vaut* $n\lambda + \lambda/2$*, n entier. En bas: les ondes interférentes sont supposées planes (haut-parleurs lointains). La figure d'interférence centrale est délimitée par deux droites (tracé continu) et est périodique suivant l'axe transversal* $\vec{y}$*, de période* $\Lambda_y = \frac{\lambda}{2\sin\phi_F}$*. La largeur* Λ_y *de l'interférence à hauteur de l'auditeur est semblable en haut et en bas.*

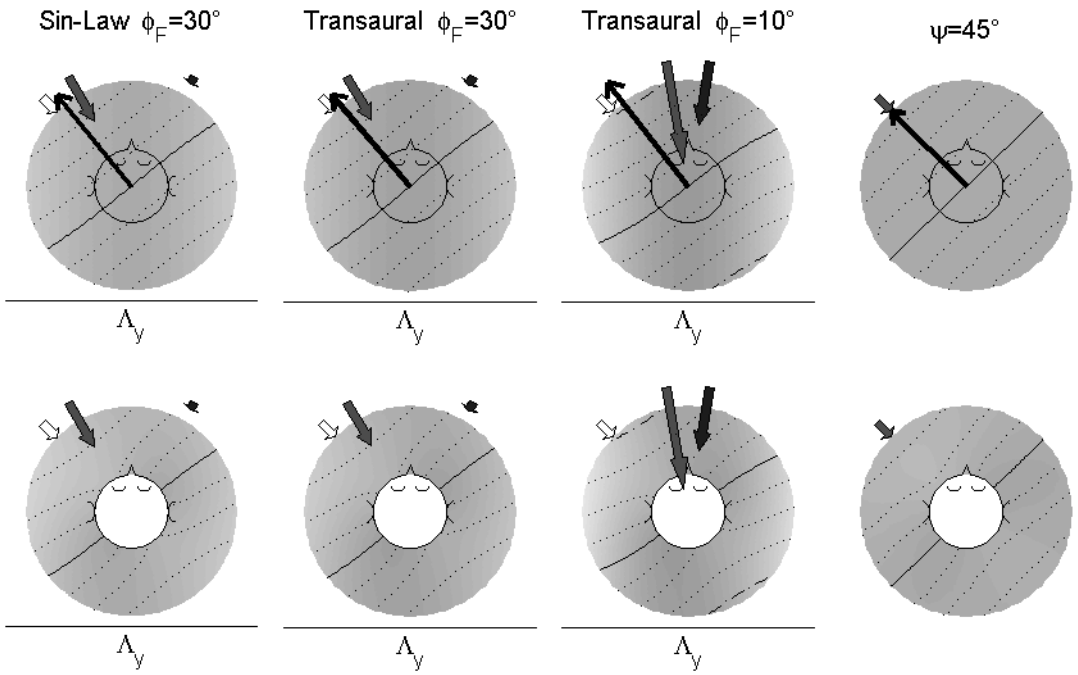

FIG. 2.21 – *Aperçu du champ acoustique synthétisé à la fréquence 300 Hz, pour la restitution d'une source virtuelle placée en* $\psi = 45°$ *à l'aide de deux haut-parleurs (ondes planes). Comparaison avec l'événement original (à droite). En haut, champ libre (tête en transparence) avec caractérisation locale par le vecteur vélocité (flèche centrée). En bas, reconstruction du champ avec diffraction de la tête (modèle sphérique). Champ d'amplitude en niveau de gris (faible amplitude en sombre). L'écart de phase est de* $\pi/8$ *entre 2 courbes iso-phase (en trait continu et en pointillés) successives.*

rection donnée, il s'avère ainsi que dans les très basses fréquences, le procédé de restitution transaural ou *stereo-dipole* s'apparente au pan-pot d'amplitude basé sur la loi des sinus (2.1) pour une tête fixe. Cette convergence basse-fréquence est particulièrement visible avec le stereo-dipole ($f_{KT} > 10$ kHz) [KNH97]. L'objet des paragraphes suivants est de fournir quelques éléments d'interprétation complémentaires concernant la stabilité et le comportement des images par mouvements de la tête – y compris de rotation –, en s'appuyant sur l'illustration des phénomènes acoustiques synthétisés à l'échelle de la tête.

Similitudes et divergences avec la loi des sinus - Interprétation complémentaire

Le comportement des systèmes et de la restitution est maintenant commenté plus spécifiquement lorsque la tâche est la création de l'effet d'une source acoustique (image d'une source virtuelle). Nous parlons alors de "pan-pot transaural" pour décrire les étapes condensées de synthèse binaurale et de filtrage transaural, en y assimilant l'approche *stereo-dipole*. A une direction ψ de source virtuelle est ainsi associée une paire de gains complexes $G_L(\psi, f)$ et $G_R(\psi, f)$ dépendant de la fréquence f. Les HRTF utilisées ici sont basées sur un modèle de tête sphérique et le filtrage transaural est défini par inversion du cross-talk dans le domaine fréquentiel. Pour cette raison et par commodité pour le calcul de diffraction, l'observation des phénomènes est faite dans le domaine fréquentiel (champs monochromatiques).

La figure 2.21 confirme la convergence des comportements des différents systèmes – pan-pot transaural et pan-pot conventionnel – en basse-fréquence. Cela nous permet d'enrichir l'interprétation de la restitution

transaurale dans ce domaine. La caractérisation de la propagation de phase par le vecteur vélocité fournit une prédiction judicieuse de l'effet de localisation. Il est facile de vérifier que lorsque la tête est fixe, c'est bien la projection du vecteur vélocité $\vec{V}$ sur le cercle unité (contour du disque, Figure 2.21) parallèlement au plan médian de l'auditeur qui détermine la position latérale apparente de la source ψ. Par légère rotation de la tête, c'est la direction $\vec{V}$ du front d'onde qui est susceptible d'être détectée. Il faut souligner que si $r_V > 1$, ce qui est le cas quand $|\psi| > |\phi_F|$, donc bien souvent avec le stereo-dipole (ϕ_F petit), les variations de l'ITD lors de la rotation ne sont pas naturelles (elles sont excessives), alors qu'elles peuvent correspondre à un effet de hauteur $\delta = \arccos r_V$ lorsque $r_V < 1$ ($|\psi| < \phi_F$). Par application de (2.1) et de (2.3), on précise facilement la loi de r_V:

$$r_V(\phi_F,\psi) = \sqrt{\cos^2\phi_F + \sin^2\psi} \tag{2.37}$$

Ainsi, la valeur maximale de r_V est $\sqrt{2} \approx 1{,}414$, observée pour une source parfaitement latérale ($\psi = 90°$) restituée avec le stereo-dipole[47]. Une autre conséquence des variations de l'ITD par rotation de la tête est le repliement frontal des images censées être situées dans le demi-espace arrière: c'est toujours un front d'onde venant du demi-espace avant qui est produit et détecté.

A titre indicatif, la coïncidence avec la loi des sinus (2.1) permet également de préciser le gain d'énergie $E^{BF} = G_L^2 + G_R^2$ "mis en jeu"[48] en basse-fréquence en fonction de ψ et de ϕ_F. La préservation de l'amplitude du front d'onde synthétisé localement ($G_L + G_R = 1$) implique directement:

$$E^{BF} = G_L^2 + G_R^2 = \frac{1}{2}\left(1 + \left(\frac{\sin\psi}{\sin\phi_F}\right)^2\right) \tag{2.38}$$

Il est clair que $E^{BF} < 1$ quand $|\psi| < \phi_F$ et que $E^{BF} \to \infty$ pour une direction $|\psi| > \phi_F$ donnée quand $\phi_F \to 0$. Mais contrairement à la tâche d'annulation du cross-talk, l'énergie mise en jeu reste bornée en fonction de la fréquence: (2.38) en donne à la fois la borne supérieure et la tendance basse-fréquence. Il est intéressant de poursuivre la comparaison entre les différents choix de ϕ_F (30°, 10°, 5°): pour une source virtuelle placée en $\psi = 45°$, ils impliquent respectivement des gains d'énergie de 1,7 dB, 9,4 dB et 15,2 dB, contre 4 dB, 12,3 dB et 18,2 dB pour une source en $\psi = 90°$. Le contraste entre le dispositif conventionnel $\phi_F = 30°$ et le *stereo-dipole* $\phi_F = 5°$ est flagrant!

Pour compléter l'analyse du comportement de l'image sonore lors d'une rotation de la tête, il faut étendre l'observation aux hautes fréquences, et finalement s'intéresser aux aspects temporels des phénomènes acoustiques. On pourrait montrer qu'à mesure que la fréquence augmente, les lois de pan-pot transaural s'éloignent de la loi des sinus, et ce d'autant plus vite que la direction de la source virtuelle est distincte de celle des haut-parleurs. Dans le même temps, l'observation de la propagation de phase perd de son intérêt: même avec le *stereo-dipole* et en-deçà de la fréquence de cross-talk, la direction du front d'onde monochromatique synthétisé à l'échelle de la tête se met à fluctuer autour de la direction frontale en fonction de la fréquence. Il devient alors plus pertinent de s'intéresser à la propagation de groupe. L'illustration des phénomènes de propagation dans le domaine temporel offerte dans [KNH97] (cas d'une impulsion de Hanning) apporte déjà des indications intéressantes: la direction de propagation de groupe s'apparente plus celle du haut-parleur le plus proche de la source virtuelle qu'à celle de la source virtuelle elle-même ($\psi = 45°$), ce qui laisse supposer que par rotation de la tête, le système perceptif *pourrait objectivement "détecter"* une direction et une vitesse de propagation anormale par rapport à l'illusion perçue en position statique, de façon similaire à nos observations basse-fréquence basée sur la propagation de phase.

47. Notons que le front d'onde synthétisé pour cela a pour angle d'incidence $\theta_V = 45°$ seulement!

48. Indication de la puissance que doit supporter le système (ampli+enceintes), et non mesure de l'énergie du champ acoustique synthétisé! Cf page 128.

Quoiqu'il en soit, il semble les indices statiques de localisation offerts par le système *stereo-dipole* soient suffisamment robustes et consistants sur une large bande de fréquence, pour que la force de l'illusion sonore survive à leurs variations anormales lors de petites rotations de la tête, bien que l'on puisse s'attendre à un effet de déplacement des sources virtuelles localisées.

Synthèse et conclusion

Après une présentation formelle du principe transaural et le rappel de quelques propriétés "signal" liées à la tâche d'annulation du cross-talk, quelques éléments d'interprétation complémentaires ont été donnés sur le comportement du système et sur la robustesse des images sonores produites selon la position et les mouvements de la tête. A partir de propriétés géométriques (demi-angle ϕ_F entre les haut-parleurs) et d'une caractérisation des phénomènes acoustiques synthétisés (vecteur vélocité $\vec{V}$) qui reste pertinente dans un domaine basse-fréquence, deux dispositifs typiques ont pu être comparés.

Des considérations géométriques simples semblent indiquer que la robustesse par translation latérale de la tête est proportionnelle à $1/\sin\phi_F$ en première approximation. D'après ce raisonnement, une reproduction utilisant la configuration *stereo-dipole* ($\phi_F = 5°$) serait presque 6 fois plus stable qu'avec une configuration conventionnelle ($\phi_F = 30°$). En contrepartie, l'effet d'amplification basse-fréquence que doit assumer le système d'un point de vue électro-acoustique – pour les sources virtuelles extérieures à l'angle des haut-parleurs et surtout pour la tâche d'annulation du cross-talk – est beaucoup important avec le stereo-dipole[49]. En principe, cet effet d'amplification ne se reporte pas (ou que peu) sur le champ acoustique synthétisé même à distance de la position d'écoute idéale (*sweet-spot*), puisque la faible distance entre les haut-parleurs limite le déphasage possible pour les grandes longueur d'ondes. L'effet de coloration, qui est une des formes de dégradation de l'image sonore attendues en position d'écoute non-idéale, risque de se manifeste principalement à travers les résonances du filtre transaural, approximativement aux fréquences multiples de la fréquence de *cross-talk* f_{XT}. Il est ainsi repoussé à une fréquence beaucoup plus élevée avec le *stereo-dipole* ($f_{XT} \approx 11$ kHz) qu'avec une configuration conventionnelle ($f_{XT} \approx 1{,}9$ kHz), quoique dans le premier cas, l'amplitude de la résonance soit également plus forte.

L'ensemble de ces arguments objectifs semble donner raison à l'engouement grandissant que connaît actuellement l'approche *stereo-dipole*.

Dans tous les cas, la restriction des moyens de reproduction à seulement deux haut-parleurs, mais aussi la limitation de la représentation intermédiaire du champ binaural à seulement deux canaux, ne permettent pas de satisfaire un comportement naturel des indices de localisation lors de légères rotations de la tête. Le caractère non-naturel des variations de l'ITD dépend à la fois de l'incidence de la source virtuelle et de la géométrie des haut-parleurs. Globalement un peu plus marqué avec le *stereo-dipole*, il peut se traduire par un déplacement des images latérales. Mais l'effet le plus regrettable est la tendance au repliement frontal des images arrière. Pour résoudre ce problème, l'extension du principe transaural requiert l'ajout de haut-parleurs à l'arrière de l'auditeur, et suppose un enrichissement de la représentation binaurale par l'introduction de nouveaux signaux, lorsque le système se base sur une représentation compacte intermédiaire. Il s'agit donc d'une évolution semblable à celle qui a été observée entre la stéréophonie conventionnelle et Ambisonic, par exemple. Cette extension est développée avec plusieurs variantes dans la section suivante. Nous y mettons en évidence que le compromis entre contrainte d'écoute et enrichissement des propriétés d'image sonore y prend l'allure d'un paradoxe, là où l'approche ambisonique se montre naturellement plus tolérante.

49. Cf note 46 en bas de page 128.

2.5.3 Avec quatre haut-parleurs: double-transaural et variantes

La problématique de restitution qui est maintenant traitée reste dédiée à un seul auditeur. Il n'est donc pas question de l'approche transaurale généralisée au sens de Cooper et Bauck [BC92], qui met aussi en jeu plus de deux haut-parleurs, mais est dédiée à plusieurs auditeurs partageant le même espace acoustique: la généralisation de l'inversion du *cross-talk* n'y a pas pour vocation d'offrir à chaque auditeur des propriétés d'images sonores plus complètes que ce qu'offrirait à chacun un système transaural simple, dans la mesure où elle reste basée sur une représentation du champ binaural à deux canaux.

Nous trouvons intéressant d'évoquer les systèmes de type double-transaural pour deux raisons: d'une part, ils représentent par rapport au transaural (et *stereo-dipole*) une évolution similaire à celle observée entre la stéréo ΔI conventionnelle et *Ambisonic*; d'autre part, on y trouve un mode de restitution potentiel du *binaural B-format* (cf 2.5.1), dont on peut pressentir un essor conjoint avec celui des techniques ambisoniques qui nous intéressent plus particulièrement. Cela étant, la lecture de cette section n'est pas indispensable à la compréhension du reste du document.

Systèmes existants

Pour résoudre le problème de l'instabilité et du repliement frontal des sources virtuelles arrière, inhérent à la restitution sur deux haut-parleurs frontaux, il paraît logique de faire intervenir des haut-parleurs arrière. Quelques stratégies susceptibles de résoudre ce problème semblent se dégager jusqu'à présent, bien qu'elles soient d'un usage encore majoritairement expérimental à notre connaissance. Pour toutes, la restitution se fait sur quatre haut-parleurs, généralement disposés en rectangle (Figure 2.23) et plus rarement en trapèze (Figure 2.24). Ces approches peuvent être décrites succinctement comme suit:

Double-Transaural[50]. Le *Double-Transaural* consiste tout simplement à dédoubler l'utilisation du principe transaural sur une paire frontale et une paire arrière de haut-parleurs ($\pm 30^\circ$), en faisant prendre en charge la spatialisation des sources situées dans le demi-espace avant (resp. arrière) par la paire de haut-parleurs avant (resp. arrière). Une transition douce (*fading*) est assurée entre les deux modes de restitution lors du déplacement d'une source d'un demi-espace à l'autre, afin d'éviter une commutation brutale entre la paire avant et la paire arrière. Ce procédé est appliqué avec succès, semble-t-il, comme l'une des modalités de restitution du *Spatialisateur* de l'Ircam. Notons que la séparation dichotomique en deux demi-espaces sonores ne peut découler d'une simple prise de son naturelle, et réserve cette approche à un contexte de spatialisation artificielle.

Les trois approches suivantes reposent sur des systèmes de prise de son naturelle (Figure 2.22), chacun d'entre eux réalisant à sa manière un encodage du champ acoustique au voisinage des oreilles.

Reproduction liée à une prise de son à quatre canaux par deux couples M-S (oreilles) à la surface d'une sphère (tête): utilisation détournée du microphone KFM360 de Bruck[51] [Bru96] (Figure 2.23). Ce microphone fournit quatre signaux M_L, M_R, S_L et S_R qui correspondent aux mesures de la pression (M) et de son gradient suivant la direction frontale (S) à l'emplacement de chaque oreille de la "tête" sphérique (Figure 2.22, gauche). Le procédé classique de *shuffling* est appliqué pour dissocier les contributions avant (somme $\Sigma = M + S$) des contributions arrière (différence $\Delta = M - S$), produisant deux paires de signaux "binauraux" (L_F, R_F) et (L_B, R_B) (Figure 2.23). De cette manière, les sons originellement frontaux se retrouvent principalement dans les signaux L_F et R_F, et les sons arrière dans les signaux L_B et R_B. Dans la version originale du système de Bruck, ces signaux sont directement diffusés sur une paire avant et une paire arrière de

50. Procédé imaginé par des chercheurs de l'Ircam (équipe de Olivier Warusfel), dont nous ne connaissons pas actuellement de publication.

51. Microphone KFM360 fabriqué par Schoeps.

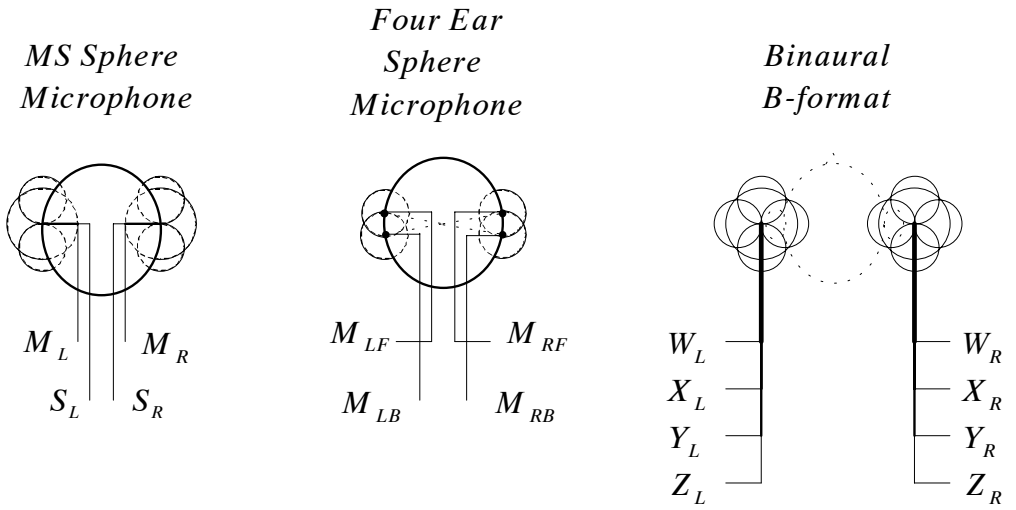

FIG. 2.22 – *Trois systèmes de microphones à la base des procédés de reproduction sur quatre haut-parleurs.*

haut-parleurs. Nous nous intéressons ici à une variante susceptible d'améliorer la qualité de l'image sonore produite, où chaque paire (L_F, R_F) et (L_B, R_B) subit un filtrage transaural adéquat (annulation du cross-talk) avant d'être diffusée sur les haut-parleurs, comme le montre la figure 2.23. Il est ainsi possible de reconstituer les signaux binauraux originaux $M_L = L_F + L_B$ et $M_R = R_F + R_B$ aux oreilles de l'auditeur. Si ces signaux sont porteurs de différences interaurales appréciables pour l'image sonore, ils pêchent en revanche par l'absence d'indices spectraux, la "tête" originale étant sphérique et ses "oreilles" dépourvues de pavillon. C'est pourquoi il est recommandé de baser la définition du filtrage transaural sur le modèle sphérique original pour éviter d'inverser les indices spectraux propres à l'auditeur. De cette façon, la coloration spectrale propre aux incidences arrière découle naturellement, quoique de façon partielle, de la simple présence de sources réelles – les haut-parleurs – à l'arrière de l'auditeur.

Reproduction multi-canal utilisant une tête artificielle à quatre oreilles (*Four Ear Sphere Microphone*) [KNKH97]. Cette méthode se base sur une prise de son (Figure 2.22) par quatre microphones (deux paires d'oreilles légèrement décalées d'avant en arrière) disposés à la surface d'une sphère (modèle de tête), la tâche du système de restitution consistant à délivrer les signaux adéquats aux haut-parleurs pour aboutir à la reconstruction contrôlée des quatre signaux captés, à leur emplacement d'origine sur la sphère. A l'échelle des basses fréquences, ce microphone fournit des informations équivalentes au microphone KFM360, la juxtaposition de deux microphones de chaque côté permettant de rendre compte approximativement du gradient de pression parallèlement à l'axe médian $(O, \vec{x})$. Le système reconstruit donc à la fois le champ de pression et son gradient au voisinage de chaque oreille – ce qui n'est pas la vocation initiale de l'approche précédente. A la restitution, les informations binaurales basse-fréquence perçues restent conformes, même par légère rotation de la tête, à l'événement acoustique original, ce qui préserve des effets d'inversion avant-arrière des sources perçues. Dans un domaine haute-fréquence en revanche, lorsque la longueur d'onde n'est plus assez grande par rapport à l'écart d'un "doublon d'oreilles" (3,3 cm), la reconstruction des signaux aux quatre oreilles n'implique plus automatiquement la recomposition d'un gradient de pression en adéquation avec l'effet attendu de la source virtuelle, et l'on peut s'attendre à un repliement alterné, en fonction de la fréquence, des participations avant-arrière des haut-parleurs (effet d'*aliasing* à partir de 5200 Hz).

Restitution transaurale du *binaural B-format* sur 4 haut-parleurs [JWL98] [JLP99]. Le *binaural B-format* est, rappelons-le (section 2.5.1), issu d'une double prise de son ambisonique, les deux microphones

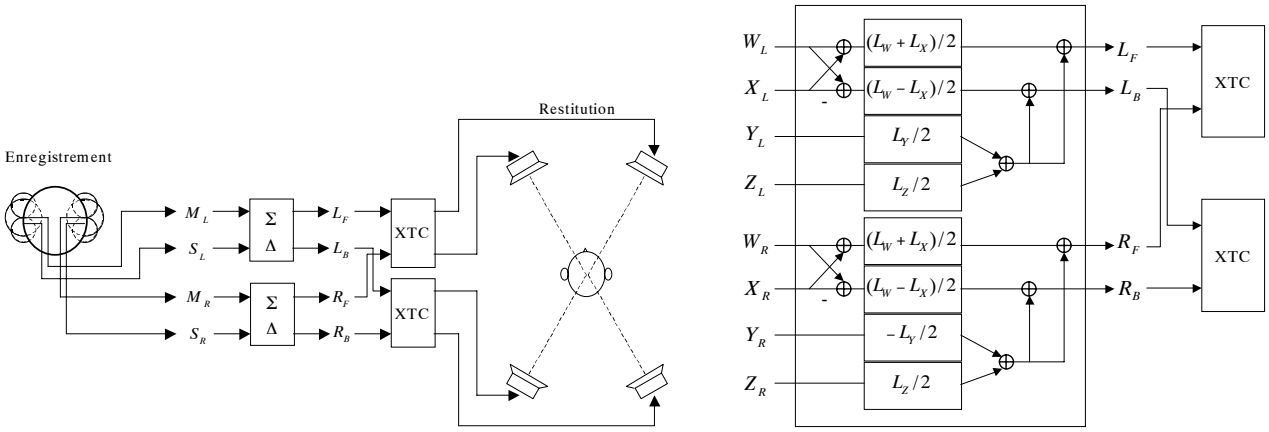

FIG. 2.23 – *A gauche: variante transauralisée du système de Bruck [Bru96]. Prise de son sur une "tête artificielle" sphérique, à l'aide de deux couples M-S situés à l'emplacement des oreilles, les capteurs de vélocité (signaux S_L et S_R) étant orientés vers l'avant. La répartition des signaux à traiter entre les paires de haut-parleurs avant et arrière, est réalisée au moyen de simples opérations de somme et de différence (*Σ *et* Δ*). Les deux paires de signaux binauraux obtenues subissent alors un filtrage transaural (XTC: pour* Cross-Talk Cancellation*) pour être finalement correctement restituées au niveau des oreilles. A droite: décodeur pour la reproduction transaurale du binaural B-format sur quatre haut-parleurs (d'après [JWL98]). Par analogie au système de gauche, le* shuffling *porte ici sur les couples* (W_L,X_L) *et* (W_R,X_R).

étant espacés approximativement d'un diamètre de tête (Figure 2.22). Des opérations de filtrage appropriées permettent de reconstituer *approximativement*, à partir de chaque groupe de composantes ambisoniques, les signaux binauraux qui auraient pu être mesurés aux oreilles d'un auditeur si celui avait pris place entre les deux microphones lors de la prise de son. Ce procédé est originellement dédié à une restitution binaurale ou encore, par extension, à une restitution transaurale sur deux haut-parleurs. La parenté frappante entre les composantes directionnelles X_L et X_R du *binaural B-format*, et les composantes S_L et S_R du système de Bruck [Bru96] évoqué plus haut, a permis aux inventeurs d'adapter le procédé décrit Figure 2.23 pour une restitution sur quatre haut-parleurs (Figure 2.23). Il a en plus l'aptitude de restituer plus fidèlement les indices spectraux dus à la présence du pavillon des oreilles, alors qu'avec la variante transaurale du système de Bruck présentée plus haut, seuls les indices spectraux correspondant aux incidences des haut-parleurs avant ou arrière son susceptibles d'être reproduits de façon judicieuse – parce que naturelle.

Commentaires et suggestions inspirées de la restitution ambisonique

L'objectif annoncé de ces quatre méthodes semble se réduire à la résolution du problème de repliement frontal des images arrière, et *vice-versa*. Un facteur d'amélioration est peut-être lié au fait que les indices spectraux des incidences arrière sont susceptibles d'être créés plus naturellement grâce à la présence des haut-parleurs arrière. On peut quand même objecter que les systèmes transauraux classiques sont en général tout à fait capables de produire ces indices spectraux, alors que l'approche *Four Ear Sphere Microphone* ne l'est pas ici à cause de son problème d'aliasing. Aussi la principale clé du succès de ces méthodes, lorsque celui-ci est validé, tient-elle probablement au fait que les indices de localisation – surtout l'ITD – varient désormais "dans le bon sens" lors d'une légère rotation de la tête, alors qu'avec le transaural simple, les variations de l'ITD ne peuvent correspondre qu'à l'effet d'incidences frontales (section 2.5.2).

Curieusement, la plupart des méthodes présentées ne semblent pas traiter jusqu'au bout cette question

du *contrôle des variations naturelles de l'ITD par légère rotation (yaw) de la tête*, alors qu'une restitution ambisonique sur un dispositif semblable assure naturellement cette propriété, tout au moins dans un domaine basse-fréquence. Pourtant les systèmes basés sur le KFM360 et sur le binaural B-format devraient pouvoir offrir ces conditions de variations naturelles en reconstruisant à la fois le champ de pression – M_L et M_R – et sa dérivée (ou son gradient) suivant $\vec{x}$ – S_L et S_R – au niveau de chaque oreille, tels qu'ils ont été mesurés à la surface de la sphère microphonique[52]. Mais telle n'est pas la vocation de la combinaison du *shuffling* et des filtres transauraux (Figure 2.23), qui a par ailleurs le mérite d'être relativement simple techniquement[53]. Au lieu de cela et à en croire [JLP99], le *shuffling* est tel que seule la paire de haut-parleurs avant participe à la création d'une image parfaitement frontale ($\theta = 0$), ce qui correspond à la synthèse d'un front d'onde de caractéristique $r_V = \cos\phi_F < 1$ susceptible d'être perçu avec un effet de hauteur exagéré ou une moindre acuité de localisation lors de légère rotation de la tête. Considéré dans un domaine basse-fréquence, le système basé sur la sphère à quatre oreilles tend quant à lui à recréer au niveau des oreilles un gradient de pression conforme à l'événement acoustique original – condition suffisante à des variations correctes de l'ITD –, mais il est mal conditionné pour la restitution haute-fréquence à cause de l'aliasing.

Une correction très simple du *shuffling* des systèmes de Bruck (version transauralisée) et de Jot *et al* est susceptible d'améliorer le naturel des images sonores compte-tenu de légères rotations de la tête. Nous nous inspirons pour cela des propriétés de la restitution ambisonique en basse-fréquence, et nous appuyons sur le constat que la prédiction de la localisation d'après le vecteur vélocité $\vec{V}$ s'applique aussi à la reproduction transaurale en basse-fréquence, comme il a été illustré précédemment (en 2.5.2). La méthode présentée ici peut être également exploitée pour le système *Double-Transaural* cité en premier, afin de définir de façon optimale la fonction de répartition (*fading*) des signaux à "transauraliser" entre la paire avant et la paire arrière de haut-parleurs. Considérons par exemple les signaux issus du *M-S Sphere Microphone* de Bruck, associés à une onde d'incidence θ. On peut alors établir de façon grossière les relations: $S_L = M_L \cos\theta$ et $S_R = M_R \cos\theta$. Le *shuffling* revient alors à répartir la prise en charge des signaux binauraux M_L et M_R entre les systèmes transauraux avant et arrière, avec des facteurs de pondération respectifs $G_F(\theta) = \alpha_F + \beta_F \cos\theta$ (opération de somme modifiée) et $G_B(\theta) = \alpha_B - \beta_B \cos\theta$ (opération de différence modifiée), où $\alpha_F = \beta_F = \alpha_B = \beta_B = \frac{1}{2}$ dans la version originale. L'opération de *shuffling* modifiée prend donc la forme suivante:

$$\begin{array}{ll} \Sigma & = \alpha_F\, M + \beta_F\, S \\ \Delta & = \alpha_B\, M - \beta_B\, S \end{array} , \tag{2.39}$$

applicable à chaque paire (M_L, S_L) et (M_R, S_R).

Considérée individuellement, chaque paire de haut-parleurs est apte à produire l'effet de latéralisation statique (pour une tête fixe) recherché. Pour chacune, cela se traduit par la synthèse d'un front d'onde caractérisé en basse-fréquence par le vecteur vélocité[54], nommé $\vec{V}_F$ pour la paire avant et $\vec{V}_B$ pour la paire arrière (Figure 2.24). Ces vecteurs vérifient $\vec{V}_F = \cos\phi_F\, \vec{u}_x + \sin\theta\, \vec{u}_y$ et $\vec{V}_B = -\cos\phi_B\, \vec{u}_x + \sin\theta\, \vec{u}_y$. Il est clair que la combinaison des deux fronts d'onde, un front d'onde synthétique global caractérisé par $\vec{V} = \frac{G_F \vec{V}_F + G_B \vec{V}_B}{G_F + G_B}$, produit le même effet de latéralisation statique: $\vec{V}.\vec{u}_y = \vec{V}_F.\vec{u}_y = \vec{V}_B.\vec{u}_y$ (revoir également la figure 1.18). Pour que ce front d'onde global conduise à un effet de latéralisation dynamique naturel, il faut qu'il ait une vitesse de propagation apparente naturelle c, c'est-à-dire que $r_V = |\vec{V}| = 1$, soit encore $\vec{V} = \cos\theta\, \vec{u}_x + \sin\theta\, \vec{u}_y$

52. Ces propos s'appliquent bien-sûr d'abord au système de Bruck, mais aussi de manière indirecte, au système basé sur le *binaural B-format* qui présente le même potentiel.

53. Effectivement, la reconstruction de chacune des quatre grandeurs M_L, M_R, S_L et S_R individuellement nécessiterait la résolution d'un système à quatre équations et quatre inconnues, plus complexe que la mise en oeuvre de deux systèmes classiques d'annulation du *cross-talk*.

54. On ne considère implicitement que sa partie réelle.

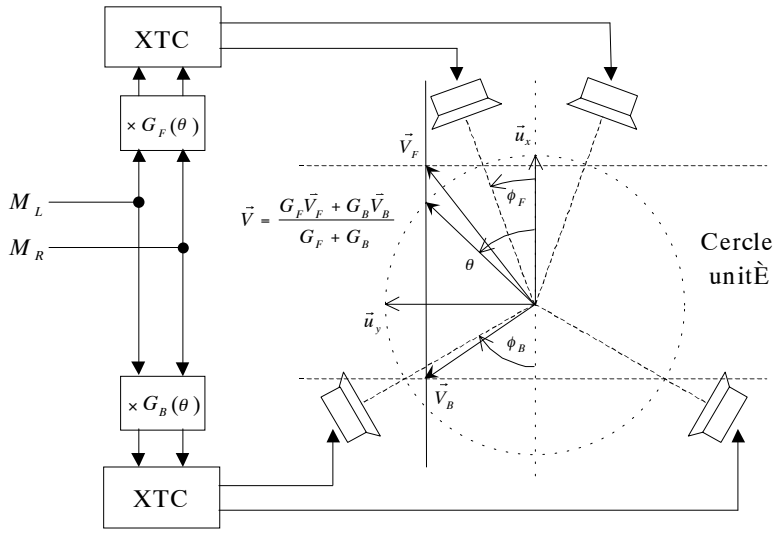

FIG. 2.24 – *Principe de répartition du filtrage transaural entre les paires de haut-parleurs avant et arrière pour respecter un effet de latéralisation dynamique naturel en basse-fréquence. L'événement acoustique global synthétisé en basse-fréquence est caractérisé par le vecteur vélocité $\vec{V}$, pondération des vecteurs $\vec{V}_F$ et $\vec{V}_B$, qui caractérisent eux-mêmes les effets respectifs des systèmes transauraux avant et arrière. Les facteurs de pondération G_F et G_B doivent être tels que le vecteur $\vec{V}$ soit de norme unité. Se reporter également aux figures 1.18 et 2.3.*

(Figure 2.24). En posant $G_F + G_B = 1$, on obtient assez rapidement:

$$\begin{aligned} \beta_F = \quad \beta_B &= \frac{1}{\cos\phi_F + \cos\phi_B} \\ \alpha_F &= \frac{\cos\phi_B}{\cos\phi_F + \cos\phi_B} \\ \alpha_B &= \frac{\cos\phi_F}{\cos\phi_F + \cos\phi_B} \end{aligned} , \tag{2.40}$$

soit, dans le cas d'une configuration rectangulaire ($\phi_F = \phi_B$):

$$\begin{aligned} \beta_F \quad &= \beta_B = \frac{1}{2\cos\phi_F} \\ \alpha_F \quad &= \alpha_B = \frac{1}{2} \end{aligned} \tag{2.41}$$

On vérifie aisément que seule la paire avant fonctionne ($\Delta = 0$) lorsque $\theta = \pm\phi_F$.

Bien que cette démonstration ne soit rigoureusement valide que dans un domaine basse-fréquence, on peut s'attendre à ce que ses résultats puissent s'extrapoler à des fréquences plus élevées. Toutefois, le *shuffling* original reste sans-doute le plus recommandable dans un domaine haute-fréquence où la reconstruction du champ au niveau des oreilles est assurée dans voisinage trop confiné par rapport à l'amplitude des rotations de la tête.

Problèmes de robustesse à la position et aux mouvements de la tête

Toutes les méthodes présentées – y compris la première – ont en commun de faire participer les quatre haut-parleurs simultanément, au moins pour la création des images latérales. Cette observation avait déjà été

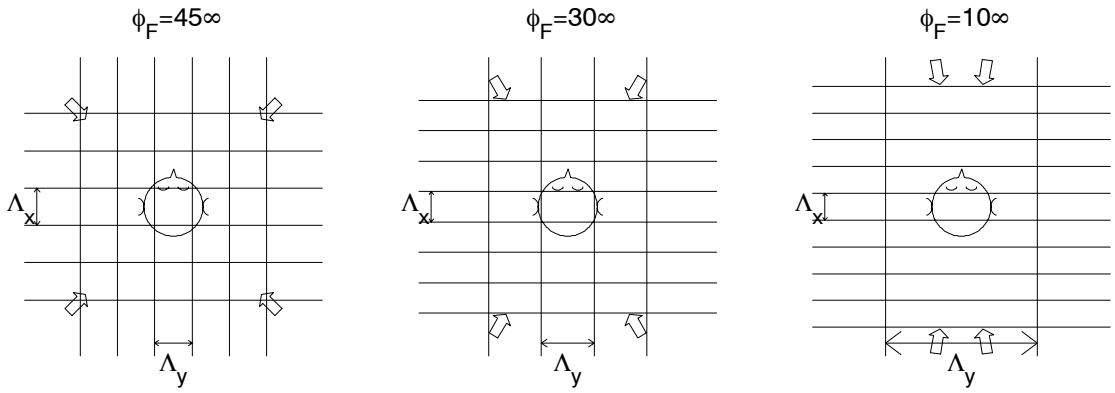

FIG. 2.25 – *Extension et périodicité spatiale* (Λ_x, Λ_y) *des figures d'interférence entre quatre ondes planes (ici, monochromatiques de fréquence* $f = 2000$ *Hz) dans trois situations: trois dispositifs de haut-parleurs. Les flèches larges indiquent la direction de propagation des ondes interférentes.*

faite à propos des systèmes ambisoniques en 2.4.4 et avait conduit à la remarque suivante: le phénomène acoustique créé – un front d'onde synthétique local – a alors une extension limitée suivant tous les axes du plan horizontal, ce qui n'est pas le cas avec l'usage de seulement deux haut-parleurs. Lorsque les ondes interférentes, supposées planes, sont émises par un dispositif rectangulaire de semi-angle frontal ϕ_F, il est facile de montrer que la figure d'interférence présente une périodicité spatiale suivant les axes $\vec{x}$ et $\vec{y}$, de périodes $\Lambda_x(f) = \frac{\lambda}{2\cos\phi_F}$ et $\Lambda_y(f) = \frac{\lambda}{2\sin\phi_F}$ proportionnelles à la longueur d'onde $\lambda = c/f$. Soulignons à nouveau que l'étalement rectangulaire de cette figure a les proportions inverses du rectangle des haut-parleurs (Figures 2.12 et 2.25).

Sur la base de simples considérations géométriques sur les phénomènes acoustiques créés en champ libre, il est facile de faire apparaître les nouveaux compromis auxquels sont confrontées ces "techniques transaurales améliorées". La figure 2.25 présente trois choix de configuration rectangulaire: "quadraphonique" ($\phi_F = 45^\circ$), "double transaural" ($\phi_F = 30^\circ$) et "double stereo dipole" ($\phi_F = 10^\circ$ au lieu de $\phi_F = 5^\circ$ pour plus de clarté). Pour garantir la robustesse des informations binaurales reconstituées jusqu'à une fréquence f donnée, les déplacements tolérés suivant les axes $\vec{x}$ et $\vec{y}$ sont proportionnels aux dimensions $(\Lambda_x(f), \Lambda_y(f))$ du rectangle central. Cette loi sommaire doit cependant être appliquée avec nuances, le degré d'instabilité dépendant sensiblement de l'*effort de reconstruction*, qui se manifeste par la participation concurrente des différents haut-parleurs quand la direction de la source virtuelle est assez éloignée des leurs.

La remarque faite à propos de la restitution ambisonique s'applique donc à nouveau ici: en cherchant à mieux reconstruire les propriétés du champ au voisinage des oreilles, on impose une contrainte de position supplémentaire. Le paradoxe s'exprime de façon plus flagrante ainsi: en voulant satisfaire plus complètement les mécanismes de localisation et tendre vers une expérience d'écoute plus naturelle, l'auditeur est bridé dans ses mouvements, la translation suivant l'axe frontal $\vec{x}$ lui devient presque interdite, et l'amplitude "autorisée" des mouvements de rotation de la tête est elle-même réduite pour la préservation des informations haute-fréquence. Partant d'une reconstruction assez fine des informations binaurales, la dégradation des qualités de l'image sonore est probablement plus sensible qu'avec Ambisonic, qui offre une qualité d'image certes moins performante ponctuellement, mais identique pour toutes les orientations de la tête[55].

Revenons à la figure 2.25: elle montre qu'en diminuant l'angle ϕ_F, il paraît possible d'augmenter la stabilité par déplacement latéral (Λ_y), ce qui diminue dans le même temps la stabilité par déplacement avant-

55. ... du moins avec une restitution sur configuration régulière de haut-parleurs.

arrière (Λ_x). On constate que $\Lambda_x = \lambda/2\cos\phi_F$ ne diminue pas énormément – il est borné inférieurement par $\lambda/2$ – lorsque $\Lambda_y = \lambda/2\sin\phi_F$ croît considérablement – en $1/\phi_F$ asymptotiquement. Ces arguments purement géométriques suggéreraient donc d'orienter le choix vers le dispositif "double stéréo-dipôle". Mais il est un autre aspect susceptible de mettre ce dispositif en défaveur par rapport aux autres: du fait que les haut-parleurs sont proches de l'axe médian, les paires avant et arrière participent de façon concurrente pour un secteur angulaire latéral plus large que les autres configurations, secteur angulaire où la variabilité des caractéristiques du champ suivant l'axe $\vec{x}$ est donc plus marquée. Dans la mesure où cela conditionne le problème d'instabilité lié à la position de la tête suivant $(O, \vec{x})$, celui-ci est donc généralisé à un plus grand nombre de sources virtuelles. Dès lors, il semble dérisoire en pratique d'offrir à l'auditeur une relativement grande liberté de mouvement de translation suivant l'axe gauche-droite et aucune ou presque suivant l'axe avant-arrière, d'autant que c'est suivant cet axe que le positionnement de la tête est généralement le plus incertain.

Cette interprétation corrobore les observations de l'étude [KNKH97] qui valide l'approche basée sur le *Four Ear Sphere Microphone*. Le dispositif *double stereo dipole* y apparaît inadapté par rapport aux dispositifs quadraphonique ($\phi_F = 45°$) et trapézoïdal (Figure 2.24 avec $\phi_F = 30°$ et $\phi_B = 70°$), entre autres à cause du problème de positionnement de l'auditeur. Dans cette étude, c'est finalement le dispositif trapézoïdal (dit "*assymetric*") qui semble apporter le plus de satisfaction. Cette disposition des haut-parleurs favorise naturellement la robustesse des images frontales et latérales, au détriment des images arrière auxquelles moins d'exigence est habituellement portée. Cette configuration pourrait également être adoptée par les autres systèmes présentés, en corrigeant le *shuffling* comme nous l'avons suggéré.

Synthèse et perspectives

Chacun des systèmes décrits se présente initialement comme une extension du principe transaural ayant pour vocation de résoudre le problème du repliement frontal des images virtuelles arrière, grâce à la participation d'une paire de haut-parleurs supplémentaire à l'arrière de l'auditeur. L'idée de base de la plupart des stratégies consiste à faire participer préférentiellement la paire la plus proche de la source virtuelle. La discrimination avant-arrière des sources virtuelles qui en résulte est surtout liée aux variations des indices de localisation – dont l'ITD en particulier – lors de légères rotations de la tête. Bien que suffisantes pour la discrimination avant-arrière, les variations de l'ITD ne sont pourtant pas naturelles pour trois des quatre méthodes. Nous en avons donc proposé une correction simple, applicable au moins dans un domaine basse-fréquence et n'imposant aucun surcoût, qui permet d'offrir les mêmes caractéristiques de restitution basse-fréquence qu'avec une reproduction ambisonique.

En définitive, ces approches peuvent être interprétées idéalement comme ayant pour objectif de contrôler la restitution des informations de localisation à la fois statiques et dynamiques au niveau des oreilles, ce qui semble assez facilement réalisable dans un domaine basse-fréquence. Malheureusement, l'augmentation des paramètres de contrôle du champ au voisinage des oreilles a aussi pour effet de diminuer les degrés de liberté de mouvement de la tête, le problème d'alignement des oreilles devenant rapidement critique dès qu'il s'agit de préserver l'information binaurale haute-fréquence (courtes longueurs d'onde). Les images latérales, en particulier, deviennent peu stables par déplacement de la tête suivant l'axe médian, ce qui n'était pas le cas avec un système transaural ou *stereo-dipole* simple.

La notion d'*effort de reconstruction* est encore présente et se reporte sur des problèmes de stabilité de l'image sonore par déplacement de la tête. Bien qu'il n'ait pas été question du vecteur énergie $\vec{E} = r_E\,\vec{u}_E$ dans ce contexte, où sa définition dépendrait d'ailleurs de la fréquence, il semble qu'il pourrait constituer un indice à la fois simple et judicieux du degré de robustesse. Considérons par exemple une source virtuelle parfaitement latérale, d'incidence $\vec{y}$. Pour des raisons de symétrie, $\vec{E}$ prend cette même direction $\vec{y}$, son module

$r_E \leq \sin\phi_F$ prend une valeur ridiculement petite avec la configuration *double stereo-dipole*, et un peu plus acceptable avec la configuration *quadraphonique*. La faible valeur de r_E coïncide avec l'instabilité constatée suivant la direction $\vec{x}$ orthogonale à $\vec{u}_E$. Par comparaison, le *stereo-dipole* simple maintenait à la fois une valeur r_E proche de 1 et une direction $\vec{u}_E$ proche de $\vec{x}$, assurant un bonne stabilité suivant la direction transversale $\vec{y}$.

Parmi les différentes techniques évoquées et indépendamment du choix de la géométrie du dispositif, c'est l'approche *binaural B-format* qui se montre *a priori* la plus séduisante en rassemblant de nombreuses propriétés: elle offre la possibilité d'une prise de son naturelle à l'aide de deux microphones ambisoniques; elle est censée être apte à reproduire les indices spectraux propres à un auditeur sans que l'encodage initial y soit particulièrement dédié; elle semble bien disposée à une séparation avant-arrière des sources sonores et à assurer un comportement naturel de l'ITD par légère rotation *yaw* de la tête, à l'aide de la correction du *shuffling* que nous proposons (au moins dans un domaine basse-fréquence). Grâce aux composantes verticales Z_L et Z_R, un autre potentiel de l'approche binaural B-format est la restitution "périphonique" sur un dispositif 3D de haut-parleurs [JWL98], susceptible d'améliorer la discrimination spatiale des sources virtuelles en terme de hauteur. Le principe du *shuffling* peut être étendu à la différenciation haut-bas des ondes incidentes afin de conserver une structure de décodeur relativement simple. En contrepartie, cette extension 3D fait apparaître une nouvelle contrainte sur le placement de l'auditeur: l'alignement vertical des oreilles. Supposons par exemple une configuration parallélépipédique, avec des haut-parleurs placés en $(\pm\phi_F, \pm\delta_{up})$ et en $(\pm\phi_F + \pi, \pm\delta_{up})$, $\delta_{up} > 0$. La figure d'interférence centrale pour une fréquence donnée a désormais une extension limitée suivant l'axe $(O,\vec{z})$, de taille $\Lambda_z(\phi_F,\delta_{up},\lambda) = \frac{\lambda}{2\sin\delta_{up}}$, cependant que ses dimensions s'élargissent dans le plan horizontal: $\Lambda_x(\phi_F,\delta_{up},\lambda) = \frac{\lambda}{2\cos\phi_F\cos\delta_{up}}$ et $\Lambda_y = \frac{\lambda}{2\sin\phi_F\cos\delta_{up}}$. Sans rentrer plus dans les détails, il faut s'attendre à ce que le problème de robustesse aux mouvements de l'auditeur soit accru. Par contre, il est probable que la robustesse de la restitution puisse être améliorée en augmentant le nombre de haut-parleurs de restitution – notamment dans le plan horizontal –, à condition d'exploiter judicieusement les composantes latérales Y_L et Y_R: il ne s'agit plus d'augmenter le nombre de paramètres de contrôle du champ au voisinage des oreilles, mais d'exploiter plus complètement les propriétés directionnelles de l'encodage pour mieux répartir l'énergie des haut-parleurs dans la direction de la source virtuelle.

2.6 Synthèse et conclusion

2.6.1 Synthèse: représentation et de restitution du champ sonore

Principales classes de stratégies et de représentations

Au cours de ce tour d'horizon, les analyses que nous avons mené par l'observation des phénomènes acoustiques ont fait émerger deux grandes classes de procédés de restitution sur haut-parleurs, chacune soutenue par des justifications théoriques solides, c'est-à-dire proposant une prédiction de l'effet et de la qualité de localisation:

– La première se caractérise par une convergence synchrone des ondes au centre du dispositif de haut-parleurs pour la création d'une image, et comprend notamment *Ambisonics* et les techniques de pan-pot d'amplitude ou d'intensité (sans ΔT entre les haut-parleurs), procédés au coût de traitement minime. Elle donne lieu à la *reconstruction locale "centralisée"* d'un front d'onde dont la propagation locale est caractérisée par un vecteur vélocité $\vec{V}$ fréquentiellement uniforme, qui permet de décrire l'effet de localisation basse-fréquence. La propagation globale (à plus large échelle) est quant à elle décrite par le vecteur énergie $\vec{E}$ ("flux d'énergie") qui prédit l'effet de localisation haute-fréquence, domaine où la consistance des indices de localisation est moindre et indiquée par $r_E = |\vec{E}|$.

– Avec la deuxième approche, la *reconstruction de l'effet* du front d'onde de référence est *focalisée sur les oreilles* – donc deux points de contrôle – et tient compte pour cela de la diffraction par la tête au moyen de procédés d'annulation du *cross-talk*, relativement coûteux. Ce sont les techniques transaurales et de type "double-transaural"[56], associées à une représentation binaurale simple (2 canaux) ou étendue (binaural B-format).

Une troisième classe, pourtant présente parmi les procédés stéréophoniques traditionnels, a été mise en marge de cette liste, parce qu'elle ne dispose pas d'outils objectifs fiables de prédiction ou de caractérisation: il s'agit des procédés "avec ΔT", caractérisés par une convergence asynchrone des ondes émises par les haut-parleurs. Ces procédés présentent en général moins d'aptitude que ceux de la première classe à créer des images individuelles stables et prédictibles, mais sont appréciés pour leurs meilleures qualités spatiales à l'issue d'une prise de son dans un champ complexe (meilleure décorrélation interaurale).

Dans chacune des deux premières classes, nous avons souligné une gradation des systèmes vers un enrichissement des conditions de l'illusion sonore, notamment par la satisfaction de variations naturelles de l'ITD lors de légères rotations de la tête, ce que réalisent *Ambisonics* et une version corrigée du *double-transaural*. La prise en compte de ces paramètres de contrôle supplémentaires exige évidemment plus de haut-parleurs (au moins quatre), et s'accompagne de nouvelles contraintes sur la position de l'auditeur: une position centrée, alors qu'une translation suivant l'axe médian était permise avec seulement deux haut-parleurs frontaux. Face à cette contrainte, les deux approches présentent des comportements fondamentalement différents dans un domaine haute-fréquence alors qu'ils convergent en basse-fréquence: *Ambisonics* offre une totale liberté de rotation de la tête, fournissant des informations de localisation haute-fréquence de qualité limitée (indiquée par r_E) mais indépendante de l'orientation de la tête; au contraire, le double-transaural n'est capable d'assurer son excellente reconstruction des informations haute-fréquence que pour une déviation minime de l'orientation de la tête. Dans ce dernier cas, la contrainte d'alignement des oreilles est en effet très problématique et rend presque paradoxale l'intention de prendre en compte la rotation de la tête.

Dans tous les cas, le compromis entre qualité de restitution, conditions naturelles d'écoute et robustesse aux déplacements, ne peut être amélioré que par une augmentation du nombre des haut-parleurs et leur exploitation judicieuse. Cette évolution se traduit pour *Ambisonics* par une extension vers les ordres supérieurs, sans augmenter de façon affolante la complexité du traitement (encodage-décodage). On pourrait imaginer une évolution parallèle, mais *a priori* beaucoup plus coûteuse, avec une stratégie basée sur le *binaural B-format*. Cette dernière, par ailleurs, est associée à une représentation intermédiaire du champ acoustique qui ne se prête pas à des manipulations (sauf la rotation *pitch* ou *tumble*, cf Figure 1.4).

D'une certaine manière et au regard d'arguments théoriques pour l'instant, *Ambisonic* (avec son extension aux ordres supérieurs) semble donc réaliser un compromis optimal entre robustesse, qualité d'image, conditions d'écoute naturelles, et concision et souplesse de représentation.

2.6.2 Atouts et potentiels de l'approche ambisonique

Nous retenons de l'approche ambisonique de nombreux atouts. Sa représentation rationnelle des informations sonores spatiales est basée sur une décomposition du champ en harmoniques sphériques, elle-même équivalente à la spécification de la propagation acoustique au voisinage du point considéré ("point de vue" de l'auditeur). Elle se prête aisément à des manipulations de type rotation ou déformation de perspective, appréciables dans un contexte interactif. Le B-format constitue un mode de représentation plus économique, en terme de données transmises, qu'un format multi-canal 5.1 habituel. Le décodage est peu coûteux et permet d'adapter la restitution à une variété de configurations de haut-parleurs (quatre, cinq ou plus). La combinai-

56. Nous incorporons dans cette dénomination l'ensemble des approches évoquées en 2.5.3

son avec les techniques binaurales et transaurales (méthode des haut-parleurs virtuels) donne la possibilité d'une restitution sur des dispositifs plus modestes, comme une paire de haut-parleurs ou des écouteurs.

Enfin, l'homogénéité qui caractérise la restitution ambisonique est la garantie de la "dématérialisation" des sources réelles (haut-parleurs), et probablement un facteur de préservation de l'illusion sonore, au moins dans un cadre de production non-supervisée.

Pour le développement de systèmes d'ordres supérieurs

Les premières investigations[57] [BV95] [Pol96a] sur le développement des systèmes ambisoniques d'ordres supérieurs ont surtout mis en avant l'extension de la zone centrale de reconstruction du champ acoustique, sans donner les moyens de juger l'amélioration de la restitution hors de ce domaine, ni de l'optimiser. Les outils et les principes présentés pour la définition les systèmes d'ordre 1 restent pourtant tout à fait pertinents. L'indice r_E – norme du vecteur énergie $\vec{E}$ – permet notamment de caractériser la manière dont sont affectées: la qualité de chaque image (latéralisation et tache de localisation), les qualités spatiales restituées pour un champ complexe (séparation latérale), la robustesse des images (sensibilité au *sweet-spot*). Ce sont tous ces aspects que l'on espère voir améliorés avec l'exploitation des ordres supérieurs. Les méthodes d'optimisation du décodage suivant les critères *max* r_E et même *in-phase*, méritent donc d'être généralisées aux ordres supérieurs.

C'est à une généralisation de tous les aspects du système ambisonique que s'intéresse le chapitre suivant: l'encodage et ses conventions, les opérations de manipulation du champ, la prise de son, le décodage.

57. ... ainsi que de plus récentes études [NE99].

Deuxième partie

Généralisation de l'approche ambisonique

Chapitre 3

Extension du formalisme et des solutions: de l'encodage au décodage

3.1 Généralisation du formalisme ambisonique

3.1.1 Intentions - Aperçu global du système

Il ressort du chapitre précédent que l'approche ambisonique – jusqu'ici abordée surtout à travers les systèmes traditionnels d'ordre 1 – présente une multitude de facettes et presque autant de qualités. Il s'agit d'abord d'une *représentation rationnelle du champ acoustique*, décrivant sans équivoque les informations directionnelles de l'espace sonore, et qui n'est pas assujettie à une configuration particulière du dispositif de restitution. Elle s'offre ainsi à une restitution adaptable à différents dispositifs, ainsi qu'à différentes conditions d'écoute: de l'écoute individuelle en position idéale à un auditoire s'étendant à proximité des haut-parleurs. *Ambisonics*, c'est aussi une *qualité de restitution*: homogénéité des images sonores les unes par rapport aux autres et continuité lors de leurs déplacements, transparence – ou dématérialisation – des haut-parleurs, et enfin un effet directionnel contrôlable et prédictible, l'approche ambisonique étant naturellement dotée d'outils objectifs de caractérisation de la restitution que sont les vecteurs vélocité et énergie.

L'encodage et la représentation ambisoniques sont fondamentalement liés à la décomposition du champ en harmoniques sphériques. Ce cadre mathématique à la fois solide et générique fournit à l'approche ambisonique un formalisme naturel pour une extension vers des degrés supérieurs de résolution spatiale du champ sonore. Tout en préservant les qualités de restitution propres à *Ambisonics* énoncées plus haut, il est attendu que les systèmes d'ordres supérieurs reculent les limites ou les faiblesses des systèmes traditionnels d'ordre 1, en termes de précision des images, de stabilité hors du *sweet-spot* et de préservation des impressions spatiales (séparation latérale). Toutefois, plus de haut-parleurs sont en général requis pour cela. En termes de format de représentation, l'extension aux ordres supérieurs consiste à enrichir le noyau initial du format B par des canaux supplémentaires. C'est alors une représentation *scalable* qu'il est possible de transmettre. En fonction des ressources disponibles à la transmission (débit) ou à la restitution (nombre de haut-parleurs et capacité du décodeur), l'ensemble ou seulement une partie des composantes ambisoniques peut être exploitée, donnant lieu à divers degrés de résolution de l'espace sonore restitué, mais avec toujours avec la garantie de sa cohérence.

Toute ces propriétés justifient l'intérêt grandissant porté à cette extension de la technologie *Ambisonics* et lui promettent de belles perspectives d'avenir(!). Bien que l'on puisse compter sur un étoffement des dispositifs de restitution (en nombre de haut-parleurs) pour le développement de systèmes d'ordres élevés, l'exploitation des ordres supérieurs peut d'ores et déjà être envisagée avec intérêt pour les dispositifs ou

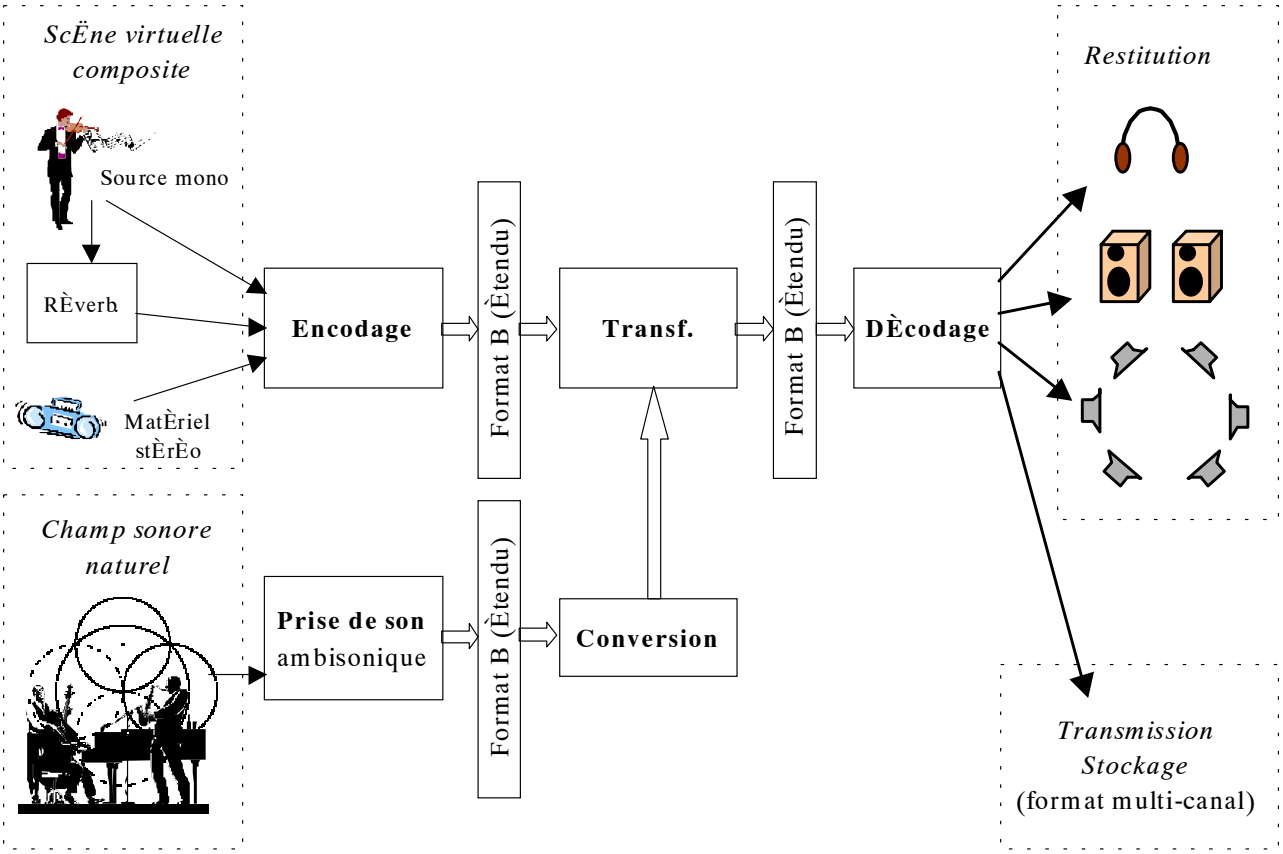

FIG. 3.1 – *Schéma global d'un système ambisonique.*

modes de restitution existants: les dispositifs dédiés à de larges auditoires, mais aussi les dispositifs multi-canal de type $3/2$, et enfin la restitution au casque ou sur une simple paire de haut-parleurs, par association avec les techniques binaurales et transaurales.

La généralisation de l'approche ambisonique aux ordres supérieurs, à laquelle ce chapitre est dédié, concerne tous les aspects du système, d'un bout à l'autre de la chaîne (Figure 3.1):

- l'*encodage* pour la composition de scènes sonores 3D,
- la définition de *microphones* pour la prise de son (encodage acoustique d'un champ sonore naturel),
- les *transformations globales du champ* (rotation, distorsion de perspective),
- les *conversions* entre les différentes conventions pour assurer le mélange et le traitement de matériel ambisonique préexistant au sein d'un système,
- le *décodage* et ses solutions selon les conditions de restitution et d'écoute.

D'un point de vue mathématique ou traitement du signal, il s'agit essentiellement d'opérations linéaires simples (produit et somme), avec éventuellement l'intervention d'un filtrage pour un traitement en sous-bande (*shelf-filtering*). Toutes ces opérations pourront être décrites sous forme vectorielle ou matricielle: matrice d'encodage $\mathbf{C}$, matrice de décodage $\mathbf{D}$, matrice de transformation $\mathbf{T}$, vecteur de conversion $\underline{\alpha}$, etc...

3.1.2 L'encodage: extension et compatibilité des conventions

Représentation tridimensionnelle

L'écriture de l'équation des ondes en coordonnées sphériques (A.10) permet d'exprimer le champ sous forme d'un développement en série de Fourier-Bessel sphérique. En tout point $\vec{r}(r,\theta_r,\delta_r)$ d'une région centrée où le champ p ne prend pas de valeur infinie, ce dernier s'écrit dans le domaine fréquentiel:

$$p(\vec{r}) = \sum_{m=0}^{\infty} (2m+1)\, j^m\, j_m(kr) \sum_{0\leq n\leq m, \sigma=\pm 1} B^{\sigma}_{mn} Y^{\sigma}_{mn}(\theta_r,\delta_r) \qquad \text{(Nombre d'onde: } k = 2\pi f/c\text{)}, \tag{3.1}$$

où apparaissent les fonctions à dépendance angulaire $Y^{\sigma}_{mn}(\theta,\delta)$, dites *harmoniques sphériques*[1] (Figure 3.2), et les fonctions de Bessel sphériques $j_m(kr)$ à dépendance radiale. Chaque $Y^{\sigma}_{mn}(\theta,\delta)$ est elle-même le produit d'une fonction du site δ (ou à dépendance polaire) et d'une fonction de l'azimut θ:

$$Y^{\sigma}_{mn}(\theta,\delta) = \tilde{P}_{mn}(\sin\delta) \times \begin{cases} \cos(n\theta) & \text{si } \sigma = 1 \\ \sin(n\theta) & \text{si } \sigma = -1 \end{cases}, \tag{3.2}$$

où les $\tilde{P}_{mn}$ sont les versions semi-normalisées au sens de Schmidt (A.13) des fonctions de Legendre associées P_{mn} (A.12). Nous dirons des fonctions Y^{σ}_{mn} qu'elles sont "semi-normalisées 3D" et nous les noterons $Y^{\sigma\,(\mathrm{SN3D})}_{mn}$ pour éviter la confusion avec les autres conventions introduites par la suite. Les fonctions nulles Y^{-1}_{m0} sont ignorées.

En arrêtant la décomposition (3.1) à un ordre M, une approximation du champ est obtenue, entièrement décrite par les signaux B^{σ}_{mn} ($0 \leq m \leq M$). Ces signaux définissent une *représentation ambisonique 3D, homogène, d'ordre M*. Il s'agit d'une extension du format B, donc on reconnaît le noyau original: $W = B^{1}_{00}$, $X = B^{1}_{11}$, $Y = B^{-1}_{11}$ et $Z = B^{1}_{10}$. On pourrait montrer qu'ils sont liés aux tenseurs d'ordres respectifs m du champ (Cf 1.2.2). Ils sont au nombre de $2m+1$ par ordre[2] m, soit au total un nombre $K = (M+1)^2$ de canaux ambisoniques. On pourra utiliser la notation vectorielle[3]:

$$\mathbf{B} = \mathbf{B}_{M(3D)} = [B^1_{00}\ B^1_{11}\ B^{-1}_{11}\ B^1_{10}\ \ldots\ \underbrace{B^1_{mm}\ B^{-1}_{mm}\ \ldots\ B^1_{mn}\ B^{-1}_{mn}\ \ldots\ B^1_{m0}}_{2m+1}\ \ldots\ \underbrace{B^1_{MM}\ B^{-1}_{MM}\ \ldots\ B^1_{M0}}_{2M+1}]^t, \tag{3.3}$$

avec en particulier $\mathbf{B}_{M=1(3D)} = [W\ X\ Y\ Z]^t$. Le vecteur des fonctions harmoniques sphériques est défini de la même manière:

$$\mathbf{y} = \mathbf{y}_{M(3D)} = [Y^1_{00}\ Y^1_{11}\ Y^{-1}_{11}\ Y^1_{10}\ \ldots\ Y^{\sigma}_{mn}\ldots]^t, \tag{3.4}$$

Les fonctions Y^{σ}_{mn} sont explicitées dans le paragraphe qui suit (Table 3.1), où elles apparaissent comme fonctions d'encodage.

Encodage d'une onde plane

L'extension des fonctions d'encodage ambisonique découle de façon naturelle de la décomposition d'une onde plane en harmoniques sphériques. Rappelons (section A.1.2, équations A.23 et A.14) que pour une onde

1. Attention au changement d'écriture par rapport à A.1.2: θ désigne désormais l'azimut, et δ le site complément à $\pi/2$ de l'angle polaire).

2. Avec les notations adoptées ici, on devrait parler de *degré m* et d'*ordre n*. Mais par un abus de langage assez fréquent, on continuera à parler d'*ordre M* ou *m*.

3. Noter l'ordonnancement que nous adoptons par défaut: indice m croissant de 0 à M et indice n décroissant de m à 0.

plane transportant un signal S (mesuré en $\vec{r}=0$) et d'incidence décrite par le vecteur unitaire $\vec{u}_S$ ou encore par le couple (θ_S,δ_S) (azimut et site), la décomposition (3.1) limitée à un ordre M s'écrit:

$$\begin{aligned} p_M(\vec{r}) &= S\sum_{m=0}^{M}(2m+1)\,j^m P_m(\vec{u}_r.\vec{u}_S)\,j_m(kr) \\ &= S\sum_{m=0}^{M}(2m+1)\,j^m \sum_{0\le n\le m,\sigma=\pm 1} Y_{mn}^{\sigma}(\theta_S,\delta_S)Y_{mn}^{\sigma}(\theta_r,\delta_r)\,j_m(kr) \end{aligned} \tag{3.5}$$

Les fonctions harmoniques sphériques $Y_{mn}^{\sigma}(\theta,\delta)$ – ici dans la version `SN3D` – définissent ainsi les *fonctions d'encodage ambisonique* généralisées. En mettant en relation les équations (3.5) et (3.1), les signaux ambisoniques B_{mn}^{σ} sont en effet décrits par:

$$B_{mn}^{\sigma} = Y_{mn}^{\sigma}(\theta_S,\delta_S).S, \qquad \text{ou encore} \quad \mathbf{B}=\mathbf{c}.S, \quad \text{avec} \quad \mathbf{c}=\mathbf{y}(\theta_S,\delta_S), \tag{3.6}$$

où l'on a introduit le *vecteur* $\mathbf{c}$ *des coefficients d'encodage* $c_{mn}^{\sigma}=Y_{mn}^{\sigma}(\theta_S,\delta_S)$, associé à la direction $\vec{u}_S(\theta_S,\delta_S)$. Naturellement, dans le cas d'un champ sonore plus complexe (somme d'ondes planes), le principe de superposition s'applique, et chaque composante B_{mn}^{σ} est calculée comme somme des différentes contributions.

A partir de l'expression des fonctions de Legendre associées (A.37), les fonctions d'encodage du premier ordre apparaissent bien conformes, à un facteur $\sqrt{2}$ près, à la définition traditionnelle du B-format (2.14) [Ger85]:

$$\left\{\begin{array}{ll} Y_{00}^{1\,(\mathtt{SN3D})}(\theta,\delta) & =1 \\ Y_{11}^{1\,(\mathtt{SN3D})}(\theta,\delta) & =\cos\theta\cos\delta \\ Y_{11}^{-1\,(\mathtt{SN3D})}(\theta,\delta) & =\cos\theta\cos\delta \\ Y_{10}^{1\,(\mathtt{SN3D})}(\theta,\delta) & =\sin\delta \end{array}\right. \Rightarrow \left\{\begin{array}{lll} W^{(\mathtt{SN3D})} & = B_{00}^{1\,(\mathtt{SN3D})} & = S \\ X^{(\mathtt{SN3D})} & = B_{11}^{1\,(\mathtt{SN3D})} & = S\cos\theta\cos\delta \\ Y^{(\mathtt{SN3D})} & = B_{11}^{-1\,(\mathtt{SN3D})} & = S\sin\theta\cos\delta \\ Z^{(\mathtt{SN3D})} & = B_{10}^{1\,(\mathtt{SN3D})} & = S\sin\delta \end{array}\right. \tag{3.7}$$

Il peut être utile d'exprimer les fonctions d'encodage en fonctions des coordonnées cartésiennes (u_x,u_y,u_z) du vecteur unitaire incidence $\vec{u}$:

$$\vec{u}=\left|\begin{array}{lcc} u_x &=& \cos\theta\cos\delta \\ u_y &=& \sin\theta\cos\delta \\ u_z &=& \sin\delta \end{array}\right. \quad\Rightarrow\quad \mathbf{y}_{M=1(3D)}^{(\mathtt{SN3D})}(\vec{u})=\begin{bmatrix}1\\ \vec{u}\end{bmatrix} \tag{3.8}$$

Ces composantes du vecteur incidence définissent donc directement les coefficients d'encodage d'ordre 1 de la convention semi-normalisée 3D (`SN3D`). Les fonctions d'encodage $Y_{mn}^{\sigma\,(\mathtt{SN3D})}$ sont explicitées en coordonnées sphériques et cartésiennes jusqu'à l'ordre 3, Table 3.1.

Il est également possible d'obtenir les coefficients d'encodage d'*ordres supérieurs* à partir des coefficients d'ordre 1 (3.7 ou 3.8) par un **calcul récursif**. Il suffit d'utiliser d'une part les relations de récurrence (A.36) sur les fonctions de Legendre associées $P_{mn}(u_z)$ et d'autre part les récurrences en n sur les fonctions $\cos(n\theta)=T_n(\cos\theta)$ (A.40) et $\sin(n\theta)$ (A.41), en posant au départ $\cos\theta=u_x/\sqrt{1-u_z^2}$ et $\sin\theta=u_y/\sqrt{1-u_z^2}$[4]. Implémenté de façon efficace (section 5.4.1), ce calcul peut se montrer plus avantageux que l'application directe des formules de la table 3.1.

Il est d'un grand intérêt, pour traiter des problèmes de décodage ou de prise de son, de définir une base orthonormée – au sens du produit scalaire (A.16) – d'harmoniques sphériques. Les fonctions[5] $Y_{mn}^{\sigma\,(\mathtt{N3D})}$ se déduisent des fonctions semi-normalisées $Y_{mn}^{\sigma\,(\mathtt{SN3D})}$ par une normalisation en énergie:

$$Y_{mn}^{\sigma\,(\mathtt{N3D})}=\sqrt{2m+1}\,Y_{mn}^{\sigma\,(\mathtt{SN3D})} \tag{3.9}$$

4. Les valeurs de $\cos\theta$ et $\sin\theta$ sont indifférentes si $1-u_z^2=0$.

5. Notées $\tilde{Y}_{mn}^{\sigma}$ en annexe A.1.2.

Ordre	B^{σ}_{mn}	$\binom{\sigma}{mn}$	$Y^{\sigma}_{mn}{}^{(\mathtt{SN3D})}(\vec{u})$	$Y^{\sigma}_{mn}{}^{(\mathtt{SN3D})}(\theta,\delta)$	$\alpha^{(\mathtt{FuMa})\mathtt{SN3D}}_{mn}$
0	W	$\binom{1}{00}$	1	1	$\frac{1}{\sqrt{2}}$
	X	$\binom{1}{11}$	u_x	$\cos\theta\cos\delta$	1
1	Y	$\binom{-1}{11}$	u_y	$\sin\theta\cos\delta$	1
	Z	$\binom{1}{10}$	u_z	$\sin\delta$	1
	U	$\binom{1}{22}$	$\sqrt{3}(u_x^2-u_y^2)/2$	$\sqrt{3}/2\cos(2\theta)\cos^2\delta$	$2/\sqrt{3}$
	V	$\binom{-1}{22}$	$\sqrt{3}u_xu_y$	$\sqrt{3}/2\sin(2\theta)\cos^2\delta$	$2/\sqrt{3}$
2	(S)	$\binom{1}{21}$	$\sqrt{3}u_xu_z$	$\sqrt{3}/2\cos\theta\sin(2\delta)$	$2/\sqrt{3}$
	(T)	$\binom{-1}{21}$	$\sqrt{3}u_yu_z$	$\sqrt{3}/2\sin\theta\sin(2\delta)$	$2/\sqrt{3}$
	(R)	$\binom{1}{20}$	$(3u_z^2-1)/2$	$(3\sin^2\delta-1)/2$	1
	-	$\binom{1}{33}$	$\sqrt{5/8}u_x(u_x^2-3u_y^2)$	$\sqrt{5/8}\cos(3\theta)\cos^3\delta$	-
	-	$\binom{-1}{33}$	$\sqrt{5/8}u_y(3u_x^2-u_y^2)$	$\sqrt{5/8}\sin(3\theta)\cos^3\delta$	-
	-	$\binom{1}{32}$	$\sqrt{15}u_z(u_x^2-u_y^2)/2$	$\frac{\sqrt{15}}{2}\cos(2\theta)\sin\delta\cos^2\delta$	-
3	-	$\binom{-1}{32}$	$\sqrt{15}u_zu_xu_y$	$\frac{\sqrt{15}}{2}\sin(2\theta)\sin\delta\cos^2\delta$	-
	-	$\binom{1}{31}$	$\sqrt{\frac{3}{8}}u_x(5u_z^2-1)$	$\sqrt{\frac{3}{8}}\cos\theta\cos\delta(5\sin^2\delta-1)$	-
	-	$\binom{-1}{31}$	$\sqrt{\frac{3}{8}}u_y(5u_z^2-1)$	$\sqrt{\frac{3}{8}}\sin\theta\cos\delta(5\sin^2\delta-1)$	-
	-	$\binom{1}{30}$	$u_z(5u_z^2-3)/2$	$\sin\delta(5\sin^2\delta-3)/2$	-

TAB. 3.1 – *Fonctions d'encodage ambisonique (versions semi-normalisées 3D) explicitées en coordonnées cartésiennes et sphériques jusqu'à l'ordre 3, avec, lorsqu'elle existe, la dénomination usuelle des canaux associés (celles récemment introduites par Furse et Malham sont entre parenthèses). Coefficients $\alpha^{(FuMa)SN3D}_{mn}$ de conversion vers la convention d'encodage de Furse-Malham.*

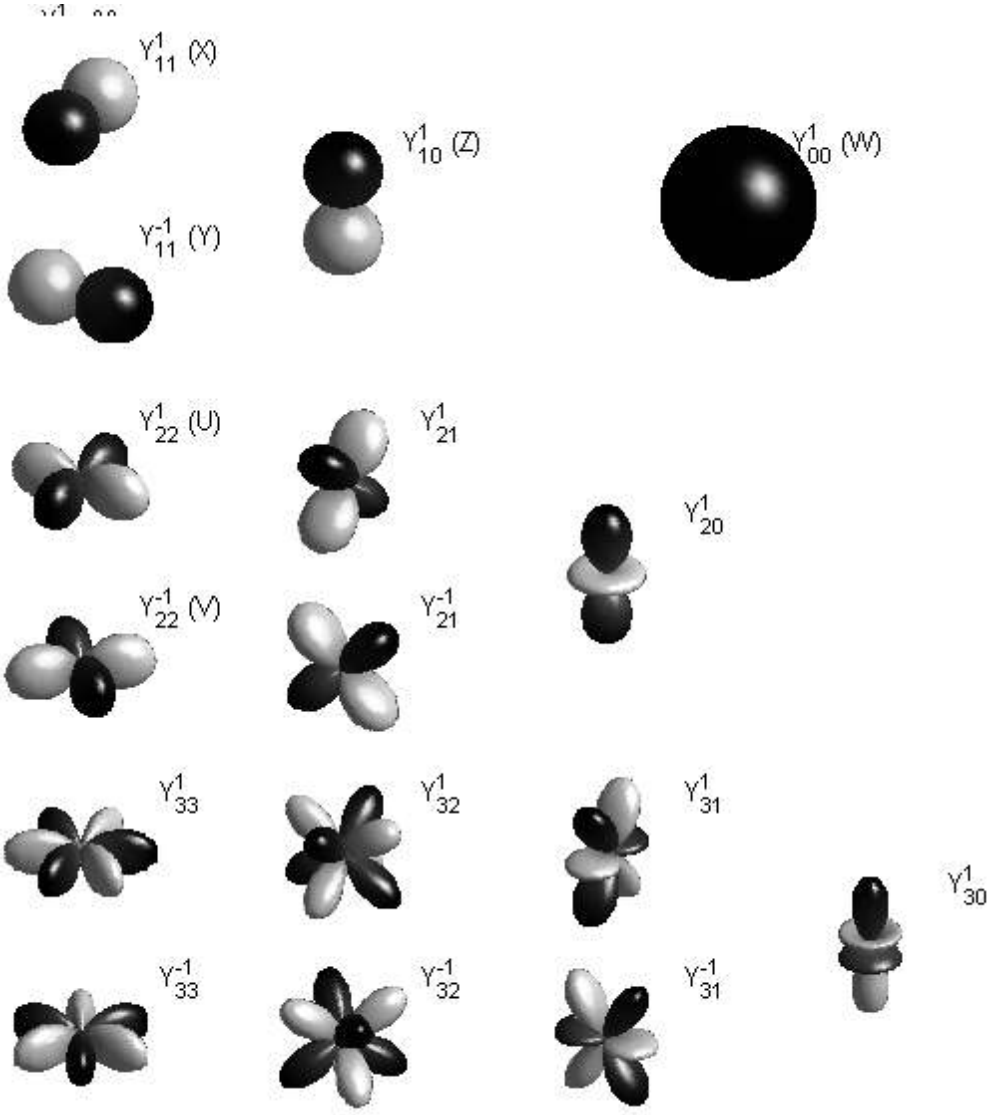

FIG. 3.2 – *Représentation des harmoniques sphériques jusqu'au degré 3. Les parties sombres et claires correspondent respectivement aux valeurs positives et négatives des fonctions harmoniques sphériques.*

Ces nouvelles fonctions d'encodage sont telles que les signaux des canaux d'encodage $B_{mn}^{\sigma\,(\mathrm{N3D})}$ associés sont de puissance égale dans le cas d'un champ diffus ou plus précisément, isotrope au lieu de mesure (contributions d'ondes réparties dans toutes les directions).

Encodage d'une source en champ proche (onde sphérique)

Pour la modélisation d'une *source en champ proche*, l'approximation d'onde plane n'est plus valable, et une correction des fonctions d'encodage est requise pour traduire un modèle de *propagation sphérique*. Si la source S est placée en $\vec{\rho} = \rho\,\vec{u}_S$, sa contribution à chaque composante d'ordre m doit être corrigée par un facteur $\Gamma_m(k\rho)$ dépendant de la fréquence, d'après (A.28):

$$B_{mn}^{\sigma} = \Gamma_m(k\rho) Y_{mn}^{\sigma}(\theta_S,\delta_S) S \qquad \text{avec:} \qquad \Gamma_m(k\rho) = kd\,(\,j_m(k\rho) - j n_m(k\rho))\,j^{-(m+1)} \tag{3.10}$$

où d est la distance de référence (par rapport au point $\vec{\rho}$) à laquelle le signal de pression S est mesuré, et n_m est la fonction de Neumann sphérique d'ordre m. En plus de l'atténuation globale en $1/\rho$ et l'expression en fréquence $e^{-jk\rho}$ du retard global ρ/c que reflète la correction Γ_0 sur la composante omnidirectionnelle W, des rapports complexes d'amplitude (donc des déphasages) apparaissent entre les composantes des différents ordres, et de façon particulièrement sensible pour les basses fréquences. Cet effet peut être pris en compte aussi bien pour la modélisation des sources virtuelles à l'encodage que pour corriger l'effet de proximité des haut-parleurs à la restitution[6] (4.2.3). Dans le premier cas, et en réservant à un traitement préalable sur le signal S encodé la question du retard et de l'atténuation, la correction des composantes d'ordres supérieurs m est réalisée à l'aide des filtres $\mathcal{F}_m^{(\rho/c)}$:

$$c_{mn}^{\sigma} = \mathcal{F}_m^{(\rho/c)}(\omega)\, Y_{mn}^{\sigma}(\theta_S,\delta_S), \tag{3.11}$$

où:

$$\mathcal{F}_m^{(\rho/c)}(\omega) = \frac{\Gamma_m(\omega\rho/c)}{\Gamma_0(\omega\rho/c)} = \sum_{n=0}^{m} \frac{(m+n)!}{(m-n)!} \left(\frac{-jc}{2\omega\rho}\right)^n \tag{3.12}$$

A la restitution, ce sont les filtres inverse $\left(\mathcal{F}_m^{(\rho/c)}\right)^{-1}$ qui sont à appliquer pour corriger l'effet de proximité des haut-parleurs. Ce phénomène de proximité est évoqué par Gerzon [Ger92f] pour les systèmes du premier ordre sous le nom d'effet *bass-boost*, et se modélise par le filtre du premier ordre:

$$\mathcal{F}_1^{(\rho/c)}(\omega) = \left(1 - \frac{jc}{\omega\rho}\right) \tag{3.13}$$

Ces filtres ont une implémentation naturelle dans le domaine analogique[7], par exemple sous la forme d'une cascade de cellules du premier ordre, après factorisation de l'expression (3.12).

Représentation bidimensionnelle

En restreignant la représentation du champ au plan horizontal, le développement d'une onde plane (d'incidence horizontale) en harmoniques cylindriques (section A.1.2, équation A.9) (2.27) fait apparaître des fonctions d'encodage Y_m^{σ}, qui sont également exhibées dans [BV95]:

$$\begin{cases} Y_m^{1}(\theta) &= \cos(m\theta) \qquad \text{pour } m \geq 0 \\ Y_m^{-1}(\theta) &= \sin(m\theta) \qquad \text{pour } m \geq 1 \end{cases} \tag{3.14}$$

6. Dans ce dernier cas, la distance de chaque haut-parleur au centre sera choisie comme distance de référence d.

7. L'obtention de filtres numériques équivalents passe par une transformation bilinéaire ou une transformée en Z.

A chaque fonction Y_m^σ est associé le canal d'encodage B_m^σ. Une *représentation 2D homogène d'ordre M* comprend donc $2M+1$ composantes. Un facteur $\sqrt{2}$ est introduit pour définir les versions normalisées en énergie au sens du produit scalaire (A.7) (se reporter en A.1.1):

$$\begin{cases} \tilde{Y}_0^1(\theta) &= Y_0^1(\theta) = 1 \\ \tilde{Y}_m^\sigma(\theta) &= \sqrt{2} Y_m^\sigma(\theta) \qquad \text{pour } m \geq 1 \end{cases} \tag{3.15}$$

Même si ce type de représentation est destiné à une reproduction 2D, il n'en est pas moins important d'y *incorporer l'information de hauteur (ou site)* δ pour les incidences non-horizontales. On choisira donc les fonctions:

$$Y_m^\sigma(\theta,\delta) = Y_m^\sigma(\theta)\cos^m\delta, \qquad \tilde{Y}_m^\sigma(\theta,\delta) = \tilde{Y}_m^\sigma(\theta)\cos^m\delta, \tag{3.16}$$

qui constituent, à un facteur près, *un sous-ensemble des fonctions d'encodage 3D*. On adoptera désormais les notations: $Y_{mm}^{\sigma\,(\mathtt{SN2D})} = Y_m$ et $Y_{mm}^{\sigma\,(\mathtt{N2D})} = \tilde{Y}_m$.

$$Y_{mm}^{\sigma\,(\mathtt{SN3D})}(\theta,\delta) = 1\times 3\times\ldots\times(2m-1)\sqrt{\frac{2}{(2m)!}}\, Y_{mm}^{\sigma\,(\mathtt{SN2D})}(\theta,\delta) = \sqrt{\frac{(2m)!}{2^{2m-1}m!^2}}\, Y_{mm}^{\sigma\,(\mathtt{SN2D})}(\theta,\delta) \tag{3.17}$$

Ou bien:

$$Y_{mm}^{\sigma\,(\mathtt{SN2D})}(\theta,\delta) = \alpha_{mm}^{(\mathtt{SN2D})\mathtt{SN3D}}\, Y_{mm}^{\sigma\,(\mathtt{SN3D})}(\theta,\delta), \quad \text{avec} \quad \alpha_{mm}^{(\mathtt{SN2D})\mathtt{SN3D}} = \sqrt{\frac{2^{2m-1}m!^2}{(2m)!}} \tag{3.18}$$

De cette façon, la représentation 2D équivaut à une *"coupe" horizontale de la représentation 3D*. On pourra constater que lors d'une restitution 2D, un effet de hauteur peut ainsi être reproduit, à une indétermination haut-bas près. En effet, la modification du rapport d'amplitude entre la composante W et les composantes X et Y se répercute sur la vitesse apparente du front d'onde reproduit horizontalement (et caractérisé par le vecteur vélocité $\vec{V}$), et se traduit par un effet de latéralisation plausible pour une source hors du plan horizontal (section 1.5 ou [DRP99] [DRP98]). Une interprétation similaire peu s'appliquer au vecteur énergie $\vec{E}$.

Représentations hybrides

On accorde généralement plus d'importance à la discrimination des événements sonores dans le plan horizontal que suivant l'axe vertical. Ce besoin peut se traduire par le choix d'une représentation non-homogène, en ne sélectionnant parmi les composantes d'ordres supérieurs que les harmoniques horizontales Y_{mm}^σ. Typiquement, on pourra ne garder que les six canaux W,X,Y,Z,U,V parmi les neuf composantes d'une représentation 3D homogène d'ordre 2. Ce choix peut évidemment se faire au moment du décodage seulement, par restriction d'une représentation homogène 3D par exemple. Comme *stratégie générale* du choix restrictif pour une préférence horizontale, nous recommandons une sélection progressive des harmoniques d'indice n décroissant (de m à 0) pour chaque nouvel indice m croissant[8].

Il faut cependant faire preuve de vigilance face à ce type de choix. En brisant l'homogénéité directionnelle de la représentation, la distribution énergétique des événements sonores et leur *définition spatiale* devient elle-même inhomogène. La question du déséquilibre de la distribution énergétique spatiale est traitée dans des conditions plus générales en 3.1.4. Elle inclut d'ailleurs le cas d'un encodage purement horizontal.

8. Rappelons que d'après les définitions mathématiques originales, l'indice n définit l'*ordre* (sous-entendu, de la dépendance azimutale), et l'indice m le *degré*, bien que par abus de langage, nous parlions généralement d'ordre m.

Conventions d'encodage, notations et conversions

Différentes conventions d'encodage sont présentes dans la littérature [Ger92b] [Mal95] [BV95], et nous venons nous-mêmes d'en introduire de nouvelles. Elles ne différent entre elles que par des coefficients de normalisation, et peuvent chacune avoir des propriétés intéressantes, qu'il s'agisse de simplifier l'écriture des fonctions d'encodage, ou bien faciliter la résolution du décodage 2D, ou bien celle du décodage 3D, etc... Ces différences posent néanmoins des problèmes de compatibilité et peuvent prêter à de dangereuses confusions, en particulier lorsqu'un même matériel ambisonique est exploité par des systèmes ou décodeurs basés sur des conventions différentes (comme c'est le cas des matrices de décodage définie en 3.3), ou encore lorsqu'un mixage est fait à partir de plusieurs formats ambisoniques. Il est donc indispensable de doter les différentes conventions de *notations claires et distinctes* (Table 3.2), et d'établir entre elles des correspondances par le biais de *coefficients* (ou, par extension, de vecteurs) *de conversion* (3.19).

Dans le tableau récapitulatif 3.2, deux autres conventions viennent s'ajouter à celles précédemment citées: la *"max-normalisation"*, telle que le maximum des fonctions d'encodage est 1, et la convention toute récente de *"Furse-Malham"* [Fur99] [Mal99c] (et [Mal99a]?), compatible avec celle de Malham et spécifiant des fonctions d'encodage 3D jusqu'à l'ordre 2. Mentionnons au passage que la convention de Bamford pour l'encodage horizontal [BV95] correspond à notre dénomination `SN2D`.

Les fonctions d'encodage de conventions respectives `conv1` et `conv2`, et de même ordre et même degré, sont liées par un coefficient multiplicatif $\alpha_{mn}^{(\texttt{conv2})\texttt{conv1}}$ tel que[9]:

$$Y_{mn}^{\sigma\ (\texttt{conv2})} = \alpha_{mn}^{(\texttt{conv2})\texttt{conv1}}.Y_{mn}^{\sigma\ (\texttt{conv1})} \tag{3.19}$$

L'opération de conversion suit les lois de réciprocité et de transitivité suivantes:

$$\begin{aligned} \alpha_{mn}^{(\texttt{conv1})\texttt{conv2}} &= 1/\alpha_{mn}^{(\texttt{conv2})\texttt{conv1}} && \text{(Réciprocité)} \\ \alpha_{mn}^{(\texttt{conv3})\texttt{conv1}} &= \alpha_{mn}^{(\texttt{conv3})\texttt{conv2}}.\alpha_{mn}^{(\texttt{conv2})\texttt{conv1}} && \text{(Transitivité)} \end{aligned} \tag{3.20}$$

La table 3.3 explicite les coefficients de quelques unes des principales conversions.

L'usage des conventions de `N2D` et `N3D` s'imposera de façon naturelle pour la résolution des problèmes de décodage respectivement 2D et 3D (section 3.3).

Extensions compatibles pour les conventions restreintes

Il faut remarquer qu'une partie des conventions recensées Table 3.2 ne spécifient pas un ensemble "exhaustif" de fonctions d'encodage, *i.e.* pouvant produire une représentation homogène 3D d'ordre quelconque: les conventions SN2D et N2D sont restreintes au plan horizontal, et celles de Gerzon et de Furse-Malham sont limitées respectivement aux ordres 1 et 2. Notons que ces dernières ne correspondent pas à des propriétés mathématiques particulières – sinon que leur restriction 2D vérifie la normalisation 2D, à un facteur $1/\sqrt{2}$ près pour Furse-Malham –, ce qui rend problématique une extension rationnelle aux ordres supérieurs. Dans la perspective d'une utilisation combinée de matériels ou d'applications ambisoniques basés sur des formats différents, il semble utile de suggérer des *spécifications* possibles *pour l'extension de ces conventions "restreintes"* – que ce soit aux ordres supérieurs ou bien au domaine 3D –, ainsi que des nouvelles passerelles entre les différents formats. De manière générale, cela revient à compléter les coefficients de conversions existants.

9. L'écriture de l'exposant de ce coefficient doit être comprise à la manière d'un *cast* (approximativement comme en langage C ou C++), le nouveau type `conv2` étant spécifié entre parenthèses et avant le type d'origine `conv1`.

Désign.	Restr.	Dénomination	Propriétés - Commentaires
$Y_{mn}^{\sigma\,(\mathtt{SN3D})}$	-	Semi-Normalisation 3D	Grande généricité: calcul récursif des coefficients d'encodage, les composantes d'ordre 1 étant celles du vecteur incidence (unitaire) $\vec{u}$.
$Y_{mn}^{\sigma\,(\mathtt{N3D})}$	-	Normalisation 3D	Base orthonormée pour la décomposition 3D. Relation simple à SN3D (facteur $\sqrt{2m+1}$). Assure une puissance égale des composantes encodées dans le cas d'un champ parfaitement diffus 3D (intérêt dans le domaine analogique). Intérêt évident pour la résolution (en 3.3) des problèmes de décodage (restitution 3D).
$Y_{mn}^{\sigma\,(\mathtt{SN2D})}$	$n=m$	Semi-Normalisation 2D	Expressions simples (fonctions trigonométriques). Conforme à [BV95] et compatible avec (i.e. "sous-ensemble de") MaxN.
$Y_{mn}^{\sigma\,(\mathtt{N2D})}$	$n=m$	Normalisation 2D	Commentaires sur N3D transposables à la 2D (restriction au plan horizontal). Facteur $\sqrt{2}$ (pour $m>0$) par rapport à SN2D.
$Y_{mn}^{\sigma\,(\mathtt{MaxN})}$	-	Max-Normalisation	Assure pour toutes les composantes une amplitude bornée par celle du signal encodé (intérêt éventuel dans le domaine numérique). Calcul des fonctions peu générique, mais expression simple jusqu'à l'ordre deux.
$Y_{mn}^{\sigma\,(\mathtt{Gerz})}$	$m\leq 1$	B-format (Gerzon)	Convention "standard" sans-doute la plus répandue (ordre 1). Fréquemment utilisée pour sa restriction horizontale, compatible avec N2D (facteur $\sqrt{2}$) dont les commentaires s'appliquent.
$Y_{mn}^{\sigma\,(\mathtt{FuMa})}$	$m\leq 2$	*Furse-Malham Set*	Récente extension (ordre 2, 3D) de la convention de Malham (ordre 1), elle-même équivalente à Gerz à un facteur global $1/\sqrt{2}$ près. Quasi-compatibilité avec MaxN, hormis le facteur $1/\sqrt{2}$ pour l'encodage de W (ordre 0).

TAB. 3.2 – *Description synthétique des différentes conventions d'encodage ambisonique: désignation symbolique, notations compacte et générique, restriction du domaine de définition, dénomination, et propriétés. Par défaut, les spécifications pour l'extension du domaine de définition sont: $m\geq 0$, $0\leq n\leq m$.*

(mn)	(00)	(11)	(10)	(22)	(21)	(20)	(33)	(32)	(31)	(30)
$\alpha_{mn}^{(\mathtt{N3D})\mathtt{SN3D}}$	1	$\sqrt{3}$	$\sqrt{3}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{7}$	$\sqrt{7}$	$\sqrt{7}$	$\sqrt{7}$
$\alpha_{mn}^{(\mathtt{MaxN})\mathtt{SN3D}}$	1	1	1	$\frac{2}{\sqrt{3}}$	$\frac{2}{\sqrt{3}}$	1	$\sqrt{\frac{8}{5}}$	$\frac{3}{\sqrt{5}}$	$\sqrt{\frac{45}{32}}$	1
$\alpha_{mn}^{(\mathtt{FuMa})\mathtt{SN3D}}$	$\frac{1}{\sqrt{2}}$	1	1	$\frac{2}{\sqrt{3}}$	$\frac{2}{\sqrt{3}}$	1	-	-	-	-
$\alpha_{mn}^{(\mathtt{N2D})\mathtt{SN3D}}$	1	$\sqrt{2}$	$(\sqrt{2})$	$\sqrt{\frac{8}{3}}$	-	-	$\frac{4}{\sqrt{5}}$	-	-	-

TAB. 3.3 – *Coefficients de conversion entre les principales conventions d'encodage mentionnées.*

Pour une extension 3D compatible avec une convention 2D (fonctions $Y_{mn}^{\sigma\,(\mathtt{N2D*})}$ et $Y_{mn}^{\sigma\,(\mathtt{SN2D*})}$, complémentaires 3D des fonctions normalisées et semi-normalisées 2D), nous proposons une *extrapolation des coefficients de conversion* $\alpha_{mm}^{(\mathtt{SN2D})\mathtt{SN3D}}$ et $\alpha_{mm}^{(\mathtt{N2D})\mathtt{N3D}}$:

$$\begin{array}{ll} \alpha_{mn}^{(SN2D*)SN3D} & = \alpha_{mm}^{(SN2D)SN3D} \\ \alpha_{mn}^{(N2D*)N3D} & = \alpha_{mm}^{(N2D)N3D} \end{array} \tag{3.21}$$

Du coup, la convention de Gerzon (B-format) devient un cas particulier de la convention étendue (`N2D*`) et peut adopter celle-ci comme sa propre extension (mais ce n'est qu'un choix possible parmi d'autres!). Les fonctions de Furse-Malham, quant à elles, sont presque un sous-ensemble des fonctions max-normalisées ($\alpha_{mn}^{(\mathtt{FuMa})\mathtt{MaxN}} = 1$ pour $m = 1{,}2$; $\alpha_{00}^{(\mathtt{FuMa})\mathtt{MaxN}} = 1/\sqrt{2}$), qui pourraient donc leur servir de complément aux ordres supérieurs. Ceci dit, nous suggérons d'adopter une *convention de référence* dont le calcul des fonctions soit le plus rationnel possible, et préférons pour cela la *semi-normalisation 3D*, qui permet le calcul récursif naturel des fonctions ou coefficients d'encodage à partir des composantes d'ordre 1 (page 150). Une option simple pour l'extension de n'importe quel format d'encodage $\mathtt{conv}_M$ (d'ordre M) existant consiste donc à utiliser ces fonctions `SN3D` comme complément: $\alpha_{mn}^{(\mathtt{conv}_M)\mathtt{SN3D}} = 1$ pour $m > M$.

Remarques sur la notion de format d'encodage

La question des *conventions d'encodage ambisonique*, dont nous venons de discuter, doit faire partie des *spécifications* d'un format d'encodage. Mais celles-ci ne s'y réduisent pas. Elles comprennent également l'ordonnancement des canaux, la présence éventuelle d'une opération de matriçage ou de filtrage (formats de types BHJ ou UHJ), et à plus bas niveau dans le domaine numérique, la fréquence d'échantillonnage, la quantification, l'entremêlement éventuel des voies, voire la présence d'une compression audio-numérique. Puisqu'aucune de ces spécifications supplémentaires n'intervient dans les développements de ce chapitre, la question du *format d'encodage* est reportée à un chapitre ultérieur (Chapitre 6), dans la partie III.

Utilisation: conversions pour les différents opérations du système

Au cours de ce chapitre, les différentes opérations du système ambisonique sont explicitées pour une convention d'encodage donnée `conv1`, le plus souvent `N2D` ou `N3D`. Nous donnons ci-dessous les formules de conversion de leur expression matricielle ou vectorielle à appliquer pour les exploiter lorsqu'une autre convention `conv2` est en vigueur. En notant le vecteur des fonctions d'encodage $\mathbf{Y} = [Y_{m_1n_1}^{\sigma_1} \ \ldots \ Y_{m_kn_k}^{\sigma_k} \ \ldots \ Y_{m_Kn_K}^{\sigma_K}]^t$, on définit le *vecteur de conversion* $\underline{\alpha} = [\alpha_{m_1n_1} \ \ldots \ \alpha_{m_kn_k} \ \ldots \ \alpha_{m_Kn_K}]^t$, par appariement des indices k.

Pour l'encodage et le transcodage, la conversion est immédiate:

$$\mathbf{c}^{(\mathtt{conv2})} = \mathrm{Diag}(\underline{\alpha}^{(\mathtt{conv2})\mathtt{conv1}}).\mathbf{c}^{(\mathtt{conv1})} \qquad \mathbf{B}^{(\mathtt{conv2})} = \mathrm{Diag}(\underline{\alpha}^{(\mathtt{conv2})\mathtt{conv1}}).\mathbf{B}^{(\mathtt{conv1})} \tag{3.22}$$

Une matrice de décodage $\mathbf{D}^{(\mathtt{conv})}$, produisant les signaux $[S_1 \ \ldots \ S_N]^t = \mathbf{S} = \mathbf{D}^{(\mathtt{conv})}.\mathbf{B}^{(\mathtt{conv})}$, est convertie suivant la relation:

$$\mathbf{D}^{(\mathtt{conv2})} = \mathbf{D}^{(\mathtt{conv1})}.\mathrm{Diag}(\underline{\alpha}^{(\mathtt{conv1})\mathtt{conv2}}) \tag{3.23}$$

Enfin, pour une transformation globale $\mathbf{T}$ du champ ambisonique (rotation ou distorsion de perspective) telle que $\mathbf{B}'^{(\mathtt{conv})} = \mathbf{T}^{(\mathtt{conv})}.\mathbf{B}^{(\mathtt{conv})}$, la conversion s'écrit:

$$\mathbf{T}^{(\mathtt{conv2})} = \mathrm{Diag}(\underline{\alpha}^{(\mathtt{conv2})\mathtt{conv1}}).\mathbf{T}^{(\mathtt{conv1})}.\mathrm{Diag}(\underline{\alpha}^{(\mathtt{conv1})\mathtt{conv2}}) \tag{3.24}$$

3.1.3 Stratégies de décodage suivant les conditions d'écoute - Structure du décodeur

Hypothèses et objectifs

Dans la suite, le champ ambisonique considéré est décrit de façon très générique par le vecteurs des signaux $\mathbf{B} = [B_{00}^1 \ \ldots \ B_{mn}^\sigma \ \ldots]^t$, auquel est associé le vecteur des fonctions d'encodage $\mathbf{y}(\theta,\delta) = [Y_{00}^1(\theta,\delta) \ \ldots \ Y_{mn}^\sigma(\theta,\delta) \ \ldots]^t$.

Le dispositif de restitution est constitué de N haut-parleurs répartis sur un cercle horizontal de rayon R_{HP} s'il s'agit d'une restitution 2D, ou sur une sphère pour une restitution 3D (Figure 1.15). Leur direction est indiquée par les vecteurs unitaires $\vec{u}_i$ ou encore les couples (θ_i, δ_i) (azimut et site). Dans un premier temps, les ondes venant des haut-parleurs sont supposés planes vues du centre O du dispositif – on ignore donc un éventuel effet de champ proche – et l'on y assimile leur contribution au champ de pression aux signaux S_i émis. En posant $\mathbf{S} = [S_1 \ \ldots \ S_i \ \ldots \ S_N]^t$, leur participation au champ ambisonique $\mathbf{B}'$ reconstruit en O s'écrit:

$$\mathbf{B}' = \mathbf{C}.\mathbf{S}, \qquad \text{avec} \quad \mathbf{C} = [\mathbf{c}_1 \ldots \ \mathbf{c}_i \ldots \ \mathbf{c}_N] \quad \text{et} \quad \mathbf{c}_i = \mathbf{y}(\vec{u}_i) \tag{3.25}$$

Plus loin, la matrice $\mathbf{C}$ est appelée "*matrice de réencodage*". L'éventuelle correction de l'effet du champ proche des haut-parleurs est traitée plus tard.

Le décodage ambisonique est une opération linéaire décrite par une matrice $\mathbf{D}$ dépendant éventuellement de la fréquence.

L'objectif du décodage est de donner lieu à la meilleure qualité d'image sonore possible *pour l'auditoire présent*. Comme il est d'usage et en s'appuyant sur la discussion 1.5.1, l'optimisation du décodage porte sur l'effet d'une onde plane, qui modélise en général bien l'onde directe et les réflexions associées à une source sonore, et qui peut être considérée en ce sens comme un événement acoustique élémentaire de référence. Dans la suite, nous considérons que le champ encodé $\mathbf{B}$ est constitué d'une onde plane d'incidence $\vec{u}_S(\theta_S, \delta_S)$ transportant un signal S, en supposant $\delta_S = 0$ dans le cas d'une restitution 2D:

$$\mathbf{B} = \mathbf{c}.S \qquad \text{avec} \quad \mathbf{c} = \mathbf{y}(\vec{u}_S) \tag{3.26}$$

Les principales conditions d'écoute pour lesquelles on cherche à optimiser le décodage sont représentées Figure 3.3. Ajoutons que la restitution au casque, basée sur la méthode des haut-parleurs virtuels, s'apparente au cas idéal d'une écoute individuelle en position centrée.

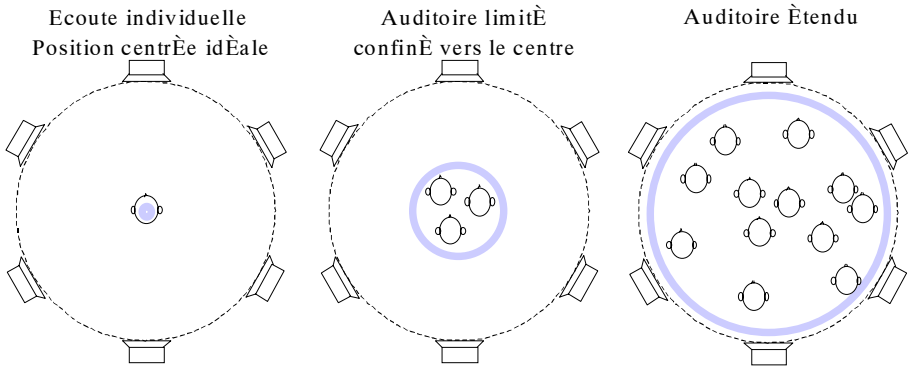

FIG. 3.3 – *Etendue de l'auditoire: trois cas de figure typiques.*

Combinaison des critères de décodage selon les conditions d'écoute

De l'étude des systèmes du premier ordre (Cf 2.4) ressortent *trois principales formes de décodage*, que nous répertorions sous les termes suivants:

- *basic*: il s'agit d'assurer la reconstruction locale du front d'onde au voisinage de la position centrale O (portée basse-fréquence). A l'ordre 1, cela se traduit par la condition sur le vecteur vélocité:$\vec{V} = \vec{u}_S$. Par ailleurs, la condition $\vec{u}_E = \vec{u}_V = \vec{u}_S$ sur le vecteur énergie $\vec{E} = r_E\,\vec{u}_E$ est une garantie de la stabilité par éloignement du centre.
- *max* r_E: la localisation haute-fréquence en position centrée étant prédite par le vecteur énergie $\vec{E}$ (Cf 1.5.3), son optimisation consiste à "optimiser le flux d'énergie" ou "concentrer les sources d'énergie" dans la direction $\vec{u}_S$, c'est-à-dire maximiser r_E tout en garantissant $\vec{u}_E = \vec{u}_S$. Gerzon préconise d'appliquer ces deux premières formes de décodage sur deux bandes basse- et haute-fréquence complémentaires pour optimiser la localisation en position d'écoute centrée (position A_0, Figure 3.4).
- *in-phase*: avec ce décodage introduit par Malham [Mal92], la participation des haut-parleurs diminue progressivement à mesure qu'ils s'éloignent de la source virtuelle, jusqu'à s'annuler (Figure 2.14). Il assure un effet de localisation plus robuste que les autres décodages pour les auditeurs excentrés, voire placés à proximité des haut-parleurs (zone d'écoute A_4, Figure 3.4).

A cela s'ajoute un critère de *normalisation*. Une normalisation *en amplitude* ($\sum G_i = 1$) est requise dans le domaine basse-fréquence où la reconstruction est assurée sur toute la zone d'écoute. Elle *devrait* donc idéalement être associée au décodage *basic*[10]. Une normalisation *en énergie* ($\sum G_i^2 = 1$) est appliquée sur le reste du domaine. La notion d'énergie globale restituée doit cependant être considérée avec précaution (remarques de la page 107).

On dispose ainsi d'un certain nombre de critères pour la définition des matrices de décodage, qui permettent d'optimiser l'effet de localisation en fonction des positions les plus critiques occupées par les auditeurs, c'est-à-dire selon l'étendue de la zone d'écoute. Il semble justifié de pouvoir appliquer ces mêmes critères de décodage aux systèmes d'ordres supérieurs.

Avant de présenter plus loin (en 3.3) une généralisation du décodage en *trois familles de solutions*[11], il convient d'exposer brièvement le principe du décodage *basique*, adopté implicitement lors des premières études sur les systèmes d'ordres supérieurs [BV95] [Pol96a]. L'écriture du champ en série de Fourier-Bessel (2.27 ou 3.1) suggère que la reconstruction du front d'onde original passe par la reconstruction des composantes ambisoniques $\mathbf{B}$, qui correspondent encore aux tenseurs d'ordres successifs du champ. Cette reconstruction implique automatiquement $\vec{V} = \vec{u}_S$. C'est ce que nous nommons ici le "*principe de réencodage*"[12], qui s'écrit $\mathbf{B}' = \mathbf{B}$, soit d'après (3.25) et (3.26):

$$\mathbf{c}.S = \mathbf{C}.\mathbf{S} = \mathbf{C}.\mathbf{G}.S \Rightarrow \mathbf{c} = \mathbf{C}.\mathbf{G} \tag{3.27}$$

La matrice de décodage basique $\mathbf{D}$ est donc définie en inversant la matrice de réencodage $\mathbf{C}$, nécessitant pour cela au moins autant de haut-parleurs que de canaux ambisoniques: $N \geq K$. Une solution générique à ce problème est la pseudo-inverse de $\mathbf{D}$ (voir plus tard, équation 3.63). Cette solution n'assure cependant la condition $\vec{u}_E = \vec{u}_S$ que si $N > K$ et si la configuration de haut-parleurs est régulière ou semi-régulière [DRP98]. Ces propriétés de régularité sont définies en 3.2. Des indications sur la portée de la reconstruction y sont également données, d'où peuvent être déduites des fréquences de transition $f^{p\to m}$ entre l'application des décodages *basic* et *max* r_E.

10. Pourtant, Gerzon applique en général la normalisation en énergie sur toute la bande de fréquence, même avec le décodage basique.

11. Les solutions de décodage *max* r_E pour les configurations régulières 2D sont également exposées dans [DRP98] (Annexe B).

12. Expression employée dans un cadre plus général dans [JLP99].

Entre la position idéale A_0 et la situation critique A_4 déjà évoquées, on peut imaginer que les situations intermédiaires requièrent d'autres combinaisons des solutions de décodage[13], ainsi que le propose la figure 3.4. En voici quelques commentaires.

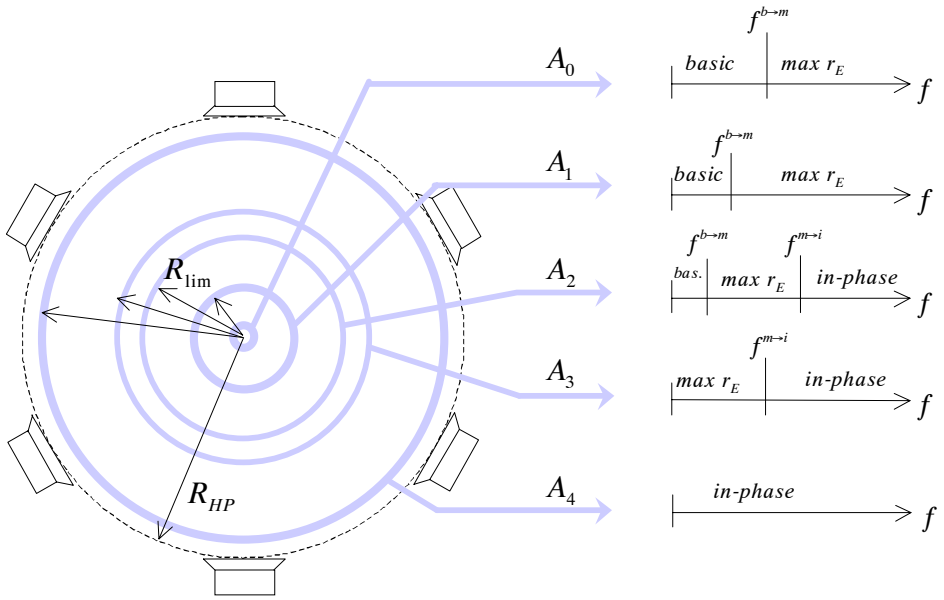

FIG. 3.4 – *Application des critères de décodage par sous-bande fréquentielle en fonction de l'étendue de la zone d'écoute (rayon R_{lim}): de la position idéale d'écoute (A_0) à un auditoire s'étendant jusqu'à proximité des haut-parleurs (A_4). Les fréquences de transition entre les solutions* basic *et* max r_E *et entre* max r_E *et* in-phase *sont notées* $f^{b\to m}$ *et* $f^{m\to i}$.

On sait que pour un décodage basique donné, la fréquence limite f_{lim} en-deçà de laquelle la reconstruction est acceptable est inversement proportionnelle au rayon R_{lim} de la zone d'écoute considérée (et *vice versa*). Sous réserve qu'on puisse l'assimiler à la fréquence de transition $f^{b\to m}$ avec le décodage *max* r_E, cette dernière diminue donc à mesure que R_{lim} augmente (Situations A_1 et A_2).

D'après 1.5.4, le vecteur énergie se révèle être un prédicteur de la localisation basse-fréquence (d'après le retard interaural de phase) dès que la cohérence de phase entre les contributions n'est plus vérifiée. Le décodage *max* r_E est donc justifié même dans un domaine basse-fréquence (A_2 et A_3) à partir du moment où la reconstruction contrôlée est hors d'atteinte.

En position excentrée, deux phénomènes viennent progressivement perturber l'effet de localisation, à cause desquels le décodage *in-phase* peut se montrer préférable au décodage *max* r_E (se reporter également à la figure 2.14):

– D'une part l'effet d'antériorité, qui dépend ici grossièrement de la *distance absolue* R_{lim}: comme il se manifeste à travers les transitoires du signal, qui sont eux-mêmes plus sensibles par les hautes fréquences que par les basses, on peut imaginer faire cohabiter les deux décodages, la fréquence de transition $f^{m\to i}$ diminuant à mesure que R_{lim} augmente (A_2 et A_3).

– D'autre part, le déséquilibre énergétique des contributions perçues, qui dépend quant à lui de la *distance relative* R_{lim}/R_{HP}: il renforce non-seulement l'effet d'antériorité mais provoque une distorsion

13. Une interpolation des solutions (ou matrices) de décodage pourrait être également envisagée.

du vecteur énergie “perçu", et peut justifier l'application pleine bande du décodage *in-phase* (situation A_4).

On peut déterminer assez facilement et de façon objective l'ordre de grandeur de $f^{b\to m}$ en fonction de l'ordre M et de R_{lim}: quelques propositions sont présentées dans [DRP98] (Table 1) et en 3.2 (Tables 3.4, 3.5 et 3.7). L'expérience montre par ailleurs que la précision de ces fréquences n'est pas très critique. La fréquence $f^{m\to i}$, si la combinaison *max* r_E*/in-phase* a lieu d'être, est en revanche inconnue pour le moment, et devrait être déterminée expérimentalement. De façon plus pragmatique, mieux vaudrait chercher dans un premier temps un rayon critique $R^{m\to i}$ au delà duquel l'application du critère *in-phase* donne “globalement"[14] un meilleur rendu qu'une solution *max* r_E. Ce rayon $R^{m\to i}$, comme la fréquence $f^{m\to i}$, dépendent *a priori* en partie du rayon R_{HP}.

Correction du champ proche des haut-parleurs

Si l'on tient compte de la proximité des haut-parleurs, le champ ambisonique réellement “recomposé" au centre O s'écrit:

$$\begin{aligned}&\mathbf{B}' = \mathrm{Diag}(\underline{\mathcal{F}}^{(R_{HP}/c)}(\omega)).\mathbf{C}.\mathbf{S}\\ &\text{avec}\quad \underline{\mathcal{F}}^{(R_{HP}/c)}(\omega) = [\mathcal{F}_0^{(R_{HP}/c)}(\omega)\ \ \mathcal{F}_1^{(R_{HP}/c)}(\omega)\ \ \mathcal{F}_1^{(R_{HP}/c)}(\omega)\ \ \ldots\ \ \mathcal{F}_m^{(R_{HP}/c)}(\omega)\ \ \ldots]\end{aligned} \tag{3.28}$$

où les composantes du vecteur $\underline{\mathcal{F}}^{(R_{HP}/c)}(\omega)$ sont appariées avec celles de $\mathbf{B}'$ d'après l'indice m. Les filtres $\mathcal{F}_m^{(R_{HP}/c)}(\omega)$ définis par (3.12) affectent les composantes ambisoniques dans un domaine basse-fréquence. Dans ce cas, la matrice de décodage basique $\mathbf{D}$ définie comme pseudo-inverse de $\mathbf{C}$ ne permet pas de recomposer précisément le champ original $\mathbf{B}$[15]. La correction du décodage consiste à le faire précéder d'un filtrage de chaque composante B_{mn}^{σ} par le filtre inverse $\left(\mathcal{F}_m^{(R_{HP}/c)}\right)^{-1}(\omega)$ de même degré m.

Cette correction n'a lieu d'être que si le décodage basique est appliqué, au moins dans le domaine basse-fréquence.

Structure du décodeur

Dans le cas où un critère de décodage est appliqué en pleine bande, le décodeur est directement assimilé à la matrice de décodage $\mathbf{D}$. Sinon, des matrices différentes doivent être appliquées sur les différentes sous-bandes fréquentielles: nommons-les par exemple $\mathbf{D}_{BF}$ et $\mathbf{D}_{HF}$ pour un décodage entre deux sous-bandes. Deux structures peuvent alors être envisagées, selon la familiarité entre les matrices, qui dépend elle-même de la régularité de la configuration de haut-parleurs.

Dans un cas très général (configuration non régulière *a priori*), la structure du décodeur en deux sous-bandes reste tout à fait similaire à celle décrite Figure 2.10 (page 109) pour le décodage d'ordre 1 d'après Gerzon, au nombre de composantes ambisoniques près. Pour une configuration régulière ou semi-régulière (au sens défini plus loin en 3.2.3), on montre ([DRP98] et en 3.3) que les différentes formes de décodage évoquées plus haut dérivent du décodage basique par une simple correction préalable des composantes ambisoniques B_{mn}^{σ}: cette correction consiste en une multiplication par des facteurs g_m. Dès lors, le décodage en sous-bandes revient à rendre les gains g_m dépendants de la fréquence par palier, que l'on nomme *Shelf-Filters*.

14. En réalité, la différence d'effet sur la localisation entre les décodages *max* r_E et *in-phase* ne dépend pas que de la distance de l'auditeur au centre R_{lim}, mais aussi très probablement de son placement par rapport aux haut-parleurs et de la position de la source virtuelle.

15. Comme nous l'illustrons plus loin (Figure 4.19, page 232), ce décodage sans correction tend à reconstruire l'effet d'une source ponctuelle placée à distance R_{HP} dans la direction $\vec{u}_S$, c'est-à-dire “projetée" sur le périmètre des haut-parleurs.

La structure du décodeur (Figure 3.5) est ainsi moins coûteuse et s'apparente à celle décrite Figure 2.11. L'éventuelle correction du champ proche des haut-parleurs pourrait être incorporée dans les filtres $g_m(f)$.

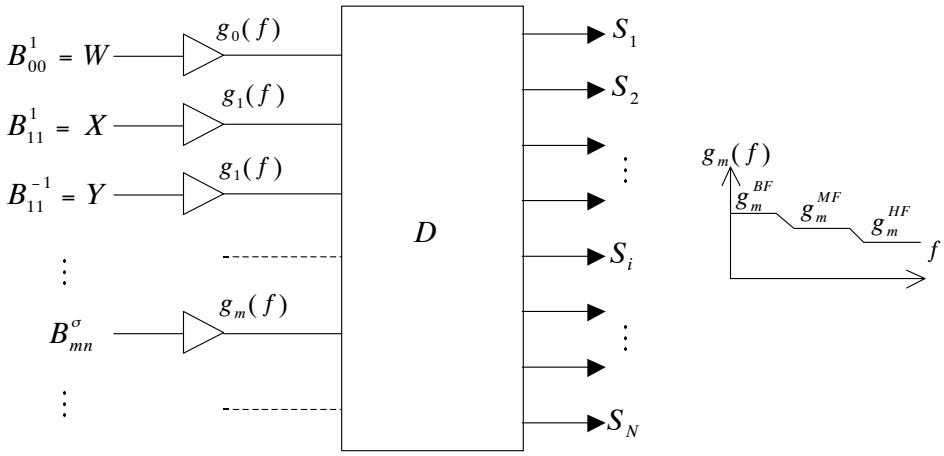

FIG. 3.5 – *Décodage ambisonique pour configuration régulière de haut-parleurs. Une opération de matriçage commune est précédée d'une opération de filtrage par palier (*Shelf-Filtering*) appliquée à chaque composante $B_m n^{\sigma}$. Exemple d'étagement du gain $g_m(f)$ en trois paliers g_m^{BF}, g_m^{MF} et g_m^{HF} correspondant à des bandes basse-, moyenne- et haute-fréquence (à droite).*

Une **restitution au casque** est possible en appliquant la méthode des haut-parleurs virtuels (Cf 2.5.1): ayant choisi un dispositif virtuel, le décodage optimisé pour une position centrée (Figure 3.4) est alors prolongé par une simulation binaurale de chaque haut-parleur. En nommant $H_L(\theta_i,\delta_i)$ et $H_R(\theta_i,\delta_i)$ les fonctions de transferts (HRTF) du haut-parleur i vers les oreilles gauche et droite, les signaux S_L et S_R reconstituées aux oreilles s'écrivent dans le domaine fréquentiel:

$$\begin{aligned} S_L &= \sum_{i=1}^{N} H_L(\theta_i,\delta_i).S_i = \mathbf{H_L}^t.\mathbf{S} = \mathbf{H_L}^t.\mathbf{D}.\mathbf{B} \\ S_R &= \sum_{i=1}^{N} H_R(\theta_i,\delta_i).S_i = \mathbf{H_R}^t.\mathbf{S} = \mathbf{H_R}^t.\mathbf{D}.\mathbf{B} \end{aligned} , \tag{3.29}$$

où l'on a constitué les vecteurs $\mathbf{H_L} = [H_L(\theta_1,\delta_1)\ \ldots\ H_L(\theta_N,\delta_N)]$ et $\mathbf{H_R} = [H_R(\theta_1,\delta_1)\ \ldots\ H_R(\theta_N,\delta_N)]$, et où la matrice $\mathbf{D}$ est fonction de la fréquence afin de représenter le décodage dans sa globalité. En écrivant cette matrice sous la forme $\mathbf{D} = [\mathbf{D}_{00}^1\ \ldots\ \mathbf{D}_{mn}^{\sigma}\ \ldots]$, on peut définir des *fonctions de transfert des composantes ambisoniques B_{mn}^{σ} vers les signaux binauraux S_L et S_R*:

$$\begin{aligned} L_{mn}^{\sigma}(f) &= \mathbf{H_L}^t(f).\mathbf{D}_{mn}^{\sigma}(f) \\ R_{mn}^{\sigma}(f) &= \mathbf{H_R}^t(f).\mathbf{D}_{mn}^{\sigma}(f) \end{aligned} \tag{3.30}$$

En pratique, il est recommandé de choisir un dispositif symétrique par rapport au plan médian de l'auditeur. On suppose de surcroît la symétrie de HRTF: $H_R(\theta_i,\delta_i) = H_L(-\theta_i,\delta_i)$. On montre alors la redondance des fonctions de transfert $L_{mn}^{\sigma}(f)$ et $R_{mn}^{\sigma}(f)$, étant donné les propriétés de parité des fonctions d'encodage $Y_{mn}^{\sigma}(\theta,\delta)$ par rapport à l'azimut θ (3.2):

$$\begin{aligned} Y_{mn}^{1}(-\theta,\delta) &= Y_{mn}^{1}(\theta,\delta) \quad &\Rightarrow R_{mn}^{1} &= L_{mn}^{1} \\ Y_{mn}^{-1}(-\theta,\delta) &= -Y_{mn}^{-1}(\theta,\delta) \quad &\Rightarrow R_{mn}^{-1} &= -L_{mn}^{-1} \end{aligned} \tag{3.31}$$

Le nombre d'opérations de filtrage requises pour le décodage peut être ainsi divisé par 2, en adoptant la structure de la figure 3.6.

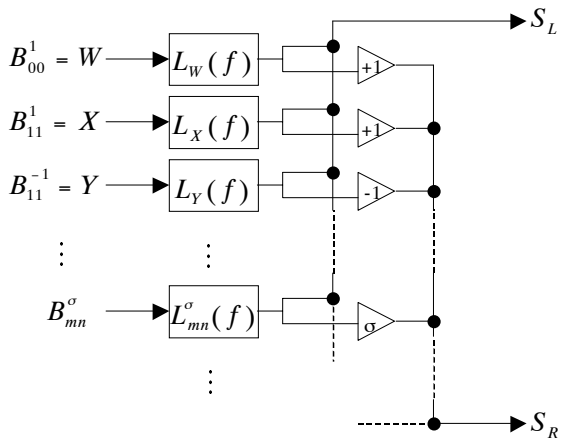

FIG. 3.6 – *Décodage ambisonique pour une restitution binaurale, avec l'hypothèse d'une symétrie par rapport au plan médian de l'auditeur.*

3.1.4 Correction de l'encodage pour une égalisation énergétique spatiale

Egalisation énergétique 3D d'un encodage 2D

Il a été remarqué plus haut (en 3.1.2) que la représentation 2D, définie comme sous-ensemble d'une représentation 3D, contient des informations de site et permet d'en restituer les effets par rotation (*yaw*) de la tête, puisque qu'on réalise une projection correcte du vecteur vélocité $\vec{V}$ dans le plan horizontal. Cependant, avec de telles fonctions d'encodage, l'énergie totale contenue dans les composantes est plus faible pour les incidences s'écartant du plan horizontal, ce qui peut être un artefact perceptible lors d'une restitution. Pour pallier ce déséquilibre énergétique, des fonctions d'encodage modifiées ont été proposées dans [PBJ98] et [JLP99] pour *reporter l'énergie des composantes verticales ignorées dans la composante omnidirectionnelle* W:

$$W = \sqrt{1+2\sin^2\delta}\,S \tag{3.32}$$

C'est l'énergie de la composante Z qui est incorporée dans (3.32). En toute rigueur, celle des composantes non-horizontales d'ordres supérieurs devrait aussi être présente pour un système d'ordre $M > 1$. Il faut pour cela considérer que par rapport à un cas d'incidence horizontale, chaque composante d'ordre m est pondérée avant décodage par un facteur $\cos^m\delta$, d'après (3.16). L'équation (3.84) établie plus loin nous enseigne la façon dont la composante d'ordre 0 doit être égalisée:

$$W = \sqrt{1+2\sum_{m=1}^{M}(1-\cos^{2m}\delta)}\,S \tag{3.33}$$

Egalisation dépendant de la correction des composantes au décodage

En réalité, l'énergie restituée dépend des solutions de décodage mises en oeuvre[16]. Dans le cas d'un décodage modifié par des facteurs $g'_m = g_m/g_0$ (voir 3.3.2), les réinjections énergétiques devraient être pondérées

16. Par contre, l'expression de la réinjection d'énergie (3.33) ou (3.34) ne dépend pas de la convention d'encodage en vigueur!

de la façon suivante:

$$W = \sqrt{1 + 2\sum_{m=1}^{M} (g'_m)^2 (1 - \cos^{2m}\delta)}\, S \tag{3.34}$$

Si, à la restitution et dans une région basse-fréquence, c'est une sommation d'amplitude et non d'énergie qu'il faut considérer au niveau de l'auditeur (voir 3.1.3), le report d'énergie n'a pas lieu d'être dans la bande basse-fréquence en question. Notons que cette compensation met en défaut l'universalité de la représentation ambisonique, et en particulier son indépendance par rapport au mode de restitution.

Soulignons que cette stratégie de compensation énergétique modifie le rapport entre W et les composantes X et Y (ainsi que les autres composantes horizontales), entraînant une diminution de l'effet de latéralisation, de même qu'une exagération probable l'effet de hauteur apparente (site). Plutôt que d'appliquer l'égalisation précédente (3.34), **nous recommandons** donc de pondérer l'ensemble des composantes ambisoniques par le même facteur de normalisation:

$$B^{\sigma}_{mm} = \frac{1}{\sqrt{E(\delta)}} Y^{\sigma}_{mm}(\theta,\delta)\, S, \qquad E(\delta) = 1 + 2\sum_{m=1}^{M} {g'_m}^2 \cos^{2m}\delta \tag{3.35}$$

Cette dernière correction permet en effet de préserver les caractéristiques horizontales des vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ [17].

Egalisation associée à un décodage et une représentation quelconques

La correction proposée ci-dessus n'est en fait valable que lorsque la restitution est homogène dans le plan horizontal, c'est-à-dire lorsque la configuration est régulière – au sens qui sera défini en 3.2.3 – pour la représentation ambisonique considérée. Dans le cas très général d'une configuration non-régulière voire d'une représentation hybride, la méthode d'égalisation consiste à utiliser directement la matrice de décodage **D** pour corriger le vecteur d'encodage **c** associé à la source à traiter. Le vecteur des gains des haut-parleurs s'écrivant $\mathbf{G} = \mathbf{D}.\mathbf{c}$, l'énergie associée a pour expression $E(\mathbf{c}) = \mathbf{G}^t.\mathbf{G} = \mathbf{c}^t.\mathbf{D}^t.\mathbf{D}.\mathbf{c}$. En faisant le choix de préserver le rapport d'amplitude entre les différentes composantes, la conservation de l'énergie de la source consiste à choisir comme nouveau vecteur d'encodage $\tilde{\mathbf{c}}$:

$$\tilde{\mathbf{c}} = \frac{\mathbf{c}}{\sqrt{E(\mathbf{c})}}, \qquad E(\mathbf{c}) = \mathbf{G}^t.\mathbf{G} = \mathbf{c}^t.\mathbf{D}^t.\mathbf{D}.\mathbf{c} \tag{3.36}$$

3.1.5 Manipulations du champ

Rotations

Des transformations du champ sonore encodé peuvent être réalisées par des opérations linéaires simples sur les composantes ambisoniques. Il s'agit en premier lieu des *rotations*, équivalentes à un changement d'axes. La représentation doit être homogène dans les plans orthogonaux aux axes de rotation, sans quoi il y a perte d'information.

Une rotation $\mathbf{B}' = \mathbf{R}.\mathbf{B}$ autour d'un axe quelconque peut être décomposée en produit de trois rotations, d'axes respectifs $\vec{x}$ (*roll*/*tilt*), $\vec{y}$ (*pitch*/*tumble*) et $\vec{z}$ (*yaw*/*rotate*) (Figure 1.4 en 1.3.1), et décrites par les matrices $\mathbf{R}_{\vec{x}}(\gamma)$, $\mathbf{R}_{\vec{y}}(\phi)$ et $\mathbf{R}_{\vec{z}}(\theta)$. Bien souvent, ce sont ces trois rotations élémentaires qui sont directement spécifiées, dans l'ordre *rotate-tumble-tilt*: la matrice de rotation globale s'écrit alors $\mathbf{R}(\theta,\phi,\gamma) = \mathbf{R}_{\vec{x}}(\gamma).\mathbf{R}_{\vec{y}}(\phi).\mathbf{R}_{\vec{z}}(\theta)$. Mais s'il s'agit de décrire la transformation du champ relativement à la tête lorsque c'est elle qui a effectué

17. ... c'est-à-dire les valeurs de leur projection dans le plan horizontal

une rotation[18] décrite par les mouvements *yaw-pitch-roll*, c'est la rotation inverse qui doit être appliquée: $\mathbf{R}^{-1}(\theta,\phi,\gamma) = \mathbf{R}_{\vec{z}}(-\theta).\mathbf{R}_{\vec{y}}(-\phi).\mathbf{R}_{\vec{x}}(-\gamma)$.

Les rotations sont ici décrites par blocs, s'appliquant à des groupes de composantes B^{σ}_{mn} de même ordre m. On notera donc $\mathbf{B}_m = [B^1_{mm}\, B^{-1}_{mm}\, \ldots\, B^1_{m1}\, B^{-1}_{m1}\, B^1_{m0}]^t$ le vecteur des composantes ambisoniques de degré m. En particulier: $\mathbf{B}_0 = [W]$, $\mathbf{B}_1 = [X\,Y\,Z]^t$ et $\mathbf{B}_2 = [U\,V\,S\,T\,R]^t$. Pour le groupe d'ordre m, la transformation s'écrit: $\mathbf{B}'_m = \mathbf{R}_m.\mathbf{B}$. Les rotations élémentaires pour les groupes d'ordres 1 et 2 sont explicitées comme suit pour la convention `SN3D`.

$$\mathbf{R}_{\vec{z},m=1}{}^{(\mathtt{SN3D})}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{R}_{\vec{y},m=1}{}^{(\mathtt{SN3D})}(\phi) = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix},$$
$$\mathbf{R}_{\vec{x},m=1}{}^{(\mathtt{SN3D})}(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{bmatrix} \tag{3.37}$$

$$\mathbf{R}_{\vec{z},m=2}{}^{(\mathtt{SN3D})}(\theta) = \begin{bmatrix} \cos 2\theta & -\sin 2\theta & 0 & 0 & 0 \\ \sin 2\theta & \cos 2\theta & 0 & 0 & 0 \\ 0 & 0 & \cos\theta & -\sin\theta & 0 \\ 0 & 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.38}$$

$$\mathbf{R}_{\vec{y},m=2}{}^{(\mathtt{SN3D})}(\phi) = \begin{bmatrix} \frac{\cos 2\phi+3}{4} & 0 & -\frac{\sin 2\phi}{2} & 0 & \sqrt{3}(1-\cos 2\phi) \\ 0 & \cos\phi & 0 & -\sin\phi & 0 \\ \frac{\sin 2\phi}{2} & 0 & \cos 2\phi & 0 & -\frac{\sqrt{3}}{2}\sin 2\phi \\ 0 & \sin\phi & 0 & \cos\phi & 0 \\ \sqrt{3}\frac{1-\cos 2\phi}{4} & 0 & \frac{\sqrt{3}}{2}\sin 2\phi & 0 & \frac{3\cos 2\phi+1}{4} \end{bmatrix} \tag{3.39}$$

$$\mathbf{R}_{\vec{x},m=2}{}^{(\mathtt{SN3D})}(\gamma) = \begin{bmatrix} \frac{\cos 2\gamma+3}{4} & 0 & 0 & \frac{\sin 2\gamma}{2} & \sqrt{3}(\cos 2\gamma-1) \\ 0 & \cos\gamma & -\sin\gamma & 0 & 0 \\ 0 & \sin\gamma & \cos\gamma & 0 & 0 \\ -\frac{\sin 2\gamma}{2} & 0 & 0 & \cos 2\gamma & -\frac{\sqrt{3}}{2}\sin 2\gamma \\ \sqrt{3}\frac{\cos 2\phi-1}{4} & 0 & 0 & \frac{\sqrt{3}}{2}\sin 2\phi & \frac{3\cos 2\phi+1}{4} \end{bmatrix} \tag{3.40}$$

On peut vérifier que ces matrices de rotation sont compatibles avec celles données par Furse [Fur99], à un changement de convention près.

La rotation du champ ambisonique autour de l'axe $(0,\vec{z})$ – rotation "dans le plan horizontal" – a une expression générique simple. Elle transforme chaque couple de composantes $\mathbf{B}_{mn} = [B^1_{mn}\, B^{-1}_{mn}]^t$ en $\mathbf{B}'_{mn} = [B^{1\prime}_{mn}\, B^{-1\prime}_{mn}]^t$ d'après la relation:

$$\mathbf{B}'_{mn} = \mathbf{R}^{mn}_{\vec{z}}(\theta).\mathbf{B}_{mn}, \quad \text{avec} \quad \mathbf{R}^{mn}_{\vec{z}}(\theta) = \begin{bmatrix} \cos(n\theta) & -\sin(n\theta) \\ \sin(n\theta) & \cos(n\theta) \end{bmatrix} \tag{3.41}$$

Cette matrice de rotation $\mathbf{R}^{mn}_{\vec{z}}$ a la même expression pour toutes les conventions évoquées plus haut.

18. Par exemple, lors d'une restitution au casque avec *Head-Tracking*.

Distorsion de la perspective

Ce qu'on pourrait appeler une *distorsion de la perspective* de la scène sonore peut être réalisée par une forme particulière de transformation de Lorentz[19], appliquée par Gerzon à une représentation ambisonique du premier ordre pour resserrer ou élargir la scène frontale (*forward* ou *backward dominance*) [Ger92a]. En adoptant *la semi-normalisation 3D* `SN3D` (équivalente à `MaxN` ou `SN2D` pour ces composantes d'ordres 0 et 1) comme convention d'encodage (3.1.2), cette opération s'écrit:

$$\mathbf{B}'^{(\mathtt{SN3D})} = \mathbf{L}_{\lambda}^{(\mathtt{SN3D})}.\mathbf{B}^{(\mathtt{SN3D})} \quad \text{où} \quad \mathbf{B}'^{(\mathtt{SN3D})} = \begin{bmatrix} W' \\ X' \\ Y' \\ Z' \end{bmatrix}, \mathbf{B}^{(\mathtt{SN3D})} = \begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix}, \mathbf{L}_{\lambda}^{(\mathtt{SN3D})} = \begin{bmatrix} \frac{\lambda+\lambda^{-1}}{2} & \frac{\lambda-\lambda^{-1}}{2} & 0 & 0 \\ \frac{\lambda-\lambda^{-1}}{2} & \frac{\lambda+\lambda^{-1}}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.42}$$

Les composantes Y et Z restent donc inchangées. Malham [Mal90] fournit des formules équivalentes, en remplaçant le paramètre de contrôle λ (compris entre 0 et ∞) par $\mu = (\lambda - \lambda^{-1})/(\lambda + \lambda^{-1})$ (compris entre -1 et 1) et en pondérant toutes les composantes par $2/(\lambda+\lambda^{-1}) = \sqrt{1-\mu^2}$. La matrice de transformation devient:

$$\mathbf{L}_{\mu}^{(\mathtt{SN3D})} = \begin{bmatrix} 1 & \mu & 0 & 0 \\ \mu & 1 & 0 & 0 \\ 0 & 0 & \sqrt{1-\mu^2} & 0 \\ 0 & 0 & 0 & \sqrt{1-\mu^2} \end{bmatrix} \tag{3.43}$$

Cette transformation a *deux effets*: une source originellement en $(\theta,\delta = 0)$ est *déplacée* en $(\theta',\delta = 0)$:

$$\cos\theta' = \frac{\mu + \cos\theta}{1 + \mu\cos\theta} \tag{3.44}$$

c'est-à-dire vers l'avant si $\mu > 0$ (si $\lambda > 1$), et voit *son gain multiplié* par un facteur $\gamma = 1 + \mu\cos\theta$ pour la version de Malham (3.43), contre un facteur $(\frac{\lambda+\lambda^{-1}}{2} + \frac{\lambda-\lambda^{-1}}{2}\cos\theta)$ pour la version de Gerzon (3.42). On peut vérifier, d'après le rapport entre les composantes, que *les caractéristiques d'onde plane sont bien préservées* (norme du vecteur vélocité égale à 1). Cette transformation est *réversible*, son inverse s'écrit: $\mathbf{L}_{\mu}{}^{-1} = \frac{1}{\sqrt{1-\mu^2}}\mathbf{L}_{-\mu}$.

L'*extension de cette transformation aux ordres supérieurs* n'est pas possible avec la loi de distorsion angulaire (3.44), sauf à détériorer les caractéristiques d'onde plane, c'est-à-dire que la transformation appliquée aux composantes d'une onde plane ne fournit pas des composantes qui puissent être générées par l'encodage d'une onde plane.

L'étude des relations entre fonctions de Legendre associées P_{mn} mériterait d'être poursuivie, dans le but de définir pour chaque ordre supérieur une loi de distorsion angulaire convenable. En considérant une distorsion suivant l'axe polaire $\vec{z}$, cela reviendrait à trouver une relation $\Gamma_{\mu}(z).\mathbf{P}(\mathcal{F}_{\mu}(z)) = \mathbf{L}_{\mu}.\mathbf{P}(z)$, où $\mathbf{P} = [P_{00} \ \ldots \ P_{M0} \ \ldots \ P_{MM}]^t$ est le vecteur des fonctions de Legendre associées jusqu'à l'ordre considéré M, $\mathbf{L}_{\mu}$ est la matrice pour la transformation linéaire cherchée, Γ_{μ} est la fonction de distorsion d'amplitude (croissante si $\mu > 0$) et $\mathcal{F}_{\mu}$ définit la loi de distorsion angulaire, croissante (si $\mu > 0$) sur $[-1, 1]$ et à valeurs $z' = \sin\delta'$ dans $[-1, 1]$.

A partir d'une transformation $\mathbf{L}_{\vec{x}}$ établie suivant l'axe $\vec{x}$ par exemple, la transformation $\mathbf{L}_{\vec{u}}$ suivant un axe quelconque $\vec{u}$ peut bien-sûr être réalisée au moyen d'un changement d'axes, c'est-à-dire en faisant précéder (3.42) ou bien (3.43) d'une rotation $\mathbf{R}$ du champ, et en lui faisant succéder la rotation inverse $\mathbf{R}^{-1}$:

$$\mathbf{L}_{\vec{u}} = \mathbf{R}^{-1}.\mathbf{L}_{\vec{x}}.\mathbf{R}, \qquad \text{où} \quad \mathbf{y}(\vec{u}) = \mathbf{R}.\mathbf{y}(\vec{x}) \tag{3.45}$$

19. Cette transformation vient de la théorie de la relativité.

Focalisation

Un cas particulier de la transformation qui vient d'être présentée consiste à fixer $\mu = 1$. Cela équivaut à un procédé de *focalisation*[20]: il s'agit d'imiter l'effet d'une prise de son directive dans le champ acoustique enregistré à l'aide d'un microphone cardioïde virtuel, lequel résulte d'une combinaison des composantes ambisoniques du format-B. Le champ résultant est calculé en encodant le signal mesuré comme une source unique dans la direction de focalisation.

Il est possible d'étendre ce principe aux *ordres supérieurs* afin d'atteindre une directivité ou une focalisation plus sélective. Pour traiter ce problème, on peut mettre à profit l'interprétation du codage/décodage ambisonique par une prise de son multicanal équivalente qui est développée plus loin (en 3.3.1), et les solutions de décodage modifiées (en 3.3.2), qui consistent à pondérer les composantes ambisoniques suivant leur ordre. En particulier, les solutions *in-phase* généralisées (développées en annexe A.4.3) ont la propriété d'accorder une importance croissante aux incidences qui se rapprochent de la direction de focalisation (Figure 3.14). C'est l'opération de focalisation suivant l'axe vertical (axe polaire) – plus "naturel" pour une représentation 3D – que nous décrivons maintenant[21]. En adoptant comme convention la semi-normalisation 3D (équivalente à la max-normalisation pour les harmoniques sphériques axiales), les composantes B'^{σ}_{mn} issues de la transformation sont telles que:

$$B'^{\sigma}_{mn} = Y^{\sigma}_{mn}(\vec{z})^{(\texttt{SN3D})}\, S' = \delta_{n0}\, S' \quad \text{avec} \quad S' = \sum_{m=0}^{M} (2m+1) g_m^{3D(M)} B^1_{m0}, \tag{3.46}$$

où les facteurs $g_m^{3D(M)}$ sont ceux définis en annexe A.4.3 ou encore donnés par (3.91).

3.2 Propriétés liées à la troncature des décompositions

Il est utile, pour définir les qualités générales à attendre d'un système ambisonique, de s'intéresser aux propriétés fondamentales de la troncature d'une décomposition en harmoniques sphériques (pour les systèmes 3D) ou cylindriques (pour les systèmes 2D).

3.2.1 Expansion radiale et fréquentielle de l'approximation

La décomposition d'une plane onde d'incidence $(\theta_S = 0, \delta_S = 0)$[22] tronquée à un ordre M réalise une approximation du champ acoustique, dont les valeurs dans le plan horizontal sont données par:

$$p_M^{2D}(kr,\theta) = S\left(J_0(kr) + 2\sum_{m=1}^{M} j^m J_m(kr)\cos(m\theta)\right) \tag{3.47}$$

pour la décomposition cylindrique, et pour la décomposition sphérique:

$$p_M^{3D}(kr,\theta) = S\sum_{m=0}^{M} (2m+1)\, j^m\, j_m(kr) P_m(\cos\theta), \tag{3.48}$$

l'onde plane de référence s'écrivant elle-même:

$$p(kr,\theta) = S e^{jkr\cos\theta} = p_\infty^{2D}(kr,\theta) = p_\infty^{3D}(kr,\theta) \tag{3.49}$$

20. Procédé, encore appelé *focus*, mis en oeuvre par Véronique Larcher pour un travail de composition sonore spatialisée de Cécile le Prado. L'analogue visuel approximatif serait l'effet d'une lampe torche balayant l'obscurité.

21. Rappelons que les composantes "polaires" sont $W = B^1_{00}$, $Z = B^1_{10}$, $R = B^1_{20}$, etc...

22. Ce choix de l'incidence ne restreint en rien la généralité des résultats qui suivent.

Ces approximations peuvent être réalisées par *superposition d'un nombre infini d'ondes planes*, dont les vecteurs incidences $\vec{u}$ décriraient respectivement le cercle unité horizontal $\mathbb{U}_2$ ou la sphère unité $\mathbb{U}_3$:

$$p_M^{2D}(kr,\theta) = \int_0^{2\pi} \left(1 + 2\sum_{m=1}^{M} \cos(m\phi)\right) S e^{jkr\cos(\phi-\theta)}\, d\phi \tag{3.50}$$

$$p_M^{3D}(k\vec{r}) = \int_{\mathbb{U}_3} \left(\sum_{m=0}^{M} (2m+1) P_m(\vec{u}.\vec{x})\right) S e^{jk\vec{u}.\vec{r}}\, d^2\vec{u} \tag{3.51}$$

Dans chaque équation, le terme entre parenthèses définit l'amplitude de l'onde plane élémentaire de direction $\vec{u}$ ou ϕ. Il s'agit en ce sens d'un *cas particulier*, bien qu'irréalisable, *de restitution ambisonique* à l'aide d'un dispositif respectivement 2D (horizontal) ou 3D de haut-parleurs. Partant de cette observation, on espère que les caractéristiques d'une troncature, en terme de qualité d'approximation, sont représentatives de ce qu'on peut attendre d'une restitution à partir d'une représentation ambisonique du même ordre, et à l'aide d'un dispositif "réaliste", c'est-à-dire utilisant un nombre fini de haut-parleurs (ce qui sera développé en 3.3). L'avantage de travailler directement sur les troncatures est de fournir des résultats génériques, indépendants d'un dispositif de restitution particulier, et de la direction d'incidence de l'onde plane. La figure 3.7 permet de visualiser l'évolution de l'approximation d'une onde plane en fonction de l'ordre de troncature.

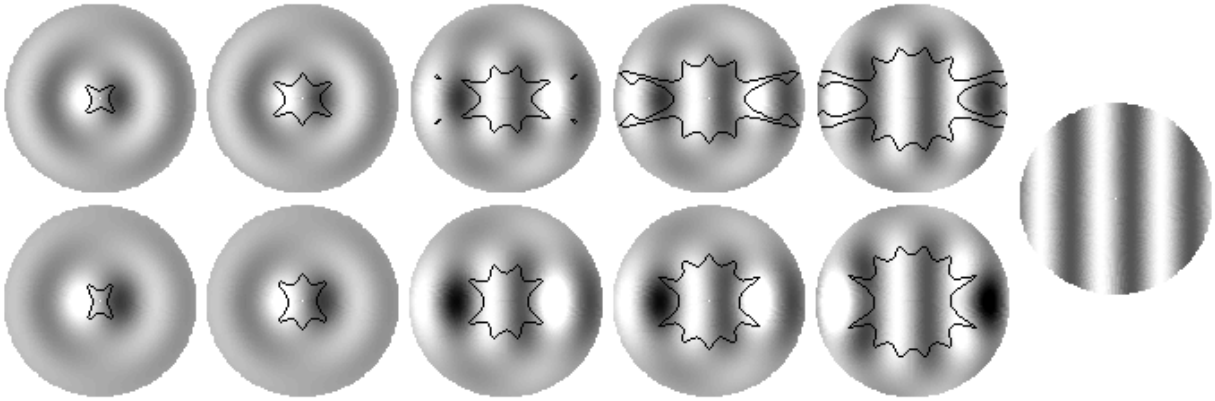

FIG. 3.7 – *Représentations (vues instantanées, coupes horizontales) du champ acoustique correspondant aux troncatures cylindriques (en haut) et sphériques (en bas) d'une onde plane (à droite) pour des ordres allant de 1 à 5 (de gauche à droite). Le niveau de gris indique la valeur de pression instantanée. Les courbes étoilées indique les limites des zones d'approximation, pour une tolérance d'erreur de 20%.*

Erreurs de reconstruction dans le plan horizontal

Afin d'estimer la qualité de l'approximation (ou de la reconstruction) de l'onde plane, deux mesures d'erreur, proposées dans [BV95] et [Pol96a], sont utilisées: la moyenne $\bar{\varepsilon}_{p_M}$ de la valeur absolue de l'erreur, calculée par intégration sur un périmètre circulaire de rayon kr autour de l'origine $\vec{r} = 0$, et son maximum ε_{p_M} sur ce même périmètre:

$$\bar{\varepsilon}_{p_M}(kr) = \frac{1}{2\pi} \int_0^{2\pi} |p_M(kr,\theta) - p(kr,\theta)|\, d\theta \tag{3.52}$$

$$\varepsilon_{p_M}(kr) = \max_{\theta} |p_M(kr,\theta) - p(kr,\theta)| \tag{3.53}$$

La figure 3.8 montre le comportement de ces deux mesures pour des ordres de 1 à 5. La table 3.4 indique les fréquences limites correspondantes en fixant un rayon $r = 8{,}75$ cm (reflétant les dimensions typiques d'une tête) et une tolérance d'erreur de 20%. Les résultats concernant les troncatures cylindriques sont du même ordre que ceux présentés par Poletti pour une reconstruction avec un système multi-canal horizontal [Pol96a][23]. La progression de la qualité de reconstruction en fonction de l'ordre apparaît de façon évidente. La reconstruction ou l'approximation par la troncature cylindrique, est, à ordre égal, légèrement meilleure que celle obtenue par troncature sphérique, ce qui apparaît plus nette avec l'erreur ε_{p_M}, globalement plus sévère.

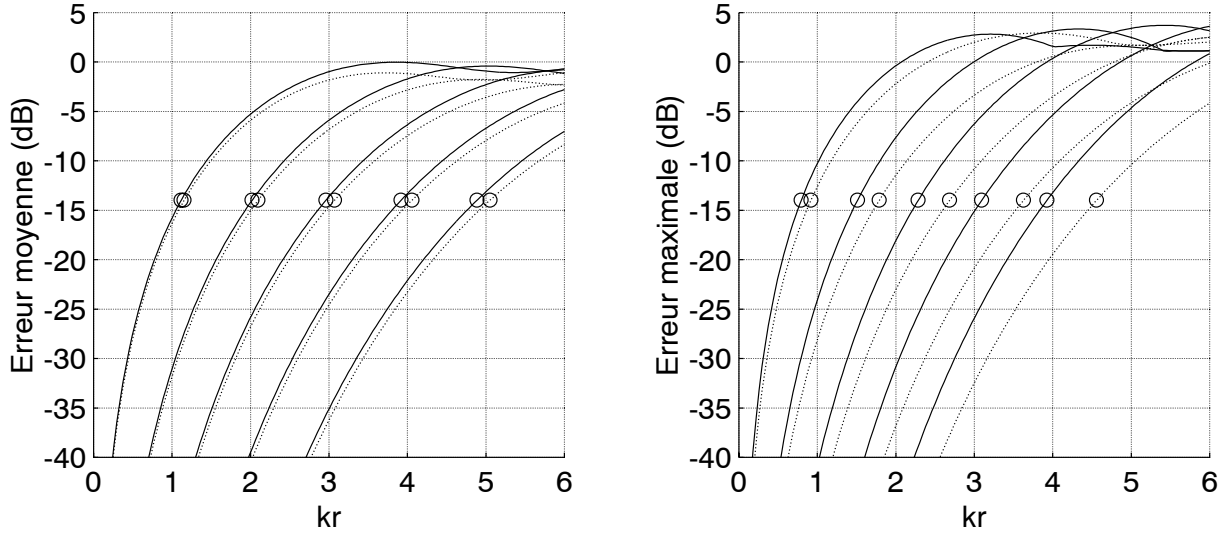

FIG. 3.8 – *Erreurs d'approximation du champ dues aux troncatures de décompositions cylindrique (pointillés) et sphérique (traits continus) du champ. Les ordres de troncature vont de 1 à 5 (indices placés près des courbes). Les cercles indiquent un seuil d'erreur de 20% (-14dB).*

ordre M	1	2	3	4	5
f_l (sphérique)	491	935	1414	1914	2428
f_l (cylindrique)	568	1106	1662	2244	2820
$\bar{f}_l$ (sphérique)	690	1250	1831	2423	3022
$\bar{f}_l$ (cylindrique)	714	1297	1899	2509	3125

TAB. 3.4 – *Fréquences limites (en Hz) de validité de l'approximation d'une onde plane réalisée par des troncatures cylindriques et sphériques, sur un disque centré de diamètre 17,5cm. Tolérance de 20% (-14dB) sur les erreurs ε_{p_M} et $\bar{\varepsilon}_{p_M}$ respectivement.*

Il est intéressant de constater que *les erreurs de reconstruction* dues aux troncatures (disons à l'ordre

23. A noter pour éviter des confusions: les courbes d'erreur dans cette référence sont exprimées en fonction de kd, où d est le diamètre critique pour une fréquence donnée, soit $kd = 2kr$, et non en fonction de kr, comme ici. Pour le calcul des fréquences limites, il est choisi légèrement inférieur au nôtre (16,5cm contre 17,5cm).

M) cylindrique et sphérique *peuvent être plus directement approchées par les fonctions* $\sqrt{2}J_{M+1}(kr)$ *et* $\sqrt{2M+3}j_{M+1}(kr)$ respectivement, en tous cas sur leur portion croissante. Comme le montre la figure 3.9, *l'erreur max est particulièrement bien approchée*, légèrement par dessous dans le cas cylindrique, et très légèrement par dessus dans le cas sphérique. Ces fonctions sont donc utilisables par exemple s'il s'agit de déterminer l'ordre de troncature nécessaire pour produire l'approximation d'un champ avec une tolérance donnée, selon la valeur maximale atteinte par kr.

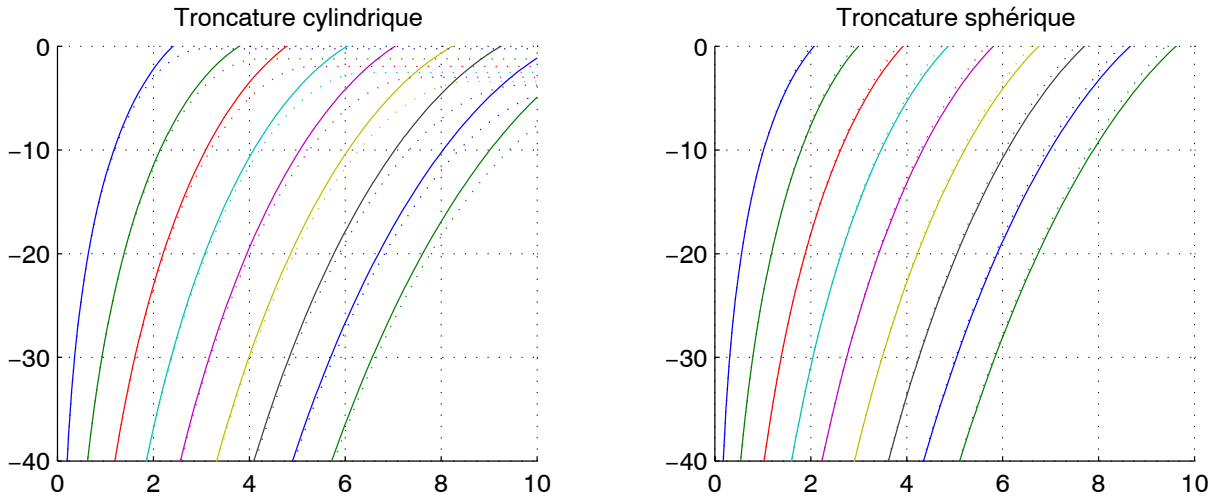

FIG. 3.9 – *Approximation des erreurs max de troncature cylindrique et sphérique (traits continus), par les fonctions* $\sqrt{2}J_{M+1}(kr)$ *et* $\sqrt{2M+3}j_{M+1}(kr)$ *(pointillés), pour des ordres de troncature* $M = 1,\dots 9$. *Pour la clarté de la figure, ces fonctions ont été rendues monotones: c'est leur portion croissante initiale qui est pertinente.*

Report de l'erreur sur la reconstruction des HRTF

Ces observations, qui portent sur la reconstruction d'une onde plane en champ libre, méritent d'être complétées par une caractérisation de l'effet de diffraction mesuré aux oreilles d'un auditeur qui serait centré en $\vec{r} = 0$. Pour simplifier les calculs, la tête est ici modélisée par une sphère de rayon $R = 8{,}75$ cm, et le reste du corps est négligé. Les HRTF $H(f,\theta)$ sont mesurables à travers les valeurs du champ en surface de la sphère. Dans le cas d'une troncature sphérique, le calcul des HRTF $H_M(f,\theta)$ à partir la décomposition sphérique limitée à l'ordre M est alors direct (voir annexe: diffraction par une sphère). Dans le cas cylindrique, en revanche, le calcul se base sur l'équation (3.50), l'effet de la diffraction devant être calculé pour chaque onde plane. De la même manière que pour l'approximation en champ libre, deux mesures d'erreur peuvent être introduites pour caractériser de façon globale les HRTF correspondant aux différentes troncatures. En se limitant aux incidences horizontales, les erreurs moyenne et max sont définies en fonction de la fréquence f par:

$$\bar{\varepsilon}_{H_M}(f) = \frac{1}{2\pi}\int_0^{2\pi} |H_M(f,\theta) - H(f,\theta)|\, d\theta \tag{3.54}$$

et:

$$\varepsilon_{H_M}(f) = \max_{\theta} |H_M(f,\theta) - H(f,\theta)| \tag{3.55}$$

Ces erreurs sont illustrées Figure 3.10 pour des ordres de 1 à 5, et les fréquences limites correspondantes sont reportées Table 3.5. Il ressort de la figure 3.10 les mêmes tendances qui pouvaient être observées Figure 3.8. Les mêmes seuils d'erreur révèlent cependant des fréquences limites plus critiques (Table 3.5). Cela dit, les mesures d'erreur globale utilisées n'ont pas la même signification selon qu'elles concernent l'approximation du champ ou celle des HRTF.

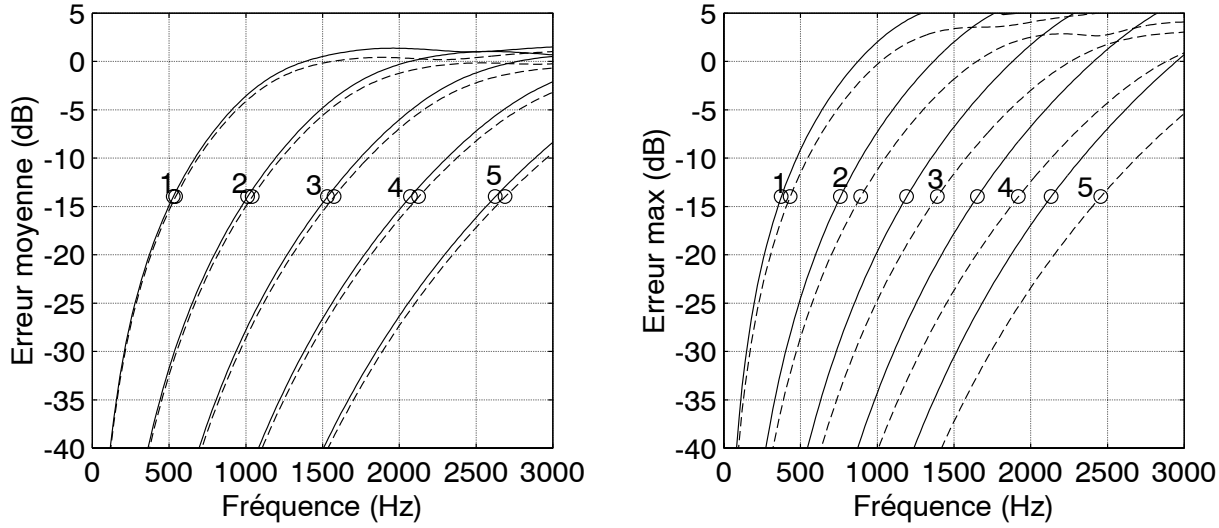

FIG. 3.10 – *Erreurs sur les HRTF dues aux troncatures de décompositions cylindrique (tirets) et sphérique (traits continus) du champ. Ordres de troncature: de 1 à 5 (indices placés près des courbes). Modèle de tête sphérique (diamètre 17,5cm).*

ordre M	1	2	3	4	5
f_l (sphérique)	375	760	1192	1653	2133
f_l (cylindrique)	433	893	1391	1920	2457
$\bar{f}_l$ (sphérique)	528	1011	1532	2072	2626
$\bar{f}_l$ (cylindrique)	543	1042	1575	2127	2691

TAB. 3.5 – *Fréquences limites (en Hz) de validité de l'approximation des HRTF associée aux troncatures cylindriques et sphériques, pour un modèle de tête sphérique de diamètre 17,5cm. Tolérance de 20% (-14dB) sur les erreurs* ε_{H_M} *et* $\bar{\varepsilon}_{H_M}$ *respectivement.*

Les limites fréquentielles de validité de la reconstruction sont données pour un seuil de tolérance à l'erreur arbitraire, qui est ici choisi relativement peu sévère par rapport à d'autres études [NE99]. Pour rendre compte d'une limite fréquentielle pertinente pour la perception, il est plus intéressant d'observer les effets de l'approximation/reconstruction sur les indices de localisation, en particulier l'ITD. C'est d'ailleurs son observation qui a paru la plus probante dans [DRP98]. En mettant en correspondance la fréquence-limite pour la reconstruction de l'ITD et l'erreur $\bar{\varepsilon}_{p_M}$ ou $\bar{\varepsilon}_{H_M}$ pour un système d'ordre 2, on détermine d'après cette étude un seuil de tolérance compris entre 14% (-17 dB) et 20% (-14 dB).

Il faut signaler que la reproduction de l'effet de diffraction, même mesuré dans le plan horizontal, dépend de l'efficacité de la reconstruction du champ (en l'absence de tête) *sur un volume* qui engloberait la tête,

ce que que ne traduisent pas les mesures d'erreur ε_M et $\bar{\varepsilon}_M$. Enfin, on peut s'attendre à ce que l'écart de "performance" entre les deux troncatures – cylindrique et sphérique – se creuse encore s'il on réintègre le reste du corps dans la modélisation des HRTF.

L'estimation d'une fréquence limite de reconstruction basse-fréquence, dont quelques méthodes viennent d'être exposées, peut servir à déterminer une fréquence de transition entre les solutions de décodage *basique* (basse-fréquence) et *max* r_E (haute-fréquence) [DRP98]. Mais ce qui semble déterminant pour définir une telle fréquence de transition, c'est finalement de savoir à partir de quel moment la deuxième solution de décodage devient "meilleur" que la première. Une méthode est proposée pour cela dans la section suivante (3.2.2).

3.2.2 Optimisation de la propagation globale: troncature avec et sans biais

Si l'approximation du champ liée à la représentation compacte (la troncature) est valable sur un voisinage du point central $\vec{r}=0$, c'est-à-dire sur la zone de convergence ou l'interférence des ondes planes (3.50) (3.51) est "contrôlée", il n'en reste pas moins intéressant de caractériser le champ hors de cette zone – où les déphasages entre les contributions sont statistiquement aléatoires. L'étude présentée en 1.5.1 montre que le vecteur énergie $\vec{E}$ traduit le transport global d'énergie acoustique, dont l'onde plane progressive constitue le cas "optimal", *i.e.* tel que $r_E=|\vec{E}|=1$. On peut définir le vecteur énergie associé aux troncatures cylindriques et sphériques d'une onde plane, à partir de leur expression sous formes de sommes intégrales d'ondes planes (équations 3.50 et 3.51). Pour une incidence $(\theta_S=0,\delta_S=0)$ (direction $\vec{x}$), une troncature cylindrique d'ordre M donne:

$$\vec{E}=\frac{\int_0^{2\pi}|a(\theta)|^2\vec{u}_\theta\,d\theta}{\int_0^{2\pi}|a(\theta)|^2\,d\theta}\quad\text{avec}\quad a(\theta)=1+2\sum_{m=1}^{M}\cos(m\theta)\quad\text{et}\quad\vec{u}_\theta=\cos\theta\vec{x}+\sin\theta\vec{y}\tag{3.56}$$

et une décomposition sphérique d'ordre M donne:

$$\vec{E}=\frac{\int_{\mathbb{U}_3}|a(\vec{u})|^2\vec{u}\,d^2\vec{u}}{\int_{\mathbb{U}_3}|a(\vec{u})|^2\,d^2\vec{u}}\quad\text{avec}\quad a(\vec{u})=\sum_{m=0}^{M}(2m+1)P_m(\vec{u}.\vec{x})\tag{3.57}$$

Les propriétés des familles de fonctions $\cos(m\theta)$ et $P_m(\cos\theta)$ sont telles que $\vec{E}$ prend la direction – ici $\vec{x}$ – de l'onde plane "tronquée". En outre, on peut dégager une relation entre l'indice $r_E=|\vec{E}|$ et l'ordre de la troncature ou par extension, de la représentation (A.4.1):

$$\vec{E}=r_E\,\vec{u},\qquad\text{où}\quad r_E=\begin{cases}2M/2M+1 & \text{(Troncature cylindrique)}\\ M/M+1 & \text{(Troncature sphérique)}\end{cases}\tag{3.58}$$

On constate ainsi qu'*augmenter l'ordre d'une représentation ambisonique n'a pas la seule propriété d'élargir la zone de reconstruction, mais que cela améliore aussi les propriétés du flux d'énergie hors de cette zone*, ce que traduisent le vecteur énergie $\vec{E}$ et sa norme (Table 3.6). Sur le plan perceptif, cette amélioration est sensible à travers les indices de localisation (ITD et ILD) haute-fréquence (section 1.5). De surcroît, on montre qu'il est possible de "biaiser" la troncature en fonction de son ordre, pour optimiser le flux d'énergie en maximisant r_E (Table 3.6)[24]:

$$\bar{p}_M^{2D}(kr,\theta)=S\left(g_0^{2D(M)}J_0(kr)+2\sum_{m=1}^{M}j^m g_m^{2D(M)}J_m(kr)\cos(m\theta)\right)\quad\Rightarrow\quad a(\theta)=g_0^{2D(M)}+2\sum_{m=1}^{M}g_m^{2D(M)}\cos(m\theta)\tag{3.59}$$

24. Il pourrait s'agir d'une forme de projection du champ acoustique sur la base des harmoniques sphériques, mais au sens d'une autre distance (ou norme).

$$\tilde{p}_M^{3D}(kr,\theta) = S\sum_{m=0}^{M}(2m+1)\,j^m g_m^{3D(M)}\,j_m(kr)P_m(\cos\theta) \quad \Rightarrow \quad a(\vec{u}) = \sum_{m=1}^{M}(2m+1)g_m^{3D(M)}P_m(\cos\theta) \quad (3.60)$$

Il s'agit donc de pondérer les composantes cylindriques ou sphériques, à l'instar des décodages modifiés présentés en 3.3.2. Les facteurs de pondération optimaux $g_m^{2D(M)}$ et $g_m^{3D(M)}$ sont donnés dans [DRP98] (Annexe B) et en A.4.2, ou encore Table 3.10.

M	formules génériques	1	2	3	4	5	6
$r_E(2D)$ (brute)	$\frac{2M}{2M+1}$	0.6667	0.8000	0.8571	0.8889	0.9091	0.9231
$r_E(3D)$ (brute)	$\frac{M}{M+1}$	0.5000	0.6667	0.7500	0.8000	0.8333	0.8571
$r_E(2D)$ (biaisée)	$\cos(\frac{\pi}{2M+2})$	0.7071	0.8660	0.9239	0.9511	0.9659	0.9749
$r_E(3D)$ (biaisée)	$P_{M+1}^{-1}(0)$	0.5774	0.7746	0.8611	0.9062	0.9325	0.9491

TAB. 3.6 – *Indices r_E du transport énergétique pour les troncatures brutes et biaisées (pour maximisation) d'une onde plane, en fonction de l'ordre M. $P_{M+1}^{-1}(0)$ désigne la plus grande racine du polynôme de Legendre de degré $M+1$.*

La table 3.6 montre la différence entre les troncatures sphériques et cylindriques encore plus clairement que l'étude des fréquences limites (Table 3.4). A ordre égal, le cas sphérique affiche toujours très nettement l'indice r_E le plus bas.

Un autre résultat relatif à ces troncatures biaisées, qui émerge de [DRP98] et de 3.3.2, est que la propagation énergétique *locale* au voisinage du point $\vec{r}=0$ (quantifiée par r_V, d'après 1.2.2) est alors équivalente à sa caractérisation globale: $r_V = r_E$. En corollaire, il semble que l'expansion radiale de la caractérisation locale $r_V = r_E$ soit alors maximale.

Conséquences sur l'ITD en présence de la tête

Il semble judicieux de vérifier la validité et la portée de la loi de prédiction (1.80) de l'ITD basse-fréquence (retard interaural de phase) d'après le vecteur vélocité $\vec{V}$. Pour plus de commodité, nous gardons le modèle d'une tête sphérique et observons uniquement l'ITD maximal, correspondant au cas d'une onde plane se propageant suivant l'axe interaural. La figure 3.11 montre le rapport $\eta_{TD}(f)$ entre l'ITD ($\text{ITD}^{2D(M)}$ ou $\text{ITD}^{3D(M)}$) correspondant à chaque troncature (biaisée ou non-biaisée, d'ordre 1 à 5) et l'ITD de référence ITD_{ref}, dû à l'onde plane originale non-tronquée.

On constate (Figure 3.11) que dans chaque cas, le rapport d'ITD $\eta_{TD}(f)$ tend bien vers la valeur r_V attendue quand la fréquence diminue: $r_V = 1$ pour les troncatures brutes et $r_V = r_E$ pour les troncatures biaisées (*max* r_E). Pour un ordre de troncature donné, la prédiction d'après r_V apparaît nettement plus stable si la troncature est biaisée pour maximiser r_E, alors que le rapport d'ITD associé à la troncature brute décroît assez rapidement. Dans le cas cylindrique (2D), ce rapport chute en dessous de celui associé à la troncature biaisée, à une certaine fréquence qui pourrait donc être choisie comme fréquence de transition entre le décodage basique et le décodage *max* r_E. Il est un peu décevant de ne pas observer un comportement semblable avec le cas sphérique (3D): chaque courbe *max* r_E reste inférieure à la courbe *basique* de même ordre, les deux convergeant l'une vers l'autre asymptotiquement. Il est probable qu'une amélioration haute-fréquence apportée par le biais *max* r_E soit sensible à travers d'autres indices comme le retard de groupe ou l'ILD.

Définition de fréquences de transition pour le décodage

S'inspirant du cas de la troncature cylindrique pour lequel le rapport η_{TD} de chaque version biaisée reste à peu près stable jusqu'à son intersection avec la courbe de la version "brute", nous suggérons d'extraire les

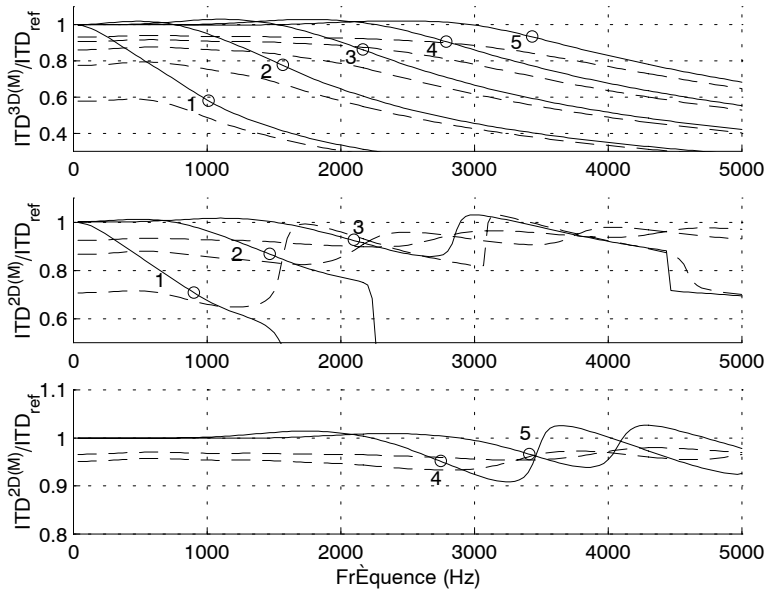

FIG. 3.11 – *Rapport entre l'ITD maximal dû aux troncatures d'onde plane, et l'ITD de référence dû à l'onde plane originale non tronquée. En haut: troncatures sphériques, d'ordres 1 à 5. Au milieu et en bas: troncatures cylindriques d'ordres 1 à 3, puis de 4 à 5. Traits continus: non-biaisées ou brutes (*basique*). Tirets: biaisées (*max r_E*). Les cercles indiquent le passage des courbes "non-biaisées" par la valeur $r_V = r_E$ des versions biaisées de même ordre.*

fréquences de transition de la façon suivante: on retient la fréquence à laquelle la valeur $r_{ITD}(f)$ de la version brute devient inférieure à l'indice $r_V = r_E$ de la version biaisée *max* r_E du même ordre (Figure 3.11). A titre indicatif, les fréquences obtenues sont reportées Table 3.7. Bien qu'on observe une évolution semblable en fonction de l'ordre M, on note des valeurs largement supérieures à celles données par les tables 3.4 et 3.5, ce qui confirme l'arbitraire des méthodes précédentes, basées sur le choix d'un seuil de tolérance à l'erreur d'approximation. Pour le décodage 2D et dans l'hypothèse d'une tête rigoureusement centrée, nous recommandons des fréquences de transition f_{trans}^{2D} proches de celles la table 3.7. Celles données pour le cas 3D sont en revanche moins fiables, d'autant qu'elles sont supérieures aux fréquences f_{trans}^{2D} à ordre égal, contrairement à ce qu'indiquent les autres tables.

M	1	2	3	4	5
f_{trans}^{2D} (Hz)	900	1475	2100	2750	3410
f_{trans}^{3D} (Hz)	1015	1570	2170	2800	3435

TAB. 3.7 – *Suggestion de fréquences de transition entre décodages basse-fréquence (*basique*) et haute-fréquence (*max r_E*) pour une position d'écoute centrée, d'après l'observation du retard interaural de phase (Figure 3.11).*

3.2.3 Echantillonnage angulaire ou directionnel

La notion d'échantillonnage des fonctions harmoniques sphériques et cylindriques est présente de façon implicite dans les problèmes duaux de la prise de son ambisonique, où le champ acoustique est mesuré à

partir d'un nombre fini de capteurs (microphones), et de la restitution, où l'on tente de reconstituer un événement acoustique par superposition d'un nombre fini d'ondes rayonnées par des haut-parleurs. Les éventuelles propriétés de régularité de l'échantillonnage – en partie liées à la régularité géométrique du dispositif de transducteurs – interviennent en particulier dans la résolution du problème de décodage et se reportent sur les qualités de la restitution (l'homogénéité par exemple).

Propriétés de régularité

Support d'échantillonnage. Il convient d'abord de distinguer deux types de support d'échantillonnage selon que l'on s'intéresse à une représentation ambisonique purement horizontale (2D) ou bien 3D:

- Le cas 2D (harmoniques cylindriques) met en jeu un échantillonnage angulaire ou azimutal du cercle unité $\mathbb{U}_2$, soit un ensemble $\Theta^{2D} = \{\theta_i\}_{i=1,\dots,N}$ de N azimuts θ_i ou de façon équivalente un ensemble $\Theta^{2D} = \{\vec{u}_i\}_{i=1,\dots,N}$ de N vecteurs $\vec{u}_i \in \mathbb{U}_2$.
- Dans le cas 3D (harmoniques sphériques), c'est un échantillonnage de la sphère unité $\mathbb{U}_3$ qui est retenu: un ensemble $\Theta^{3D} = \{\vec{u}_i\}_{i=1,\dots,N}$ d'incidences $\vec{u}_i(\theta_i,\delta_i) \in \mathbb{U}_3$.

Considérons une base orthonormée de K fonctions harmoniques cylindriques[25] $\mathbf{Y} = [\tilde{Y}^{\sigma_1}_{m_1} \dots \tilde{Y}^{\sigma_K}_{m_K}]^t$. Ces fonctions sont orthonormées au sens du produit scalaire défini par (A.7). Le support Θ^{2D} réalise un échantillonnage de chaque fonction $\tilde{Y}^{\sigma}_{m}$, représenté par le vecteur $\mathbf{y}^{\sigma}_{m} = [\tilde{Y}^{\sigma}_{m}(\theta_1) \dots \tilde{Y}^{\sigma}_{m}(\theta_N)]^t$. On peut dès lors introduire la matrice-base $\mathbf{C} = [\mathbf{y}^{\sigma_1}_{m_1} \dots \mathbf{y}^{\sigma_K}_{m_K}]^t$.

De façon analogue, considérant une base orthonormée – au sens de (A.16) cette fois-ci – d'harmoniques sphériques[26] $\mathbf{Y} = [\tilde{Y}^{\sigma_1}_{m_1 n_1} \dots \tilde{Y}^{\sigma_K}_{m_K n_K}]^t$, le support Θ^{3D} réalise un échantillonnage $\mathbf{y}^{\sigma}_{mn} = [\tilde{Y}^{\sigma}_{mn}(\theta_1,\delta_1) \dots \tilde{Y}^{\sigma}_{mn}(\theta_N,\delta_N)]^t$ de chaque fonction, dont on déduit la matrice $\mathbf{C} = [\mathbf{y}^{\sigma_1}_{m_1 n_1} \dots \mathbf{y}^{\sigma_K}_{m_K n_K}]^t$ qui représente la base échantillonnée.

Régularité. Le support d'échantillonnage Θ est dit *régulier* pour la base $\mathbf{Y}$ s'il *préserve la propriété d'orthonormalité* de la base échantillonnée, c'est-à-dire si les vecteurs $\mathbf{y}^{\sigma}_{m}$ sont orthonormés au sens du produit scalaire $\langle \mathbf{x}|\mathbf{y}\rangle_N = (\mathbf{x}^t.\mathbf{y})/N$:

$$\begin{array}{ll} \langle \mathbf{y}^{\sigma}_{m}|\mathbf{y}^{\sigma'}_{m'}\rangle_N & = \delta_{mm'}\delta_{\sigma\sigma'} \qquad \text{(Echantillonnage 2D)} \\ \langle \mathbf{y}^{\sigma}_{mn}|\mathbf{y}^{\sigma'}_{m'n'}\rangle_N & = \delta_{mm'}\delta_{nn'}\delta_{\sigma\sigma'} \qquad \text{(Echantillonnage 3D)} \end{array} \tag{3.61}$$

La propriété de *régularité* se traduit de manière plus compacte par:

$$\frac{1}{N}\mathbf{C}.\mathbf{C}^t = \mathbf{I}_K, \tag{3.62}$$

où K est le nombre de fonctions harmoniques échantillonnées et $\mathbf{I}_K$ est la matrice identité de rang K. On rappelle que $K = 2M+1$ pour une représentation 2D homogène et $K = (M+1)^2$ pour une représentation 3D homogène.

Semi-régularité. Un échantillonnage est dit *semi-régulier* s'il préserve *l'orthogonalité* de la base échantillonnée, et non plus nécessairement l'orthonormalité. Ceci se traduit par le fait que *la matrice* $\mathbf{C}.\mathbf{C}^t$ *est diagonale*.

Exemples et spécimens

Dans la suite, nous confondons le support d'échantillonnage du cercle unité (resp. de la sphère unité) avec le polygone (resp. le polyèdre) dont il définit les sommets, et nous ne considérons que des représentations homogènes 2D ou 3D.

25. Rappel: $\tilde{Y}^{\sigma}_{m}(\theta) = Y^{\sigma}_{mm}{}^{(\mathtt{N2D})}(\theta,0)$.
26. Rappel: $\tilde{Y}^{\sigma}_{mn} = Y^{\sigma}_{mn}{}^{(\mathtt{N3D})}$.

Dans le cas des représentations 2D, les supports d'échantillonnage réguliers sont typiquement les polygones réguliers dont le nombre N de sommets est au moins égal au nombre $2M+1$ de composantes ambisoniques. La concaténation de deux supports réguliers constitue un support régulier, cette propriété étant transposable à celle de semi-régularité. Un cas typique de semi-régularité pour le premier ordre est l'échantillonnage rectangulaire.

La question de l'existence des configurations régulières pour l'échantillonnage sphérique (3D) est en revanche beaucoup moins évidente. D'une part, l'existence de polyèdres réguliers est elle-même très limitée: ceux qui existent ont 4, 6, 8, 12 ou 20 sommets (Figure 3.12 et Table 3.8). D'autre part, la régularité géométrique, même accompagnée de la condition $N \geq 2M+1$, ne semble pas impliquer nécessairement la régularité pour l'échantillonnage des harmoniques sphériques: par exemple, les 20 sommets d'un dodécaèdre régulier (Figure 3.12) ne constituent pas un support rigoureusement régulier pour une représentation d'ordre $M=3$, qui ne comprend pourtant que $K=16$ composantes. Un échantillonnage plus affiné de la sphère unité peut être obtenu par triangulation des faces d'un polyèdre régulier [Wei99], et il est vraisemblable que cette expansion géodésique des polyèdres réguliers donne lieu à une approximation de plus en plus correcte de la propriété de régularité. Il resterait à vérifier si l'on peut ou non définir de cette manière un support d'échantillonnage régulier ou même semi-régulier pour l'ordre 3. Par exemple, les 32 sommets du dodécaèdre et de l'icosaèdre réunis (Figure 3.12) ne constituent toujours pas un échantillonnage rigoureusement régulier ni semi-régulier pour l'ordre 3. Cependant, étant donné les faibles valeurs des coefficients hors-diagonale de $\frac{1}{N}\mathbf{C}.\mathbf{C}^t$ pour une matrice-base $\mathbf{C}$ d'ordre $M=3$ ou 4, nous dirons de cet échantillonnage qu'il est *quasi-régulier pour les ordres 3 et 4*. Citons, comme cas de configuration semi-régulière pour l'ordre $M=2$, la concaténation des sommets d'un cube avec ceux d'un octaèdre ($8+6=14$ sommets, Figure 3.12). On peut chercher des modèles de configurations semi-régulières parmi les solides archimédiens et les solides catalans (polyèdres canoniques) [Wei99].

Nom	Nombre de sommets	Nombre de faces	Régularité (ordre)
Tétraèdre	4	4	$M=1$ ($K=4$)
Octaèdre	6	8	$M=1$ ($K=4$)
Cube	8	6	$M=1$ ($K=4$)
Icosaèdre	12	20	$M=2$ ($K=9$)
Dodécaèdre	20	12	$M=2$ ($K=9$)

TAB. 3.8 – *Polyèdres réguliers (convexes), réalisant chacun un échantillonnage de la sphère unité. L'ordre maximal de la base d'harmoniques sphériques (3D) pour lequel l'échantillonnage est régulier est indiqué dans la colonne de droite.*

3.2.4 Conclusions

En s'intéressant à une troncature d'ordre M de la décomposition du champ en harmoniques cylindriques ou sphériques, on a pu caractériser partiellement l'*information spatiale contenue* dans une représentation ambisonique homogène 2D ou 3D d'ordre M, à travers les *qualités potentielles de la restitution*, en considérant que la troncature constitue un cas très particulier de reconstruction du champ acoustique original. Ces qualités couvrent non seulement l'expansion radiale ou fréquentielle de la *reconstruction locale*, mais aussi les *caractéristiques globales* de la propagation, que traduisent le vecteur énergie$\vec{E}$ et son module r_E. Il a été montré que la troncature pouvait être *biaisée* – en pondérant les termes du développement de Fourier-Bessel tronqué – pour optimiser les caractéristiques de propagation au sens du vecteur énergie, le cas idéal étant celui d'une onde plane ($r_E = |\vec{E}| = 1$). Cette correction se retrouve dans un contexte de restitution sous la

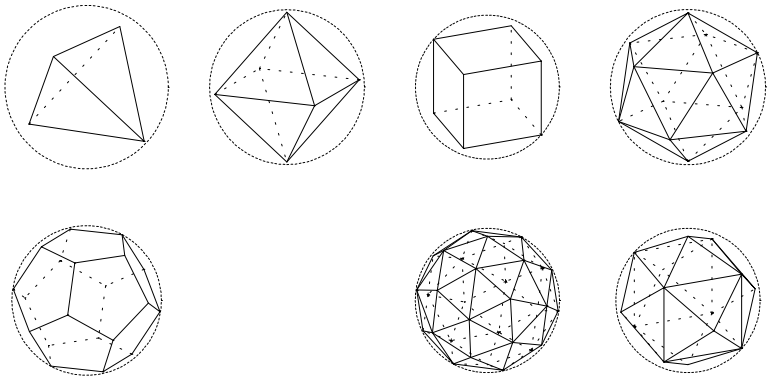

FIG. 3.12 – *De haut en bas et de gauche à droite. Les cinq polyèdres réguliers convexes existants (encore appelés solides platoniques): tétraèdre, octaèdre, cube, icosaèdre, dodécaèdre. Un exemple d'expansion géodésique: "concaténation" de l'icosaèdre et du dodécaèdre (ou encore triangulation de l'un des deux), polyèdre à 32 sommets, régulier pour l'échantillonnage d'ordre 2 maximum, et seulement quasi-régulier pour les ordres 3 et 4. "Concaténation" de l'octoèdre et du cube (ou encore triangulation de l'un des deux), polyèdre à 14 sommets, semi-régulier pour l'échantillonnage d'ordre 2.*

forme du décodage *max* r_E, traditionnellement appliqué dans un domaine haute-fréquence. L'estimation de fréquences limites de validité de la reconstruction locale (pour une tête centrée) a fait l'objet de plusieurs méthodes, avec l'idée de déterminer des fréquences de transition entre les décodages *basique* et *max* r_E.

Au-delà de l'amélioration des propriétés spatiales en fonction de l'ordre M, il est intéressant de retenir qu'*à ordre égal, la résolution spatiale dans le plan horizontal s'annonce meilleure avec une restitution 2D* (troncature cylindrique) *qu'avec une restitution 3D* (troncature sphérique), aussi bien d'après l'expansion locale de la reconstruction que d'après les caractéristiques globales (r_E).

Enfin, la notion d'échantillonnage directionnel de la base des harmoniques sphériques est présente de façon implicite dans les questions de restitution et de prise de son, contextes dans lesquels le support échantillonnage est défini par l'ensemble des directions des transducteurs. La propriété de régularité de l'échantillonnage, qui permet une résolution facile du problème de décodage et dont dépend les propriétés d'homogénéité de la restitution, a été définie de façon formelle. L'existence des supports réguliers se montre beaucoup plus limitée dans le cas d'un échantillonnage 3D que d'un échantillonnage 2D, notamment pour les ordres supérieurs à 2, où seule une *quasi-régularité* semble pouvoir être atteinte.

3.3 Généralisation des solutions de décodage

Choix des conventions pour la résolution des problèmes de décodage, méthodes de conversion

Dans la suite, les fonctions d'encodage utilisées sont implicitement les fonctions normalisées $Y_{mm}^{\sigma\,(\mathtt{N2D})}$ quand il s'agit de décodage 2D, et $Y_{mn}^{\sigma\,(\mathtt{N3D})}$ quand il s'agit de décodage 3D. Les matrices de décodage qui résultent des différentes méthodes sont notées respectivement $\mathbf{D}^{(\mathtt{N2D})}$ et $\mathbf{D}^{(\mathtt{N3D})}$. Pour obtenir le même décodage dans un système où une autre convention d'encodage `conv` est en vigueur, il faut appliquer la formule de

conversion (3.23). La matrice de décodage à utiliser est alors par exemple:
$\mathbf{D}^{(\texttt{conv})} = \mathbf{D}^{(\texttt{N2D})}.\text{Diag}(\underline{\alpha}^{(\texttt{N2D})\texttt{conv}})$.

On note K le nombre de composantes ambisoniques: pour une représentation homogène d'ordre M, $K = 2M+1$ dans le cas 2D et $K = (M+1)^2$ dans le cas 3D.

3.3.1 Configurations régulières: décodage basique

La définition et des illustrations des configurations régulières 2D et 3D sont données en 3.2.3.

Résolution des équations de décodage

Lorsque la configuration de haut-parleurs est régulière pour la représentation (au sens défini en 3.2.3), ce qui implique $N \geq K$, la résolution du problème de décodage prend une expression triviale. En adoptant implicitement la convention `N2D` pour une configuration 2D et `N3D` pour une configuration 3D, la matrice de "réencodage" $\mathbf{C}$ est en effet telle que $\mathbf{C}.\mathbf{C}^t = NI_K$. Dès lors, la matrice de décodage $\mathbf{D}$ étant définie comme **pseudo-inverse** de la matrice $\mathbf{C}$:

$$\mathbf{D} = \mathbf{D}_{\text{pinv}} = \mathbf{C}^t.(\mathbf{C}.\mathbf{C}^t)^{-1} \tag{3.63}$$

elle s'écrit encore [DRP98]:

$$\mathbf{D} = \mathbf{D}_{\text{proj}} = \frac{1}{N}\mathbf{C}^t, \tag{3.64}$$

d'après (3.62). Le décodage revient donc ici à une **projection** des composantes de l'onde plane encodée sur le vecteur-composantes associé à chaque haut-parleur:

$$S_i = G_i.S = \langle \mathbf{c}_i | \mathbf{c}.S \rangle_N = \frac{1}{N}\mathbf{c}_i^t.\mathbf{c}.S, \tag{3.65}$$

ce que traduit de façon plus compacte l'expression de la matrice $\mathbf{D}_{proj}$.

De surcroît, lorsque $N > K$, le vecteur énergie associé à la restitution de la source S est colinéaire à $\vec{u}_S$ ($\vec{u}_E = \vec{u}_S$) et de module r_E constant[27], l'énergie globale restituée étant elle-même indépendante de l'incidence. Cela signifie aussi que plusieurs haut-parleurs (tous en général) participent à la fois, quelle que soit la direction de la source. Ce n'est plus vrai quand $N = K$, où un seul haut-parleur émet si la source virtuelle est placée dans sa direction, ce que traduit l'égalité $\mathbf{D}.\mathbf{C} = \mathbf{I}_N$. Si ce dernier cas définit une *restitution optimale pour Poletti* [Pol96a] (également recommandée par Nicol et Emerit [NE98][NE99]), nous préférons appuyer la recommandation contraire $N > K$, plus largement répandue dans la littérature relative à *Ambisonic* [Ger77] [Malham (sursound)]: par souci d'homogénéité de l'image sonore et dans une quête de "dématérialisation" des haut-parleurs, ceux-ci ne doivent pas émerger comme sources réelles.

Loi de pan-pot et prise de son équivalentes

L'opération de codage/décodage peut se résumer en une loi de pan-pot équivalente, fonction de la direction $\vec{u}_S$ de la source virtuelle. Pour une restitution 2D (azimut θ_S):

$$G_i(\theta_S) = \frac{1}{N}\mathcal{G}^{M(2D)}(\theta_S - \theta_i) \tag{3.66}$$

27. Assertion démontrée pour le cas 2D dans [DRP98], et admise (vérifiée) pour le cas 3D (annexe A.4.2). Cette propriété est établie par Gerzon pour les systèmes 2D et 3D du premier ordre, et énoncée sous les noms de "*Regular Polygon Theorem*" et de "*Regular Polyhedron Theorem*" [Ger92b].

où l'on a introduit la fonction de directivité générique:

$$\begin{aligned} \mathcal{G}^{M(2D)}(\theta) &= 1+2\sum_{m=1}^{M}\cos(m\theta) \\ &= 1+2\sum_{m=1}^{M}T_m(\cos\theta) \end{aligned} \tag{3.67}$$

T_m étant le polynôme de Chebychev de degré m (Annexe A.2.3). Pour une restitution 3D:

$$G_i(\vec{u}_S) = \frac{1}{N}\mathcal{G}^{M(3D)}(\theta), \qquad \theta = \arccos(\vec{u}_S.\vec{u}_i), \tag{3.68}$$

avec comme fonction de directivité:

$$\mathcal{G}^{M(3D)}(\theta) = \sum_{m=0}^{M}(2m+1)P_m(\cos\theta) \tag{3.69}$$

où les P_m désignent les polynômes de Legendre (Annexe A.2.2). Ces diagrammes de directivité sont illustrés Figure 3.14 (page 186).

Les équations (3.66) et (3.68) suggèrent la possibilité d'une prise de son directe pour chaque signal S_i. Dans [Pol96a], Poletti définit un système de prise de son multicanal 2D par un ensemble de microphones coïncidents, orientés, dans l'espace de prise de son, suivant le même schéma que les haut-parleurs associés dans l'espace de restitution (Figure 3.13). Ce principe peut naturellement être étendu aux systèmes de restitution périphonique ou 3D. Cependant, il ne s'agit que d'un procédé équivalent *virtuel*, présenté ici à titre d'illustration, puisqu'il n'existe pas de capteur réalisant les directivités (3.69) et (3.67) pour les ordres M supérieurs à 1.

Conservation de l'énergie

L'énergie totale associée à la restitution d'une source S est $E.S^2$, où:

$$E = \sum_{i=1}^{N} G_i^2 \tag{3.70}$$

Si $N > K$, E est indépendante de la direction $\vec{u}_S$ et prend la valeur (voir [DRP98] et 3.3.2):

$$E = \frac{K}{N} = \begin{cases} \frac{2M+1}{N} & \text{(cas 2D)} \\ \frac{(M+1)^2}{N} & \text{(cas 3D)} \end{cases} \tag{3.71}$$

Définir un décodage basique *qui préserve l'énergie* de la source virtuelle, revient donc à diviser l'expression des gains G_i (3.66) (3.68) par $\sqrt{E}$, ou encore à utiliser, au lieu de $\mathbf{D}$, la matrice de décodage $\tilde{\mathbf{D}}$:

$$\tilde{\mathbf{D}} = \sqrt{\frac{N}{K}}\mathbf{D} = \frac{1}{\sqrt{KN}}\mathbf{C}^t \tag{3.72}$$

Indice r_E

Comme le gain d'énergie E, et toujours sous réserve que $N > K$, l'indice r_E est constant pour toute les directions d'incidence. Il prend la valeur $\frac{2M}{2M+1}$ dans le cas 2D et $\frac{M}{M+1}$ dans le cas 3D (Table 3.10).

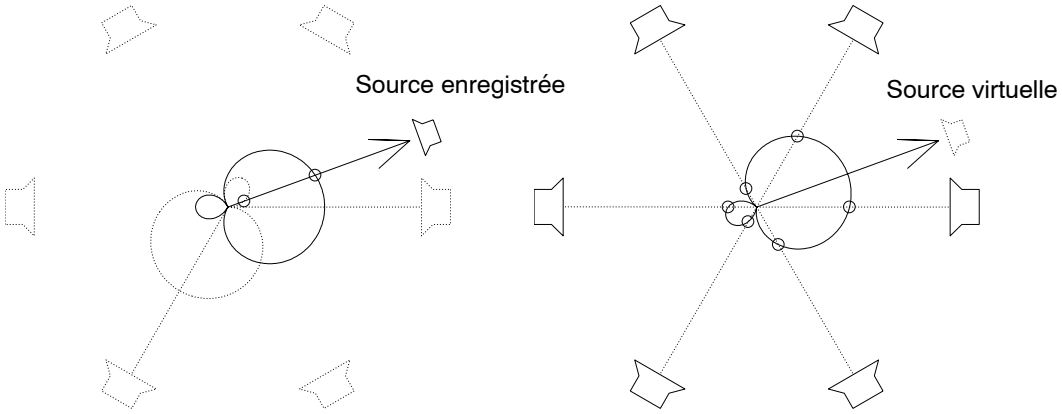

FIG. 3.13 – *A gauche: prise de son multicanal équivalente à un système d'encodage/décodage ambisonique pour une configuration régulière. Les microphones (deux seulement, pour des raisons de clarté) sont représentés par leurs diagrammes de directivité, dirigés vers les haut-parleurs associés. Chaque source enregistrée est captée par chaque microphone avec un gain d'amplitude défini par l'intersection (petit cercle) de la demi-droite orientée vers une source avec le diagramme de directivité. A droite: présentation du principe de pan-pot équivalent. Pour chaque source virtuelle, les gains associés aux haut-parleurs sont déterminés en fonction de leur direction relativement à un même diagramme polaire dirigé vers la source. Cas équivalent à un codage/décodage basique du premier ordre.*

Lois de pan-pot "minimales" et "super-minimales"

La restitution ambisonique sur $N = K$ haut-parleurs est considérée comme optimale par Poletti [Pol96a] dans le sens où elle minimise l'erreur de reconstruction, par rapport à une configuration utilisant plus de haut-parleurs pour un ordre de représentation égal. Nous qualifierons quant à nous une telle configuration de "*minimale*". En fait, il apparaît dans [Pol96a] qu'à ordre égal, la différence de qualité de reconstruction entre une configuration minimale et une configuration non-minimale n'est sensible[28] qu'à travers l'expression de l'erreur moyenne $\bar{\varepsilon}_{p_M}$ (intégrée sur un périmètre (3.52)), et essentiellement grâce aux positions pour lesquelles la reconstruction est exacte, et non par l'erreur maximale ε_{p_M} (3.53). En partant de la contrainte du nombre N de haut-parleurs (configuration régulière) et en se limitant aux systèmes horizontaux, comme le fait Poletti, une loi de pan-pot minimale est donc définie par (3.66) lorsque N est impair, en choisissant un ordre $M = (N-1)/2$. Les fonctions $\mathcal{G}^{M(2D)}$ sont désignées dans [Pol96a] comme les *fonctions* sinc[29] *circulaires* ou fonctions csinc (3.74). Pour offrir au cas N pair une loi de pan-pot ayant les mêmes propriétés – reconstruction optimale, et exacte pour les sources placées sur les haut-parleurs –, Poletti propose d'exploiter les composantes d'une représentation d'ordre $M = N/2$, pour laquelle la configuration n'est plus régulière (au sens défini dans 3.2.3):

$$\mathcal{G}^{M(2D)}_{\text{supermin}}(\theta) = 1 + 2\sum_{m=1}^{M-1}\cos(m\theta) + \cos(M\theta) \tag{3.73}$$

28. Et encore!
29. sinc: sinus cardinal.

Nous parlerons dans ce cas de *configuration "super-minimale" pour l'ordre M*. Poletti rassemble ces lois optimales sous le nom de *fonctions sinc angulaires* (asinc):

$$\text{asinc}_N(\theta) = \begin{cases} \dfrac{\sin(N\theta/2)}{\sin(\theta/2)} = \text{csinc}_N(\theta) & N \text{ impair} \\ \dfrac{\sin[(N-1)\theta/2]}{\sin(\theta/2)} + \cos(\dfrac{N\theta}{2}) = \text{csinc}_{N-1}(\theta) + \cos(\dfrac{N\theta}{2}) & N \text{ pair} \end{cases} \tag{3.74}$$

On constate par ailleurs que si une configuration super-minimale à $N = 2M$ haut-parleurs en comprend un en $\theta = 0°$, aucun d'eux ne participe à la reconstruction de l'harmonique en $\sin(M\theta)$. De manière générale, on pourrait associer à un décodage super-minimal pour l'ordre $M+1$, la qualité de reconstruction d'ordre $M+1/2$, par comparaison à un décodage "basique" d'ordre M.

Nous reprenons le formalisme du décodage ambisonique afin d'étendre le principe de "super-minimalité" à la restitution 3D, pour laquelle l'existence des configurations minimales se réduit au tétraèdre régulier et à l'ordre 1. Si $N < K$, le système d'équations de réencodage (3.27) est sur-déterminé. Dans ce cas, la pseudo-inverse de $\mathbf{C}$ s'écrit:

$$\mathbf{D} = (\mathbf{C}^t.\mathbf{C})^{-1}.\mathbf{C}^t \tag{3.75}$$

Avec une telle matrice de décodage, on est assuré que si la source est placée dans la direction d'un haut-parleur, seul celui-ci est alimenté (ce que traduit l'égalité $\mathbf{D}.\mathbf{C} = \mathbf{I}_N$), et la "reconstruction" de l'onde plane est alors parfaite. En revanche, on ne peut pas espérer, dans un cas général, réaliser l'approximation ambisonique d'ordre M du champ dans l'espace de restitution, c'est-à-dire reconstruire l'ensemble des K composantes ambisoniques.

En choisissant le plus petit ordre M pour lequel le nombre K des composantes ambisoniques est supérieur au nombre de haut-parleurs N, et sous la condition d'une *configuration régulière pour l'ordre* $M-1$, l'encodage/décodage ambisonique en utilisant (3.75) réalise[30] une loi de pan-pot ayant les propriétés de minimalité ou de super-minimalité définies plus haut. L'application à la restitution 2D redonne aux fonctions panoramiques (3.73). Les résultats établis pour quelques cas de configurations régulières 3D sont présentés Table 3.9.

M	K	N	$\mathcal{G}_{\text{supermin}}^{M(3D)}(\gamma)$
1	4	4	$1+3P_1(\cos\gamma)$
2	9	6	$1+3P_1(\cos\gamma)+2P_2(\cos\gamma)$
2	9	8	$1+3P_1(\cos\gamma)+3P_2(\cos\gamma)$
3	16	12	$1+3P_1(\cos\gamma)+5P_2(\cos\gamma)+3P_3(\cos\gamma)$

TAB. 3.9 – *Lois de pan-pot "super-minimales" pour la restitution 3D sur configurations tétraédrique(N=4) , octoédrique(6), cubique(8), et icosaédrique(12). Se référer à l'annexe A.2.2 pour l'expression des polynômes de Legendre P_m, et à la figure 3.12 pour un aperçu des polyèdres.*

Complément: quasi-régularité du polyèdre à 32 sommets

Le polyèdre rassemblant les sommets du dodécaèdre et de son dual l'icosaèdre (Figure 3.12) ne définit un échantillonnage rigoureusement régulier pour une base d'harmoniques sphériques que jusqu'à l'ordre $M = 2$. Cependant, si $\mathbf{C}$ désigne la matrice d'encodage associée pour un ordre M donné (convention `N3D`),

30. Dans le domaine 3D, il s'agit d'une conjecture: nous n'en faisons pas de démonstration générale, mais la vérifions pour les cas présentés.

on constate que la matrice $\frac{1}{N}\mathbf{C}.\mathbf{C}^t$ est “presque" diagonale – et même presque égale à la matrice identité $\mathbf{I}_K$ – pour $M = 3$ ou 4 (Cf 3.2.3). On peut vérifier la quasi-régularité de cette configuration pour ces ordres en observant les distorsions du vecteur énergie $\vec{E}$ et du gain d'énergie E à l'issue de l'encodage/décodage d'une source virtuelle. On note $\mathbf{D}_{\text{pinv}} = \mathbf{C}^t.(\mathbf{C}.\mathbf{C}^t)^{-1}$ et $\mathbf{D}_{\text{proj}} = \frac{1}{N}\mathbf{C}^t$ les matrices de décodage calculées par pseudo-inverse (3.63) et par projection (3.64). Des simulations numériques montrent que:

- La distorsion directionnelle de $\vec{E}$ est nulle avec $\mathbf{D}_{\text{pinv}}$ et inférieure à 1° avec $\mathbf{D}_{\text{proj}}$ (ordres 3 et 4).
- Le gain d'énergie E a une amplitude de variation de 0,11 dB pour $M = 3$ et de 0,26 dB pour $M = 4$ ($\mathbf{D}_{\text{pinv}}$ et $\mathbf{D}_{\text{proj}}$).
- L'indice r_E fluctue autour de sa valeur moyenne (conforme à celle prédite par la table 3.10) avec un écart d'environ 1% ou moins.

En pratique, cette configuration à 32 sommets peut donc être considérée comme quasi-régulière pour les ordres 3 et 4, qui comprennent respectivement $K = 16$ et $K = 25$ composantes ambisoniques. En tronquant cette configuration pour ne retenir que les 16 sommets supérieurs, on obtient un exemple de dôme géodésique, utilisable pour une restitution hémisphérique (section 3.3.5).

Par comparaison, notons que les 20 sommets du dodécaèdre – polyèdre pourtant régulier – ne correspondent pas à un échantillonnage régulier pour l'ordre 3 ($K = 16$), même approximativement. A l'issue d'un encodage/décodage, d'importantes distorsions de $\vec{E}$ portent soit sur l'angle ($15{,}5^\circ$ avec $\mathbf{D}_{\text{proj}}$), soit sur r_E ($\approx 25\%$ avec $\mathbf{D}_{\text{pinv}}$). Le gain d'énergie varie dans un intervalle de 3,5 dB ($\mathbf{D}_{\text{pinv}}$) ou de 2,4 dB ($\mathbf{D}_{\text{proj}}$).

3.3.2 Configurations régulières: décodages modifiés

Méthode générale

Dans [DRP98], il est montré que dans le cas d'une configuration régulière non-minimale ($N > 2M+1 = K$), la modification du décodage basique par correction des composantes ambisoniques par un facteur g_m propre à chaque ordre m, avant application de la matrice de décodage, maintient le vecteur énergie $\vec{E}$ dans la direction $\vec{u}_S$ de la source virtuelle. L'optimisation du décodage suivant le critère “max r_E" et la contrainte de préservation de l'énergie consiste alors en un choix judicieux des paramètres g_m.

Ce principe de *décodage modifié* ou *corrigé* est maintenant repris et développé pour exhiber les solutions de décodage optimisées suivant le critère “max r_E" ou la contrainte "*in-phase*" généralisée, dans le cas des *configurations régulières non-minimales* 2D et 3D. La matrice de décodage dérive de la matrice $\mathbf{D}$ définie en (3.64) par la correction:

$$\mathbf{D}_{\{g_m\}} = \mathbf{D}.\mathbf{\Gamma}_{\{g_m\}}, \tag{3.76}$$

où $\mathbf{\Gamma}_{\{g_m\}}$ désigne la matrice diagonale $\mathbf{\Gamma}^{2D}_{\{g_m\}}$ ou $\mathbf{\Gamma}^{3D}_{\{g_m\}}$, selon le cas:

$$\mathbf{\Gamma}^{2D}_{\{g_m\}} = \text{Diag}([g_0\ g_1\ g_1\ g_2\ g_2\ \dots\ g_M\ g_M]^t) \tag{3.77}$$

$$\mathbf{\Gamma}^{3D}_{\{g_m\}} = \text{Diag}([g_0\ \dots\ \underbrace{g_m \dots g_m}_{2m+1}\ \dots\ \underbrace{g_M \dots g_M}_{2M+1}]^t) \tag{3.78}$$

Pour décrire la correction correspondant à un décodage optimisé *max* r_E ou *in-phase*, 2D ou 3D d'ordre M, on pourra utiliser les notations plus explicites $\mathbf{\Gamma}^{M(2D)}_{\max\ r_E}$, $\mathbf{\Gamma}^{M(2D)}_{\text{in-phase}}$, etc...

La loi de pan-pot équivalente utilise la fonction de directivité $\mathcal{G}^{M(3D)}_{\{g_m\}}$ ou $\mathcal{G}^{M(2D)}_{\{g_m\}}$ selon le cas:

$$\mathcal{G}^{M(2D)}_{\{g_m\}}(\theta) = g_0 + 2\sum_{m=1}^{M} g_m \cos(m\theta) \tag{3.79}$$

$$\mathcal{G}^{M(3D)}_{\{g_m\}}(\theta) = \sum_{m=0}^{M} (2m+1) g_m P_m(\cos\theta) \tag{3.80}$$

Normalisation en amplitude ou en énergie

En introduisant l'*énergie réduite* $\mathcal{E}$, associée au diagramme de directivité $\mathcal{G}$:

$$\mathcal{E}^{2D} = \tfrac{1}{2\pi}\int_0^{2\pi} [\mathcal{G}^{2D}(\theta)]^2 \, d\theta \tag{3.81}$$

$$\mathcal{E}^{3D} = \tfrac{1}{2}\int_0^{\pi} [\mathcal{G}^{3D}(\theta)]^2 \sin\theta \, d\theta, \tag{3.82}$$

on peut montrer ([DRP98] et annexe A.4.2) que, dans le cas d'une *configuration régulière non-minimale* ($N > K$), le gain en énergie résultant du décodage est constant et s'écrit:

$$E = \frac{\mathcal{E}}{N} \tag{3.83}$$

Il doit être égal à 1 pour un décodage préservant l'énergie. L'énergie réduite pour un décodage modifié s'écrit plus précisément:

$$\mathcal{E}^{M(2D)}_{\{g_m\}} = g_0^2 + 2\sum_{m=1}^{M} g_m^2 \tag{3.84}$$

$$\mathcal{E}^{M(3D)}_{\{g_m\}} = \sum_{m=0}^{M} (2m+1) g_m^2 \tag{3.85}$$

De manière générale, l'optimisation permet de définir de façon univoque les paramètres $g'_m = g_m/g_0$ (où $g'_0 = 1$). La détermination complète des $g_m = g_0\, g'_m$ dépend du choix entre la préservation de l'amplitude du champ de pression reconstruit au point central $\vec{r} = 0$ et celle de l'énergie totale restituée $E.\mathcal{S}$, ce qui se traduit respectivement par les équations:

$$g_0 = 1 \qquad \text{Préservation de l'amplitude} \tag{3.86}$$

$$g_0 = \sqrt{N/\mathcal{E}_{\{g'_m\}}} \qquad \text{Conservation de l'énergie totale} \tag{3.87}$$

Optimisation "max r_E"

L'*optimisation suivant le critère "max r_E"* est réalisée dans [DRP98] (annexe B) pour les systèmes 2D (horizontaux) et en annexe A.4.2 pour la restitution 3D. Nous en donnons les coefficients résultants g'_m, ainsi que le paramètre g_0 pour la conservation de l'énergie:

Cas 2D:
$$\begin{cases} g'_m = \cos\frac{m\pi}{2M+2} = T_m(r_E), \qquad g_0 = \sqrt{\frac{N}{M+1}} \\ r_E = \cos\frac{\pi}{2M+2} = g'_1 \text{ est la plus grande racine de } T_{M+1}. \end{cases} \tag{3.88}$$

Cas 3D:
$$\begin{cases} g'_m = P_m(r_E), \qquad g_0 = \sqrt{\frac{N}{\mathcal{E}_{\{g'_m\}}}} \\ r_E = g'_1 \text{ est la plus grande racine de } P_{M+1}. \end{cases} \tag{3.89}$$

La table 3.10 en donne les valeurs numériques pour les premiers ordres, et les diagrammes de directivité équivalents sont illustrés Figure 3.14. L'indice r_E converge vers 1 quand M croît, plus rapidement avec une restitution 2D qu'avec une restitution 3D.

Type	ordre	$r_V (= g'_1)$	r_E	$\mathcal{E}\left(g_0 = \sqrt{\frac{N}{\mathcal{E}}}\right)$	$\{g'_1, \ldots, g'_m, \ldots, g'_M\}$
Basique (2D)	M	1	$\frac{2M}{2M+1}$	$2M+1$	$\{g'_m = 1\}$
	1	1	0.667	3	$\{1\}$
	2	1	0.800	5	$\{1,1\}$
	3	1	0.857	7	$\{1,1,1\}$
	4	1	0.889	9	$\{1,1,1,1\}$
Basique (3D)	M	1	$\frac{M}{M+1}$	$(M+1)^2$	$\{g'_m = 1\}$
	1	1	0.5	4	$\{1\}$
	2	1	0.667	9	$\{1,1\}$
	3	1	0.750	16	$\{1,1,1\}$
	4	1	0.800	25	$\{1,1,1,1\}$
Max r_E (2D)	M	$r_V = r_E$	$\cos\frac{\pi}{2M+2}$	$M+1$	$\{g'_m = \cos(\frac{m\pi}{2M+2})\}$
	1	0.707	0.707	2	$\{0.707\}$
	2	0.866	0.866	3	$\{0.866\ 0.500\}$
	3	0.924	0.924	4	$\{0.924\ 0.707\ 0.383\}$
	4	0.951	0.951	5	$\{0.951\ 0.809\ 0.588\ 0.309\}$
Max r_E (3D)	M	$r_V = r_E$	${P_{M+1}}^{-1}(0)$	(*)	$\{g'_m = P_m(r_E)\}$
	1	0.577	0.577	2	$\{0.577\}$
	2	0.775	0.775	3.6	$\{0.775\ 0.400\}$
	3	0.861	0.861	5.75	$\{0.861\ 0.612\ 0.305\}$
	4	0.906	0.906	8.441	$\{0.906\ 0.732\ 0.501\ 0.246\}$
In-phase (2D)	M	$\frac{M}{M+1}$	$\frac{2M}{2M+1}$	(*)	$\{g'_m = \frac{M!^2}{(M+m)!(M-m)!}\}$
	1	0.500	0.667	1.5	$\{0.500\}$
	2	0.667	0.800	1.944	$\{0.667\ 0.167\}$
	3	0.750	0.857	2.310	$\{0.750\ 0.300\ 0.050\}$
	4	0.800	0.889	2.627	$\{0.800\ 0.400\ 0.114\ 0.014\}$
In-phase (3D)	M	$\frac{M}{M+2}$	$\frac{M}{M+1}$	$\frac{(M+1)^2}{2M+1}$	$\{g'_m = \frac{M!(M+1)!}{(M+m+1)!(M-n)!}\}$
	1	0.333	0.500	1.333	$\{0.333\}$
	2	0.500	0.667	1.8	$\{0.500\ 0.100\}$
	3	0.600	0.750	2.286	$\{0.600\ 0.200\ 0.029\}$
	4	0.667	0.800	2.778	$\{0.667\ 0.286\ 0.071\ 0.008\}$

TAB. 3.10 – *Paramètres et caractéristiques des principales solutions de décodage ambisonique 2D et 3D (pour configurations régulières non-minimales uniquement). (*): expressions génériques peu simples.*

L'optimisation peut être envisagée avec la contrainte de préserver l'intégrité des premières composantes ambisoniques à la reconstruction, à un facteur de normalisation près. Il s'agit de réaliser un compromis entre restitution basse-fréquence/centrée et haute-fréquence/excentrée. Cette contrainte appliquée jusqu'à l'ordre Q se traduit par $g_0 = g_1 = \ldots = g_Q$. L'optimisation est effectuée dans [DRP98] (Annexe B) pour le cas 2D. Nous n'insistons pas sur ces solutions qui ne sont pas d'un intérêt majeur.

Solutions "*in-phase*" généralisées ou "à loi de pan-pot monotone"

Pour la *généralisation des solutions "in-phase"* aux ordres supérieurs, le critère essentiel retenu consiste à minimiser les gains des haut-parleurs à mesure que leur direction s'éloigne de celle de la source virtuelle, jusqu'à assurer un gain nul dans la direction diamétralement opposée. Le choix des paramètres g_m doit donc être tel que la fonction $\mathcal{G}_{\{g_m\}}^{M(2D,3D)}(\theta)$ soit décroissante sur l'intervalle $[0\,\pi]$ et nulle en $\theta = \pi$. Les résultats des calculs développés en annexe A.4.3 sont rappelés Table 3.10 et ci-dessous:

$$\text{Cas 2D:} \qquad \begin{cases} g'_m = \frac{M!^2}{(M+m)!(M-m)!}, \\ r_E = \frac{2M}{2M+1} \end{cases} \qquad g_0 = \sqrt{\frac{N}{\mathcal{E}_{\{g'_m\}}}} \tag{3.90}$$

$$\text{Cas 3D:} \qquad \begin{cases} g'_m = \frac{M!(M+1)!}{(M+m+1)!(M-n)!}, \\ r_E = \frac{M}{M+1} \end{cases} \qquad g_0 = \frac{\sqrt{N(2M+1)}}{M+1} \tag{3.91}$$

Notons qu'à ordre M égal, la loi de pan-pot équivalente est la même pour les deux cas 2D et 3D (Figure 3.14), alors que le jeu de paramètre $\{g_m\}$ n'est pas le même. Des solutions obéissant à des critères combinés (*in-phase* et *max* r_E, ou *in-phase* et *max* r_V) sont également reportées dans A.4.3 à titre indicatif, mais ne sont pas retenues comme solutions pertinentes pour notre étude. On remarque l'indice r_E pour un ordre M donné est identique à celui des solutions *basiques* (respectivement 2D et 3D). Il tend vers 1 quand M croît, mais beaucoup moins vite que les solutions *max* r_E (Table 3.10), ce qu'on observe également à travers l'évolution des diagrammes de directivité (Figure 3.14). Enfin, la convergence vers 1 est deux fois plus lente avec une restitution 3D qu'avec une restitution 2D.

Signalons au passage que la solution *in-phase* au second ordre est utilisée depuis peu par Furse et Malham [Fur99]. Furse en propose la dénomination générique "*controlled-opposites*", ce qui se conforme finalement assez bien à la méthode de résolution présentée en annexe A.4.3.

Effet de régularisation pour des configurations non-régulières

Nous avons pu généraliser les solutions de décodage 3D à tout ordre, alors même que l'existence des configurations 3D rigoureusement régulières pour l'échantillonnage d'ordre supérieur est très limitée. La configuration à 32 haut-parleurs (Figure 3.12) peut être assimilée aux configurations régulières pour les ordres 3 et 4, dans le sens où les caractéristiques de restitution $(\vec{E}, E)$ présentent de très faibles distorsions par rapport aux caractéristiques idéales.

Avec une configuration définie par les 20 sommets du dodécaèdre et pour l'ordre $M = 3$, les distorsions observées à la restitution sont les suivantes:

- Moyenne de r_E: conforme à la table 3.10 ou quasiment.
- Amplitude de variation de r_E: 0,117 (*max* r_E) et 0,015 (*in-phase*), soit une erreur maximale d'environ 7% et 1% respectivement.
- Ecart angulaire maximal entre $\vec{u}_E$ et la direction d'encodage $\vec{u}$: 6,51° et 0,55° respectivement.
- Amplitude de variation du gain d'énergie E: 0,5774 dB et 0,0125 dB.
- Distorsion du vecteur vélocité (par rapport à une configuration vraiment régulière): infime dans tous les cas.

Ces résultats sont à comparer aux résultats de la fin de la section précédente (page 181). Les décodages *max* r_E et surtout *in-phase* ont pour effet notable de réduire les distorsions, donc d'*homogénéiser ou régulariser la restitution*.

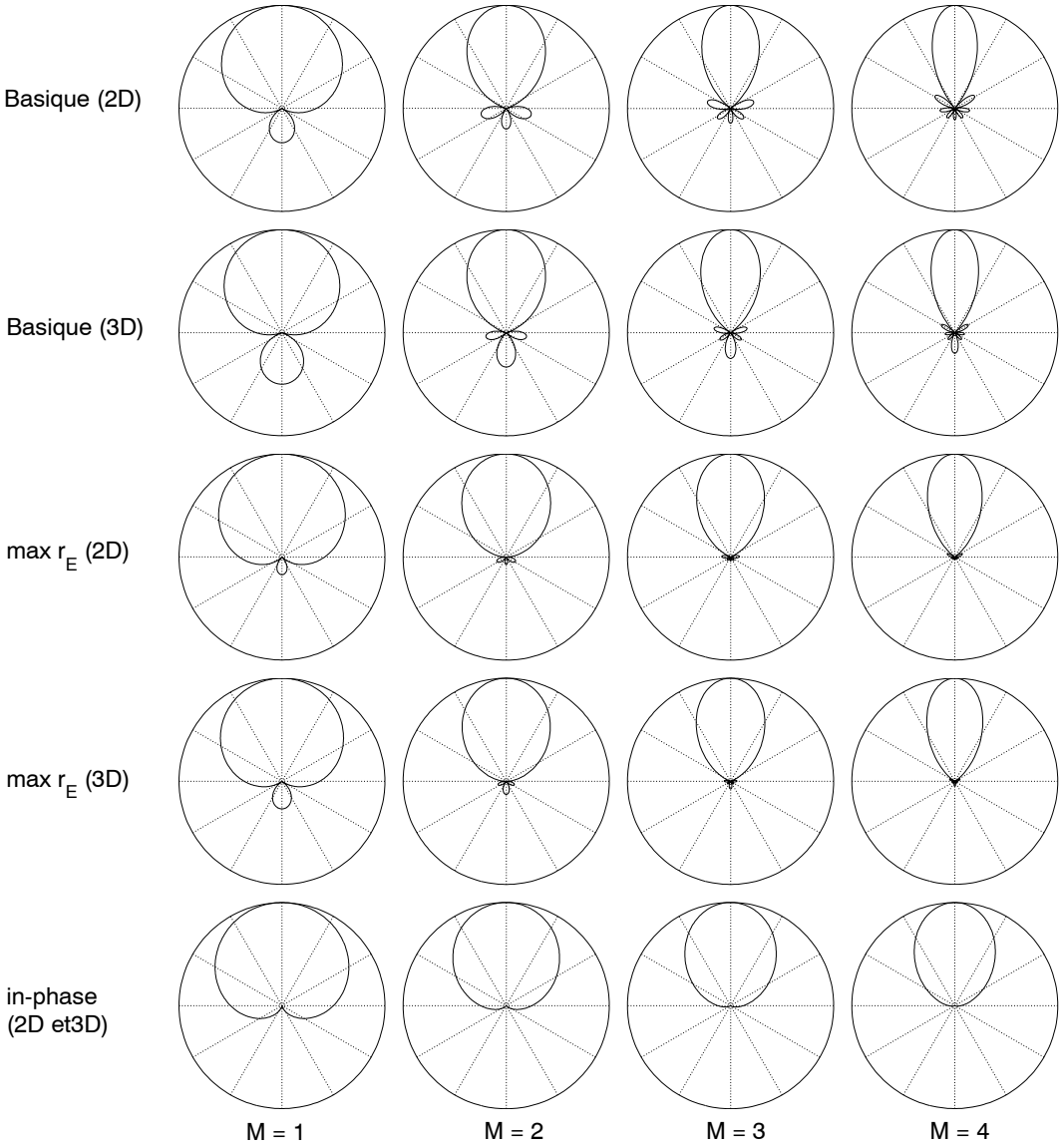

FIG. 3.14 – *Diagrammes de directivité des microphones virtuels associés aux différents décodages ambisoniques, ou définissant différentes lois de pan-pot ambisonique (systèmes 2D et 3D). Echelle radiale linéaire. Les échelles sont modifiées pour montrer des maxima équivalents. Noter qu'à l'ordre 1, le diagramme "in-phase" a une directivité cardioïde, et les autres ont des directivités hyper-cardioïdes, les lobes arrière semblant plus prononcés pour les cas 3D que pour les cas 2D, à ordre égal et à critère de décodage identique.*

L'effet de régularisation peut s'expliquer par la réduction de la participation des composantes d'ordres supérieurs (pondération par les gains g_m). En raisonnant en termes de prise de son équivalente, l'interprétation peut se compléter ainsi: les lobes secondaires sont atténués (voire éliminés avec *in-phase*) et le lobe primaire est élargi (un peu moins avec *max* r_E), donc un meilleur recouvrement est assuré entre microphones (ou haut-parleurs) adjacents.

3.3.3 Configurations semi-régulières

Définition du décodage

Rappelons que, d'après 3.2.3, les configurations semi-régulières sont telles que la matrice $\mathbf{C}.\mathbf{C}^t$ est diagonale:

$$\frac{1}{N}\mathbf{C}.\mathbf{C}^t = \mathrm{Diag}(\underline{\mu}), \qquad \text{avec} \quad \underline{\mu} = [\mu_1 \ \ldots \ \mu_K]^t \tag{3.92}$$

Les plus évidentes sont les formes rectangulaires pour les systèmes 2D d'ordre 1, et les parallélépipèdes rectangles pour les systèmes 3D d'ordre 1. La réunion des sommets de plusieurs polyèdres réguliers peut définir aussi une configuration semi-régulière pour les systèmes 3D d'ordres supérieurs: c'est le cas de la réunion d'un cube (8 sommets) et d'un octaèdre (6 sommets) pour l'ordre 2 (Figure 3.12, page 177).

La résolution du problème de décodage est presque aussi simple que dans le cas régulier, puisqu'il suffit de substituer une matrice diagonale à la matrice identité dans (3.63):

$$\mathbf{D}_{\mathrm{pinv}} = \mathbf{C}^t.\mathrm{Diag}([\mu_1^{-1}\,\mu_2^{-1}\ \ldots\ \mu_K^{-1}]^t) \tag{3.93}$$

Ces configurations ont aussi l'avantage de bénéficier des résultats d'optimisation "max r_E" développé pour le configurations régulières. On vérifie en effet dans les cas connus[31] que la colinéarité de $\vec{E}$ avec $\vec{u}_S$ est automatiquement assurée et que l'optimisation "max r_E" consiste en la même correction des composantes ambisoniques par les coefficients g_m:

$$\mathbf{D}_{\max r_E} = \mathbf{D}_{\mathrm{pinv}}.\mathbf{\Gamma}_{\max r_E} \tag{3.94}$$

Quelques propriétés

La différence essentielle avec le cas régulier réside dans l'inhomogénéité directionnelle de la restitution, que traduisent les variations des valeurs de E et de r_E en fonction des directions d'incidence. De manière générale, r_E augmente dans les zones de plus forte densité angulaire de haut-parleurs, alors que E suit la tendance inverse (Figure 3.15). A noter également que si l'optimisation "max r_E" offre une *maximisation globale* de l'indice r_E, celui-ci peut être localement inférieur à celui produit par le décodage basique, comme l'atteste la figure 3.15 dans le cas d'une configuration rectangulaire: cela apparaît, quoique de façon peu marquée, dans les voisinages angulaires où r_E atteint un maximum. Le décodage "max r_E" tend à réduire l'amplitude des variations de E comme de r_E.

La normalisation en énergie peut être effectuée pour une incidence de référence, ou bien pour un champ diffus. Dans ce dernier cas, le gain d'énergie E à considérer pour une matrice de décodage $\mathbf{D}$ est donné par $E = \mathrm{tr}(\mathbf{D}.\mathbf{D}^t)$, où tr dénote la trace de la matrice. La matrice de décodage normalisée en énergie devient alors:

$$\tilde{\mathbf{D}} = \frac{1}{\sqrt{E}}\mathbf{D} = \frac{1}{\sqrt{\mathrm{tr}(\mathbf{D}.\mathbf{D}^t)}}\mathbf{D} \tag{3.95}$$

31. Mais on ne le montre pas ici! Gerzon [Ger92b] le montre pour les systèmes du premier ordre.

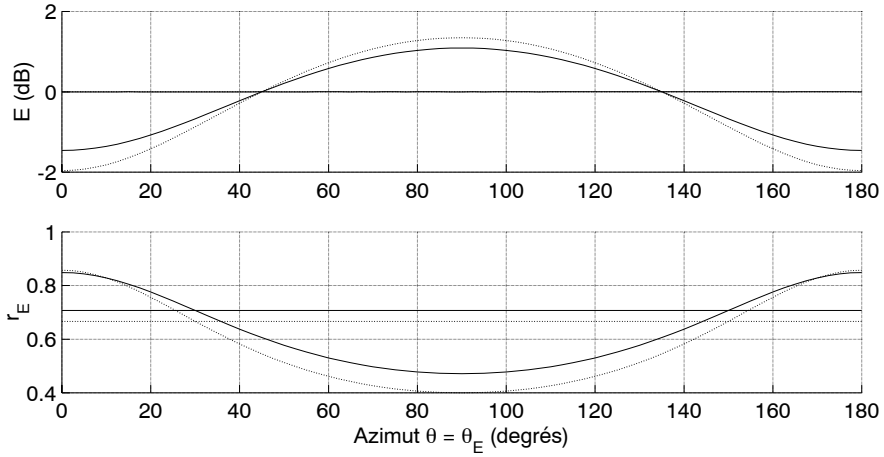

FIG. 3.15 – *Gain énergétique E et indice r_E en fonction de l'azimut $\theta = \theta_E$ de la source virtuelle. Décodages basique (pointillés) et "max r_E" (trait continu) avec normalisation en énergie pour une configuration rectangulaire (haut-parleurs placés en $\pm 30^\circ$ et $\pm 150^\circ$) et pour une configuration carrée (courbes constantes).*

Au contraire de r_E, l'indice r_V est quant à lui constant et indépendant de la direction d'incidence. Il prend les mêmes valeurs qu'avec un décodage pour configuration régulière ($\kappa_V = g_1'$). Il en est donc de même pour la vitesse apparente de propagation au voisinage du centre. Par contre la qualité de rendu basse-fréquence n'est pas la même pour toutes les directions d'incidence. L'expansion de la reconstruction acoustique est en effet favorisée lorsque la direction de la source virtuelle se situe dans un secteur de plus forte densité angulaire de haut-parleurs (Figure 3.16), de façon corrélée avec l'indice r_E.

3.3.4 Configurations non-régulières

Solutions optimisées

Pour une configuration non-régulière, comme celles recommandées par exemple par l'ITU [CCI92] [ITU94] pour la restitution multi-canal, le décodage défini par pseudo-inverse (3.63) de la matrice de ré-encodage ne vérifie plus la propriété de colinéarité de $\vec{V}$ et de $\vec{E}$, de même la matrice définie par projection (3.64). La définition de matrice de décodage satisfaisant cette contrainte, et maximisant κ_V ou r_E selon le domaine fréquentiel, est un problème non-trivial nécessitant des méthodes d'optimisation non-linéaire. Il a été réalisé par Gerzon pour un certain nombre de configurations [Ger92a][32]. Un travail similaire a été réalisé au CCETT par Louis-Cyrille Trébuchet [Tre97], et a abouti à des résultats analogues, au moins pour le décodage basse-fréquence. En conséquence du relâchement de la contrainte $\kappa_V = 1$, l'optimisation haute-fréquence comporte quant à elle des degrés de liberté qui reflètent les compromis dus à une maximisation non-uniforme de l'indice r_E (Figure 3.17).

De même qu'avec les configurations semi-régulières, la restitution est rendue inhomogène. La qualité de restitution (définition des images sonores) est meilleure dans les directions où la densité angulaire des haut-parleurs est plus grande, et c'est d'ailleurs de cette façon que les dispositions 3/2 recommandées (Figure 2.6) privilégient la scène frontale. En revanche, l'effort pour reproduire les directions conformément aux sources sonores encodées entraîne un déséquilibre de la balance énergétique dans le sens contraire: le champ sonore

32. les décodeurs définis dans cette référence sont souvent évoqués sous le nom de "*Vienna Decoders*".

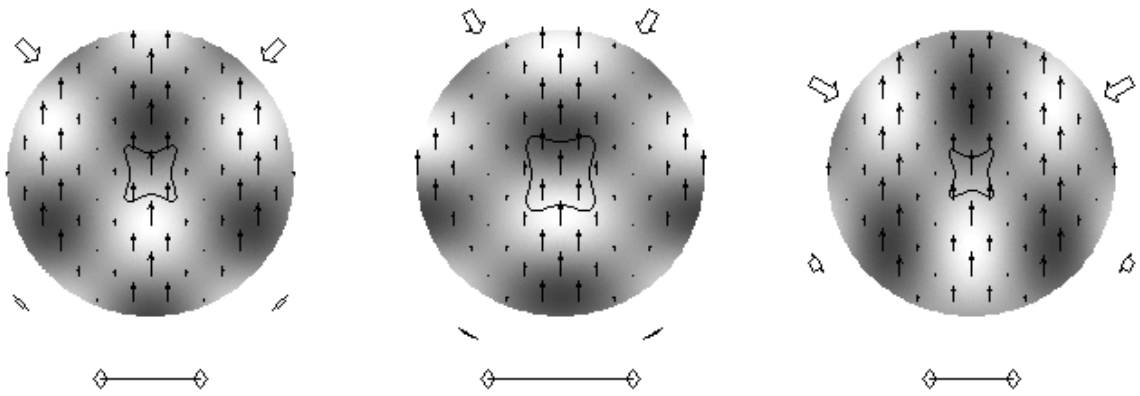

FIG. 3.16 – *Représentation du champ de pression instantané (niveau de gris, vue du dessus) pour la restitution ambisonique (décodage basique) sur dispositifs carré (à gauche) et rectangulaires ($\phi_F = 30°$ au centre et $\phi_F = 60°$ à droite). Les flèches fines représentent le vecteur vélocité pondéré par l'énergie du champ au même lieu, soit encore l'opposé de l'intensité active $\vec{I}$. Les segment horizontaux (en dessous) indiquent la périodicité Λ_y des figures d'interférence suivant l'axe (Oy). La faible longueur des vecteurs $-\vec{I}$ en bordure des figures d'interférences (à la verticale des extrémités des segments) atteste des creux d'énergie du champ de pression. La courbe étoilée indique les limites de reconstruction de l'onde plane, pour une erreur tolérée de 20%.*

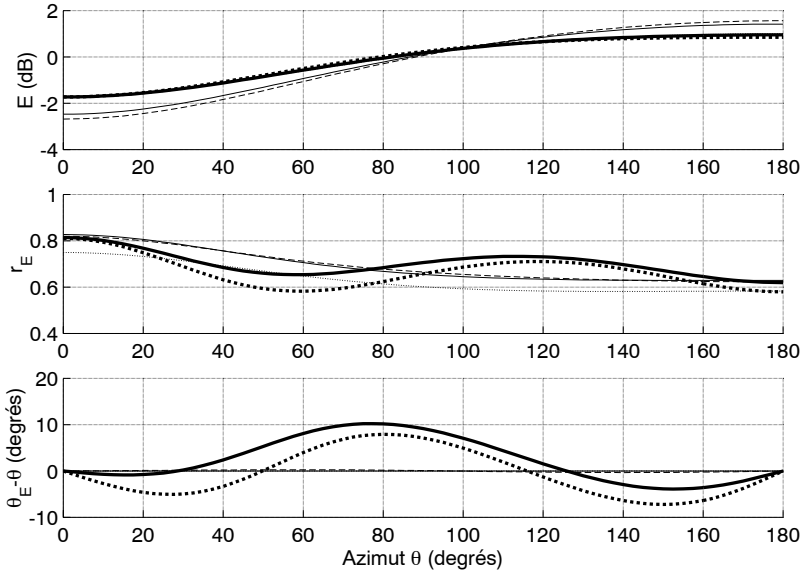

FIG. 3.17 – *Gain d'énergie E, indice r_E et erreur $\theta_E - \theta$ en fonction de l'azimut θ de la source virtuelle. Configuration 3/2 avec $\phi_F = 45°$ et $\phi_B = 50°$. Décodages optimisés par Gerzon (traits fins, basse-fréquence: pointillés, haute-fréquence: continu), par Trébuchet (basse-fréquence: id. Gerzon, haute-fréquence: tirets fins), et d'après pseudo-inverse (telle quelle: pointillés gras, avec correction "max r_E": trait continu gras).*

correspondant aux directions de faible densité de haut-parleurs (typiquement vers l'arrière) est restitué avec un excès de puissance. Gerzon [Ger92a] propose le compromis d'une distorsion préalable du champ ambisonique: en appliquant la *Forward Dominance* (3.42), la balance énergétique est à peu près rétablie en même temps que les directions apparentes sont déplacées vers l'avant, tout en préservant la cohérence des indices directionnels $\vec{V}$ et $\vec{E}$.

Pseudo-inverse et dérivée

A défaut de ces solutions optimisées, l'*usage de la pseudo-inverse* peut se révéler être un compromis acceptable. Des évaluations objectives de la reconstruction basse-fréquence ne montrent en fait que des différences très minimes (à peine sensibles) avec les solutions "optimisées basse-fréquence". Au mépris du critère $\theta_E = \theta_V = \theta_S$, la *correction préalable des composantes ambisoniques* – celle de l'optimisation "max r_E" pour configurations régulières et semi-régulières – semble même, dans certains cas, constituer une solution avantageuse. En contrepartie d'une distorsion entre θ_E, θ_V et θ_S, la valeur de r_E est améliorée, bien que n'ayant pas une évolution monotone entre la direction frontale et l'arrière. Par contre, le problème de balance énergétique est à nouveau présent. La figure 3.17 permettent de constater de plus que le gain d'énergie E, lors du déplacement panoramique d'une source, n'a pas le comportement monotone observé pour les solutions optimisées.

"Régularisation" de la configuration pour le décodage

Parmi les résolutions "approximatives" ou "non-rigoureuses" du problème du décodage, on peut encore citer celle-ci: définir comme *configuration de décodage* une configuration *régulière*, *voisine de la configuration de restitution* (par exemple, un pentagone régulier pour un dispositif 3/2). Dans ce cas, la balance énergétique des événements sonores est préservée, au prix d'une distorsion directionnelle. Notons que cela peut constituer une solution voisine du décodage optimisé par Gerzon doublé de la *Forward Dominance*, en particulier si la configuration de restitution peut dériver d'une configuration régulière par la loi de distorsion d'une *Forward Dominance*. D'après simulations, cette dernière condition ne suffit cependant pas à assurer la propriété $\theta_V = \theta_E$, au contraire des solutions de Gerzon.

Méthode de projection sur les directions des haut-parleurs (directivité des microphones équivalents unique)

Le "détournement" des propriétés du décodage pour configuration régulière conduit – après l'usage de la pseudo-inverse et la régularisation de la configuration – à une dernière option: appliquer le *principe de prise de son équivalente avec une directivité unique pour les microphones*, illustré Figure 3.13, *comme si la configuration était régulière*. D'un point de vue mathématique, cela correspond à exploiter le *principe de projection*, normalement propre aux échantillonnages réguliers: la matrice de décodage à utiliser est alors de la forme[33] $\mathbf{C}^t.\mathbf{\Gamma}_{\{g_m\}}$. Ce principe simplifié est apparemment utilisé dans certains décodeurs ambisoniques commercialisés pour configurations classiques 3/2 (Cf sursound), à la place des matrices optimisées par Gerzon (*Vienna Decoders*). Soulignons qu'un tel choix ne permet pas d'assurer la validité des critères basés sur $\vec{V}$ et $\vec{E}$. En plus d'une distorsion directionnelle probable, la balance énergétique des événements sonores selon leur localisation n'est en général pas respectée. En revanche, c'est la bonne approche pour pouvoir appliquer le critère *in-phase*, c'est-à-dire le principe du *pan-pot à loi monotone* (Figure 3.18, droite). Il peut être toutefois nécessaire de corriger individuellement le gain associé à chaque haut-parleur

33. Pour une configuration 2D, on suppose ici que la convention en vigueur est `N2D`.

pour compenser l'irrégularité de la densité angulaire des haut-parleurs et rétablir au mieux l'équilibre panoramique de l'énergie. Sans cette compensation, l'énergie "totale" pour une source virtuelle en θ s'écrit: $E(\theta) = \sum_i \mathcal{G}^2(\theta - \phi_i)$. En corrigeant la sortie de chaque haut-parleur par un gain d'énergie γ_i, on cherche à ce que la fonction d'énergie totale $E_\gamma(\theta) = \sum_i \gamma_i \mathcal{G}^2(\theta - \phi_i)$ soit la plus constante possible, par exemple égale à 1. Il s'agit donc de minimiser $\int_0^{2\pi} \left(1 - E_\gamma(\theta)\right)^2 d\theta$, et en définitive, de résoudre le système d'équations $\sum_k \gamma_k \int \mathcal{G}^2(\theta - \phi_i)\mathcal{G}^2(\theta - \phi_k)\, d\theta = \int \mathcal{G}^2(\theta) d\theta, \forall i$.

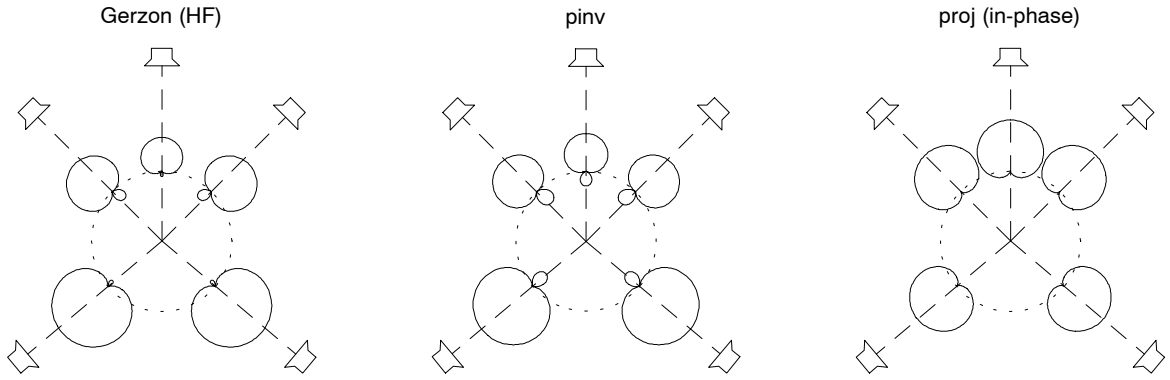

FIG. 3.18 – *Prises de son équivalentes à des décodages pour une configuration 3/2 ($\phi_F = 45°, \phi_B = 50°$). Pour des raisons de clarté, les directivités des microphones virtuels sont représentées de façon excentrée, mais ces microphones sont coïncidents. Décodage haute-fréquence d'après Gerzon, décodage par pseudo-inverse, et décodage par principe de projection modifié pour obtenir des directivités* in-phase.

Bilan - Vers une exploitation partielle des directivités d'ordres supérieurs

A défaut de méthode générique pour définir un décodage qui satisfasse les critères de Gerzon, plusieurs approches "approximatives" ont été proposées, dont on pourrait encore dériver des variantes. Des simulations (non-présentées ici) montrent que ces méthodes donnent des résultats acceptables tant que la répartition angulaire des haut-parleurs reste assez équilibrée. Mais pour les dispositions 3/2 les plus fréquemment recommandées ($\phi_F = 30°$ et $\phi_B \approx 70°$) où le secteur angulaire arrière est largement déserté à l'inverse du secteur frontal, de grosses distorsions du vecteur énergie peuvent être observées. Mais il faut remarquer que Gerzon lui-même ne propose pas de solution de décodage pour ce type de configuration. Celles abordées dans [Ger92a] sont relativement équilibrées dans le sens où ϕ_F et ϕ_B y sont en général du même ordre de grandeur. En supposant que l'on puisse définir un décodage "rigoureux" pour ($\phi_F = 30°, \phi_B \approx 70°$), on peut s'attendre à ce que l'effort de reconstruction, compte-tenu des contraintes de colinéarité $\vec{u}_E = \vec{u}_V = \vec{u}_S$, soit tel qu'un important déséquilibre énergétique soit observé entre l'avant et l'arrière.

La description du décodage sous la forme d'une prise de son équivalente permet d'en donner une interprétation intuitive et de dessiner de nouvelles pistes. Dans le décodage version Gerzon par exemple (Figure 3.18, gauche), il semble que la sélectivité limitée des directivités d'ordre 1 soit la cause d'un recouvrement inadéquat entre des secteurs angulaires qui sont de largeurs différentes, ainsi que d'une amplification des directivités des microphones arrière par effet de compensation. Il en résulte ainsi, outre une "précision" r_E sans-doute sous-optimale au seul regard de la géométrie, un déséquilibre énergétique entre les images avant et les images arrière. Un objectif – surtout avec la configuration ($\phi_F = 30°, \phi_B \approx 70°$) – serait d'affiner la directivité des microphones qui correspondent aux zones de plus forte densité angulaire de haut-parleurs, afin d'obtenir un

recouvrement entre secteurs angulaires qui soit plus homogène sur 360°. Pour ce faire, il faut exploiter les composantes ambisoniques d'ordres supérieurs[34]. Dans une première approche simplifiée, on pourrait imaginer définir les microphones virtuels frontaux avec des directivités d'ordre 2 (de type *max* r_E ou *in-phase* par exemple, cf figure 3.14) et les micros arrière avec une directivité d'ordre 1. Pour une définition optimale du décodage, un grand nombre de paramètres sont à prendre en compte: directivité et orientation de chaque microphone virtuel (si la figure de directivité est symétrique!) et amplitude.

Une étude sans-doute passionnante mériterait d'être poursuivie dans cette direction. Grâce à cette *exploitation hybride* de la représentation ambisonique, on peut espèrer obtenir une *meilleure homogénéité énergétique des images sonores* malgré la non-régularité du dispositif. En revanche, il est normal que la précision des images reste meilleure là où les haut-parleurs sont les plus rapprochés, même si elle est globalement améliorée.

3.3.5 Configurations de type hémisphérique

La restitution ambisonique 3D telle qu'elle a été abordée jusqu'ici met en jeu des dispositifs de haut-parleurs réguliers ou semi-réguliers qui englobent l'auditeur (Figure 3.12 ou autre), et exigent la présence de haut-parleurs en-dessous du plan de l'auditeur (ou disons de ses oreilles). Indépendamment du nombre de haut-parleurs requis, les configurations de ce type se heurtent à plusieurs problèmes pratiques évidents: le placement de haut-parleurs sous les pieds de l'auditoire, qu'exigeraient certaines configurations, et le problème du masquage interpersonnel. C'est pourquoi ce sont des configurations de type hémisphérique qui sont le plus raisonnablement envisagées pour la restitution ambisonique 3D, en dehors des configurations cubiques ou parallélépipédiques, réservées à un auditoire réduit et à un système d'ordre 1.

Principales formes de configurations

Parmi les formes principales de configurations hémisphériques, on trouve:

- Les dômes géodésiques (Figure 3.19)[35]: la moitié supérieure d'un polyèdre régulier ou d'une expansion géodésique. Il faut noter que leur couronne la plus basse est en général surélevée par rapport à la base de l'hémisphère.
- Les pyramides (Figure 3.19): constituées d'une couronne horizontale de haut-parleurs – de préférence disposés en polygone régulier – et d'un haut-parleur au zénith ($\delta = 90°$).
- Plus généralement, les étagements de couronnes horizontales de haut-parleurs.

Stratégies de base pour le décodage

Il est difficile de trouver des descriptions satisfaisantes sur les systèmes de décodage ambisonique qui ont été mis en oeuvre pour les dispositifs de ce type. La présentation du projet de Kimmo Vennonen pour une restitution ambisonique d'ordre 1 sur un dôme [Ven] (Figure 3.19) n'en livre pas les détails, et l'on comprend d'ailleurs mal comment les critères classiques de décodage qui y sont mentionnés, comme la cohésion des vecteurs vélocité et énergie ($\vec{u}_V = \vec{u}_E = \vec{u}$), peuvent être vérifiés dans ces conditions. En l'absence de solution vérifiant rigoureusement ces critères, il est intéressant de comparer les méthodes génériques introduites dans

34. Attention, il ne s'agit plus d'utiliser la pseudo-inverse de la matrice d'encodage d'ordre 2, le résultat serait catastrophique!

35. Depuis le projet de l'ACAT [Ven] utilisant un dispositif à 16 haut-parleurs comme celui de la figure 3.19, la conception et la réalisation de dôme géodésiques portatifs sont poursuivies par David Worral [Sur].

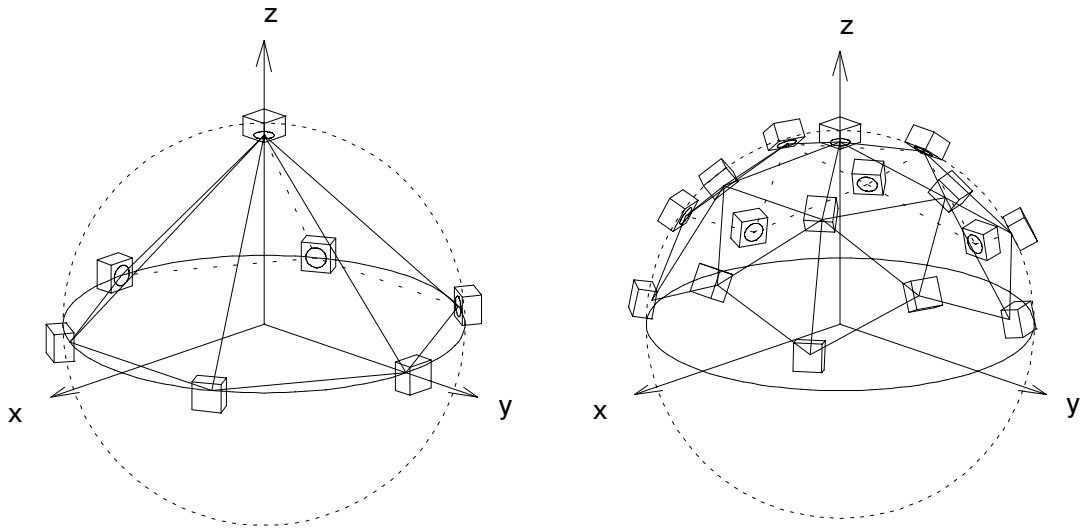

FIG. 3.19 – *Exemples de configurations hémisphériques pour la restitution ambisonique: pyramide (ici, de base hexagonale régulière) et dôme géodésique (moitié supérieure du polyèdre à 32 sommets de la figure 3.12) tel celui utilisé par l'ACAT [Ven]. Les trois couronnes de la configuration en dôme sont des pentagones réguliers en quinconce disposés approximativement aux sites* 10,8°, 26,6° *et* 52,6°.

les sections précédentes pour la définition de la matrice de décodage. Les deux approches de base consistent à:

- Définir la matrice de décodage **D** comme la pseudo-inverse de la matrice de réencodage **C** (3.63).
- Appliquer la méthode de projection (3.64) ou l'une de ses variantes (décodage modifié *max* r_E ou *in-phase*): $\mathbf{D}^{(\mathrm{N3D})} = \frac{1}{N}\mathbf{C}^{(\mathrm{N3D})t}.\mathbf{\Gamma}_{\{g_m\}}$ (Equations 3.76 et 3.78).

Des variantes peuvent ensuite être imaginées à partir de ces deux schémas. Nous portons notre attention sur deux exemples de configurations: la pyramide à base hexagonale régulière et le dôme géodésique à 16 sommets (Figure 3.19). La figure 3.20 illustre le schéma de prise de son équivalant à chacune des deux approches de base évoquées pour un système d'ordre 1. Dans la suite, nous nous focalisons sur une caractérisation de la restitution par le vecteur énergie $\vec{E}$ et le gain d'énergie E, ce que décrit la figure 3.21 pour les cas évoqués ci-dessus.

Le décodage par pseudo-inverse[36] a pour vocation la reconstruction des caractéristiques locales du front d'onde (*i.e.* des composantes ambisoniques) au centre du dispositif. Son cas impose plusieurs commentaires:

- Dans le schéma de prise de son équivalent (Figure 3.20), les microphones virtuels ne sont en général pas orientés vers leurs haut-parleurs respectifs, et l'on constate une prédominance des lobes orientés vers le bas, même quand il s'agit d'un lobe secondaire négatif. Il y a donc un déséquilibre énergétique (Figure 3.21) en faveur des sources virtuelles du demi-espace inférieur:cela traduit un *effort de reconstruction excessif*.
- On pourrait vérifier que la *reconstruction des fronts d'onde venant du bas* n'est réalisée que sur une zone centrale très petite, ou encore à l'échelle de la tête, sur un domaine basse-fréquence très réduit,

36. Cette stratégie de décodage ambisonique a par exemple été appliquée par Véronique Larcher à une configuration pyramidale à base carrée, lors d'un travail collaboratif avec la compositrice Cécile Le Prado en 1998-1999.

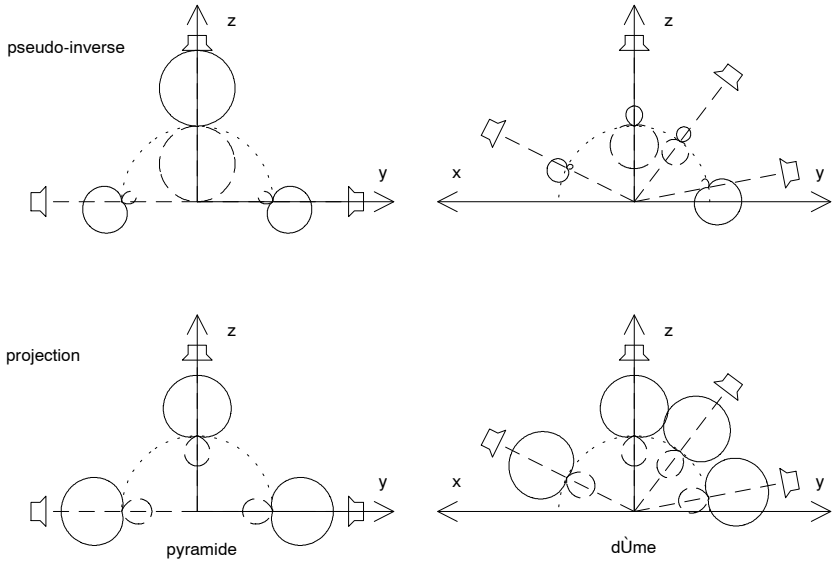

FIG. 3.20 – *Prises de son équivalentes aux décodages par pseudo-inverse* $\mathbf{D}_{pinv}$ *et par projection* $\mathbf{D}_{proj}$ *(système d'ordre 1) pour les configurations de la figure 3.19 (pyramide et dôme). Les diagrammes de directivité des microphones associés aux haut-parleurs sont montrés en coupe verticale (lobes négatifs en tirets) et excentrés pour plus de clarté (les microphones équivalents sont en réalité coïncidents).*

même pour une position d'écoute centrale. Cette remarque est appuyée par le fait que le vecteur énergie $\vec{E}$ est alors très différent du vecteur incidence espéré $\vec{u}$. La reconstruction locale et sa caractérisation par le vecteur vélocité $\vec{V}$ sont donc *perceptivement peu pertinentes* pour ces incidences négatives.

– On observe un phénomène de *repliement* de l'hémisphère inférieur sur l'hémisphère supérieur, que traduit la courbe du site δ_E du vecteur énergie (Figure 3.21).

– Dans le cas de la configuration pyramidale, la figure de directivité du microphone zénithal (figure "en 8") indique que seule la couronne horizontale participe à la création des images de site $\delta = 0$. Ce *découplage vertical-horizontal* permet de traiter les images horizontales comme dans le cas d'une restitution purement 2D, et d'espérer pour celles-ci une meilleure résolution spatiale. Cette idée est utilisée plus loin pour définir des variantes de décodage (décodages hybrides) exploitant les solutions optimisées 2D, voire des composantes horizontales d'ordres supérieurs (U, V), selon le nombre de haut-parleurs de la couronne. En contrepartie, la qualité (r_E) des images est peu homogène en fonction du site.

Cette stratégie de décodage est donc peut recommandable, à moins de ne pas craindre le déséquilibre et le repliement bas-haut. Son usage peut à la rigueur être envisagé s'il est restreint à un domaine basse-fréquence et à une écoute individuelle centrée. Appliquée aux ordres supérieurs et dans le cas du dôme, l'effort de reconstruction pour les incidences négatives et le déséquilibre haut-bas se montre encore plus excessif (excroissance démesurée des lobes inférieurs). La pyramide est quant à elle impropre à la reconstruction des harmoniques sphériques d'ordres supérieurs non-horizontales (Y_{mn}^{σ} telles que $n < m$).

La méthode par projection (Figure 3.20, en bas) met en jeu des microphones orientés vers les haut-parleurs, de directivité "*basique*" au sens du décodage basique 3D (voir Figure 3.14). Des variantes de décodage consistent à substituer à cette directivité l'une de celles correspondant aux décodages modifiés *max*

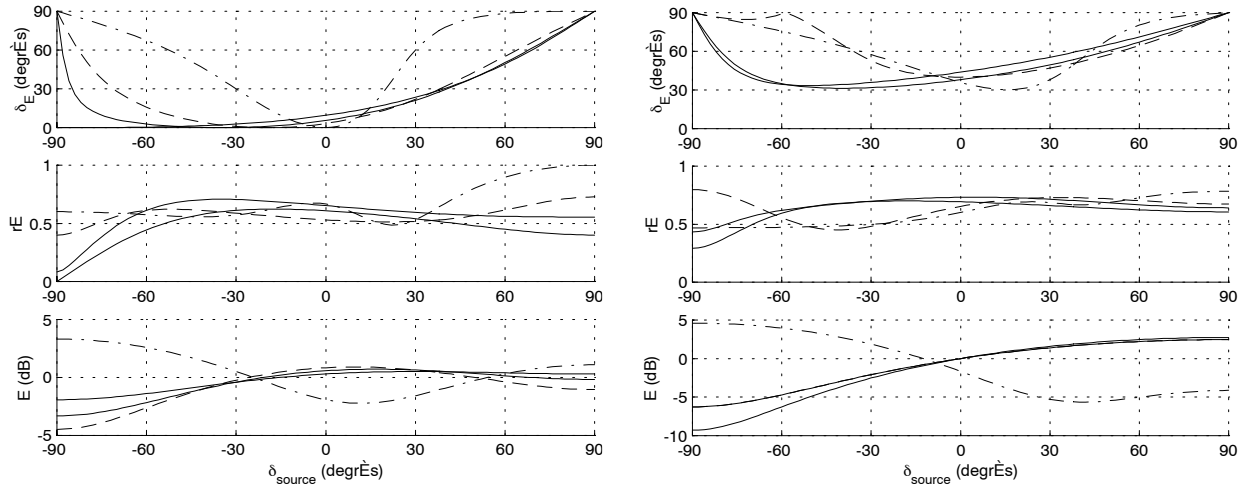

FIG. 3.21 – *Caractéristiques de restitution pour la configuration pyramidale (à gauche) et le dôme (à droite) de la figure 3.19. Site δ_E et module r_E du vecteur énergie, ainsi que le gain d'énergie E en fonction du site δ_{source} de la source virtuelle. Décodages (ordre $M = 1$) par pseudo-inverse (tirets-pointillés) et par projections "basique" (tirets), "max r_E" (trait continu fin) et "in-phase" (trait continu gras). On ne montre pas la loi de l'azimut ($\theta_E = \theta_{source}$): la régularité azimutale est assurée à l'ordre 1.*

r_E ou *in-phase* (cardioïde) du même ordre (Figure 3.14). Ces trois formes de décodage sont présentes Figure 3.21. Quelques nuances sont observées entre le cas de la pyramide et celui du dôme.

Cas de la pyramide. L'effet de renversement "bas-haut" observé avec la pseudo-inverse est ici largement gommé, grâce à la réduction des lobes secondaires des microphones. Il est même absent avec la variante *in-phase* (directivité cardioïde), pour laquelle le haut-parleur au zénith ne participe pas à l'image diamétralement opposée ($\delta_{source} = -90°$) et la courbe $\delta_E(\delta_{source})$ est croissante. L'effet de verticalité pour les sites δ_{source} négatifs passe par une diminution de l'indice r_E, c'est-à-dire de la latéralisation, particulièrement bien réalisée avec la version *in-phase*: $r_E = 0$ pour $\delta_{source} = -90°$. On note enfin un effet d'atténuation des sources virtuelles s'enfonçant vers les sites négatifs, au contraire du décodage par pseudo-inverse. Il faut souligner que la version *in-phase* ne doit son comportement caractéristique qu'à la directivité cardioïde du microphone orienté vers le zénith[37]. Tout en conservant des propriétés semblables vis-à-vis de la dimension verticale, d'autres directivités pourraient être choisies pour la couronne horizontale dans le but d'optimiser la résolution azimutale. Cette nouvelle option de décodage hybride est abordée un peu plus loin.

Cas du dôme. On constate pour cette configuration un effet de surélévation ($\delta_E > 30°$) beaucoup plus évident qu'avec la pyramide. Cette effet a deux causes: d'une part la couronne la plus basse est elle-même surélevée par rapport à la base de l'hémisphère (site $\delta \approx 11°$), et d'autre part la couverture plus homogène de l'hémisphère par les haut-parleurs ne permet plus de découplage horizontal-vertical, même partiel. Une autre conséquence est un effet de repliement bas-haut, confiné à un secteur relativement réduit pour les versions *in-phase* et *max* r_E, mais beaucoup plus marqué pour la version *basique* (courbe δ_E, Figure 3.21) du fait d'un lobe secondaire plus important (Figure 3.20). Par contre, l'imagerie sonore dans l'hémisphère supérieur est plus consistante et homogène qu'avec le dispositif pyramidal, d'après l'indice r_E.

37. Notons au passage que son gain est ici fixé de façon arbitraire, et mériterait un réglage pour une égalisation de l'énergie en fonction du site.

Restitution d'ordres supérieurs pour le dôme

Le dôme étudié ici constitue la moitié d'un polyèdre (à 32 sommets) quasi-régulier pour la restitution ambisonique jusqu'à l'ordre 4, comme nous l'avons signalé en 3.2.3 et en 3.3.1. Il est donc envisageable d'utiliser ce dôme pour des systèmes d'ordres supérieurs à 1. Nous écartons ici la méthode de décodage par pseudo-inverse, et nous intéressons à la méthode par projection et ses variantes. Pour un ordre M donné, si $\mathbf{C}^{(\mathtt{N3D})}$ est la matrice de réencodage associée à la configuration (cf équation?), la matrice de décodage prend la forme $\mathbf{D}^{(\mathtt{N3D})} = \frac{1}{N}\mathbf{C}^{(\mathtt{N3D})}$ ou bien $\frac{1}{N}\mathbf{C}^{(\mathtt{N3D})}.\mathbf{\Gamma}^{M(3D)}_{\max r_E}$, ou encore $\frac{1}{N}\mathbf{C}^{(\mathtt{N3D})}.\mathbf{\Gamma}^{M(3D)}_{\text{in-phase}}$, selon la variante de décodage et à un coefficient de normalisation près. Le schéma de prise de son équivalent revient alors à remplacer, dans la figure 3.20 (en bas à droite), le diagramme de directivité par celui d'ordre M correspondant au décodage 3D considéré (Figure 3.14).

Il est clair que la restitution hémisphérique ne peut pas jouir de toutes les propriétés de cohérence et d'homogénéité qu'offrirait la configuration polyédrale régulière non-tronquée. Reste à préciser quelles sont les propriétés dégradées, et quelles sont celles qui sont préservées. La distorsion des caractéristiques de restitution (site δ_E, indice r_E et énergie E) suivant la dimension verticale a déjà été observée avec l'ordre 1. D'autres aspects se manifestent plus spécifiquement avec les ordres supérieurs.

Si l'on regarde le décodage sous l'angle d'une prise de son équivalente, la *distribution relativement dense et homogène des haut-parleurs* reste *a priori* suffisante pour **tirer profit des directivités microphoniques d'ordres supérieurs** – plus sélectives – sans craindre d'effet de "trou" ou d'irrégularité entre les haut-parleurs: les microphones virtuels assurent correctement la couverture de l'hémisphère supérieur au moins jusqu'à l'ordre 4. La figure 3.22 l'atteste en grande partie: pour les sources virtuelles assez nettement incluses dans l'hémisphère supérieur, les caractéristiques de restitution (direction apparente $\vec{u}_E$, "qualité" r_E et énergie E) sont relativement homogènes et tendent vers celles qu'offrirait une configuration régulière[38]. Dans cet hémisphère, la qualité de restitution augmente donc avec l'ordre M, au moins jusqu'à l'ordre 4.

En revanche, l'absence de haut-parleurs dans l'hémisphère inférieur est responsable d'un **effet de bord**. Il se manifeste par les fluctuations du vecteur énergie $\vec{E}$ et de l'énergie E lors du déplacement azimutal des sources en bordure ou à l'extérieur de l'hémisphère supérieur (Figure 3.22a), et ceci d'autant plus que l'ordre M est élevé. Ces fluctuations n'étaient pas présentes avec les encodages/décodages d'ordre 1, pour lesquelles la régularité azimutale est assurée indépendamment par chacune des couronnes horizontales[39]. On note que cet effet de bord moins marqué avec le décodage *max* r_E, et très largement gommé par le décodage *in-phase*. L'augmentation de l'ordre M accroît également le phénomène de *repliement bas-haut* pour les décodages basique et, dans une moindre mesure, *max* r_E (Figure 3.22)[40]. Le décodage *in-phase*, au contraire, y résiste bien. De manière générale, on note que le site minimal apparent δ_E s'abaisse avec les ordres supérieurs, de même que l'effet d'atténuation augmente pour les sites négatifs.

On a pu souligner certaines vertus du décodage *in-phase*, présentes à tout ordre: lissage des effets de bords, relative cohérence (anti-repliement), continuité et homogénéité. Quelques simulations complémentaires semblent indiquer que ces propriétés survivent assez bien – ou ne se dégradent que lentement – en augmentant l'ordre du système au-delà de 4, alors que le polyèdre dont est issu le dôme n'est quasi-régulier que jusqu'à cet ordre! Mais il faut aussi remarquer que la résolution spatiale associée à ce décodage n'augmente elle-même que lentement, comparée au décodage *max* r_E (Table 3.10): le décodage *in-phase* d'ordre 5 ($r_E = 0{,}833$) n'offre pas la précision du décodage *max* r_E d'ordre 3 ($r_E = 0{,}861$), par exemple. Le choix du

38. On note cependant, vers le zénith, un comportement de l'indice r_E un peu plus singulier de la part du décodage basique, pour lequel l'absence de contre-poids dans l'hémisphère inférieur est plus sensible.

39. $N = 5$ haut-parleurs par couronne suffisent en effet pour l'ordre $M = 1$. De manière générale, il en faut au moins $2M + 2$ pour assurer la régularité azimutale au sens du vecteur énergie.

40. Les irrégularités de la courbe $\delta_E(\delta_{\text{source}})$ sont liées au nombre et à l'importance des lobes secondaires des directivités microphoniques associées.

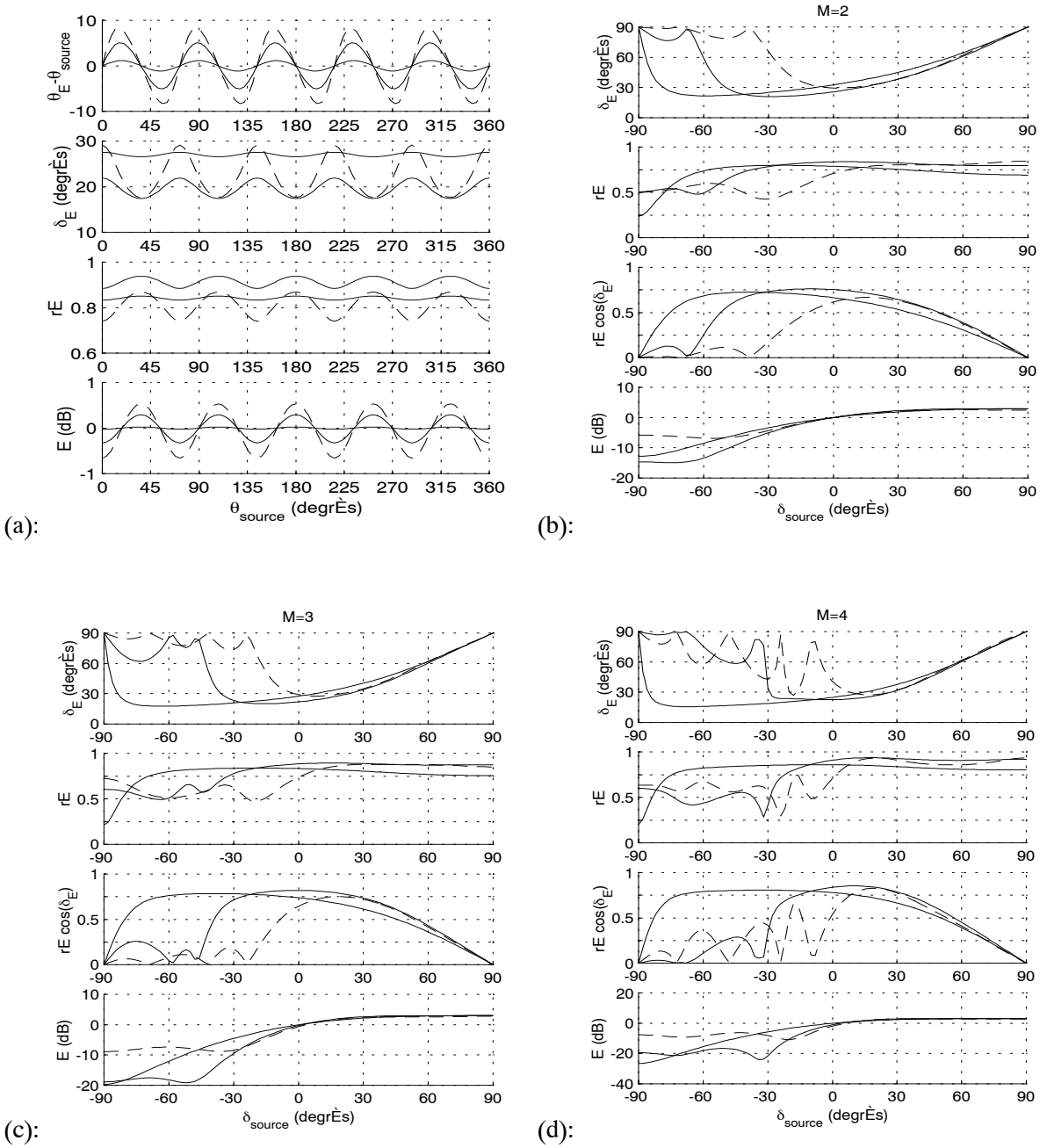

FIG. 3.22 – *(a): restitution d'ordre 3 en fonction de l'azimut* θ_{source} *et pour un site* $\delta_{source} = 0$*. (b), (c) et (d): restitutions d'ordres 2, 3 et 4 sur le dôme géodésique en fonction du site* δ_{source} *et pour un azimut* $\theta_{source} = 0$*. Légende similaire à la figure 3.21, mais en l'absence du décodage par pseudo-inverse.*

décodage à un ordre M donné reste donc une affaire de compromis entre précision et cohérence générale.

Enfin, un autre aspect de la restitution mérite de retenir l'attention: la *discrimination* des sources virtuelles du demi-espace inférieur en terme d'**impression de verticalité**. Compte-tenu de mouvements de rotation *yaw* de la tête (Cf 1.3.3), cette impression est antagonique de la qualité de latéralisation dynamique dans le plan horizontal, ce que traduit l'indice $r_E \cos\delta_E$ (module de la projection du vecteur énergie sur le plan horizontal). Pour observer un effet de verticalité pour les sources de site δ_{source} négatif sans avoir recours à l'artefact du renversement bas-haut, il faudrait donc voir diminuer progressivement l'indice r_E associé à leur restitution. Force est de constater (Figure 3.22) que la discrimination offerte par le décodage *in-phase* est de ce point de vue très faible sur une large tranche de l'hémisphère inférieur, et ce d'autant plus que l'ordre M est élevé. La seule discrimination se fait par effet d'atténuation (courbe d'énergie E). Une question corollaire est celle de l'impression d'horizontalité[41] maximale: observée pour des δ_{source} franchement négatifs avec le décodage *in-phase*, il serait préférable de la voir réservée à des sources virtuelles originellement plus proches du plan horizontal, ce que réalisent mieux les autres décodages, mais au prix de l'effet de repliement.

On peut espérer améliorer cette discrimination vertical/horizontal des sources de l'hémisphère inférieur (plan horizontal compris). Il suffirait par exemple de diminuer la sélectivité azimutale des microphones virtuels – donc diminuer r_E – pour les incidences négatives. Des moyens empiriques peuvent être proposés pour cela: abaisser les orientations des microphones virtuels[42] par exemple, ou bien injecter à leur directivité une dépendance en $-Z$. Ces deux mesures auraient par ailleurs pour effet de rééquilibrer le gain d'énergie E en fonction du site.

Configuration pyramidale: exemples de décodage hybride

La configuration pyramidale a la particularité d'avoir ses haut-parleurs répartis suivant deux sous-espaces directionnels orthogonaux: le plan horizontal (couronne polygonale) et l'axe vertical (haut-parleur au zénith). Il semble donc possible de différencier les critères de décodage selon ces deux sous-espaces, et mettre à profit les résultats des sections précédentes sur le décodage 2D. Concernant la dimension verticale, deux options extrêmes se sont dessinées au cours de cette section, qui se traduisent par la figure de directivité du microphone virtuel associé au zénith:

1. soit accepter le repliement du demi-espace sonore inférieur sur l'hémisphère supérieur et pouvoir contrôler de façon optimale la qualité de l'image dans le plan horizontal (bidirectivité ou "figure en 8"),
2. soit assurer une évolution cohérente et monotone des caractéristiques de restitution suivant la dimension verticale (site de la source encodée), au prix d'une surélévation des sources virtuelles horizontales (directivité cardioïde).

A titre d'illustration, il est utile de montrer de façon formelle comment définir les matrices de décodage qui correspondent à ces propositions. Nous prenons pour cela l'exemple d'une représentation hybride, d'ordre $M_{\text{hor}} = 2$ suivant le plan horizontal et d'ordre $M_{\text{ver}} = 1$ suivant la dimension verticale, et d'une optimisation du décodage horizontal suivant le critère *max* r_E. Les canaux ambisoniques W, X, Y, Z, U, V sont encodés d'après la convention `N3D`. Des opérations de conversion s'imposent donc pour exploiter les résultats du décodage 2D. Dans le cas présent, le vecteur d'encodage associé à une direction d'incidence $\vec{u}$ a ses composantes ordonnées de la façon suivante: $\mathbf{c}(\vec{u}) = [Y_{00}^1(\vec{u}), Y_{11}^1(\vec{u}), Y_{11}^{-1}(\vec{u}), Y_{10}^1(\vec{u}), Y_{22}^1(\vec{u}), Y_{22}^{-1}(\vec{u})]^t$ (toujours avec la convention implicite `N3D`), et associées respectivement aux canaux W, X, Y, Z, U, V. La configuration pyramidale étant décrite par les incidences $\{\vec{u}_z, \vec{u}_1, \vec{u}_2, \ \ldots \ \vec{u}_{N_{\text{hor}}}\}$, où $N_{\text{hor}} = N - 1$ est le nombre de haut-parleurs

41. Proportionnelle à l'indice $r_E \cos\delta_E$.

42. Ce qui revient à définir la matrice de réencodage $\mathbf{C}$ sur la base de directions $\vec{u}_i$ abaissées par rapport à celles des haut-parleurs, avant d'en déduire la matrice de décodage $\mathbf{D} = \frac{1}{N}\mathbf{C}^t$.

horizontaux[43], la matrice de réencodage associée peut se décomposer ainsi:

$$\mathbf{C}^{(\mathtt{N3D})} = [\mathbf{c}_z \; \mathbf{C}_{\text{hor}}^{(\mathtt{N3D})}], \quad \text{avec:} \left\{ \begin{array}{ll} \mathbf{c}_z & = \mathbf{c}(\vec{u}_z) \\ \mathbf{C}_{\text{hor}}^{(\mathtt{N3D})} & = [\mathbf{c}(\vec{u}_1) \; \ldots \; \mathbf{c}(\vec{u}_{N_{\text{hor}}})] \end{array} \right. \tag{3.96}$$

De même, la matrice de décodage peut s'écrire:

$$\mathbf{D}^{(\mathtt{N3D})} = \begin{bmatrix} \mathbf{D}_z \\ \mathbf{D}_{\text{hor}}^{(\mathtt{N3D})} \end{bmatrix} \tag{3.97}$$

Selon l'option choisie, la matrice-ligne de décodage associée au haut-parleur zénithal prend la forme:

$$\begin{array}{lll} \mathbf{D}_z & = \gamma_z [0\;0\;0\;1\;0\;0] & \text{figure "en 8" (option 1)} \\ \mathbf{D}_z & = \gamma_z [1\;0\;0\;\frac{1}{\sqrt{3}}\;0\;0] & \text{cardioïde (option 2)} \end{array} \tag{3.98}$$

Pour optimiser la partie horizontale du décodage, il est nécessaire de passer par la convention `N2D`:

$$\mathbf{D}_{\text{hor}}^{(\mathtt{N3D})} = \mathbf{D}_{\text{hor}}^{(\mathtt{N2D})} . \text{Diag}(\underline{\alpha}^{(\mathtt{N2D})\mathtt{N3D}}), \tag{3.99}$$

où les coefficients du vecteur de conversion $\underline{\alpha}^{(\mathtt{N2D})\mathtt{N3D}} = [\alpha_{00}\;\alpha_{11}\;\alpha_{11}\;\alpha_{10}\;\alpha_{22}\;\alpha_{22}]^t$ sont définis en 3.1.2: $\alpha_{00} = 1$, $\alpha_{11} = \alpha_{10} = \sqrt{2/3}$, $\alpha_{22} = \sqrt{8/15}$. La résolution se poursuit par:

$$\mathbf{D}_{\text{hor}}^{(\mathtt{N2D})} = \frac{1}{N_{\text{hor}}} \left(\mathbf{C}_{\text{hor}}^{(\mathtt{N2D})} \right)^t . \mathbf{\Gamma}_{\max r_E}^{M=2(2D)} \quad \text{où:} \left\{ \begin{array}{ll} \mathbf{C}_{\text{hor}}^{(\mathtt{N2D})} & = \text{Diag}(\underline{\alpha}^{(\mathtt{N2D})\mathtt{N3D}}) . \mathbf{C}_{\text{hor}}^{(\mathtt{N3D})} \\ \mathbf{\Gamma}_{\max r_E}^{M=2(2D)} & = \text{Diag}([g_0\;g_1\;g_1\;0\;g_2\;g_2]^t) \end{array} \right. \tag{3.100}$$

Notons qu'un coefficient nul, associé à la composante Z, a dû être rajouté à la diagonale de $\mathbf{\Gamma}_{\max r_E}^{M=2(2D)}$, par rapport à sa définition originale "purement 2D" (3.77). Les facteurs g_m sont définis d'après la table 3.10: $g_1/g_0 = \sqrt{3}/2$, $g_2/g_0 = 1/2$ et $g_0 = \sqrt{N_{\text{hor}}/3}$. Finalement:

$$\mathbf{D}_{\text{hor}}^{(\mathtt{N3D})} = \frac{1}{N_{\text{hor}}} \left(\mathbf{C}_{\text{hor}}^{(\mathtt{N3D})} \right)^t . \text{Diag}([g_0.\alpha_{00}^2\;\; g_1.\alpha_{11}^2\;\; g_1.\alpha_{11}^2\;\; 0\;\; g_2.\alpha_{22}^2\;\; g_2.\alpha_{22}^2]^t) \tag{3.101}$$

Naturellement, d'autres critères de décodage peuvent être appliqués à la place du critère *max r_E* pour le décodage horizontal: il suffit pour cela de changer le jeu de gains $\{g_0, g_1, g_2\}$. Quant au gain γ_z (3.98) pondérant le haut-parleur placé au zénith, il joue à la fois sur l'évolution du site apparent δ_E et de l'énergie restituée E en fonction du site δ_{source} de la source encodée. Il est clair que l'option 2 entraîne un effet de déséquilibre énergétique bas-haut similaire au cas du décodage par projection (Figure 3.21). Comme proposé plus haut pour le dôme, une compensation pourrait être obtenue par abaissement de l'orientation des microphones virtuels horizontaux.

Il faut voir dans ces propositions une partie seulement des solutions de décodage possibles, et l'occasion de montrer comment exploiter formellement les solutions de décodage 2D dans ce contexte. De nombreuses variantes pourraient être envisagées.

43. On suppose ici que la couronne horizontale définit un polygone régulier, et que $N_{\text{hor}} \geq 2M_{\text{hor}} + 2$.

Conclusions

Les stratégies de décodage qui viennent d'être présentées pour la restitution sur dispositif hémisphérique s'inspirent largement des méthodes utilisées pour les configurations régulières: il s'agit basiquement des décodages par pseudo-inverse et par projection[44]. Privilégiant la seconde approche, il s'avère tout à fait envisageable de définir des systèmes ambisoniques 3D d'ordres supérieurs (jusqu'à l'ordre 4 par exemple) pour des dispositifs, comme le dôme géodésique à 16 haut-parleurs (Figure 3.19)[45], qui ne relèvent pas nécessairement de la science-fiction! De plus, l'application des solutions de décodage *max r_E* et *in-phase* généralisées en 3.3.2 s'y révèle très bénéfique, tant avec le dôme (solutions 3D) qu'avec le dispositif pyramidal (décodage hybride utilisant les solutions 2D).

Il y a matière à un approfondissement des méthodes de décodage pour les configurations de ce type. Des variantes de décodage sont pressenties, qui pourraient réduire les artefacts de la restitution, et notamment concilier la limitation du repliement bas-haut avec l'amélioration de la discrimination verticale des sources de l'hémisphère inférieur et la réduction de leur atténuation. Les pistes privilégiées vont à une différenciation des critères de décodage, voire de l'ordre de décodage, selon les dimensions verticale et horizontale.

3.4 Microphones ambisoniques d'ordres 1 et supérieurs

Les solutions existantes pour la prise de son ambisonique d'ordre 1 (B-format) ont été évoquées en 2.4.2.

L'approche dont nous développons ici quelques tenants théoriques, s'appuie sur la notion d'*échantillonnage directionnel* (sphérique) de la base orthonormée des fonctions harmoniques sphériques utilisées pour l'encodage, et sur une *opération de projection "discrète"* qui résulte de la *propriété de régularité* du support d'échantillonnage, par laquelle l'orthonormalité de la base des harmoniques échantillonnées est préservée (définition en 3.2.3).

L'opération de "projection du champ acoustique" est d'abord appliquée sur des fonctions harmoniques sphériques "continues", *i.e.* non-échantillonnées, afin de mettre en évidence l'intérêt d'utiliser une directivité de type cardioïde pour la mesure du champ, pour résoudre le problème de l'égalisation, s'appliquant comme une correction fréquentielle aux résultats de la projection.

3.4.1 Intégration sur une sphère: utilisation de mesures à directivité cardioïde

Rappelons que le champ de pression à la surface d'une sphère imaginaire de rayon R exempte de source, peut s'écrire –en régime harmonique– comme l'expansion en série de Fourier-Bessel:

$$p(R,\theta,\varphi) = \sum_{m\geq 0, 0\leq n\leq m, \sigma=\pm 1} j^m A^{\sigma}_{mn}(\omega)\tilde{Y}^{\sigma}_{mn}(\theta,\delta)\, j_m(kR), \tag{3.102}$$

où l'on utilise les fonctions harmoniques sphériques normalisées $\tilde{Y}^{\sigma}_{mn}(\theta,\delta) = Y^{\sigma}_{mn}{}^{(\mathrm{N3D})}(\theta,\delta)$. Le problème posé est l'extraction des gains $A^{\sigma}_{mn}(\omega)$ représentant chaque composante ambisonique dans le domaine fréquentiel. Grâce aux propriétés d'orthonormalité de la base des fonctions harmoniques sphériques utilisée, on peut écrire, pour tout triplet $(m \geq 0, 0 \leq n \leq m, \sigma = \pm 1)$:

$$A^{\sigma}_{mn}(\omega) = \frac{1}{j^m\, j_m(kR)} \frac{1}{4\pi} \oint p(R,\theta,\varphi)\tilde{Y}^{\sigma}_{mn}(\theta,\delta)\, d\theta\, \sin\delta\, d\delta \tag{3.103}$$

44. Méthodes qui sont équivalentes pour les configurations régulières.

45. Depuis le projet de l'ACAT [Ven] utilisant un dispositif à 16 haut-parleurs comme celui de la figure 3.19, la conception et la réalisation de dôme géodésiques portatifs sont poursuivies par David Worral [Sur].

Or les fonctions $j_m(kR)$ varient en passant par des zéros lorsque la fréquence $f = \omega/2\pi = kc/2\pi$ parcourt l'ensemble des valeurs réelles (Figure 3.23a), ce qui rend critique l'estimation des A^{σ}_{mn}. Substituer des mesures de vélocité radiale $v(R,\theta,\delta) \sim \partial p/\partial r(R,\theta,\delta)$ aux mesures de pression $p(R,\theta,\delta)$ revient à remplacer en substance $j_m(kR)$ par $j'_m(kR)$ dans (3.103) mais pose les mêmes problèmes d'estimation (Figure 3.23b). Une solution intéressante consiste à utiliser une mesure $c(R,\theta,\delta)$ du champ par des capteurs à directivité cardioïde, et orientés radialement (direction centrifuge). Une telle mesure peut s'écrire, à une constante multiplicative près:

$$s(R,\theta,\delta) = p(R,\theta,\delta) - j\frac{1}{k}\frac{\partial p}{\partial r}(R,\theta,\delta) \tag{3.104}$$

Le facteur de pondération mis en jeu devient $c_m(kR) = j_m(kR) - j\,j'_m(kR)$, il ne s'annule qu'en 0 et pour les ordres $m \geq 2$ (Figure 3.23c).

$$A^{\sigma}_{mn} = \frac{(-j)^m}{j_m(kR) - j\,j'_m(kR)}\frac{1}{4\pi}\oint s(R,\theta,\delta)\tilde{Y}^{\sigma}_{mn}(\theta,\delta)sin\delta\,d\theta\,d\delta \tag{3.105}$$

Il peut être intéressant de noter que les fonctions c_m ont la même tendance asymptotique à un déphasage près:

$$c_m(kR) \underset{kR\to\infty}{\longrightarrow} \frac{1}{kR}e^{j(kR-\frac{m+1}{2}\pi)} \tag{3.106}$$

Ce sont effectivement des capsules à directivité cardioïde qui sont utilisées pour la réalisation du microphone *SoundField* [CG77]. Comme il est suggéré dans ce brevet, il est naturellement possible d'envisager des directivités hypo- ou hyper-cardioïdes éventuellement variables en fonctions de la fréquence, ce qui donne un facteur de pondération de la forme: $c_m(kR) = j_m(kR) - j\alpha(f)\,j'_m(kR)$. Cette possibilité doit s'avérer appréciable, en pratique, étant donné que les capsules réelles ont rarement une directivité idéale et constante sur toute la bande de fréquence.

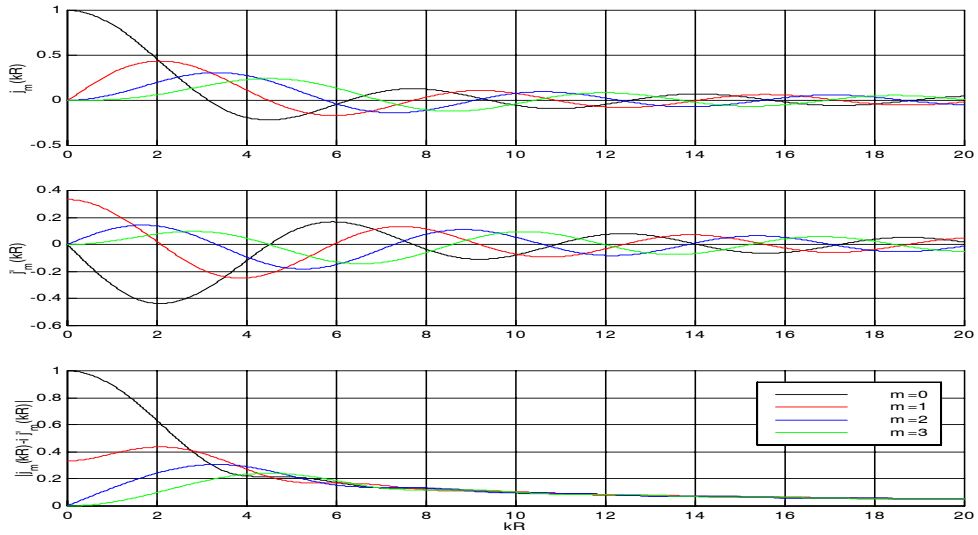

FIG. 3.23 – *De haut en bas: (a) les fonctions de Bessel sphériques $j_m(kR)$ pour les premiers ordres; (b) leurs dérivées $j'_m(kR)$; (c) le module des fonctions complexes $c_m(kR) = j_m(kR) - j\,j'_m(kR)$.*

3.4.2 Principe du système microphonique: projection discrète et égalisation

L'échantillonnage: disposition géométrique des capsules

En pratique, la prise de son ne peut reposer que sur un nombre fini de capsules élémentaires. Il s'agit donc de réaliser un échantillonnage de la sphère considérée plus haut.

Considérons donc un ensemble discret de positions sur une sphère centrée de rayon R, décrites par les couples $\{(\theta_q,\delta_q)\}_{q=1,\dots,N}$ (azimuts et sites). On dispose en chacun de ces points une capsule microphonique, supposée acoustiquement transparente, à directivité cardioïde, et orientée radialement vers l'extérieur de la sphère. La mesure du champ fourni par chaque capsule s'écrit:

$$s_q(\omega) = s(R,\theta_q,\delta_q) = \sum_{m,n,\sigma} j^m A^{\sigma}_{mn}(\omega)\tilde{Y}^{\sigma}_{mn}(\theta_q,\delta_q)c_m(kR) \tag{3.107}$$

On introduit le vecteur des mesures captées $\mathbf{s} = [s_1 \ \dots \ s_N]^t$, qui constitue en quelque sorte un "*format A étendu*". Lorsque les capsules sont au nombre de quatre et montées en tétraèdre, c'est bien le format A original du microphone Soundfield qui est obtenu (cf 2.4.2). On cherche à estimer les composantes ambisoniques A^{σ}_{mn} du champ jusqu'à un ordre M. On suppose que le support $\{(\theta_q,\delta_q)\}_{q=1,\dots,N}$ réalise un *échantillonnage régulier* des fonctions harmoniques sphériques pour une base d'ordre M, c'est-à-dire que les vecteurs $\tilde{\mathbf{Y}}^{\sigma}_{mn} = [\tilde{Y}^{\sigma}_{mn}(\theta_1,\delta_1),\dots,\tilde{Y}^{\sigma}_{mn}(\theta_N,\delta_N)]^t$ ($m \leq M$) définissent une base orthonormée (Cf 3.2.3). Dans la suite, nous utilisons également cette notation pour des degrés $m > M$.

Les principaux supports d'échantillonnage réguliers existants ont été discutés en 3.2.3. Notons que dans le contexte présent de prise de son, il est d'usage de considérer que les capsules sont placées *au centre des faces* d'un polyèdre imaginaire, alors que pour un échantillonnage directionnel équivalent mais dans un contexte de restitution, les haut-parleurs sont illustrés comme étant situés *aux sommets du polyèdre dual*. Les géométries les plus typiques retenues pour la définition de microphones sont: le tétraèdre (4 faces) pour l'ordre $M = 1$ ($K = 4$ composantes), le dodécaèdre (12 faces) pour l'ordre 2 ($K = 9$), et le polyèdre (à 32 faces) dual de celui à 32 sommets (Figure 3.12), régulier jusqu'à l'ordre 2 et *quasi*-régulier pour les ordres 3 ($K = 16$) et 4 ($K = 25$).

Projection discrète et *aliasing*

L'équation (3.103) suggère une première méthode d'estimation des A^{σ}_{mn} pour $m \leq M$, à savoir l'opération de projection:

$$\hat{A}^{\sigma}_{mn}(\omega) = \frac{(-j)^m}{c_m(kR)}\frac{1}{N}\sum_{q=1}^{N} s_q(\omega)\tilde{Y}^{\sigma}_{mn}(\theta_q,\delta_q) = \frac{(-j)^m}{c_m(kR)}\langle \mathbf{s}(\omega)|\tilde{\mathbf{Y}}^{\sigma}_{mn}\rangle_N \tag{3.108}$$

Cette estimation introduit une erreur résiduelle $\varepsilon^{\sigma}_{mn}(\omega)$:

$$\begin{aligned}\hat{A}^{\sigma}_{mn}(\omega) &= \sum_{m',n',\sigma'} \frac{c_{m'}(kR)}{c_m(kR)} A^{\sigma'}_{m'n'}(\omega)\,\frac{1}{N}\tilde{\mathbf{Y}}^{\sigma'}_{m'n'}{}^{t}.\tilde{\mathbf{Y}}^{\sigma}_{mn} \\ &= A^{\sigma}_{mn}(\omega) + \underbrace{\sum_{m'>M} \frac{c_{m'}(kR)}{c_m(kR)} A^{\sigma'}_{m'n'}(\omega)\,\frac{1}{N}\tilde{\mathbf{Y}}^{\sigma'}_{m'n'}{}^{t}.\tilde{\mathbf{Y}}^{\sigma}_{mn}}_{\varepsilon^{\sigma}_{mn}(\omega)}\end{aligned} \tag{3.109}$$

L'erreur d'estimation $\varepsilon^{\sigma}_{mn}$ vient du fait que les produits scalaires $\tilde{\mathbf{Y}}^{\sigma'}_{m'n'}{}^{t}.\tilde{\mathbf{Y}}^{\sigma}_{mn}$ ne sont *a priori* pas nuls pour $m \leq M$ et $m' > M$. Elle est donc interprétable comme l'effet d'un *aliasing spectral* (repliement du spectre harmonique sphérique). Il est possible de la réduire en effectuant un sur-échantillonnage directionnel, donc

en mettant en jeu plus de capteurs que le nombre nécessaire à l'échantillonnage d'ordre M. Par exemple, choisir une géométrie dodécaédrique pour une prise de son d'ordre 1 élimine complètement le repliement des harmoniques d'ordre 2 et réduit celui des harmoniques d'ordres supérieurs. L'erreur d'aliasing s'annule pour les basses fréquences et croît avec kR. En effet, le facteur $c_{m'}(kR)/c_m(kR)$ tend vers 0 quand $kR \to 0$ et vers $j^{m-m'}$ quand $kR \to \infty$.

Il serait donc possible de définir pour chaque ordre m une fréquence d'aliasing f_m en-deçà de laquelle l'extraction des composantes $A^{\sigma}_{mn}(f)$ peut être jugée satisfaisante, et au-delà de laquelle l'égalisation en $\frac{(-j)^m}{c_m(kR)}$ doit être remise en cause. Pour une configuration donnée, $f_m > f_{m'}$ si $m < m'$. Cette fréquence augmente en multipliant le nombre N des capsules utilisées et en réduisant le rayon R.

3.4.3 Conclusion

Les pistes théoriques pour la prise de son d'ordre supérieur que nous avons évoquées ici ont été l'occasion d'illustrer la notion sous-jacente d'échantillonnage directionnel introduite en 3.2.3. Pour donner lieu à la conception d'un microphone, il faudrait encore approfondir le problème de l'égalisation dans le domaine haute-fréquence où l'*aliasing* du spectre harmonique sphérique est présent. Enfin, il faudrait inclure dans l'étude l'influence d'autres paramètres, comme la perturbation du champ par les capsules et leurs caractéristiques de directivité non-idéales.

Chapitre 4

Evaluation selon les conditions écoutes - Conclusion

4.1 Evaluation en conditions d'écoute idéales

4.1.1 Objectifs de l'étude

Par "conditions idéales d'écoute", on entend ici une écoute individuelle, l'auditeur étant placé au centre du dispositif de haut-parleurs. L'écoute au casque par le biais d'une simulation binaurale (méthode des haut-parleurs virtuels) en constitue donc un cas de figure, avec cependant la particularité que la tête reste fixe par rapport au dispositif virtuel. Dans ces conditions, sur lesquelles repose l'optimisation du décodage d'après Gerzon (cf 2.4.3 et 3.1.3), il est d'usage de caractériser les performances de restitution à travers les vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ (considérés au centre du dispositif) et des indices r_V et r_E qui en dérivent. C'est ce que présente la table 3.10 du chapitre précédent pour les solutions de décodage que nous avons généralisées aux ordres supérieurs.

Au delà de ces indices et des propriétés de reconstruction du champ promises par l'étude 3.2, l'étude qui suit a pour objectif d'apporter une validation plus substantielle de l'apport des ordres supérieurs et des décodages optimisés. Par ailleurs, bien que nous ayons explicité et argumenté des relations de prédiction entre les grandeurs $\vec{V}$ et $\vec{E}$ et l'effet supposé de localisation (en 1.5), il semble utile, à la fois pour leur accorder crédit et à titre d'illustration, de les confronter à des mesures objectives du rendu sonore et à des expériences d'écoute subjective. Un des objectifs particuliers est de corréler les indices r_V et r_E aux lois d'ITD et d'ILD mesurées aux oreilles de l'auditeur.

Cette étude poursuit donc celle entreprise dans [DRP98], de façon plus complète et plus illustrative: l'éventail des décodages testés est plus large, une évolution des mesures est donnée pour un balayage panoramique de la source virtuelle, et des méthodes originales sont employées, notamment pour l'estimation globale de l'ITD haute-fréquence (Cf 1.4). Le *principe des haut-parleurs virtuels* évoqué en 2.5.1 (et en 3.1.3) est appliqué à la restitution ambisonique – en utilisant des mesures de HRTF fournies par le MIT [GM94] – pour calculer des réponses binaurales synthétiques et en extraire les spectres et indices de localisation habituels (ITD, ILD, d'après 1.4).

L'évaluation objective est ensuite confrontée aux résultats d'expériences d'écoutes informelles, réalisées au casque – en utilisant encore la méthode des haut-parleurs virtuels – et avec des haut-parleurs réels.

4.1.2 Analyse objective, validité et portée des prédictions théoriques

Extension du domaine de reconstruction et indices spectraux

Ainsi que le suggère l'étude sur les troncatures cylindrique et sphérique du champ en 3.2, la reconstruction acoustique à l'échelle de la tête – et plus particulièrement au niveau des oreilles – est restreinte à une région basse-fréquence qui s'élargit avec l'ordre M du système ambisonique, en supposant qu'un décodage basique soit appliqué. Cela a été illustré dans [DRP98] (Annexe B) pour une direction d'incidence isolée. La figure 4.1 permet de vérifier cette tendance pour un balayage panoramique de la source virtuelle et avec des configurations de haut-parleurs plausibles. On note au passage que la reconstruction est meilleure lorsque la source virtuelle est placée au voisinage d'un haut-parleur.

Ce que montre aussi la figure 4.1, c'est que les systèmes ambisoniques d'ordres limités sont peu aptes de reproduire les indices spectraux, qui commencent à participer à la localisation à partir de 5 kHz environ. Le spectre résultant est en effet issu d'un lissage des spectres correspondant aux directions des haut-parleurs. Il semble que le "relief spectral" soit reconstitué avec un meilleur contraste[1] à mesure que l'ordre M augmente, même au-delà du domaine basse-fréquence de reconstruction. Cela peut s'interpréter par le fait que les systèmes d'ordres supérieurs tirent meilleur profit de la densité angulaire des haut-parleurs.

ITD basse fréquence (retard de phase): portée de la prédiction d'après $\vec{V}$

La figure 4.2 montre, pour des décodages basiques, que le retard de phase est correctement restitué sur une bande basse-fréquence qui s'élargit en fonction de l'ordre. Il est intéressant de noter que la fréquence limite est plus élevée, à ordre égal, que celle calculée à partir de l'erreur de reconstruction des réponses binaurales, où le seuil de tolérance fixé à 20% ne semblait pourtant pas très sévère au regard d'autres études [NE99]. Cette observation a déjà été faite en 3.2.

En observant en outre les autres cas de décodage *max* r_E et *in-phase* (Figure 4.3), il est intéressant de constater que la loi de prédiction de l'ITD basse-fréquence d'après le vecteur vélocité$\vec{V}$ (1.80) est remarquablement bien vérifiée par les simulations. Nous la reportons ici:

$$\mathrm{ITD}_{LF}^{r_V}(\vec{u}_V) = r_V \,\mathrm{ITD}_{LF}^{\text{réf}}(\vec{u}_V), \tag{4.1}$$

avec $\vec{V} = r_V\,\vec{u}_V$ et où l'ITD de référence $\mathrm{ITD}_{LF}^{\text{réf}}(\vec{u}_V)$ est celui correspondant à l'effet d'une onde plane d'incidence $\vec{u}_V$. Cependant, la *largeur de la bande basse-fréquence* sur laquelle cette loi s'applique varie selon l'ordre du système, et aussi selon le type de décodage en jeu, ou plus généralement *selon la dissemblance entre les indices* r_V *et* r_E: à ordre égal, la limite fréquentielle la plus élevée apparaît pour les solutions "max r_E" (pour lesquelles $r_E = rV$), elle est sensiblement inférieure pour les solutions basiques ($r_E < r_V = 1$), et globalement plus basse pour les solutions "in-phase" (r_V très inférieur à r_E) (Figure 4.3).

L'indice r_V donne donc une information fiable de l'effet de latéralisation basse-fréquence qu'on peut attendre d'un système. Schématiquement et d'après le contenu basse-fréquence des images sonores, un indice $r_V < 1$ peut se traduire par un angle latéral maximal perçu de $\theta = \arccos r_V$ ou bien par un effet de hauteur (site $\delta = \arccos r_V$) par rotation "yaw" de la tête, ainsi qu'il est commenté en 1.5 et reporté à la suite d'expériences écoutes (en 4.1.3).

1. A défaut de mesure objective et synthétique de caractérisation du spectre, nous nous contentons de cette observation qualitative. Une mesure de "contraste" a été récemment employée dans [LJGW00].

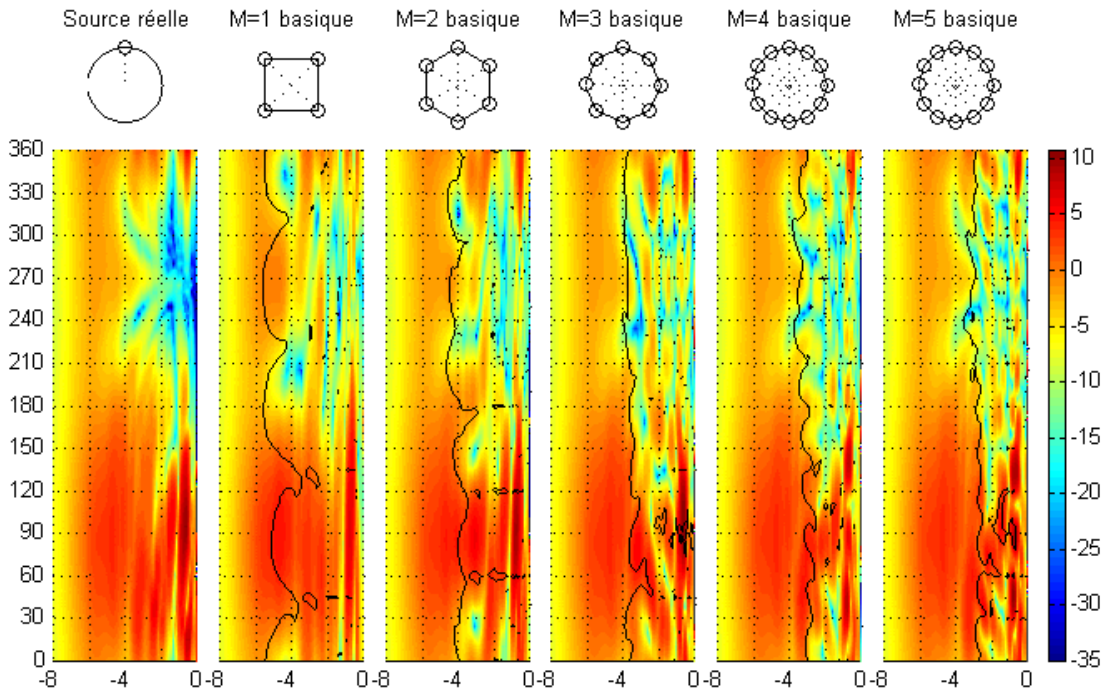

FIG. 4.1 – *Spectre d'énergie (en dB) de la réponse monaurale gauche en fonction de l'azimut (en ordonnée) de la source virtuelle. Le spectre est ici calculé par sommation en énergie des spectres correspondant à chaque contribution de haut-parleur (allure très similaire à celui obtenu par sommation en amplitude). Fréquences (en abscisse) exprimées en octave (échelle logarithmique en base 2): octave 0 à 20 kHz, octave -4 à 1250 Hz et octave -8 à* 78,125 *Hz. Les rendus ambisoniques (configuration et décodage spécifiés au-dessus) sont à comparer à la simulation binaurale d'une source unique (à gauche). Les courbes de niveau (trait plein) indiquent une erreur de* 20% *(-14dB) sur le spectre complexe.*

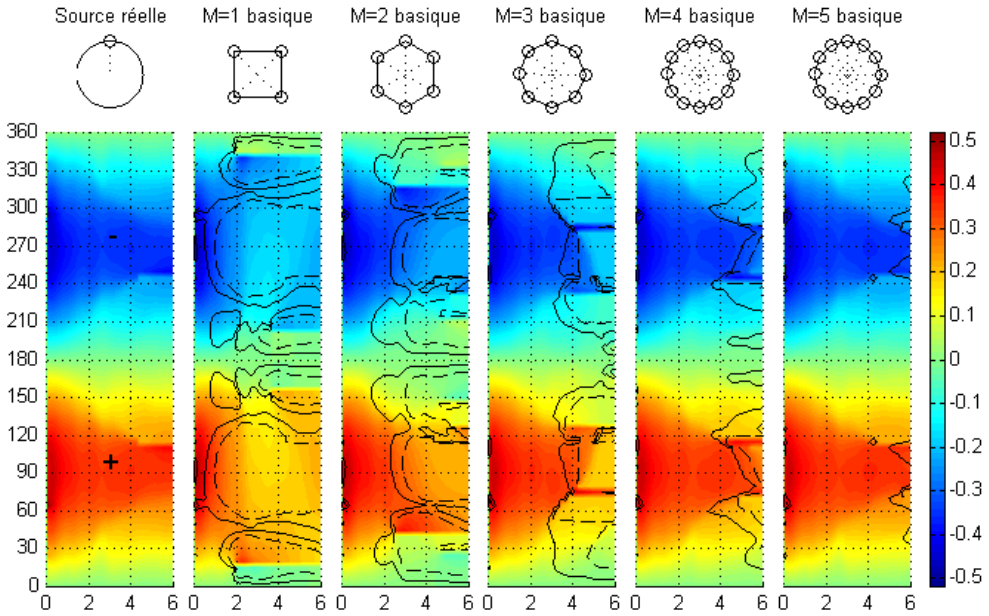

FIG. 4.2 – *Retard interaural de phase (ITD basse-fréquence) en fonction de la fréquence (exprimée en kHz) pour les mêmes situations que la figure 4.1. Les courbes de niveau indiquent des erreurs de* 0,02 *ms (trait plein) et* 0,05 *ms (tirets) par rapport à la restitution binaurale de référence (à gauche), soit respectivement* 1/25 *et* 1/10 *du retard maximal.*

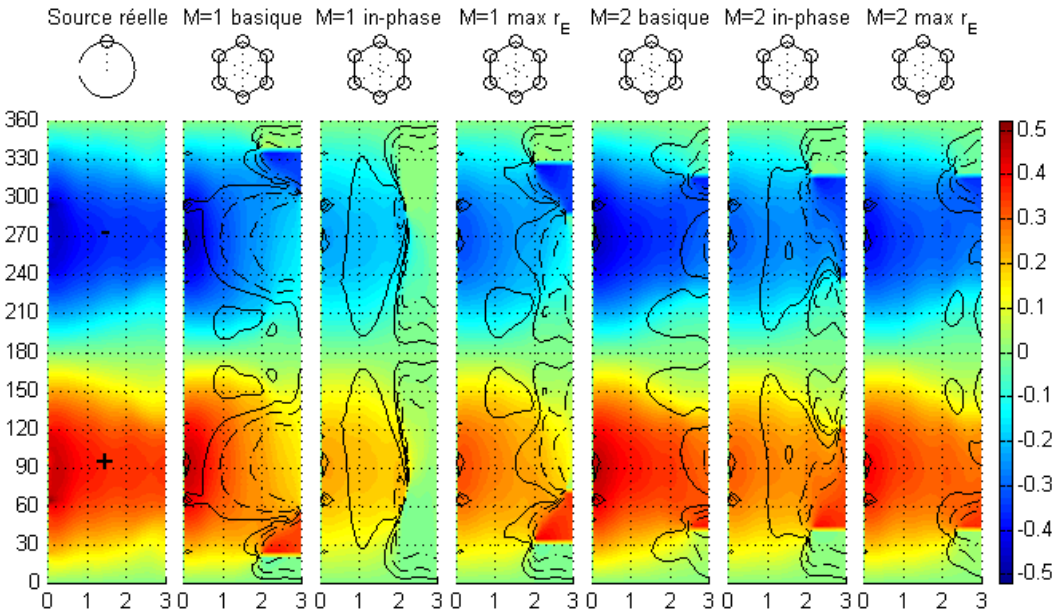

FIG. 4.3 – *ITD basse-fréquence (retard de phase en ms) estimé en fonction de la fréquence (en kHz) pour des restitutions ambisoniques "basiques", "in-phase" et "max r_E" d'ordres 1 et 2, avec pour référence une restitution binaurale directe (ITD_{ref} à gauche). Les courbes de niveau (tracés continu, tireté, tireté-pointillé) indiquent des écarts $|ITD/r_V - ITD_{ref}|$ de* 0.02*ms,* 0.05*ms et* 0.1*ms respectivement.*

ILD: évolution avec l'ordre *M* et les solutions de décodage

L'ILD, qui se manifeste habituellement au-delà de 2 ou 3 kHz, participe également à l'effet de latéralisation, bien qu'il soit reconnu [Bla83] [Beg94] que l'ITD prédomine en cas de conflit. Il est donc instructif d'observer à travers cet indice les répercussions des décodages selon leur type et l'ordre du système. L'ILD, reporté Figure 4.4 en fonction de la fréquence, est estimé par soustraction (en dB) des spectres d'énergie gauche et droit, eux-mêmes calculés par sommation en énergie des contributions élémentaires. Ce type de sommation – discutable car il signifierait que les sources contributives sont décorrélées – est choisi pour donner un aspect plus lisse à l'illustration. Pour traduire l'ILD réellement mesurable au niveau des oreilles, une sommation en amplitude aurait été plus juste. Quoiqu'il en soit, des tendances similaires sont observées dans un cas ou dans l'autre (voir la figure 8 de [DRP98], annexe B). On note un net renforcement de l'ILD, donc de la latéralisation, en passant de l'ordre 1 à l'ordre 2, et en passant du décodage basique au décodage *max* r_E. Pour ces quatre décodages, l'ILD est globalement plus manifeste pour l'incidence latérale $\theta = 90°$, contrairement au cas de référence (à gauche) et au décodage super-minimal (à droite). Rappelons en effet que pour une restitution "idéale" (source unique), l'ILD connaît un minimum local aux positions parfaitement latérales (azimut $\pm 90°$ et site nul), au contraire de l'ITD. Ce "défaut de masquage" est expliqué en 1.3.2.

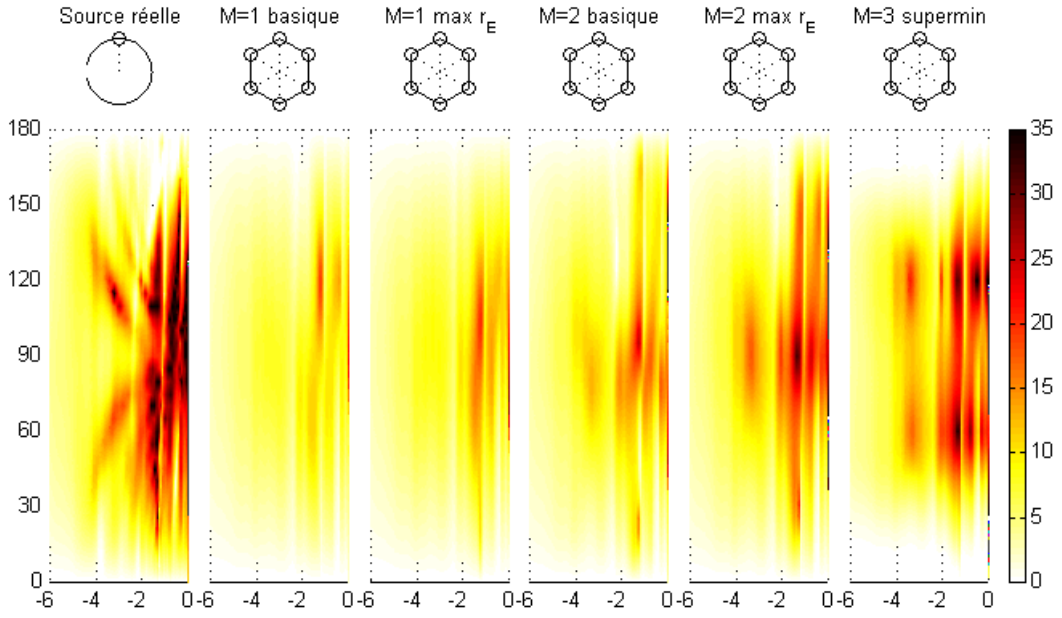

FIG. 4.4 – *ILD (échelle de gris, en dB) en fonction de la fréquence (abscisses: comme pour la figure 4.1 mais sans les deux octaves inférieures -8 à -6) et de l'azimut.*

Une évaluation plus globale de l'ILD, qui se prête mieux à une analyse comparée, se base sur le rapport des énergies totales – *i.e.* intégrées sur tout le spectre – des réponses binaurales[2] (Figure 4.5). Comme dans [DRP98] (Annexe B), mais à plus large échelle, on vérifie qu'à ordre égal, les solutions *max* r_E fournissent des valeurs plus proches de l'ILD de référence que les solutions basiques. A ordre égal, le décodage *in-phase* présente un ILD du même ordre que le décodage basique, et même légèrement supérieur. Cette légère différence, alors que les deux types de solutions ont le même indice de "concentration énergétique"

2. Ces réponses sont cette fois-ci obtenues par sommation en amplitude.

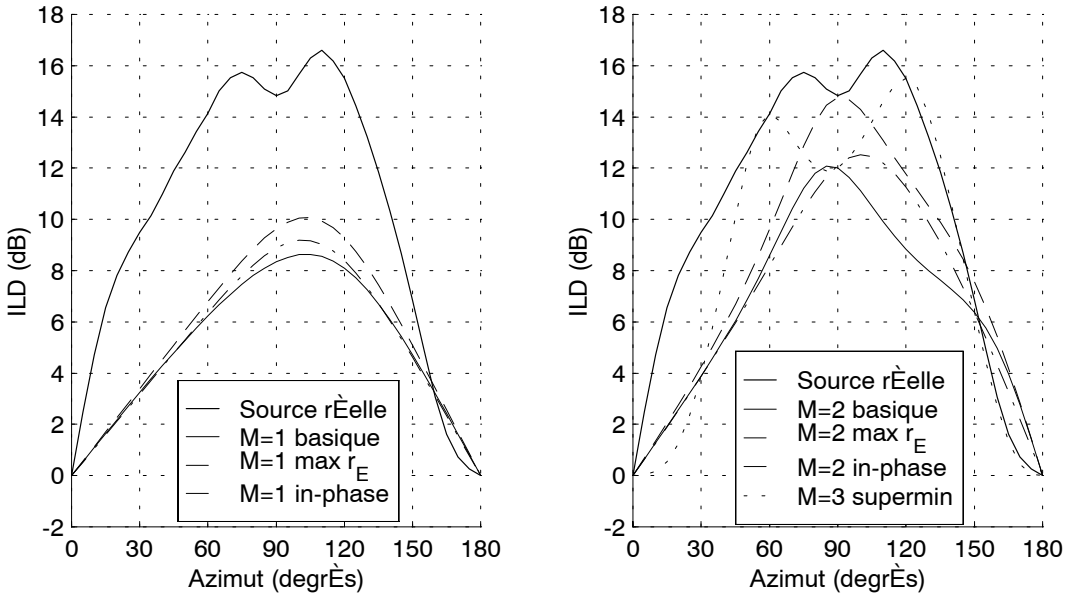

FIG. 4.5 – *Estimation globale de l'ILD (dB) en fonction de l'azimut pour la même configuration que la figure 4.4: décodages "basique" (−), "max r_E" (− −), "in-phase" (−·) et "super-minimal" (···).*

r_E (Table 3.10), peut s'expliquer par l'absence de lobe arrière dans la directivité équivalente associée à la solution *in-phase* (Figure 3.14). Pour les différents types de solution, le renforcement de l'ILD maximal en passant du premier au second ordre est de 3 à 5 dB environ. En considérant successivement les solutions *basic* ($M = 1$), *max* r_E ($M = 1$), *basic* ($M = 2$) et *max* r_E ($M = 2$), on constate une analogie frappante entre l'évolution de l'amplitude des courbes d'ILD et celle des indices r_E associés aux différentes restitutions, soit respectivement: 0,667; 0,707; 0,8; 0,866. Quant à la restitution "super-minimale" et toujours avec la configuration de la figure 4.4, elle offre un ILD idéal quand la source est placée sur un haut-parleur, mais moins bon que la restitution *max* r_E du second ordre pour une incidence purement latérale où aucun haut-parleur n'est présent.

Il apparaît plus loin que l'allure des courbes, à décodage identique, peut varier de façon très nette en fonction du choix du placement des haut-parleurs, même si la géométrie reste régulière. Le cas présent est par exemple très particulier pour le décodage *max* r_E d'ordre 2: pour une image parfaitement latérale $\theta = 90°$, seuls deux haut-parleurs latéraux (en $60°$ et $120°$) participent. Cependant, l'évolution des courbes en fonction de l'ordre du système et des solutions de décodage reste similaire pour les différentes configurations, pourvu qu'elles soient régulières.

ITD haute-fréquence

Le retard de groupe estimé comme dérivée de la phase en fonction de la fréquence (équation 1.54) est assez difficilement exploitable [DRP98] (Annexe B). Une estimation globale de l'ITD haute-fréquence est plus souhaitable. Plusieurs méthodes ont été recensées à cet effet en 1.4. Parmi elles, la méthode basée sur les segments quasi-linéaires d'excès de phase est proposée et utilisée dans [JLP99] pour la comparaison entre des restitutions ambisoniques d'ordres 1 et 2 et d'autres techniques de pan-pot (VBAP et VBIP). Cette

méthode, ainsi que la méthode classique basée sur le maximum de l'IACC (équations 1.49 et 1.50), ont l'inconvénient de fournir des estimations discontinues en fonction de l'azimut, particulièrement dans le cas présent où l'on a affaire à des réponses impulsionnelles "composites". L'expérience auditive montre que de telles discontinuités ne sont pas perçues comme telles. Nous préférons donc faire appel à des méthodes (Cf 1.4.3) basées sur l'estimation d'une moyenne, accompagnée d'une variance (ou d'un écart type) qui donne une information très appréciable sur l'acuité de l'estimation et de façon plus indirecte, sur la tache de localisation.

Dans un premier temps, on choisit de se baser sur l'enveloppe d'énergie des réponses binaurales[3] (Cf 1.4), chacune d'entre elles étant préalablement filtrée (passe-haut, ici de fréquence de coupure 3 kHz). La corrélation interaurale des enveloppes d'énergie est présentée en fonction de l'azimut (Figure 4.6), et ses maxima (pics) pourraient servir à prédire l'ITD haute-fréquence perçu, fournissant cependant des estimations discontinues et peu fiables. On préfère utiliser comme mesure globale de l'ITD la différence δ_τ des époques moyennes des enveloppes d'énergie (1.59), et l'écart-type σ_τ associé (1.61). On peut constater (Figure 4.6) que ces mesures reflètent de façon continue l'évolution des lieux des maxima de la fonction d'inter-corrélation, tandis que l'écart-type donne une information d'incertitude sur la mesure, fourchette délimitée par les deux courbes $\delta_\tau + \sigma_\tau$ et $\delta_\tau - \sigma_\tau$. Sur les figures 4.6 et 4.9, elles épousent approximativement les courbes de niveau de la fonction d'inter-corrélation à la valeur 0,5. L'écart-type, dû à l'étalement temporel des enveloppes d'énergie, donne aussi une indication du flou attaché à la localisation. Pour une disposition identique des haut-parleurs, la figure 4.6 montre clairement que l'ITD estimé suit la même tendance que l'ILD (Figures 4.4 et 4.5) en fonction de l'ordre M et du type de décodage. Le minimum local observé vers l'azimut 90° pour le décodage super-minimal signifierait une régression de l'effet de latéralisation.

A titre de comparaison, la figure 4.7 présente les estimations de l'ITD par d'autres méthodes évoquées en 1.4.3, avec la même disposition hexagonale que la figure 4.6 et pour un cas de restitution choisi arbitrairement: le décodage basique d'ordre 2. Les différences d'allure des courbes et les écarts-type – non-nuls mêmes dans le cas d'une source réelle – rappellent que l'estimation univoque d'*un* ITD global est sans-doute une quête vaine. La méthode par modélisation gaussienne des enveloppes d'énergie (GMD), déjà présentée plus haut (Figure 4.6), fournit des courbes d'ITD globalement plus amples et plus continues que les autres méthodes, ainsi qu'un plus grand écart-type, qu'il s'agisse du rendu ambisonique ou du cas de référence (source réelle ou onde plane). L'estimation par intégration pondérée du retard de groupe interaural (IRGD) s'accompagne quant à elle du plus faible écart-type, qui devient nul lorsque la source se trouve dans le plan médian de l'auditeur (situation symétrique). De manière générale, le fait que l'écart-type pour le rendu ambisonique est supérieur à celui de référence, est à relier à l'élargissement de la tache de localisation sur le plan perceptif.

Comme il semble difficile de se fier à une mesure absolue de l'ITD haute-fréquence, on s'intéressera plutôt aux relations d'ordre entre les différents type de restitution, au vu de l'ITD estimé. De nombreuses simulations montrent que ces relations d'ordre sont globalement préservées d'une méthode d'estimation à l'autre, en dépit de leurs différences. Afin de faciliter l'observation, c'est la première méthode évoquée (GMD) qui sera utilisée, dont la régularité des courbes correspond d'ailleurs mieux à l'effet perçu lors d'un balayage panoramique de la source virtuelle.

Confrontation aux prédictions d'après $\vec{E}$ et aux formules empiriques

Dans l'étude 1.5, nous avons cherché à établir des formules de prédiction de l'ITD haute-fréquence et/ou son effet de latéralisation, à partir du vecteur énergie $\vec{E}$ ou d'informations plus précises sur la distribution spatiale de l'énergie. L'étude présente est l'occasion de les illustrer sur des exemples de restitution ambisonique

3. Là aussi, précisons: réponses binaurales obtenues par sommation en amplitude.

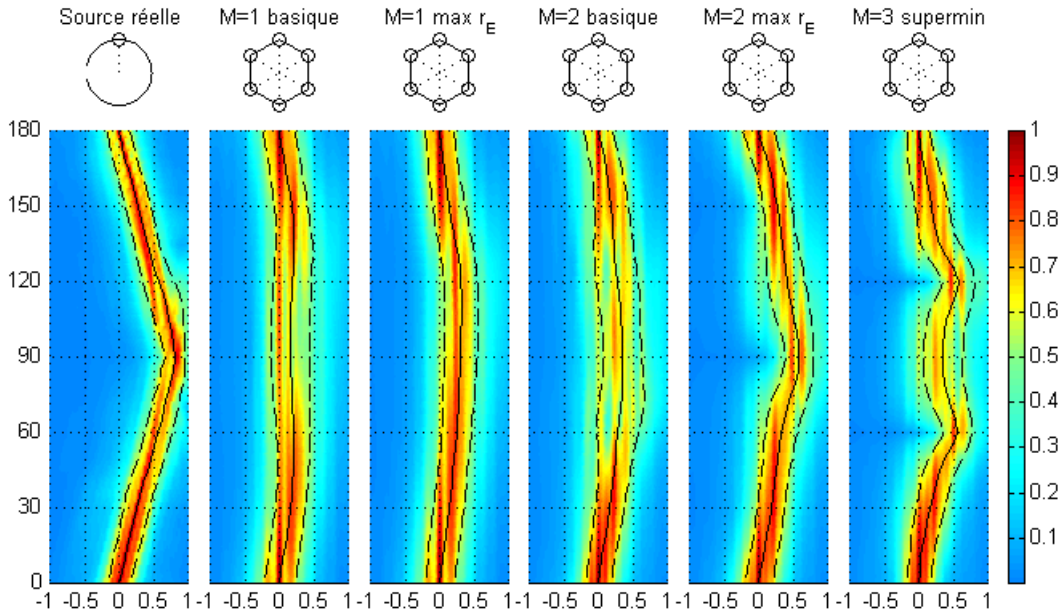

FIG. 4.6 – *ITD hautes-fréquences (en ms) évalué par l'inter-corrélation normalisée des enveloppes d'énergie (échelle de couleurs ou de gris) et la différence (trait plein) de leurs époques moyennes (écart type: tirets). Décodages basiques et "max r_E" (ordre 1 et 2) et décodage "super-minimal" (ordre 3) pour une même configuration hexagonale comprenant un haut-parleur frontal.*

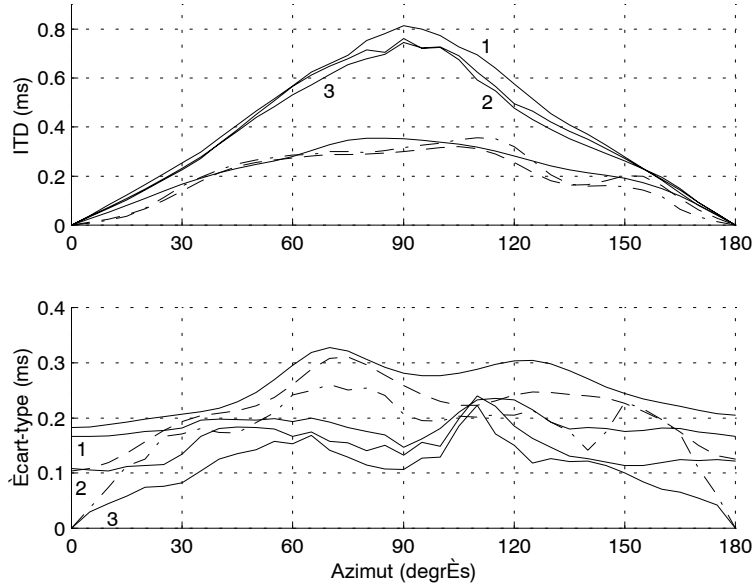

FIG. 4.7 – *ITD haute-fréquence (en haut) estimé par différentes méthodes, et accompagné d'une mesure d'incertitude (écart-type en bas). En traits fins, estimations pour une restitution basique du second ordre sur le même dispositif hexagonal que la figure 4.6 par les méthodes suivantes (se reporter à la figure 1.14): GMD (–), GMIACC (- -), IRGD (-.), avec un filtrage passe-haut préalable (> 3 kHz) des réponses binaurales. En traits gras et indexées respectivement de 1 à 3, les estimations par les mêmes méthodes appliquées au cas d'une source réelle unique (cas de référence).*

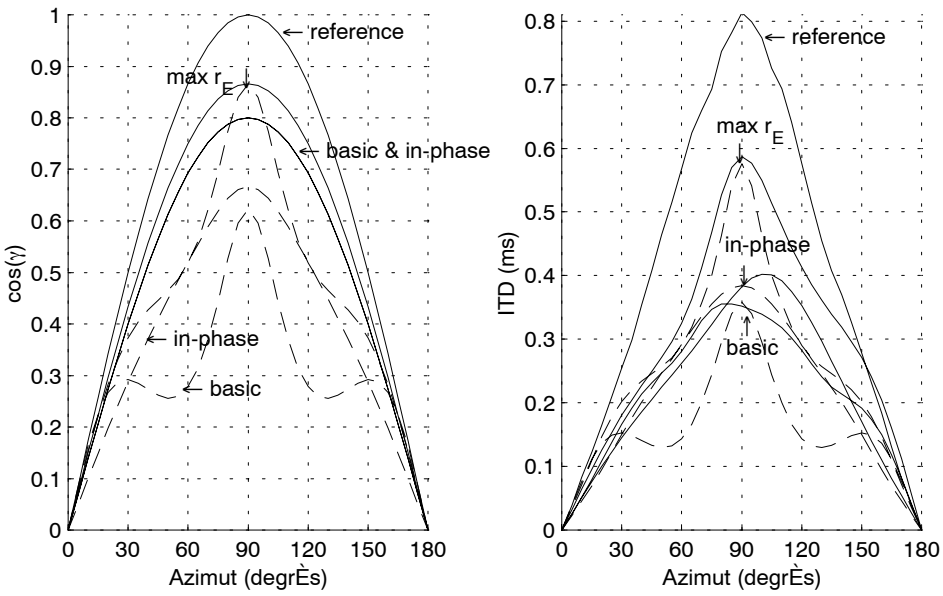

FIG. 4.8 – *Effet de latéralisation dans le cas de restitutions ambisoniques d'ordre 2. A gauche, le degré de latéralisation prédit est indiqué par le cosinus de l'angle d'ouverture* γ *du cône d'ambiguïté: modèle de tête acoustiquement transparente (trait continu) et modèle de masquage simplifié (tirets). A droite: ITD estimé (méthode GMD) à partir des simulations binaurales (trait continu) et ITD équivalent à la deuxième prédiction dans la figure de gauche (tirets).*

et de les confronter aux mesures issues des simulations binaurales.

Seul un modèle de tête acoustiquement transparente permet d'obtenir une loi de prédiction haute-fréquence (1.87) d'après $\vec{E}$ aussi simple que la loi de prédiction basse-fréquence d'après le vecteur vélocité$\vec{V}$ (équations 1.79, 1.79 ou 4.1): d'après ce modèle, la réduction de l'effet de latéralisation par rapport au cas d'une onde plane serait directement traduite par le facteur $r_E \leq 1$. En prenant en compte l'effet de diffraction par la tête, comme le fait la formule empirique (1.89) basée sur un modèle simplifié de masquage, l'effet de latéralisation prédit est diminué dès lors que les sources contributives ne sont pas placées sur le même cône d'ambiguïté. C'est ce que confirme la figure 4.8 (gauche) pour différentes restitutions ambisoniques du second ordre. Il faut remarquer que la latéralisation prédite pour le décodage *in-phase* est alors le plus souvent supérieure à celle prédite pour le décodage *basique*, malgré un indice r_E identique.

La figure 4.8 (droite) montre l'ITD qui correspondrait au cône d'ambiguïté donné par la deuxième prédiction – c'est-à-dire l'ITD obtenu par projection de l'angle prédit $(\pi/2 - \gamma)$ sur la courbe de l'ITD de référence – et le met en rapport avec l'ITD estimé directement à partir des simulations binaurales. Des différences sont observées entre la prédiction et l'estimation, surtout pour le décodage basique, mais l'évolution des valeurs maximales en fonction du décodage est assez bien respectée.

Effet de la disposition des haut-parleurs

La figure 4.9, comparée à la figure 4.6, montre que la disposition des haut-parleurs – ou bien encore l'orientation de la tête – a une franche influence sur les indices objectifs de localisation, ainsi que sur leur effet supposé sur la latéralisation.

La présence de haut-parleurs sur l'axe interaural (Figure 4.9) rend la restitution "super-minimale" très

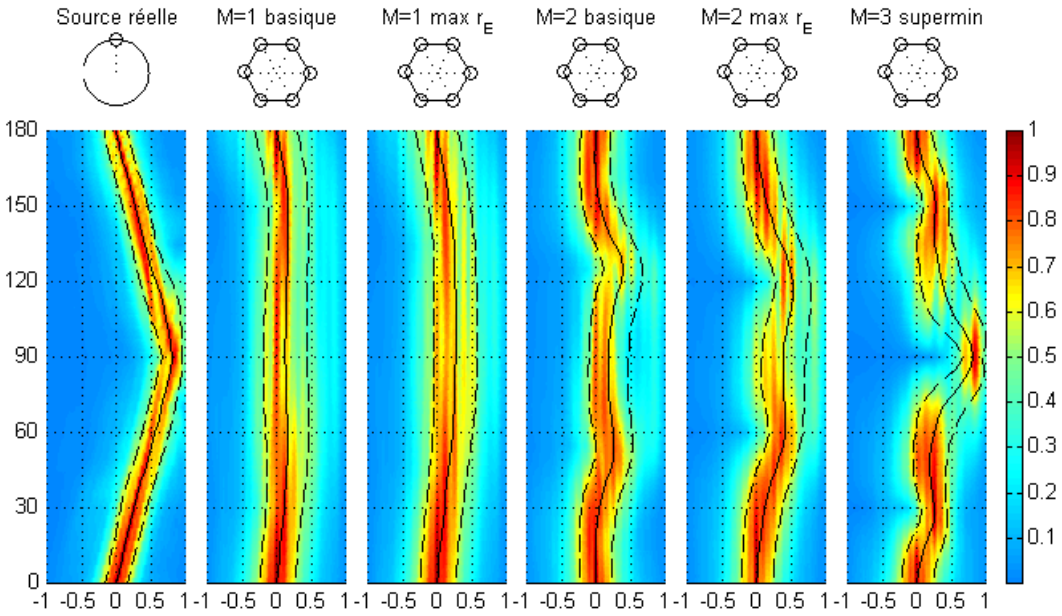

FIG. 4.9 – *ITD hautes-fréquences (abscisses: ms) évalué par l'inter-corrélation des enveloppes d'énergie (niveaux de couleur ou de gris) et la différence (trait plein) de leurs époques moyennes (écart type: tirets). Décodages basiques et "max r_E" (ordre 1 et 2) et décodage "super-minimal" (ordre 3) pour une même configuration hexagonale (avec deux haut-parleurs latéraux).*

avantageuse pour les images purement latérales, au contraire de la configuration *(F)* – celle de la figure 4.6 – avec laquelle l'ITD régresse anormalement pour les incidences latérales. La tendance inverse s'affiche avec les décodages *basique* et *max* r_E d'ordre 2, au regard de l'ITD comme de l'ILD (voir également Figure 4.10). L'explication est simple: dans la configuration *(L)* – celle de la figure 4.9 –, l'effet de latéralisation quand la source est en $\theta = 90^\circ$ est atténué par la participation du haut-parleur diamétralement opposé (voir les diagrammes de directivité équivalents Figure 3.14). Ce n'est plus le cas avec le décodage *in-phase* au vu de l'ITD, même si la configuration *(L)* reste défavorable à l'ILD. Ajoutons que les courbes d'ITD et d'ILD résultant d'une restitution *in-phase* gardent une allure régulière et monotone quelle que soit la configuration.

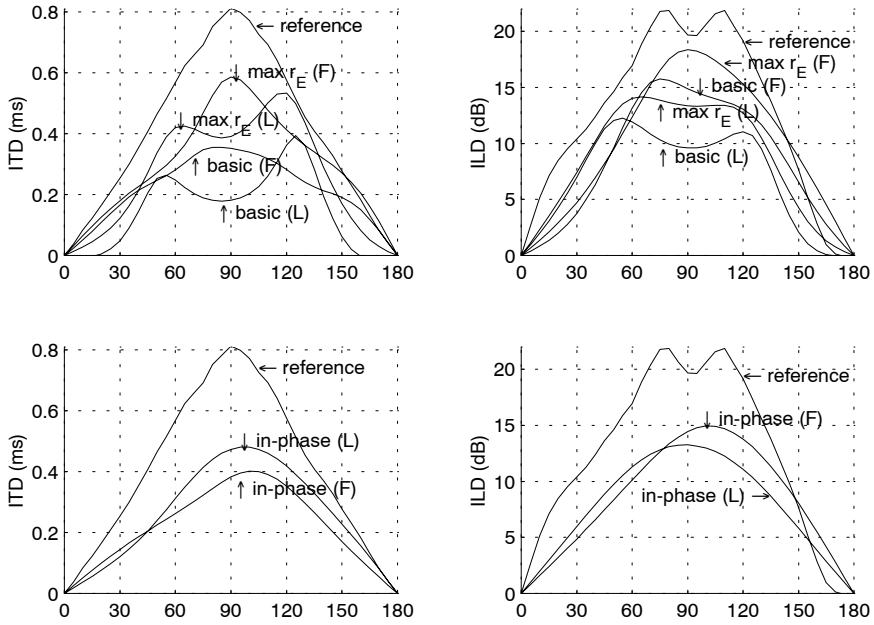

FIG. 4.10 – *Estimations de l'ITD et de l'ILD pour différentes restitutions ambisoniques du second ordre sur deux configurations hexagonales: l'une (F) comprenant un haut-parleur frontal (Figure 4.6) et l'autre (L) comprenant des haut-parleurs parfaitement latéraux (Figure 4.9).*

Au regard des estimations de l'ITD et de l'ILD et à supposer que la tête ait toujours la même orientation frontale, la configuration hexagonale *(F)* offre un effet de latéralisation "optimal" pour le décodage *max* r_E d'ordre 2. Pour un ordre M quelconque, la configuration "optimale" serait le polygone régulier à $N = 2M+2$ haut-parleurs, tel que l'axe interaural soit l'axe médian de deux haut-parleurs adjacents (Figure 4.11, ligne du haut). La latéralisation maximale est alors caractérisée par un cône d'ambiguïté d'angle d'ouverture $\pi/N = \arccos r_E$, sur lequel sont placés les deux seuls haut-parleurs en fonctionnement quand $\theta = 90^\circ$. Avec un décodage *in-phase* d'ordre M ou *super-minimal* d'ordre $M+1$, c'est la disposition en quinconce qui est au contraire la plus favorable. Ces propriétés sont à prendre en compte lors du choix d'un dispositif virtuel pour simulation binaurale.

On ne saurait manquer de formuler quelques réserves, cependant, quant à l'interprétation de tous ces résultats: les différences ou les fluctuations qui sautent aux yeux sur les courbes ne se reportent pas forcément de façon aussi nette sur le plan perceptif lors d'une expérience d'écoute. D'une part, toutes les informations

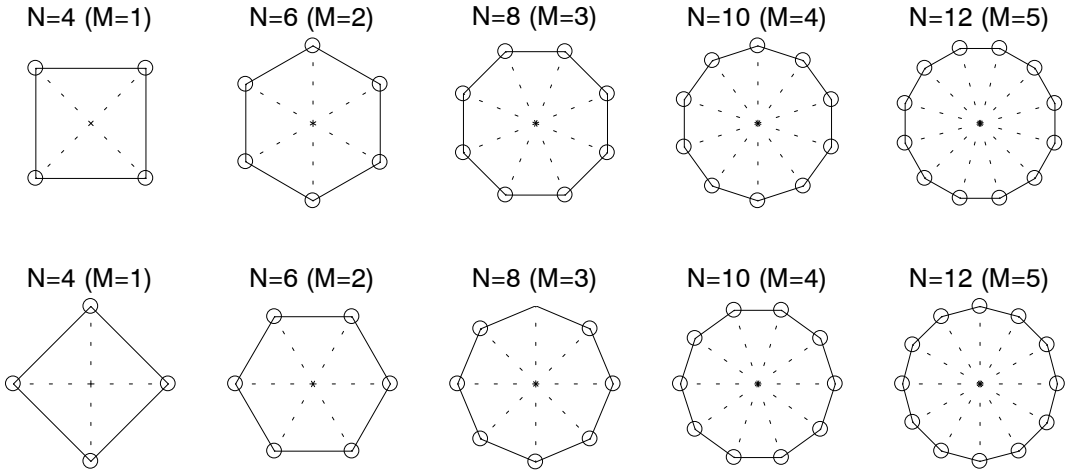

FIG. 4.11 – *En haut: configurations "optimales" pour le décodage* max r_E*, au sens d'une latéralisation supposée maximale au vu des mesures objectives. Les haut-parleurs sont représentés par des ronds et l'axe* $\vec{x}$ *(direction frontale) est dirigé vers le haut. En bas: configurations en quinconce, supposées "optimales" pour les décodages* in-phase *et* super-minimal *au sens des mêmes critères.*

sonores sont loin d'être caractérisées ici. D'autre part, la donnée de l'écart-type, qui traduit en quelque sorte le "flou" de l'estimation de l'ITD, permet de comprendre que certaines différences soient gommées au niveau perceptif.

Evolution en fonction de l'ordre et du décodage

Pour clore cette évaluation objective de la restitution ambisonique en position d'écoute idéale, la figure 4.12 trace l'évolution de l'ITD haute-fréquence et de l'ILD en fonction de l'ordre M du système[4], et pour chacun des trois types de décodage: *basique*, *max* r_E et *in-phase*. L'accroissement de l'amplitude des courbes en fonction de l'ordre est manifeste, quoique des fluctuations assez marquées apparaissent avec le décodage basique à partir de l'ordre 3, ce qu'il faut rapprocher de la présence des lobes secondaires sur les figures de directivités équivalentes (Figure 3.14). A ordre égal et donc à indice r_E égal, les courbes *in-phase* sont d'amplitude semblable ou supérieure à celle des courbes *basic*. Celle des courbes *max* r_E leur est toujours assez largement supérieure. On note au passage, pour les ordres élevés et avec les décodages *in-phase* et *max* r_E, que l'ILD reconstitué dépasse l'ILD de référence.

Pour résumer, l'amélioration des indices de latéralisation haute-fréquence apparaît très clairement lors du passage aux ordres supérieurs, mais aussi à ordre égal, lors du passage du décodage *basique* ou *in-phase* au décodage *max* r_E. Il semble que l'évolution de l'ITD puisse être partiellement corrélée à celle de l'indice r_E. Cela étant, les différences entre le cas *basic* et le cas *in-phase* montrent qu'à indice r_E égal – soit le même "taux de concentration énergétique" – les indices haute-fréquence reconstitués (ITD et ILD) ne sont pas indifférents à la *"qualité" de la distribution spatiale des sources d'énergie*: le décodage *in-phase*, qui propose une répartition plus progressive de l'énergie vers la source virtuelle, induit globalement une meilleure

4. La simulation pour l'ordre $M = 3$ et utilisant $N = 8$ haut-parleurs (Figure 4.11, ligne du haut) est légèrement erronée, puisque nous ne disposons pas des HRTF correspondant exactement aux azimuts des haut-parleurs.

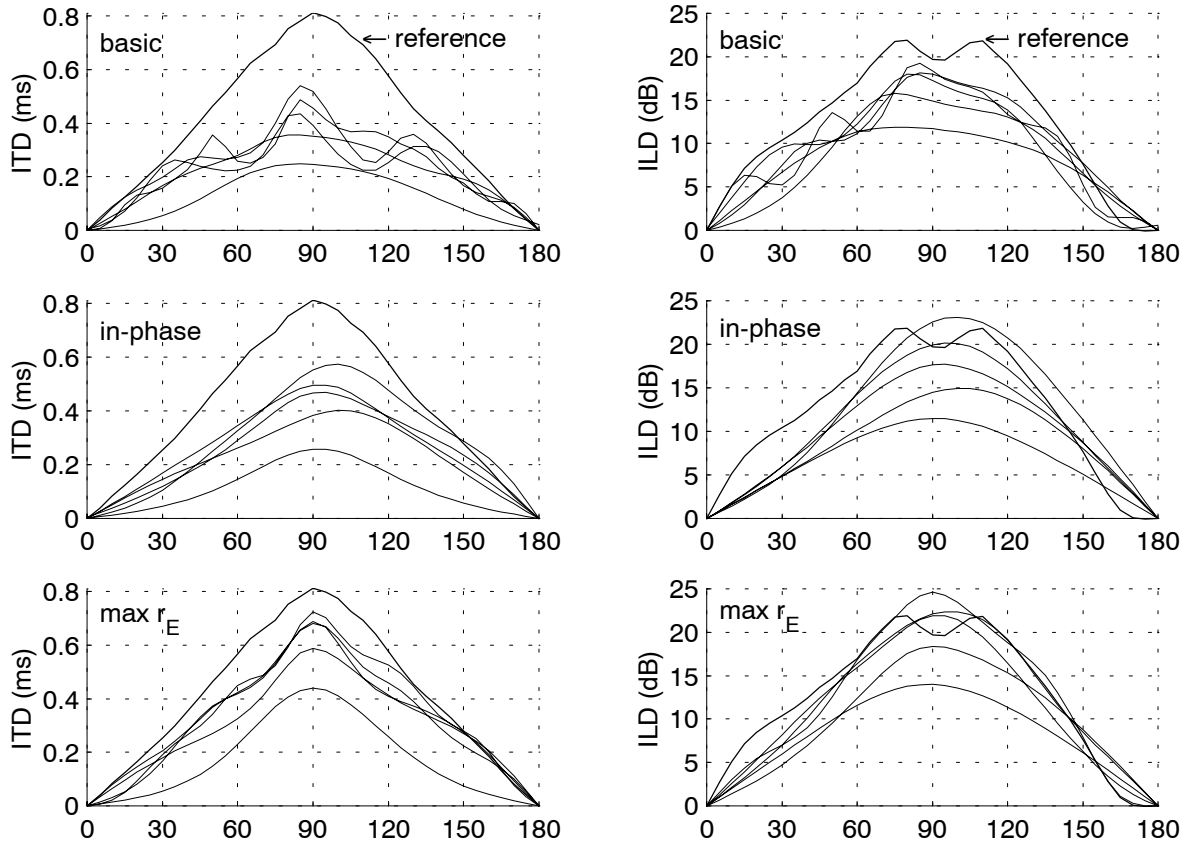

FIG. 4.12 – *Estimations de l'ITD et de l'ILD pour des restitutions ambisoniques d'ordres 1 à 5, les courbes étant ordonnées de bas en haut (non numérotées faute de place, l'ordonnancement étant respecté au moins au voisinage de* 90° *lorsqu'il y a ambiguïté). Pour chaque ordre, la configuration choisie est celle qui est "optimale" pour le décodage* max r_E *(Figure 4.11, ligne du haut).*

latéralisation que le décodage *basique*.

4.1.3 Confrontation aux expériences d'écoute

Les solutions de décodage du premier et du second ordre ont pu être testées au cours de quelques expériences d'écoute, réalisées d'une part "grandeur nature" sur haut-parleurs, et d'autre part au casque, avec le recours des techniques binaurales (méthode des haut-parleurs virtuels). Le logiciel développé à cet effet est présenté au chapitre 5. Bien que de caractère informel, ces expériences ont permis de corroborer ou bien de modérer certains des résultats objectifs présentés plus haut, mais aussi de mettre le doigt sur des aspects complémentaires.

Description sommaire du dispositif expérimental

Dans l'interface utilisée, les diverses solutions de décodage peuvent être appliquées en pleine-bande ou bien combinées en deux sous-bandes, la fréquence de transition étant elle-même ajustable. Les gains correcteurs g_m caractérisant chaque solution de décodage peuvent par ailleurs être contrôlés directement par l'utilisateur (Figure 6 de [DRP98], annexe B). L'écoute au casque offre la possibilité d'évaluer le rendu ambisonique sur haut-parleurs virtuels avec comme référence l'effet d'une source virtuelle unique obtenu par simulation binaurale directe. Lors d'une écoute sur haut-parleurs, le rendu ambisonique peut-être éventuellement comparé avec un pan-pot par paire de haut-parleurs, le VBAP horizontal [Pul97] présenté en 2.3.2 et en 5.2.2. L'utilisateur peut effectuer une commutation instantanée entre deux solutions de décodage, entre l'ordre 1 et l'ordre 2, ou entre *ambisonic* et un autre mode de restitution.

Le matériel sonore testé est de nature variée: parole, musique, bruitage, bruit (blanc ou bande étroite). Pour ces tests, une seule source virtuelle est simulée à la fois et sans effet de salle ajouté. Sa position (azimut et distance) peut être contrôlée par l'utilisateur ou bien se voire assignée une trajectoire circulaire autour de l'auditeur. Enfin, les configurations de haut-parleurs testées "en grandeur nature" sont surtout les dispositions carrée et hexagonales. L'écoute au casque a aussi permis de tester des dispositifs avec un plus grand nombre de haut-parleurs virtuels (jusqu'à 72).

Expériences et résultats

Que ce soit au casque ou sur haut-parleurs, une franche amélioration de la *précision* et – le cas échéant – de la *latéralisation* des images est remarquée dans les cas suivants:

- Au premier ordre: entre l'application pleine-bande de la solution *basique* et l'application des décodages *basique* et *max* r_E dans leur sous-bande respective (basse- et haute-fréquence), c'est-à-dire le décodage optimisé selon Gerzon.
- Entre une restitution d'ordre 1, même optimisée, et une restitution d'ordre 2 (par exemple avec un décodage *basique* pleine-bande).

Lors de la restitution *sur haut-parleurs* d'une source à trajectoire circulaire, la qualité d'homogénéité et de continuité du rendu ambisonique est remarquable à l'ordre 1 comme à l'ordre 2, alors qu'avec le VBAP la qualité de l'image est alternativement floue et précise, au passage des haut-parleurs (effet "pâquerette").

La suite décrit les expériences d'écoute *au casque*, qui permettent une appréciation plus affinée des différents types de restitution ambisonique.

Il est tout d'abord intéressant de vérifier auditivement la reconstruction des informations binaurales: en soustrayant à la simulation ambisonique (décodage basique) la simulation binaurale directe de la source

virtuelle, un effet coupe-bas très net est observé, plus large avec l'ordre 2 qu'avec l'ordre 1. Les fréquences en-deçà desquelles le contenu sonore basse-fréquence disparaît sont de l'ordre de 700 Hz et 1200 Hz[5].

Quelques expériences[6] simples permettent de constater la bonne corrélation entre les indices r_V et r_E et l'effet de latéralisation perçu. Ayant placé en une position latérale fixe la source sonore virtuelle – un bruit basse-fréquence (respectivement haute-fréquence) – celle-ci paraît se déplacer vers le plan médian lorsque l'on passe d'une simulation binaurale directe à une restitution ambisonique avec un $r_V < 1$ (resp. $r_E < 1$) ou lorsque l'on abaisse encore l'indice r_V (resp. r_E) en changeant de décodage. L'expérience est également probante avec des extraits sonores plus ordinaires (musique, parole) et un décodage en deux sous-bandes, la commutation entre solutions de décodage se faisant sur la bande haute-fréquence pendant qu'une solution basique est appliquée en basse-fréquence.

Force est de reconnaître que les différences perçues entre les différentes solutions de décodage sont beaucoup moins sensibles au sein de l'ordre 2 que de l'ordre 1. Partant d'une restitution basique d'ordre 2 et avec une écoute attentive, la commutation vers le décodage *max* r_E (d'ordre 2) ne produit qu'un très léger déplacement latéral. Le changement est plutôt perçu en terme de coloration, voire d'impression spatiale[7]. Pour les quelques sujets qui y ont été sensibles, il semble que la préférence aille au décodage *basique* plutôt qu'aux décodages *in-phase* et *max* r_E, en raison d'un caractère diffus plus plaisant, d'une sonorité moins sourde, plus "ouverte" et plus proche de la référence (simulation binaurale directe).

Ces observations invitent donc à relativiser les résultats de l'évaluation objective (section 4.1.2) qui semblaient recommander à tout ordre la solution *max* r_E pour le décodage haute-fréquence, les indices objectifs (ITD et ILD) promettant notamment une amélioration de la latéralisation aussi sensible pour l'ordre 2 que pour l'ordre 1. Il faut croire que les mesures effectuées – principalement focalisées sur ces indices de latéralisation – sont loin de résumer la réalité perceptive. Même la mesure d'incertitude attachée à l'ITD haute-fréquence ne semble pas suffire à interpréter le fait que deux expériences de restitution puissent être presque confondues (ordre 2) quand d'autres présentent des qualités nettement distinctes (ordre 1). Elle n'offre finalement qu'une évocation grossière de la qualité (largeur) de la tache de localisation[8], qui découle d'une analyse beaucoup plus fine des informations binaurales, ce que nous n'avons pas les moyens de quantifier objectivement. Enfin, au-delà de la seule qualité de localisation, la préservation de la coloration ou de la "sonorité" – sensible au casque – apparaît comme un autre critère de choix de décodage. L'expérience suivante confirme le devoir de prudence dans l'interprétation des indices de latéralisation.

Grâce à la simulation d'une source à trajectoire circulaire, il est possible de confronter l'effet de balayage panoramique perçu aux courbes de la figure 4.10, en choisissant alternativement comme dispositif virtuel l'une des deux configurations hexagonales. Au cours des différentes écoutes, il n'a pas été fait mention d'une régression anormale de la latéralisation (aux alentours de $\pm 90°$), comme l'aurait fait supposer la figure 4.10 pour la configuration (L) et les décodages *basique* et *max* r_E du second ordre. Aucune différence flagrante n'a été notée entre les deux choix de configuration, l'interface utilisée obligeant *il est vrai* à interrompre la simulation pour changer de configuration. Pour se rendre compte plus sûrement des conséquences du choix du dispositif sur les images sonores et estimer la pertinence des courbes de la figure 4.10, il faudrait pouvoir commuter instantanément entre les deux configurations (avec une source virtuelle fixe), et observer un éventuel déplacement latéral ou un changement de coloration.

5. Vérifié sur des bruits à bande limitée.

6. Expériences mettant en jeu le plus souvent une configuration virtuelle hexagonale, l'une ou l'autre des deux présentées plus haut.

7. Il s'agit pourtant d'une simulation sans effet de salle ajouté!

8. On le constate sur le cas d'une source seule (cas de référence), pour lequel l'écart-type associé à l'ITD est déjà relativement important.

Le choix du nombre N de haut-parleurs virtuels utilisés a quant à lui des conséquences beaucoup plus perceptibles. On note un effet de coloration de plus en plus marqué à mesure que N augmente au-delà du nombre $N = 2M + 2$ juste suffisant pour un décodage régulier. Lorsque le même critère de normalisation (page 159) est appliqué sur toute la bande de fréquence, la multiplication des contributions (haut-parleurs) se traduit par un effet passe-bas, à cause du fait qu'elles s'additionnent en amplitude en basse-fréquence et globalement en énergie en haute-fréquence. Lorsqu'une normalisation différenciée est appliquée (amplitude en BF, énergie en HF) afin de rétablir l'équilibre spectral, une coloration de plus en plus désagréable apparaît[9], assimilable à une sorte d'effet de peigne. Cet effet de peigne est déjà visible sur les spectres (versions "sommation en amplitude") de la figure 7 de [DRP98] (Annexe B): il est plus marqué pour la restitution d'ordre 1 que pour celle d'ordre 2, pour une même configuration à six haut-parleurs . Dans ces conditions d'écoute en position centrée et fixe, il n'est donc pas recommandé de trop augmenter le nombre de haut-parleurs au-delà du minimum requis $N = 2M + 2$.

4.1.4 Conclusions

Résumé

Les conclusions de cette étude touchent plusieurs aspects. Tout d'abord, on a pu vérifier objectivement l'effet des ordres supérieurs et/ou des diverses solutions de décodage sur les indices de localisation, et par la même occasion valider les relations de prédiction entre des grandeurs $\vec{V}$ et $\vec{E}$ et la qualité de la latéralisation. Il aurait été intéressant de compléter ces mesures par une estimation de l'effet des rotations de la tête sur les indices de localisation.

La confrontation aux expériences d'écoute informelles a permis de confirmer l'apport – en termes de précision d'image et de latéralisation – de l'ordre 2 sur l'ordre 1 ainsi que du décodage *max* r_E en haute-fréquence, surtout à l'ordre 1. Elle a aussi mis en lumière d'autres aspects qui ont échappé à nos prédictions et mesures objectives: notamment les problèmes de coloration du signal perçu (altération de la sonorité subjective de la source), et les limites de sensibilité à l'amélioration objective des indices de latéralisation (ITD haute-fréquence et ILD). En soulignant les limites des mesures objectives réalisées, cette confrontation incite à modérer l'extrapolation qu'on voudrait en faire sur le plan subjectif. Cela étant, une validation subjective plus complète mériterait des tests d'écoute formels de plus grande envergure.

Avant de tirer des conclusions définitives de ces expériences, il faut rappeler les conditions d'écoute très particulières imposées par la restitution au casque, et des contraintes que cela implique: tête centrée et immobile par rapport au dispositif de restitution, restitution sans effet de salle, HRTF non-individualisées. C'est pourquoi la légère préférence pour le décodage basique constatée avec une restitution d'ordre 2, ne signifie pas que le décodage *max* r_E n'est pas recommandable (en haute-fréquence) pour une écoute sur haut-parleurs!

Recommandations sur le choix des décodages et des configurations

De manière générale, le décodage *in-phase* n'a pas lieu d'être appliqué dans les conditions d'écoute considérées. Le décodage *basique* est évidemment requis, au moins dans une bande basse-fréquence. Nous recommandons d'appliquer le décodage *max* r_E dans la bande haute-fréquence complémentaire dans les cas suivants:

– à l'ordre 1, qu'il s'agisse d'une écoute au casque ou sur haut-parleurs;

9. Cela laisse à penser que la différenciation binaire du critère de normalisation "amplitude/énergie" ne donne pas lieu à une égalisation spectrale optimale.

– aux ordres supérieurs pour une écoute sur haut-parleurs (à voir selon les préférences pour une écoute au casque), bien que ce soit moins “indispensable" qu'à l'ordre 1.

Le nombre de haut-parleurs virtuels recommandé pour la simulation binaurale est $N = 2M + 2$, avec de préférence les dispositions de la figure 4.11 (haut), bien que ce choix ne semble pas très critique. Pour une écoute sur haut-parleurs, les désagréables effets de coloration dus à un N élevé sont *a priori* moins sensibles qu'à l'écoute au casque, mais il semble préférable de garder un nombre raisonnable de haut-parleurs.

Vers une optimisation spécifique pour la présentation binaurale

La particularité de la présentation binaurale, par rapport à une restitution sur haut-parleurs, réside dans le contrôle des conditions de reconstruction du champ au niveau des oreilles: *la tête est fixe par rapport au dispositif virtuel* de haut-parleurs[10]. Donc le seul *mode de latéralisation* à considérer est *statique et non dynamique* (Cf 1.3.3 et 1.5). Même dans les applications avec *Head-Tracking*, tenant compte des rotations de la tête, c'est le champ ambisonique qui subira une transformation (rotation), et non l'orientation de la tête par rapport au dispositif virtuel. En somme, il faut considérer que les fonctions de transferts des composantes ambisoniques vers les signaux binauraux sont fixes.

L'idée de base de l'optimisation consiste donc à maximiser les facteurs de latéralisation statique (l'ITD et l'ILD), puisque c'est l'un des principaux défauts de la restitution ambisonique. Mais rappelons-nous aussi que l'ITD et l'ILD ne sont pas à eux-seuls garants de la qualité d'une “image sonore spatialisée": la préoccupation de leur optimisation ne doit pas occulter le problème de la détérioration des indices spectraux, et plus généralement de l'équilibre du spectre d'énergie et des phénomènes de coloration.

Partant de la méthode des haut-parleurs virtuels, un moyen simple permettrait d'améliorer l'ITD et l'ILD en haute-fréquence. Il s'agit d'anticiper le manque de latéralisation (en hautes-fréquences) en étirant latéralement la configuration de restitution par rapport à la configuration régulière de décodage. La figure 4.13 illustre l'exemple d'une distorsion elliptique des angles d'un hexagone régulier. Pour un coefficient de distorsion α, l'angle ϕ_i de chaque haut-parleur est changé en ϕ_i' tel que $\tan\phi_i' = \alpha \tan\phi_i$.

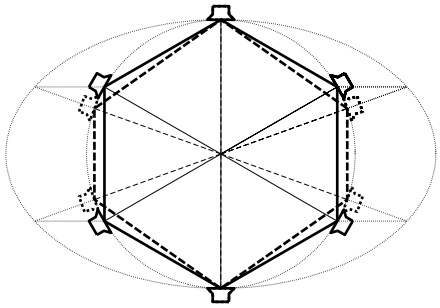

FIG. 4.13 – *Distorsion elliptique d'une configuration hexagonale régulière (traits continus) produisant une configuration étirée latéralement (tirets).*

En contrepartie, on peut s'attendre à ce que l'effet de coloration soit accru pour les images proches du plan médian, à cause de l'écartement des haut-parleurs (exagération du *cross-talk*). Par ailleurs, cette distorsion

10. ... si la méthode des haut-parleurs virtuels est bien l'approche adoptée.

ne devrait être appliquée que dans le domaine haute-fréquence (elle n'a pas lieu d'être appliquée dans le domaine de reconstruction valide), ce qui pose question de la transition fréquentielle du dispositif virtuel.

Avec ou sans cet artifice, des procédés d'égalisation spectrale adaptés mériteraient d'être développés, en remplacement de l'application dichotomique des critères de normalisation en amplitude et en énergie.

4.2 Effet des décodages en conditions non-idéales ou étendues

4.2.1 Critère de localisation basse-fréquence en position excentrée

Nous avons montré en 1.5.4 que dans l'hypothèse où les contributions s'ajoutent avec des relations de phase statistiquement aléatoires au point d'écoute considéré, la moyenne statistique $\langle\vec{V}\rangle_E$ du vecteur vélocité pondérée par l'énergie (1.76) s'identifie avec le vecteur énergie $\vec{E}$ (1.78). On en a déduit que $\vec{E}$ donne une prédiction de la localisation d'après l'ITD basse-fréquence (retard de phase), pour les positions d'écoute excentrées, en dehors de la zone de reconstruction contrôlée (Cf 3.1.3).

L'hypothèse d'une distribution uniforme des valeurs du vecteur des phases $\underline{\varphi}$ reste à être confrontée aux valeurs (1.74) qu'il prend effectivement sur une bande de fréquence $[f_1\ f_2]$ donnée, pour des points de mesure $\vec{r}$ s'éloignant du centre. Il est surtout intéressant d'en illustrer les conséquences sur $\langle\vec{V}\rangle_E$ (ou $\vec{\mathcal{V}}_f(\vec{r})$, équation 1.73), c'est-à-dire montrer comment sa valeur converge vers $\vec{E}$. On suppose ici que les différentes contributions restent assimilables à des *ondes planes sur toute la zone considérée*, ce qui suppose que le rayon R_{HP} du dispositif est assez grand. Pour le calcul de la moyenne pondérée, on fait ici le choix (arbitraire) d'une intégration suivant une échelle de fréquence linéaire:

$$\langle\vec{V}\rangle_E(\vec{r}) = \frac{\int_{f_1}^{f_2} E(\varphi(\vec{r},f))\vec{V}(\varphi(\vec{r},f))\,df}{\int_{f_1}^{f_2} E(\varphi(\vec{r},f))\,df} \tag{4.2}$$

La figure 4.14 montre les valeurs de $\langle\vec{V}\rangle_E$ pour quelques positions d'écoute au cours de la restitution ambisonique d'une source virtuelle, estimées d'après (4.2) en fixant $f_1 = 100$ Hz et $f_2 = 800$ Hz. La convergence

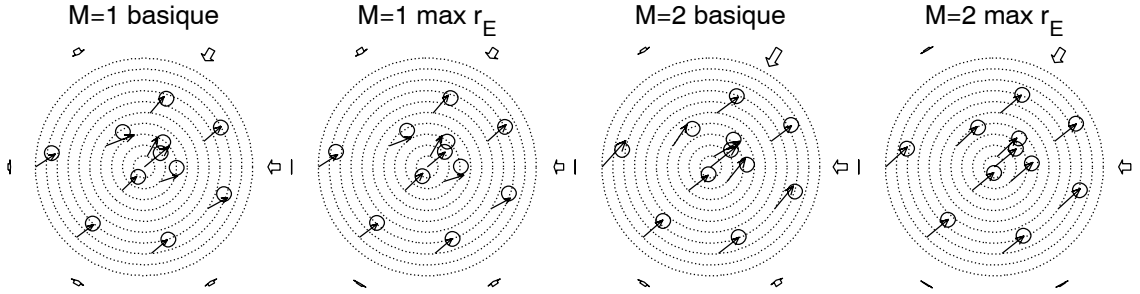

FIG. 4.14 – *Effet directionnel basse fréquence représenté par la moyenne pondérée $\langle\vec{V}\rangle_E$ du vecteur vélocité (flèches fines), mesurée en différentes positions pour quatre type de restitution ambisonique. Pour chaque position, un cercle (clair) indique le lieu que $\langle\vec{V}\rangle_E$ devrait pointer s'il égalait le vecteur énergie $\vec{E}$. La distance au centre des positions d'écoute s'étale de 0 à 2,5 mètres par pas de 25 cm (cercles concentriques).*

de $\langle\vec{V}\rangle_E$ vers $\vec{E}$ est assez bien vérifiée, malgré quelques petits écarts.

Cette identification $\vec{\mathcal{V}}_f(\vec{r}) = \langle \vec{V} \rangle_E \simeq \vec{E}$ pour les positions excentrées justifie donc l'application du décodage *max* r_E à un domaine relativement basse-fréquence dans les situations A_2 et A_3 présentées Figure 3.4.

4.2.2 Auditoires s'étendant à proximité des haut-parleurs et hors contrôle de reconstruction

Rappelons rapidement (Cf 3.1.3) qu'en position d'écoute excentrée, les conditions de création d'image sonore sont altérées de deux manières: par la convergence désormais asynchrone des contributions au point considéré, avec le risque d'une distorsion par effet d'antériorité, et par une distorsion directionnelle et énergétique des contributions perçues, donc une distorsion du vecteur énergie (Figure 4.15). L'objectif de cette section est d'illustrer la robustesse de la restitution dans ces conditions en fonction des différentes solutions de décodage développées au chapitre précédent, avec une attention particulière pour les solutions *in-phase* qui sont spécialement dédiées à ces conditions critiques.

En admettant un modèle de propagation sphérique et un rayonnement non-directif de la part des haut-parleurs, la contribution du haut-parleur i placé en $\vec{R}_i = R_{HP}\vec{u}_i$, vue du point excentré $\vec{r}$, est une onde d'incidence $\vec{u}'_i(\vec{r}) = \vec{d}_i/|\vec{d}_i|$, d'amplitude $G'_i(\vec{r}) = G_i R_{HP}/|\vec{d}_i|$, où $\vec{d}_i = \vec{R}_i - \vec{r}$, et arrivant avec un retard $\tau_i = |\vec{d}_i|/c$ (Figure 4.15). L'étalement temporel des contributions dépend ainsi des distances absolues, alors que la distorsion du vecteur énergie[11] $\vec{E}(\vec{r})$ dépend des distances relatives:

$$\vec{E}(\vec{r}) = \frac{\sum_{i=1}^{N} \left(G'_i(\vec{r})\right)^2 \vec{u}'_i(\vec{r})}{\sum_{i=1}^{N} \left(G'_i(\vec{r})\right)^2}, \qquad G'_i(\vec{r}) = \frac{R}{|R\vec{u}_i - \vec{r}|} G_i \tag{4.3}$$

La valeur prédictive du vecteur énergie sur l'effet de localisation a été montrée en 1.5.3 – au regard de l'ITD haute-fréquence et de l'ILD – dans l'hypothèse d'une convergence synchrone des ondes. Sans cette hypothèse, la prédiction d'après $\vec{E}$ se réduit aux effets des signaux de nature stationnaire, c'est-à-dire les effets sur l'ILD et le retard de phase (ITD basse-fréquence). Le seul vecteur énergie n'offrirait qu'une interprétation incomplète des mécanismes de localisation dans l'exemple de la figure 4.15: on y note que le vecteur énergie en position X reste plutôt orienté vers la source virtuelle, quoiqu'étant de petite norme r_E, mais que la contribution venant de la direction opposée (en bas) est la plus précoce et risque d'imposer la direction perçue, selon l'échelle temporelle de l'étalement (donc du paramètre R_{HP}) et la nature du signal.

On ne dispose malheureusement pas de modèle de prédiction prenant en compte l'étalement temporel – au-delà de 1 ms, ce qui correspond à une différence de marche de 34 cm – et traduisant les éventuels effets d'antériorité, de pondération ou d'inhibition des fronts d'onde successifs. Dans l'analyse qui suit, on se contente donc d'observer le comportement du vecteur énergie pour comparer les différentes solutions de décodage, présumant qu'il donne malgré tout une idée pertinente de l'évolution des propriétés de restitution, d'un décodage à l'autre.

Comparaison des décodages d'après le comportement du vecteur énergie

La simulation qui suit met en lice les principaux pan-pots ambisoniques envisageables pour un dispositif (2D) hexagonal régulier, à savoir: "*basique*", "*max* r_E" et "in-phase" aux ordres 1 et 2, ainsi que le pan-pot "*super-minimal*" (ordre 3). Ces décodages sont décrits en 3.3.1 et en 3.3.2. Pour simplifier le discours, on assimile abusivement l'effet de localisation perçu au vecteur énergie $\vec{E}$.

11. Dans la suite de cette section, $\vec{E} = r_E.\vec{u}_E$ désignera implicitement le vecteur défini par (4.3) pour un lieu donné $\vec{r}$, et plus nécessairement le vecteur énergie théorique estimé en $\vec{r} = 0$.

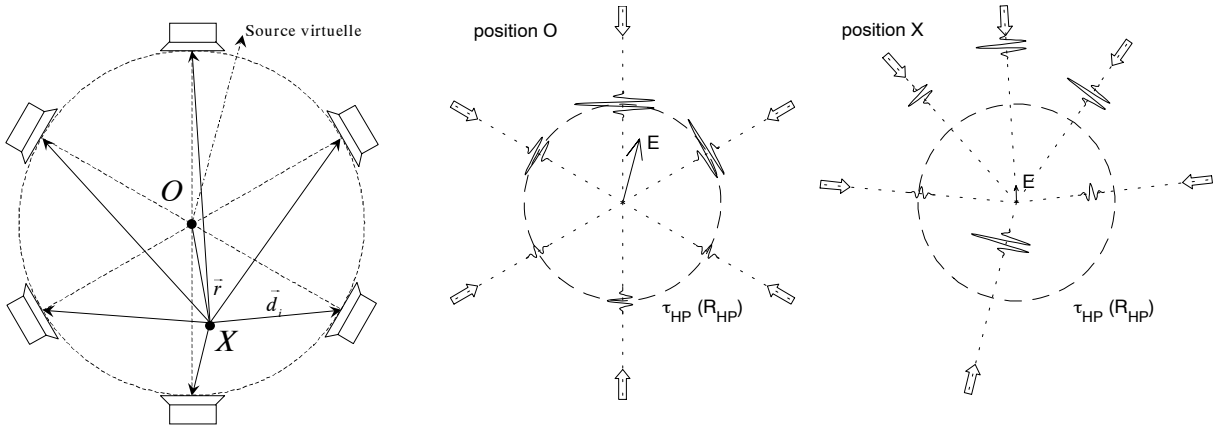

FIG. 4.15 – *Pour une position excentrée $\vec{r}$ de l'auditeur (notée X), la contribution venant du haut-parleur i est perçue comme une onde d'incidence $\vec{u}_i = \vec{d}_i/|\vec{d}_i|$, d'amplitude relative $R_{HP}/|\vec{d}_i|$ par rapport à l'amplitude vue du centre, et avec un retard $\tau_i = |\vec{d}_i|/c$. Au centre et à droite: l'arrivée des différentes contributions (impulsions gaussiennes), vue des positions centrale O et excentrée X, est illustrée pour la restitution basique d'ordre 1 d'une source placée en $\psi = -15°$. Le cercle en tiret sert à la fois de référence pour le temps de parcours τ_{HP} correspondant au rayon R_{HP} du dispositif, et de cercle unité pour le vecteur énergie $\vec{E}$ (flèche centrée). Vu de X ($r_X = 0{,}6R_{HP}$ et $\theta_X = -170°$), le vecteur énergie subit une distorsion et l'effet d'antériorité rentre en jeu (la convergence des ondes est asynchrone).*

La figure 4.16 illustre la perception du balayage panoramique d'une source sonore en différents points d'écoute. Il y apparaît clairement que si l'effet de balayage panoramique est reproduit de façon satisfaisante au centre (cohérence et homogénéité directionnelles, en tout cas pour les configurations non-minimales), il est perçu avec des distorsions pouvant être sévères pour les positions excentrées, l'image sonore ayant tendance à se resserrer autour des haut-parleurs. Dans les cas critiques (position H_2 par exemple, voir également la figure 4.17), des effets de *rebroussement* de la trace du vecteur énergie apparaissent – sous forme de "crochets" – avec les pan-pots ambisoniques basiques (ordres 1 et 2) et super-minimal (ordre 3). Cela signifie que le balayage panoramique est perçu avec des retours en arrière, ou encore que dans une scène complexe, les positions relatives de certaines images peuvent être inversées, créant en quelque sorte un *aliasing directionnel* ou *panoramique*. Pire, on note que pour un auditeur placé en H_2, par exemple, les décodages basiques sont inaptes à reproduire des images sonores dans toutes les directions: il y a des *secteurs angulaires "aveugles"* (des "trous") dans le rendu de la scène sonore (Figure 4.17, milieu et bas). Le rendu directionnel associé aux solutions de décodage *max* r_E semblent beaucoup moins catastrophique pour les positions illustrées, mais il faut signaler qu'il est sujet aux mêmes artefacts en des positions plus critiques (si l'on éloigne H_2 tel que $OH_2 = 0{,}8R$ par exemple). Qu'il s'agisse du décodage basique, *max* r_E, ou super-minimal, on constate des *fluctuations des caractéristiques directionnelles* (r_E et θ_E)[12], d'autant plus nombreuses que l'ordre M du système est élevé. Le nombre et l'amplitude de ces fluctuations peuvent d'ailleurs être directement corrélés au *nombre et à l'importance des lobes secondaires que présentent les directivités microphoniques associées à ces solutions de décodage* (Figure 3.14). En définitive, seules les solutions *in-phase* offrent un effet de balayage panoramique continu au regard de la trace du vecteur énergie, et ceci, quel que soit l'ordre considéré

12. Ces fluctuations se traduisent graphiquement (Figure 4.16) par les concavités de la trace de $\vec{E}$.

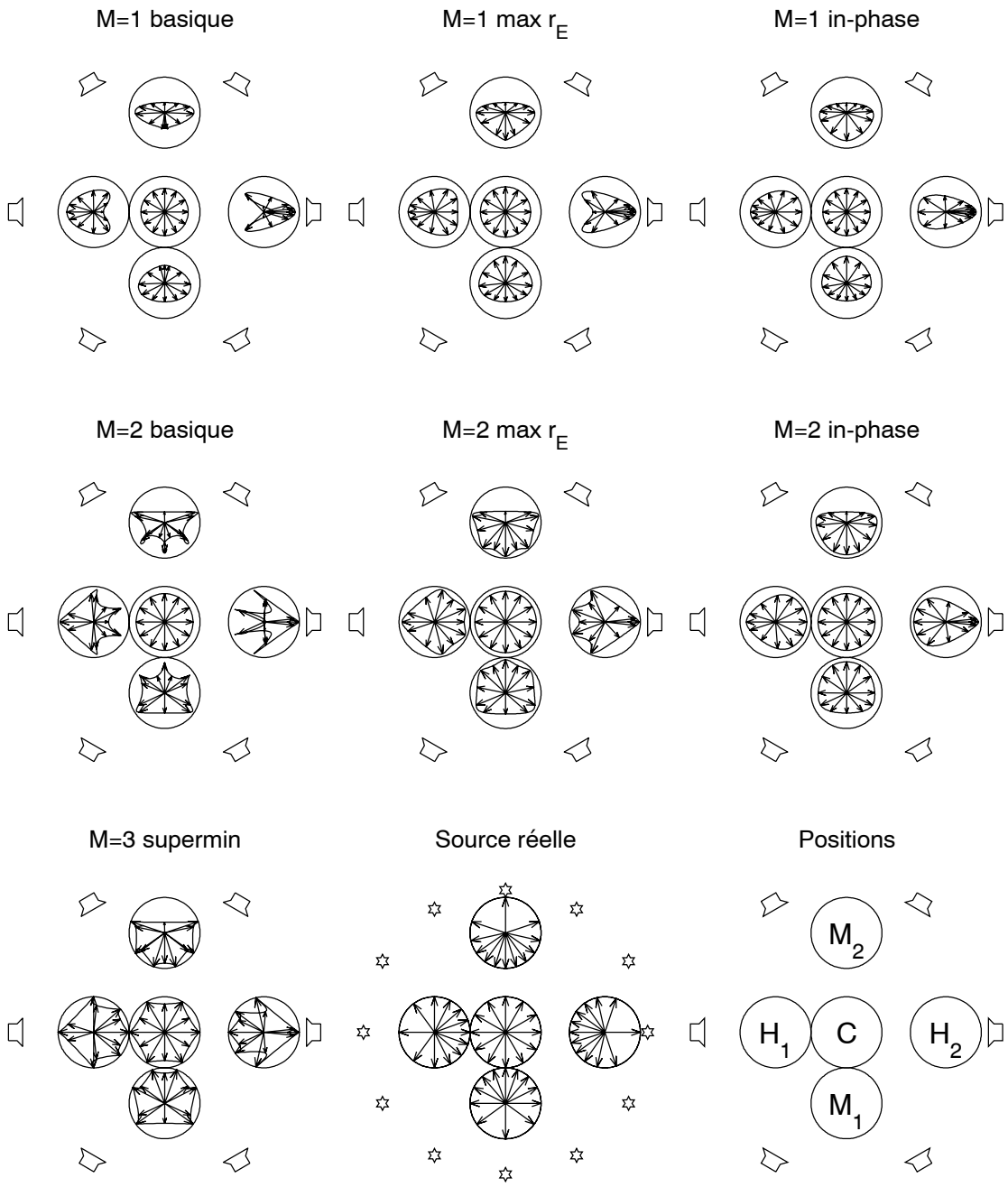

FIG. 4.16 – *Effet directionnel reflété par la trace du vecteur énergie, en cinq positions sur la zone d'écoute, pour un balayage panoramique d'une source virtuelle par pas de 30°, et représenté pour sept cas de codage/décodage ambisonique différents, plus le cas idéal d'une source réelle se déplaçant sur le cercle des haut-parleurs. (C): centre; (M_i): positions médianes entre deux haut-parleurs; (R_i): positions sur l'axe d'un haut-parleur; distances au centre: 0,5 R pour M_1 et H_1, et 0,7 R pour M_2 et H_2. En chaque position d'écoute, le vecteur énergie est représenté par une flèche pour chaque état de la source virtuelle. Un cercle indique le rayon unité que le vecteur énergie devrait atteindre pour définir une image précise. Une courbe continue indique la trace du vecteur énergie lors d'un balayage panoramique continu.*

et en tout point de la zone d'écoute, sans toutefois pouvoir s'affranchir d'une certaine distorsion directionnelle vers les haut-parleurs les plus proches. De manière générale, augmenter l'ordre du système permet bien sûr d'obtenir des images plus précises, c'est-à-dire des valeurs de r_E plus proches de 1.

Il est également intéressant d'observer l'*énergie perçue*, associée aux images selon leur direction. Là encore, les solutions *in-phase* sont les seules à offrir une évolution monotone (sur un demi-périmètre, Figure 4.17) en tout lieu. Il est normal de constater une amplitude de variation relativement importante pour les positions excentrées: elle serait d'ailleurs maximale si la source virtuelle était "idéalement projetée" (sur le périmètre des haut-parleurs), c'est-à-dire restituée par un seul haut-parleur à la fois dans chaque direction souhaitée[13]. La répartition de l'énergie est plus uniforme pour les décodages de bas ordre, mais correspond aussi à une moindre définition de l'image.

Cette analyse d'après le vecteur énergie pourrait être appliquée au cas de haut-parleurs directifs ainsi qu'aux cas de dispositifs non-concentriques.

Vers une topographie de la zone d'écoute: critères pour la définition de périmètres critiques

A ordre M égal, le décodage *in-phase* est sous-optimal pour une écoute en position centrée par rapport au décodage *max* r_E, mais il apparaît être le plus recommandable dans la situation critique d'un auditoire très élargi. Il existe donc probablement un périmètre critique délimitant l'auditoire, à partir duquel le décodage *in-phase* est préférable au décodage *max* r_E. Par ailleurs, on peut observer (Figure 4.16) qu'à même distance du centre, les propriétés de restitution ne sont pas les mêmes pour un auditeur placé sur l'axe médian de deux haut-parleurs ou sur l'axe d'un haut-parleur.

Pour ces raisons, il serait utile d'établir une topographie de la zone d'écoute pour chaque type de restitution, c'est-à-dire attribuer à chaque position d'écoute une note qui rende compte de façon synthétique de la préservation – ou inversement de la dégradation – des qualités spatiales de la scène sonore perçue. Dans de telles conditions, la distorsion directionnelle absolue ne semble pas être le critère le plus pertinent: elle est en effet inéluctable. Il faudrait d'ailleurs préférer comme direction de référence, celle de l'image "idéalement projetée" sur le périmètre des haut-parleurs, localisée en $\vec{r}$ dans la direction $\vec{u}_S(\vec{r}) = (R\vec{u}_S - \vec{r})/|R\vec{u}_S - \vec{r}|$ (angle ψ', figure 4.17). Il semble plus crucial de porter attention à des propriétés plus globale comme l'équilibre et la cohérence de la scène restituée, ainsi que la précision et la régularité des images sonores en mouvement.

Si l'on basait l'évaluation sur le seul vecteur énergie, la définition d'indices pertinents pourrait reposer sur les spécifications suivantes:

- Pénalisation des "trous" ou secteurs angulaires aveugles dans le panorama sonore.
- Pénalisation des interversions positionnelles entre images sonores, soit encore des retours en arrière de l'image lors d'un balayage panoramique continu ($\partial\theta_E/\partial\psi < 0$).
- Pénalisation des variations de la vitesse $\partial\theta_E/\partial\psi$ du balayage apparent, accélérations ou décélérations d'autant plus sensibles que l'image est précise (r_E proche de 1). On peut proposer une fonction "de coût" du type $\int r_E \left(\frac{\partial^2\theta_E}{\partial\psi^2}\right)^2 d\psi$.
- Pénalisation des variations de r_E (précision de l'image). Pourrait se combiner avec le critère précédent.

Mais comme il a été souligné plus haut, le vecteur énergie ignore des aspects temporels (asynchronisme) qui, dans la situation présente, ont des conséquences certaines sur la localisation. Ce sont donc des tests subjectifs qu'il faudrait mener pour établir notre topographie, ou bien directement pour comparer les différents décodages. Voici pour suggestion d'expérience: répartir les auditeurs en évitant les problèmes de masquage

13. Cette amplitude de variation maximale se reporte dans le cas "super-minimal" (Figure 4.17): environ 15dB en H_2.

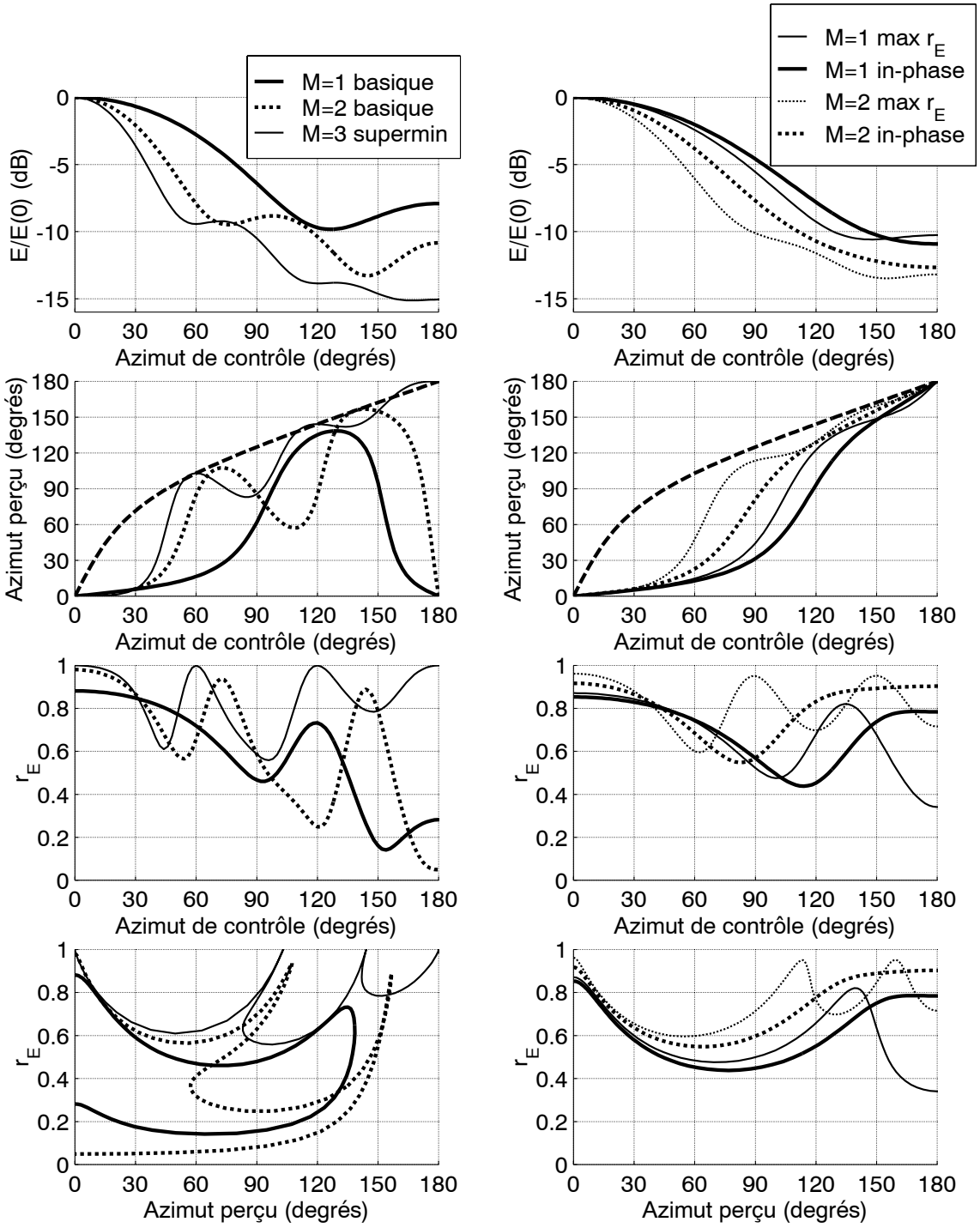

FIG. 4.17 – *Caractérisation du balayage panoramique perçu à la position H_2 (Figure 4.16). De haut en bas: énergie perçue en fonction de l'azimut de la source virtuelle (azimut de contrôle); azimut perçu en fonction de l'azimut de contrôle, accompagné de l'azimut de référence ψ en tirets gras (celui qui serait perçu si l'image était idéalement projetée sur le périmètre des haut-parleurs, i.e. diffusée par un seul haut-parleur); indice r_E en fonction de l'azimut de contrôle; indice r_E en fonction de l'azimut perçu (version "linéarisée" de la trace de $\vec{E}$ en polaire dans la figure 4.16).*

inter-personnel (éventuellement surélever les haut-parleurs), restituer une source en mouvement à trajectoire circulaire (balayage panoramique), et recueillir de chaque auditeur une note sur chacun des critères suivants:

- Précision générale de l'image sonore.
- Régularité (ou continuité) de la qualité de l'image, avec implicitement les critères énoncés au-dessus (pénalisations).

Pour établir une topographie, la notation devrait se baser sur une écoute au centre. Pour une comparaison entre décodage, on peut se contenter de notes données par le même auditeur à la même position. Il faut avoir conscience du nombre de paramètres dont dépendent les résultats:

- La nature du signal: présence de transitoires, dynamique des attaques, etc...
- Le nombre de haut-parleurs, qui joue sur la densité temporelle des contributions et sur leur fusion perceptive.
- Le rayon R_{HP} du dispositif: parce que l'échelle temporelle est incompressible sur le plan perceptif, la topographie de la zone d'écoute ne saurait être proportionnelle à R_{HP}, contrairement à ce qui pourrait être obtenu par la seule observation du vecteur énergie.

Ces tests pourraient être réalisés avec beaucoup plus de souplesse en simulation binaurale, par une extension de la méthode des haut-parleurs virtuels (Cf 5.5.2), simplifiant le problème de l'écoute de référence.

Conclusions

L'analyse objective de la restitution basée sur le vecteur énergie montre de façon très claire la robustesse qu'assure le décodage *in-phase*: il permet d'éviter en tout point certains artefacts attendus avec les autres décodages, comme l'interversion positionnelle des sources, et d'assurer une certaine régularité dans la qualité des sources en mouvement. Les caractéristiques de la loi de pan-pot associée[14] permettent d'attendre que ces propriétés restent vraies avec un modèle de perception plus complet, et soient confirmées par l'expérience subjective. Le décodage *max* r_E étant optimal, à ordre M égal, sur une zone d'écoute plus réduite, il doit exister un périmètre critique délimitant l'auditoire, à partir duquel un décodage devient préférable à l'autre. Partant d'un périmètre limite imposé, on peut supposer que le décodage optimal est intermédiaire, résultant d'une interpolation[15] entre les décodages *in-phase* et *max* r_E, ou entre les coefficients correcteurs g_m associés. Le décodage *basique* apparaît quant à lui le moins recommandable dans ces conditions d'écoute.

Des tests subjectifs mériteraient d'être réalisés pour valider et préciser ces résultats. Ils pourraient être l'occasion d'approfondir la connaissance sur les mécanismes de localisation dans les cas de convergence asynchrone d'ondes cohérentes, voire d'en dégager des modèles mathématiques.

4.2.3 Reconstruction étendue: correction ou non du champ proche des haut-parleurs

Dans les spécifications pour la définition du décodage ambisonique énoncées en 3.1.3, le décodage basique, dédié à une reconstruction locale du front d'onde original, est tout d'abord basé sur l'hypothèse – généralement adoptée – de haut-parleurs assez distants (R_{HP} grand) pour assimiler leurs contributions à des ondes planes sur la zone d'écoute. Dans un second temps (page 161), nous avons proposé une correction de l'effet de champ proche des haut-parleurs afin de reconstituer fidèlement les composantes originales du champ ambisonique.

14. Monotonie et annulation des gains dans la direction opposée à la source.
15. On peut également envisager une transition fréquentielle, comme suggéré en 3.1.3.

La simulation d'un cas de reconstruction étendue, par le biais d'une restitution ambisonique d'ordre élevé (Figure 4.19)[16] sur un dispositif de rayon R_{HP} modéré, nous permet maintenant d'illustrer les conséquences de la présence ou de l'absence de cette correction, et d'en discuter la pertinence dans une stratégie globale de restitution.

Interprétations de la reconstruction

D'après l'équation (3.28) page 161, le champ ambisonique reconstitué après un décodage basique *sans correction* prend la forme $\mathbf{B}' = \text{Diag}\left(\underline{\mathcal{F}}^{(R_{HP}/c)}(\omega)\right).\mathbf{B}$. Un effet de champ proche est ainsi appliqué par le filtre $\mathcal{F}_m^{(R_{HP}/c)}(\omega)$ (équation 3.12, page 153) à chaque composante originale B_{mn}^{σ} (Figure 4.18). Supposons que le champ encodé $\mathbf{B}$ se réduit à une onde plane d'incidence $\vec{u}_S$. Dès lors, le champ recomposé prend les caractéristiques d'une onde sphérique de même incidence $\vec{u}_S$ vue du point central O, mais de source placée à distance R_{HP}. Comme l'illustre la figure 4.19 (partie gauche), l'image sonore est donc en quelque sorte *"projetée" sur le périmètre de haut-parleurs* en un même point $\vec{r} = R_{HP}\vec{u}_S$ pour tous les auditeurs placés dans la zone de reconstruction, au lieu d'être *"projetée à l'infini"*, avec la même direction d'incidence pour tous les auditeurs, comme le voudrait le cas d'une onde plane (Figure 4.19, partie droite).

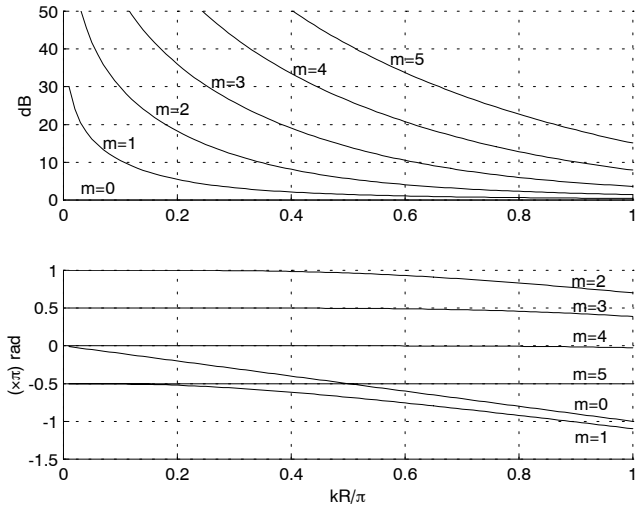

FIG. 4.18 – *Réponses fréquentielles des filtres $\mathcal{F}_m^{(R/c)}(\omega)$ traduisant l'effet du champ proche (source à une distance R) sur les composantes harmoniques sphériques, c'est-à-dire leur altération par rapport à l'effet d'une onde plane de même direction. A "rayon-fréquence" égal, l'effet d'amplification et de déphasage croît de façon considérable avec l'ordre M. Pour conversion: en posant $R = 1$ m (et $c = 340$ m/s), $f = 170(kR/\pi)$ (en Hz); en posant $f = 100$ Hz, $R = 1{,}7\,(kR/\pi)$ (en m).*

Une interprétation plus intuitive peut être apportée à ce phénomène. La modélisation des *contributions des haut-parleurs comme ondes planes* étant maintenue pour la définition du décodage, la restitution "idéale" ou "optimale" d'une *image sonore encodée comme onde plane* est acquise lorsque la reconstruction du champ est réalisée par un seul haut-parleur. Lorsqu'on augmente à la fois le nombre de haut-parleurs et l'ordre du système, on s'approche de cette situation idéale: les haut-parleurs qui contribuent le plus à la reconstruction

16. En pratique, il faudrait tenir compte de la taille des haut-parleurs (étendue des sources). Cependant ce paramètre est surtout sensible à l'échelle des petites longueurs d'ondes et à faible distance des haut-parleurs.

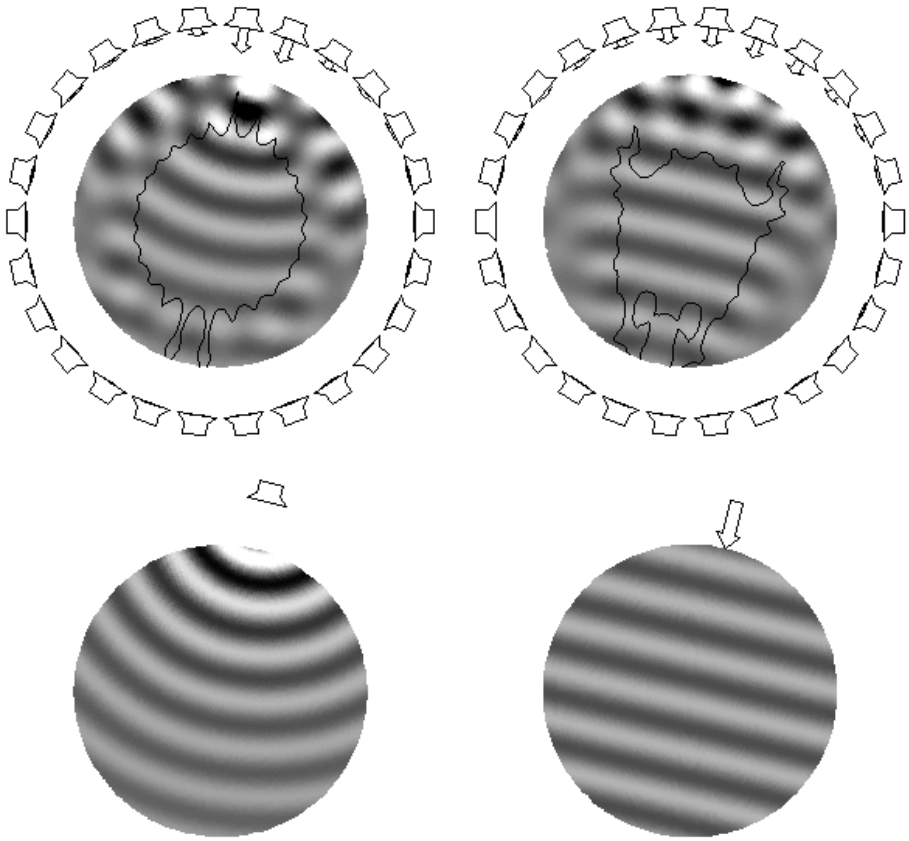

FIG. 4.19 – *Restitution ambisonique (ordre $M = 12$, $N = 26$ haut-parleurs, décodage basique) d'un champ issu de l'encodage d'une onde plane (en bas à droite): sans (en haut à gauche) et avec (en haut à droite) correction de la propagation sphérique des contributions des haut-parleurs pour le décodage. En l'absence de correction, le champ reconstruit tend vers une onde sphérique (en bas à gauche) dont la source serait placée sur le périmètre de haut-parleurs dans la même direction (vue du centre) que l'onde plane encodée. Dans l'autre cas, on tend bien vers un reconstruction de l'onde plane originale. Simulation réalisée pour une onde monochromatique de fréquence $f = 700Hz$, avec un rayon de $R = 2$ m pour le périmètre de haut-parleurs, et de $1{,}5$ m pour la région visualisée du champ. Les limites de bonne reconstruction par rapport aux tendances asymptotiques (en bas) sont représentées (en haut) pour une erreur de 20%.*

sont confinés dans un secteur angulaire de plus en plus réduit. Comme en réalité les haut-parleurs ne sont pas très lointains mais placés sur un cercle rapproché, on tend vers l'effet d'une source placée sur ce cercle, et le champ d'onde effectivement créé tend vers une onde sphérique.

La correction suggérée (page 161) pour préserver les caractéristiques de l'onde plane, consiste à appliquer à chaque composante ambisonique B^{σ}_{mn} le filtrage inverse $\left(\mathcal{F}_m^{(R_{HP}/c)}\right)^{-1}(\omega)$ préalablement au décodage basique. La figure 4.19 (partie droite) montre que l'onde plane est grâce à cela bien reconstruite, sur une zone du même ordre de grandeur que pour la reconstruction de l'onde sphérique (à gauche).

Ambisonic et Holophonie

Abordée comme une technique ayant pour vocation de tendre vers une reconstruction exacte du champ acoustique, *Ambisonic* est décrite dans [NE98] [NE99] [Nic99] comme un cas particulier de l'*Holophonie* au sens défini par Jessel [Jes73], en usant pour cela de l'hypothèse d'un rayon R_{HP} tendant vers l'infini. La correction du champ proche que nous avons présentée permet d'illustrer la reconstruction asymptotique du champ par le système ambisonique sans avoir recours à cet artifice.

Il reste que l'approche ambisonique se distingue fondamentalement des autres systèmes de type holophonique, comme le *Wave Field Synthesis* (WFS) [BVT95], par la façon dont le champ reconstruit converge vers le champ original en fonction de l'ordre du système (ou du nombre de haut-parleurs). Avec *Ambisonic*, la progression de la reconstruction s'exprime en termes d'expansion radiale à partir du centre, expansion proportionnelle à la longueur d'onde. Avec le WFS (par exemple), la qualité de reconstruction reste au contraire homogène sur toute la zone de restitution, et la progression se manifeste par l'élargissement de la bande basse-fréquence où la reconstruction est correcte, c'est-à-dire exempte d'*aliasing* spatial[17].

Sans la correction du champ proche, la convergence asymptotique entre Ambisonic et le WFS ne se rencontre que si les sources sonores simulées par WFS sont placées sur le périmètre des haut-parleurs, lorsqu'elles sont encodées comme ondes planes pour Ambisonic (Figure 4.19).

Quel choix de décodage?

La correction du champ proche proposée est indispensable à la reconstruction "véritable" de l'événement acoustique original, dans les limites (rayon-fréquence) permises par l'ordre du système. Mais faire le choix ou non d'une telle correction reste une question qui mérite d'être débattue, indépendamment de la faisabilité et du coût de calcul supplémentaire investi.

Quel type de projection? Si l'on s'intéresse au cas d'une source sonore encodée comme une onde plane et dans l'hypothèse d'une reconstruction assez étendue (ordre M élevé), cette alternative revient à choisir, comme on l'a vu plus haut, entre une projection de l'image sonore "à l'infini" (dans la direction $\vec{u}_S$ pour tous les points "de vue") *ou bien* sur le périmètre des haut-parleurs (le point $R_{HP}\vec{u}_S$ pour tous les points de vue). Dans une application audio-visuelle où l'image visuelle est elle-même projetée sur un écran, il peut être bienvenu que l'image sonore coïncide avec elle en étant "projetée" sur le même écran, ce qui laisserait préférer l'absence de correction. Précisons que cet effet de projection n'empêche pas l'impression de profondeur et de relief sonore, qui est quant à elle essentiellement due à la présence des réflexions reproduites en plus de l'onde directe, et devrait *a priori* être préservée pour les différentes positions de la zone d'écoute[18].

17. La limite fréquentielle est définie par la fréquence de Nyquist [NE98] qui dépend de l'espacement entre les haut-parleurs.

18. De même que les informations visuelles de profondeur et de relief découlent des effets de perspectives et d'ombrage présents dans l'image projetée. La *sensation* de relief peut être ajoutée (lunettes stéréoscopiques). Mais c'est le même point de vue (celui de la caméra) qui est offert aux spectateurs.

Cohérence de l'ensemble des informations spatiales. Pour décrire de façon plus complète les conditions de localisation en une position *excentrée* donnée (à distance r_{pos}), il faudrait théoriquement distinguer les informations spatiales perçues en deux ou trois domaines fréquentiels, dont les limites dépendent de l'ordre M du système, de r_{pos} et de R_{HP}:

- un domaine basse-fréquence où la reconstruction acoustique est contrôlée (avec ou sans correction du champ proche),
- en cas de correction, un domaine intermédiaire où la reconstruction n'est plus contrôlée, mais où le décodage (ou la loi de pan-pot) est encore affecté par la correction,
- un domaine haute-fréquence hors de portée de la reconstruction et de l'éventuelle correction, et où un décodage modifié (*max* r_E ou *in-phase*) reste à envisager.

Avec cette distinction se pose la question de la cohérence entre les informations spatiales recueillies dans chacun de ces domaines. Qu'un décodage basique ou bien modifié y soit appliqué, les informations du troisième domaine (haute-fréquence) se conforment de plus en plus au cas d'une image projetée sur le périmètre des haut-parleurs à mesure que l'ordre M augmente (ainsi que la densité angulaire des haut-parleurs). Mais parallèlement, ce domaine est de plus en plus confiné vers les hautes fréquences, puisque la reconstruction s'étend (en "rayon-fréquence") ainsi que la portée fréquentielle de la correction du champ proche (Figure 4.18[19]) le cas échéant. Avec un système d'ordre M limité, il faudrait approfondir l'étude proposée précédemment (en 4.2.2) pour connaître l'effet de localisation au regard du domaine haute-fréquence et du domaine intermédiaire.

Cas des sources *encodées* en champ proche. Lorsque le décodage sans correction est appliqué à un champ constitué d'une onde sphérique et non plus plane, les composantes ambisoniques reconstruites au centre du dispositif sont caractérisées par des facteurs de correction d'encodage $\mathcal{F}_m^{(\rho/c)}\,\mathcal{F}_m^{(R_{HP}/c)}$ par rapport à l'encodage d'une onde plane, ρ étant la distance de la source de l'onde sphérique. L'événement acoustique reconstruit "ressemble" à une onde sphérique, avec un foyer compris à l'intérieur du périmètre des haut-parleurs, mais n'en est pas une! Resterait à évaluer l'effet perceptif d'un tel artefact acoustique.

Alors: correction du champ proche ou pas? Gerzon la recommandait pour les systèmes d'ordre 1. Nous avons exposé des arguments divergents à partir de l'exemple d'un système d'ordre élevé. La question reste finalement à soumettre à l'expérience! Cette correction n'a pas été incluse dans les décodeurs ambisoniques réalisés au cours de cette thèse (chapitre 5).

4.2.4 Effet de la salle sur la restitution

Dans la plupart des cas, la restitution ne se fait pas dans une salle anéchoïque. A l'onde directe rayonnée par chaque haut-parleur s'ajoutent des réflexions précoces, qui lui confèrent une coloration spécifique. Si l'effet de salle est assez important, il est particulièrement recommandé de toujours mettre en jeu plusieurs haut-parleurs à la fois pour la création d'une image sonore, surtout si elle est mobile, pour éviter qu'elle soit empreinte d'une coloration artificiellement changeante. Pour D.Malham, c'est un des grands avantages qu'offre *Ambisonic* de pouvoir multiplier le nombre de haut-parleurs afin de mieux noyer le caractère individuel de chaque haut-parleur. Ces conditions de restitution mettent donc en défaveur les configurations ou lois de pan-pot "minimales" et "super-minimales" (3.3.1).

19. Attention: pour se rendre compte de l'effet de la correction sur les composantes ambisoniques, il faut inverser ces courbes: pour un kR donné, les composantes B_{mn} sont d'autant plus atténuées que l'ordre m est élevé.

4.3 Mise en défaut de l'hypothèse d'onde plane pour le champ encodé

Les principaux critères de décodage ambisonique qui ont été exposés portent sur l'optimisation de l'effet de localisation d'images sonores, avec l'hypothèse que celles-ci sont encodées sous forme d'ondes planes. Quelle est alors la pertinence des différentes solutions de décodage (basique, *max r_E* et *in-phase*) lorsque le champ ambisonique encodé est plus complexe et met en défaut cette hypothèse? C'est ce qui est discuté dans les paragraphes suivants, où sont abordés différents cas de figures: champ proche, prise de son non-idéale, effet de salle.

4.3.1 Onde directe associée à une source

Source encodée en champ proche: effet sur la localisation hors contrôle de reconstruction

Lors de l'enregistrement d'une scène sonore, il est tout à fait envisageable qu'une source sonore se trouve à proximité du microphone, à une distance ρ disons de l'ordre de 0,5 m (par exemple). L'onde directe mesurée est donc de nature sphérique et non plus plane. Le figure 4.18 montre la façon dont l'effet de champ proche se traduit sur les composantes ambisoniques mesurées, par comparaison à celles qui seraient mesurées dans le cas d'une onde plane de même incidence: les composantes elles-mêmes, mais surtout le rapport d'amplitude et de phase entre les composantes B^{σ}_{mn} d'ordres m différents, se trouvent affectés de manière d'autant plus significative que la fréquence est basse et/ou la distance ρ est petite. Ces relations de phase sont explicitées par les facteurs $\mathcal{F}_m^{(\rho/c)}(\omega) = \Gamma_m(k\rho)/\Gamma_0(k\rho)$ (3.12) que nous noterons encore $\mathcal{F}_m(k\rho)$.

Dans le domaine de reconstruction contrôlée (bande basse-fréquence, auditoire peu étendu), le décodage basique[20] permet de reconstituer les caractéristiques de cette onde sphérique. Ces conditions d'écoute ne causent pas d'inquiétude. Ce qui nous intéresse ici en revanche, c'est de *caractériser la restitution dans des conditions où la reconstruction est hors de portée* (position excentrée r_{pos}), même dans un domaine basse-fréquence, et d'évaluer la robustesse des différents décodage. Nous nous plaçons dans le cas de dispositifs de restitution réguliers.

Le champ proche pondère les composantes ambisoniques B^{σ}_{mn} par des "gains" $\mathcal{F}_m(k\rho)$ suivant leur ordre m, de la même manière que les décodages modifiés (*max r_E* et *in-phase*) consistent à les pondérer par des gains g_m préalablement au décodage basique. On peut donc définir une loi de pan-pot équivalente à l'encodage/décodage modifié de l'onde sphérique reposant sur la fonction désormais complexe:

$$\mathcal{G}_{k\rho}^{M(2D)}(\theta) = g_0 + 2\sum_{m=1}^{M} g_m\,\mathcal{F}_m(k\rho)\cos(m\theta) \qquad \text{(Cas 2D)} \tag{4.4}$$

$$\mathcal{G}_{k\rho}^{M(3D)}(\theta) = g_0 + \sum_{m=1}^{M} (2m+1) g_m\,\mathcal{F}_m(k\rho)\,P_m(\cos\theta) \qquad \text{(Cas 3D)} \tag{4.5}$$

Présentées sous forme de diagramme polaire, ces lois se montrent très utiles pour caractériser la restitution. La figure 4.20 montre qu'elles ont une forme d'autant plus perturbée que l'ordre M augmente et que la fréquence diminue (pour un ρ fixé): l'harmonique d'ordre le plus élevé[21] devient prépondérante à mesure $k\rho$ diminue et imprime ainsi sa forme au diagramme de façon exacerbée. Les lois *in-phase* résistent beaucoup mieux que les autres à cette dégénérescence.

Dans les cas critiques, la forme "en fleur" du diagramme indique que l'énergie émises par les haut-parleurs est répartie à peu près dans toutes les directions, et non plus principalement dans la direction $\vec{u}_s$ de la

20. A condition, au besoin, de compenser l'effet de proximité des haut-parleurs (section 4.2.3).

21. $Y^1_{M0} \propto P_M(\cos\theta)$ pour le cas 3D ou $Y^1_{MM} \propto \cos(M\theta)$ pour le cas 2D.

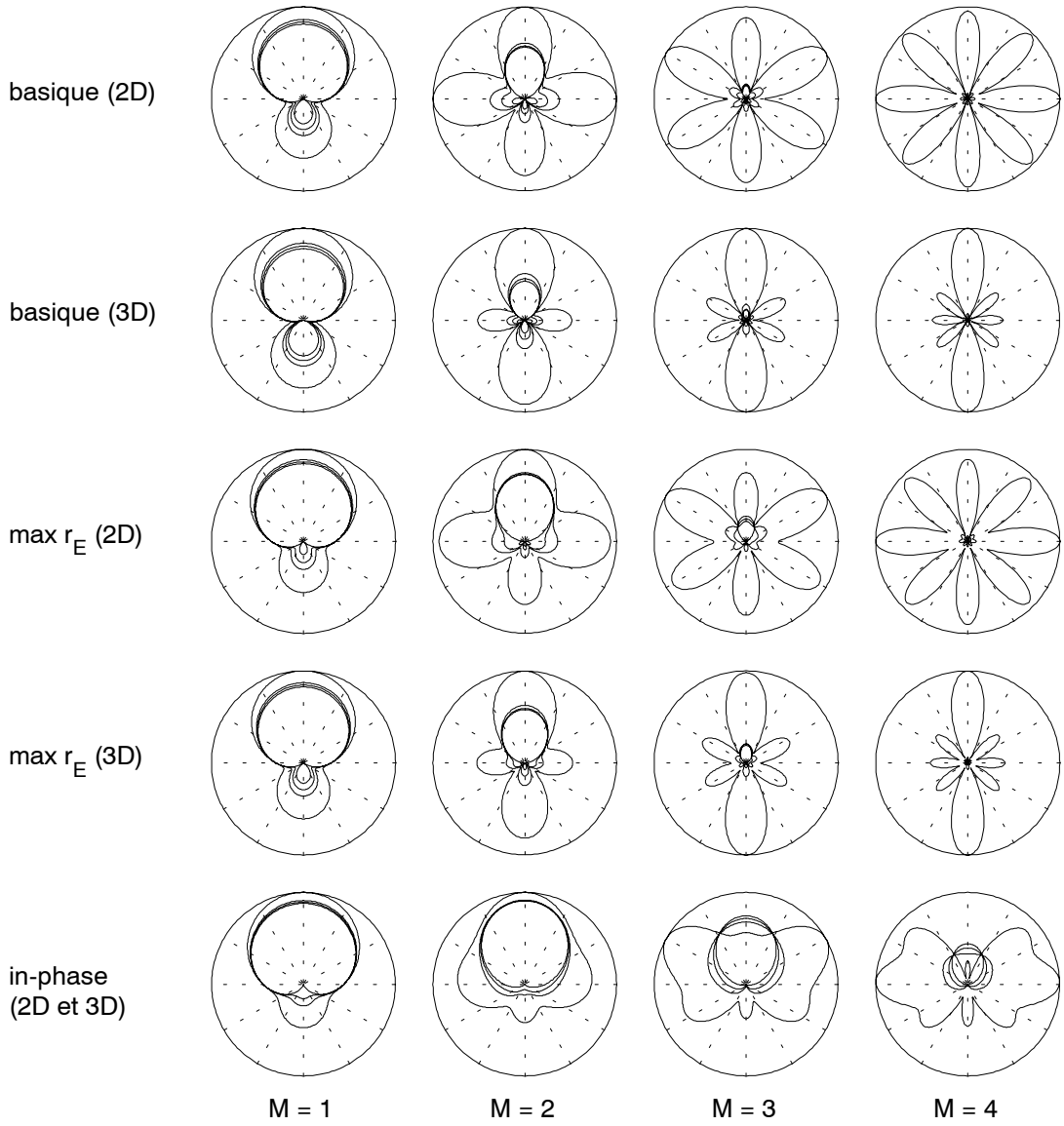

FIG. 4.20 – *Lois de pan-pot (gain en valeur absolue représentés en diagramme polaire) équivalant à l'encodage/décodage d'une onde sphérique (se reporter également à la figure 3.14). Pour chaque type de décodage présenté sont superposées, en plus de la loi de référence correspondant au cas d'une onde plane ($k\rho = \infty$), les lois correspondant aux paramètres $k\rho = 0.92$, 1.84, 2.8 (par amplitude décroissante), soit par exemple aux fréquences* 100 *Hz,* 200 *Hz et* 300 *Hz en supposant la source distante de* $\rho = 0.5$ *m. Les lois sont normalisées par rapport à celle de plus grande amplitude ($k\rho = 0.92$), de sorte que la loi de référence va presque jusqu'à disparaître dans les cas* $M = 4$.

source virtuelle. Elle indique aussi une fluctuation de cette répartition en fonction de la direction $\vec{u_s}$. On peut en attendre comme conséquence théorique une dégradation de l'effet de localisation, et pour les positions d'écoute excentrée, le risque d'un effet d'attraction par le haut-parleur le plus proche (cf 4.2.2). Mais comme ces "dégénérescences" ne se manifestent que dans un domaine basse-fréquence (selon la proximité ρ et l'ordre M), il est difficile d'affirmer l'effet négatif de cette dispersion directionnelle de l'énergie. Notons tout de même que l'effet d'amplification des harmoniques d'ordres supérieurs (Figure 4.18) se reporte directement sur l'énergie globale restituée à la fréquence donnée: ceci peut être dommageable en revanche.

Afin de tracer une évolution des performances de localisation *a priori* en fonction de la fréquence, il est bon d'utiliser un indice synthétique de la dispersion directionnelle. L'indice r_E associé à chaque type de décodage peut être réestimé en fonction de la fréquence en substituant les coefficients $\mathcal{F}_m(k\rho)g_m$ aux coefficients g_m dans (A.56) (A.64):

$$r_E^{2D} = \frac{2\sum_{m=1}^{M} g_{m-1}g_m\Re(\mathcal{F}_{m-1}\mathcal{F}_m^*)}{{g_0}^2 + 2\sum_{m=1}^{M} {g_m}^2|\mathcal{F}_m|^2} \qquad r_E^{3D} = \frac{2\sum_{m=1}^{M} m g_{m-1}g_m\Re(\mathcal{F}_{m-1}\mathcal{F}_m^*)}{\sum_{m=0}^{M} (2m+1){g_m}^2|\mathcal{F}_m|^2} \tag{4.6}$$

La figure 4.21 montre l'évolution de r_E en fonction de $k\rho$ pour un certain nombre de décodages. Ce qui frappe en considérant chaque type de solution individuellement, c'est que la restitution directionnelle est d'autant plus "robuste", dans le domaine basse-fréquence (i.e. pour les bas k, à ρ constant), que l'ordre du système est bas. En comparant les familles de solutions de décodage entre elles (Figure 4.22), il apparaît qu'à ordre égal et dimension identique (2D ou 3D), les solutions *in-phase* sont les plus robustes en basse-fréquence – et présentent toujours un meilleur r_E que les solutions basiques –, les solutions *max* r_E occupant une position intermédiaire. Il existe une interprétation assez simple de ces résultats: comme déjà mentionné plus haut, le champ proche perturbe plus particulièrement les composantes ambisoniques d'ordres élevés, qui sont liées aux tenseurs du champ d'ordres supérieurs. En modérant progressivement la participation des composantes d'ordres supérieurs (par pondération avant décodage basique, Table 3.10), les solutions *max* r_E et surtout *in-phase* réduisent l'effet "négatif" du champ proche[22], comme il advient aussi en choisissant un ordre inférieur, auquel cas cependant les caractéristiques haute-fréquence sont moins bonnes. Bien que présentant à ordre égal le même indice r_E en champ lointain, le décodage *in-phase* est donc plus recommandable que le décodage basique dans la mesure où les conditions de reconstruction contrôlée en basse-fréquence n'est pas assurée pour la majeure partie des auditeurs.

Avant de conclure, rappelons-nous que les critères observés ne sont pertinents que dans un domaine lieu-fréquence où la reconstruction n'est plus contrôlée. Or, si le champ proche de la source enregistrée affecte un domaine basse-fréquence d'autant plus large que l'ordre M est élevé, le domaine "rayon-fréquence" (kr_{pos}) où la reconstruction acoustique est contrôlée s'étend lui-même avec l'ordre M. Par ailleurs, il faudrait déterminer par l'expérience si, et dans quelles conditions (ρ,r_{pos},f), l'effet d'amplification et de dispersion des contributions observé en basse-fréquence est perçu de façon nuisible, ou bien participe éventuellement à l'enveloppement et l'effet de proximité (effet *bass-boost*). Enfin, soulignons que ce problème ne se rencontre pas avec les sources virtuelles encodées artificiellement sans simulation du champ proche (filtrage $\mathcal{F}_m(k\rho)$)! Dans les cas où l'on sait la présence de sources virtuelles en champ proche et qu'il semble souhaitable d'éviter la perturbation en basses fréquences, on peut imaginer éliminer ou fortement réduire la participation des composantes ambisoniques d'ordres supérieurs en basse-fréquence, mais en les préservant en haute-fréquence afin de ne pas pénaliser la qualité globale de l'image sonore.

22. On considère qu'ici, l'effet du champ proche (à l'encodage ou enregistrement) devient négatif dans la mesure où l'on n'est pas capable d'assurer sa reconstitution pour la majorité des auditeurs, et qu'à défaut, c'est bien le vecteur énergie "qui fait loi".

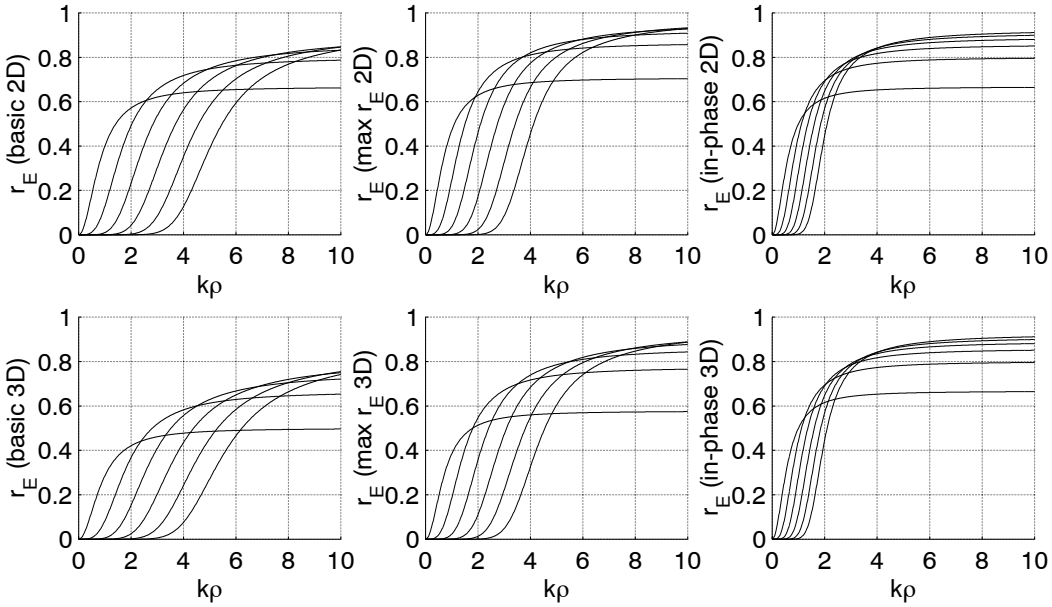

FIG. 4.21 – *Indice r_E associé à la restitution, avec différents types de décodages, d'une source (placée à une distance* ρ *du centre) encodée comme onde sphérique. r_E est donc dépendant de la fréquence, ou de k*ρ*. Pour chaque type de décodage, les courbes correspondant aux ordres de 1 à 6 sont ordonnées de gauche à droite (pour les k bas) puis de bas en haut (pour les k hauts). Précisons par exemple que pour* $\rho = 50$ *cm,* $k\rho = 1$ *correspond à la fréquence 108 Hz.*

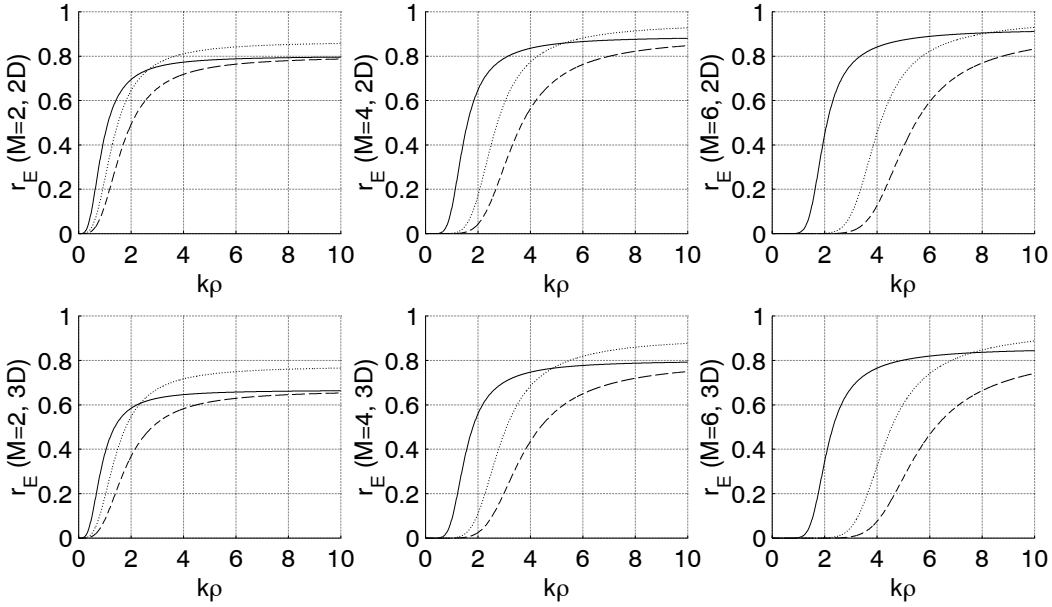

FIG. 4.22 – *Courbes déjà présentées Figure 4.21, mais réorganisées par ordre (2, 4 et 6) et dimension (2D et 3D) pour comparer les solutions basiques (tirets),* max *r_E (pointillés) et* in-phase *(traits continus).*

Prise de son non strictement conforme aux équations d'encodage théoriques

Il est montré en 3.4 que les systèmes de prise de son ambisonique basés sur l'utilisation de microphones non-coïncidents, comme le micro *SoundField*, recomposent les harmoniques sphériques du champ avec un *aliasing* du spectre harmonique sphérique qui croît avec les fréquences. Selon l'incidence de l'onde plane, l'encodage acoustique réalisé par le microphone s'écarte donc plus ou moins des équations théoriques, de sorte que le rapport d'amplitude entre les composantes encodées peut ne plus être conforme à des caractéristiques d'onde plane. Nous ne disposons pas actuellement d'assez de connaissance – d'autant que les microphones d'ordres supérieurs n'existent pas encore – pour mener une analyse sur la pertinence des différentes solutions de décodage dans le cas d'une prise de son réelle. Une vérification méritera d'être faite à ce sujet, à l'avenir.

4.3.2 Champ complexe, présence d'effet de salle

Localisation de source avec effet de salle très présent

Même dans le cas d'une source unique, le champ mesuré en un point est le résultat de l'onde rayonnée directement par la source et de son interaction avec l'environnement, sous forme de réflexions – spéculaires ou diffuses – et d'un champ réverbéré diffus, plus tardif. En fonction de la structure modale du lieu et de la stationnarité du signal émis, le champ mesuré peut se révéler de nature stationnaire et s'éloigner des caractéristiques de propagation d'une onde plane. Cela se produit rapidement dans les lieux clos de petit volume, en particulier les habitacles de voiture. On peut dès lors se poser la question[23] de la pertinence du décodage optimisé *max* r_E, qui cherche à améliorer l'image associée à une onde plane progressive.

Un tel décodage reste avantageux si l'on observe les événements sonores sur le plan temporel, le système auditif tirant des informations capitales pour la localisation lors des phénomènes transitoires. En effet, à chaque fois que le signal présente un caractère modulant ou transitoire, le caractère stationnaire du champ mesuré disparaît au profit d'un caractère propagatif entre l'instant d'arrivée de l'onde directe et celui des autres contributions (réflexions précoces). En supposant la source assez lointaine ou en restreignant l'observation à un domaine haute-fréquence, l'hypothèse d'onde plane reste donc valable pour les transitoires et justifie le décodage optimisé *max* r_E.

Séparation latérale, décorrélation interaurale, impressions spatiales

Dans toute cette partie II, la qualité de restitution a été discutée surtout à travers les performances de localisation. Mais il est une autre grande qualité attendue d'un système de restitution: c'est son aptitude à reproduire chez l'auditeur des impressions spatiales satisfaisantes, et notamment un effet d'enveloppement (Cf 1.3.4). Cet aspect a été abordé en 2.3.3 dans le cadre général de la restitution multi-canal en lien avec la question de la prise de son, puis en 2.4.4 concernant les systèmes ambisoniques d'ordre 1. D'après les arguments qui y sont exposés, il faut attendre que les systèmes d'ordres supérieurs permettent d'atteindre un degré d'impression spatiale de plus en plus proche de celui qui serait éprouvé dans la scène sonore originale: en améliorant la résolution angulaire des contributions élémentaires restituées, on améliore en particulier la séparation latérale, nécessaire à la préservation de la décorrélation interaurale, donc des impressions spatiales. Ce problème de séparation est surtout critique dans le domaine haute-fréquence où se manifeste le *cross-talk*, *i.e.* hors contrôle de reconstruction. Son amélioration peut être appréciée à travers celle de la directivité des microphones équivalents (Figures 3.13 et 3.14), mettant en faveur pour un ordre M donné les versions *max* r_E

23. Objection de Angelo Farina.

et même *in-phase*[24]. Il devrait être possible de quantifier la dégradation de la décorrélation interaurale dans des conditions de champ diffus en fonction du type de restitution ambisonique.

4.4 Conclusion

4.4.1 Développements théoriques, notions et propriétés émergentes

Bilan

Toute cette partie II a été consacrée à l'extension de l'approche ambisonique aux ordres supérieurs, abordant les différentes opérations de la chaîne, de l'encodage au décodage. Outre une description de l'encodage et des questions de conversion entre les différentes conventions (section 3.1.2), ce sont surtout les problèmes de décodage et de restitution que nous avons approfondis. Dans le souci de proposer de solutions de décodage adaptées aux différentes conditions d'écoute envisageables – d'une écoute individuelle en position optimale centrée à un auditoire s'étendant à proximité des haut-parleurs –, nous avons étendu aux ordres supérieurs les trois formes de décodage qui existaient jusque lors pour les systèmes d'ordre 1, à savoir:

- Un décodage basé sur le critère d'une reconstruction locale du champ acoustique (décodage dit "*basique*"), impliquant la spécification classique sur la propagation du front d'onde synthétisé localement: $\vec{V} = \vec{u}_S$ (critère dit de Makita).
- Un décodage basé sur un critère d'optimisation du flux d'énergie (propagation globale) associé à la restitution de chaque événement élémentaire (onde plane): $\vec{u}_E = \vec{u}_S$ et maximisation de r_E (décodage nommé "*max* r_E").
- Un décodage tel que les gains des haut-parleurs soient décroissants à mesure qu'ils sont écartés de la source virtuelle jusqu'à devenir nuls dans la direction opposée (décodage dit "*in-phase*", d'après la version proposée à l'ordre 1 par Malham).

Les deux critères relatifs aux vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ ont été argumentés en détail en 1.5. Ce sont ainsi trois familles de solutions de décodage que nous avons définies de façon générique, pour les configurations régulières et semi-régulières, 2D et 3D (sections 3.3.1, 3.3.2, 3.3.3). Les cas des configurations non-régulières (3.3.4) ou hémisphériques (3.3.5) ont été également abordés.

En préalable à la résolution des problèmes de décodage et de prise de son – traitée de façon moins approfondie –, il a semblé important d'expliciter des notions mathématiques qui leur sont sous-jacentes. Il s'agit notamment de la notion d'échantillonnage de la base d'harmoniques sphériques utilisée pour la représentation ambisonique et des propriétés de régularité de cet échantillonnage (section 3.2.3).

Les systèmes de restitution 2D ont été soumis à différentes méthodes d'évaluation objective au cours du chapitre 4. Etayées par quelques expériences d'écoute informelles, elles ont permis de valider partiellement l'intérêt des ordres supérieurs et la pertinence des critères de décodage. Lors de ces études, plusieurs suggestions ont été formulées qui pourraient orienter des tests subjectifs plus poussés, dans l'optique d'une validation complète.

Notions émergentes

De nombreuses notions ont émergé au cours de cet approfondissement de l'approche ambisonique. Elles se traduisent bien souvent sous la forme de propriétés duales.

24. Malgré un indice r_E identique à celui de version *basique*, l'absence de lobe arrière chez les directivités in-phase permet un meilleur effet de séparation latérale (Cf 4.1.2), notamment avec des configurations comportant des haut-parleurs parfaitement latéraux (Figure 4.11, ligne du bas).

Reconstruction contrôlée *versus* non-contrôlée, ou encore "**expansion locale *versus* caractérisation globale**". La qualité de la reconstruction locale se mesure à son expansion radiale *kr*, et les propriétés du champ hors contrôle de reconstruction s'estiment en termes de propagation globale de l'énergie[25], caractérisée de façon synthétique par le vecteur énergie $\vec{E}$. Ces deux mesures de la qualité du champ restitué sont intimement liées à la représentation ambisonique, c'est-à-dire la troncature de la décomposition en harmoniques sphériques (Section 3.2), et croissent avec son ordre M. Ce sont *deux manifestations acoustiques du degré de résolution spatiale* de la représentation/restitution, *perçues en termes de précision de l'image sonore*.

Restitution 2D *versus* 3D, ou encore "troncature cylindrique *versus* sphérique" (Section 3.2 et Table 3.10). Il apparaît qu'à ordre M égal, la résolution spatiale offerte dans le plan horizontal est moindre dans le cas d'une représentation/restitution homogène 3D (troncature sphérique) que dans le cas 2D (troncature cylindrique, restriction horizontale). Et ceci malgré un nombre de composantes 3D supérieur au nombre de composantes 2D: $K = (M+1)^2$ contre $K = 2M+1$. Cela étant, il manque une dimension à la restitution 2D.

Restitution: minimale (ou super-minimale) *versus* non-minimale (section 3.3.1). Les décodages que nous disons minimal ($N = K$) et super-minimal ($N < K$) sont "optimaux" (pour Poletti [Pol96a]) du point vue de la reconstruction locale, mais ne préservent pas les propriétés directionnelles de la propagation globale ($\vec{u}_E \neq \vec{u}_S$ en général): cela brise la cohérence des informations de localisation. Les sources virtuelles émergent sporadiquement comme sources réelles au passage des haut-parleurs: cela brise l'homogénéité de la restitution. Ce n'est plus le cas avec les configurations non-minimales ($N > K$), moyennant un décodage adéquat.

Configurations non-minimales: régulières, semi-régulières, non-régulières. La propriété de régularité doit être définie au sens de l'échantillonnage de la base d'harmoniques sphériques utilisée (cf 3.2.3), et rend triviale la définition du décodage (cf 3.3.2). S'accompagnant de la condition de non-minimalité ($N > K$), elle signifie que les propriétés locales (reconstruction) et globales (r_E) de la troncature du champ[26] au même ordre sont préservées. Le cas de semi-régularité entraîne une définition presque aussi simple du décodage, permettant de préserver que les propriétés directionnelles ($\vec{u}_E = \vec{u}_V = \vec{u}_S$), mais plus l'uniformité de la qualité (expansion locale *kr* et r_E) en fonction de la direction, ni d'ailleurs l'uniformité de l'énergie restituée, qui traduit en quelque sorte l'effort de reconstruction. Pour les configurations non-régulières, la résolution du décodage sous la contrainte ($\vec{u}_E = \vec{u}_V = \vec{u}_S$) n'est pas triviale et n'est peut-être pas toujours possible. [...]

Comportements asymptotiques de la reconstruction: *Ambisonics versus* Holophonie (section 4.2.3). Alors qu'avec les systèmes holophoniques comme le WFS (*Wave Field Synthesis*), la reconstruction du champ est de qualité uniforme sur la zone de restitution et progresse en terme de fréquence d'aliasing, la reconstruction se fait par expansion radiale avec Ambisonic. Ainsi l'approche ambisonique ne se montre pas comme une approche très "démocratique" (ou égalitaire), la position centrale étant toujours privilégiée. C'est d'ailleurs un "point de vue" unique – celui de la prise de son – qui est imposé et "extrapolé". Reste à préciser la façon dont ce point de vue peut ou doit être extrapolé pour les auditeurs excentrés...

Projection de l'image sonore: à l'infini ou sur le périmètre de haut-parleurs. Cette question, qui indique deux possibilités idéales pour l'extrapolation du point de vue central, a été discutée en 4.2.3.

Prise de son (section 3.4) ***versus* restitution.** Ces deux problèmes sont apparemment duaux: s'agissant pour l'un de mesurer le champ ambisonique ou pour l'autre de le reconstituer, mais utilisant tous deux un nombre fini de transducteurs répartis sur une sphère, ils font intervenir la notion d'échantillonnage du champ ambisonique et tirent profit de la propriété de régularité le cas échéant. Une propriété les distingue cependant, qui vient de la distance entre les transducteurs: la mesure du champ (prise de son) est affectée,

25. ... en considérant l'onde plane comme événement acoustique élémentaire et de référence.

26. Troncature biaisée ou non selon le type de décodage appliqué (*max* r_E ou *basique*), et sphérique ou cylindrique selon la dimension (3D ou 2D).

dans un domaine haute-fréquence, par un phénomène de repliement du spectre harmonique sphérique.

4.4.2 A suivre...

Validations et définition d'une échelle subjective

Dans le but d'une validation complète des solutions de décodage développées, il faudrait mener des tests d'écoute formels avec un nombre représentatif de sujets et une quantification des jugements par des notes. L'outil d'expérimentation et de démonstration présenté au chapitre 5, incorporant les techniques ambisoniques parmi d'autres techniques de spatialisation, pourrait servir et être encore adapté à cet effet. Son utilisation et quelques suggestions méthodologiques sont proposées en 5.5.2.

On peut attendre d'une telle étude qu'elle permette d'établir des correspondances entre une échelle objective de caractérisation de la restitution (notamment par l'indice r_E) et une échelle subjective (précision de la localisation). Elle devrait permettre également de préciser les domaines d'application des solutions de décodage en fonction des conditions d'écoute (Figure 3.4 et suggestions section 4.2.2).

Axes de recherche

L'optimisation du décodage pour des configurations non-régulières, en particulier celles qui privilégient la scène frontale (restitution 2D), et les configurations (3D) de type hémisphérique ou pyramidal, reste un problème pour lequel il n'a pas été dégagé de méthode à la fois générique et rigoureuse au sens des critères classiques. La prise en compte des composantes ambisoniques d'ordres supérieurs suscite depuis peu un vif intérêt, y compris dans le contexte de la prise de son multi-canal pour configuration "3/2". La *notion de représentation et de décodage hybride 2D ou 2D/3D pour les configurations non-régulières ou à symétrie partielle* (Cf 3.3.4 et 3.3.5) semble émerger comme une question-clé qui pourrait définir un axe de recherche passionnant pour les temps à venir.

A l'autre extrémité de la chaîne, la prise de son du champ ambisonique (section 3.4) constitue en quelque sorte le problème dual de la restitution. Là encore, l'intérêt pour les ordres supérieurs devrait bientôt se concrétiser dans ce domaine (note[24] page 104).

Troisième partie

Mise en oeuvre, tests et applications

Chapitre 5

Implémentation et incorporation dans une interface

5.1 Objectifs et conditions de réalisation

5.1.1 De l'utilité et de l'usage des développements techniques

Les outils logiciels présentés dans ce chapitre ont été développés pour répondre à plusieurs nécessités dans le contexte de cette thèse.

Il s'agit d'abord de démontrer et tester l'efficacité des techniques existantes, individuellement, par l'écoute et la manipulation temps-réel des sources virtuelles.

Conjointement et indissociablement d'une démarche de recherche et de prospection théorique, les expérimentations permettent de se faire une idée des qualités et carences de ces techniques et de dégager des aspects caractéristiques de chaque technique qui sont pertinents ou frappants sur le plan subjectif (à l'écoute). Dans bien des cas, de tels aspects ne se seraient pas imposés de façon évidente à la conscience au seul regard de préoccupations théoriques d'origine (notamment les critères de décodage ambisonique d'après Gerzon) ou d'évaluations objectives quantitatives (cf 4.1). Ces expériences d'écoute ont donc pu enrichir les évaluations objectives, voire corroborer les attentes théoriques par des interprétations subjectives, en même temps qu'orienter les préoccupations et investigations théoriques.

Ces outils doivent servir à réaliser des *tests subjectifs* – jusqu'ici informels – pour l'évaluation critique et comparative des décodages ambisoniques et des techniques de restitution en général, sur haut-parleurs comme au casque (haut-parleurs virtuels). Les écoutes au casque ont permis en particulier la comparaison du rendu ambisonique d'une source virtuelle (suivant l'ordre et le décodage) avec l'effet d'une source "réelle" unique (simulation binaurale directe).

Dans un cadre plus général, il s'agit de démontrer l'utilité et le potentiel des techniques de spatialisation – et tout spécialement ambisoniques – et de leur usage combiné, dans le contexte de la *navigation 3D* dans des scènes virtuelles et composites. Cela inclut, en complément des techniques présentées dans les chapitres précédents, la synthèse d'*effets de salle* et effets sonores "spéciaux" (effet Doppler), qui participe à l'immersion de l'auditeur dans l'espace sonore virtuel et augmente le "réalisme" de la restitution, tout en enrichissant les informations de positionnement spatial des sources (effet de distance, statique (réverbération) et dynamique (Doppler)). Un objectif particulier est de mettre à l'épreuve un potentiel offert par les techniques ambisoniques: le *mélange et la manipulation de sources audio de différentes natures, pour une restitution sur un dispositif quelconque*: sources monophoniques spatialisées, matériel stéréo conventionnel (deux canaux),

matériel multi-canal, et bien-sûr enregistrements ambisoniques originaux au format B[1].

Que ce soit à des fins d'expérimentation ou de démonstration, le développement d'interfaces, pour la manipulation des objets sonores ou pour le contrôle de paramètres des différentes techniques, représente un travail souvent au moins aussi conséquent que le développement des techniques de spatialisation.

«*A quoi cela pourrait encore servir, et que faudrait-il faire de plus?*»: cette question est abordée en conclusion (5.5)

5.1.2 Conditions de développement et d'expérimentation

Précisons tout d'abord que les développements sont réalisés dans un domaine purement numérique – et non analogique[2] – et de façon logicielle.

Il est évident que l'évolution des développements – et à plus fortes raisons des expérimentations – a été fortement conditionnée par les ressources matérielles et logicielles disponibles. C'est ainsi que les travaux présentés plus loin dans les sections 5.2, 5.3 et 5.4, correspondent aux différentes "générations" ou phases de développement suivantes.

Conditions d'origine pour les développements "première génération"

Le développement a d'abord eu lieu sur des stations de travail Sun (modèle?) et Silicon Graphics (Indigo2), sous l'environnement Unix. Il s'est fait exclusivement en langage C, les programmes étant exécutés en mode "console" (sans interface graphique) donc sans possibilité d'interaction temps-réel. Les entrées-sorties se faisant d'ailleurs sous forme de fichiers son, l'écoute avait obligatoirement lieu "en différé", après le traitement complet du ou des fichiers-son d'entrée. La carte son de l'Indigo2 n'offrant qu'une sortie stéréo, la restriction à une écoute au casque a requis l'artifice des "haut-parleurs virtuels" pour l'expérimentation des techniques ambisoniques. Seules quelques écoutes multi-canal "grandeur nature" (configuration 3/2) ont pu être réalisées à l'époque au studio-son voisin, moyennant un transfert des fichiers-son par le réseau informatique et une conversion des formats pour Macintosh[3].

C'est dans ces conditions qu'a été implémenté l'essentiel des techniques présentées en 5.2.

Seconde étape: interfaçage et manipulation temps-réel

Deux éléments clés ont offert la possibilité de tester de manière plus approfondie les techniques de spatialisation développée, grâce à la manipulation et l'écoute temps-réel du champ sonore:

- Un PC Pentium Pro 200 MHz (Hewlett Packard) et l'environnement *Visual C++* [Kru96], qui a permis le développement relativement rapide d'interfaces graphiques et de contrôle.
- Une carte numérique multi-sons de haute-qualité, conçue au CCETT pour des fins expérimentales et de démonstration par Pierre Urcun et Jean-Christophe Rault: 8 entrées mono / 8 sorties mono (UER/AES), échantillonnées sur 32 bits à 48 kHz par défaut (horloge interne, synchronisation possible avec les entrées, rééchantillonnage au besoin), transfert par le bus PCI.

Par la même occasion, le matériel sonore utilisable, jusque là restreint à des fichiers-son lus sur disque dur, a été enrichi grâce à l'accès à des sources extérieures:

- Lecteur CD externe à sortie numérique, pour apport de sources mono ou stéréo en entrée de la carte multi-son.

1. Ces enregistrements ont été gracieusement fournis par Dave Malham, Département des technologies de la musique, Université de York, Royaume-Uni.

2. Cela dit, nombre de décodeurs ambisoniques présents sur le marché sont analogiques.

3. Avec l'assistance de Jean-Yves Leseure, du CCETT.

– Lecteur-enregistreur multi-piste Tascam, pour apport de sources mono, mais surtout multi-canal, en entrée de la carte multi-son (usage occasionnel). Enregistrements[4] stockés sur cassette Hi8 (jusqu'à 8 voies), à 48 kHz.

– Lecteur JAZ (lecture sous forme de fichiers) pour apport d'enregistrements ambisoniques au format B[1] (usage très tardif).

Pour l'écoute temps-réel, le matériel utilisé en sortie de la carte multi-son a consisté en:

– Convertisseurs numérique/analogique: d'abord plusieurs convertisseurs disparates, puis par une console numérique (*Yamaha 03D*).

– Haut-parleurs: petites enceintes Genelec (modèle 10-29A) amplifiées (enceintes de proximité), de 2 à 6 selon les disponibilités (et avec le secours occasionnel de 2 petites Fostex).

– Sortie stéréo de la console vers un casque, pour l'écoute binaurale.

Développements ultérieurs en C++

Un langage orienté objet comme le C++ [référence?] offre au code développé une plus grande modularité et réutilisabilité que le langage C. Une fois leur incorporation effectuée au sein de l'interface *Visual*, le portage des techniques de spatialisation en C++ a donc été envisagé et entamé. Pour un certain nombre d'entre elles dont l'implémentation initiale est déjà assez bien structurée, le portage peut consister en une simple "encapsulation" des routines existantes au sein des nouvelles *classes* (Cf 5.4). D'autres techniques font l'objet d'une refonte plus radicale: c'est notamment le cas avec l'extension des techniques ambisoniques aux ordres supérieurs (sans limitation de l'ordre *a priori*), pour laquelle une certaine généricité est requise. Par ailleurs, une librairie de géométrie 3D (objets et opérations élémentaires) a été constituée. Conçue pour servir à terme à une synthèse d'effet de salle par modèle géométrique (réflexions précoces), elle est pour l'instant exploitée pour la visualisation 3D et la manipulation d'objets sonores au coeur d'une seconde interface (section 5.4.2), que nous avons couplée avec une version du spatialisateur de l'IRCAM portée en C++ par M.Emerit (CNET Lannion).

5.2 Implémentation des techniques de spatialisation

L'ensemble des techniques présentées ici correspond en quelque sorte au développement "première génération" réalisé durant le travail de thèse, utilisant essentiellement (et même exclusivement dans un premier temps) le langage C [Référence: Korn et Ritchie?]. On rappelle que ces développements "première génération" sont restreints à une restitution 2D (dans le plan horizontal). Ils ont été dans un second temps incorporés au sein d'une interface *Visual* sur PC (voir la section 5.3 suivante), et ont été à l'occasion étendus ou corrigés.

Le principe de la plupart de ces techniques a déjà été évoqué au cours de ce document, à l'exception de la synthèse d'effet de salle. Ce sont donc surtout des compléments techniques que nous apportons ici. De manière générale, tous les traitements évoqués sont effectués par blocs d'échantillons[5].

4. Un bon nombre de ces enregistrements nous proviennent de Radio-France.

5. Typiquement, 1024 échantillons, ce qui représente 21 ms à 48 kHz, le traitement étant cadencé par la carte multi-son (section 5.3).

5.2.1 Binaural et transaural

Filtrage binaural par convolution rapide

Comme cela a été évoqué à diverses reprises déjà, les HRTF que nous avons utilisées pour la synthèse binaurale sont celles proposées par Martin et Gardner [GM94], mesurées sur un mannequin KEMAR. L'échantillonnage original ayant été réalisé à 44.1 kHz, il a été nécessaire de rééchantillonner les réponses impulsionnelles (HRIR) à 48 kHz (pour la carte multi-son du CCETT) et à l'occasion à 32 kHz (pour l'utilisation d'autres cartes son). Pour chaque fréquence d'échantillonnage, on dispose d'un jeu de HRTF égalisées en champ diffus et d'un jeu de HRTF non-égalisées.

L'implémentation du filtrage binaural est basée sur un algorithme de convolution rapide, et s'inspire largement du code C mis à disposition par Martin et Gardner[6]. Cet algorithme consiste à remplacer l'opération de convolution discrète dans le domaine temporel:

$$y(n) = \sum_{k=0}^{N-1} h(k)\,x(n-k), \tag{5.1}$$

pour une réponse impulsionnelle h à N échantillons, par l'opération de multiplication terme à terme dans le domaine fréquentiel dual (Figure 5.1), moyennant des FFT et FFT inverse pour passer d'un domaine à l'autre. Cette méthode, appelée *overlap-discard* et connue depuis les années 70 (Rabiner), possède une variante nommée *overlap-add* plus couramment utilisée.

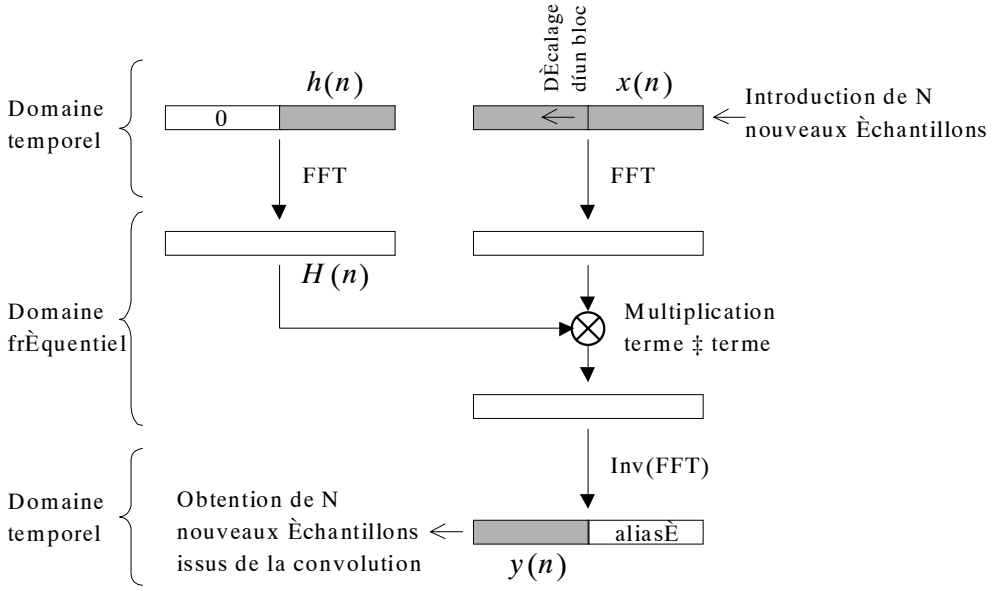

FIG. 5.1 – *Algorithme de convolution rapide (d'après [Gar96]). Les blocs grisés comprennent chacun N échantillons. Le résultat "brut" de la convolution (en bas) comprend un bloc affecté d'un aliasing temporel, dû au fait qu'elle est effectuée de manière circulaire sur les deux blocs de $x(n)$ (en haut). Ce bloc est donc ignoré.*

Dans notre application, les FFT (notées ici $H_L(\theta,\delta)$ et $H_R(\theta,\delta)$) correspondant aux différentes paires de HRIR sont calculées au départ – de la même manière que H est calculé à partir de la réponse $h(n)$ dans

6. `http://www.sound.media.mit.edu/KEMAR.html`

la figure 5.1 – et stockées une fois pour toutes. Le traitement binaural de chaque source virtuelle nécessite d'appliquer l'algorithme au signal s en question pour chacune des réponses gauche $H_L(\theta,\delta)$ et droite $H_R(\theta,\delta)$ correspondant à la position courante, fournissant une paire de signaux (s_L, s_R). Il est nécessaire de garder en mémoire le bloc du signal s injecté la fois précédente. Lors d'un changement de position qui entraîne un changement de paire de HRTF[7], l'opération est appliquée de surcroît avec la paire de HRTF correspondant à la nouvelle position, et un fondu (ou *fading*) temporel est réalisé entre les deux paires de signaux (s_L, s_R) obtenues.

En pratique, le traitement est effectué par blocs de $N = 128$ échantillons, et met donc en jeu des FFT à 256 points.

Filtrage transaural simplifié

Pour l'écoute transaurale sur deux haut-parleurs, un traitement chargé d'annuler le *cross-talk* est appliqué sur les signaux issus de la synthèse binaurale qui vient d'être présentée.

C'est une version simplifiée du filtrage transaural (section 2.5.2) que nous avons mise en oeuvre. Elle reprend la modélisation du *cross-talk* présentée dans [Gar95] d'après des suggestions de D. Griesinger[8]. La réponse ipsi-latérale $H_i(z)$ y est assimilée à l'identité, et la réponse contra-latérale $H_c(z)$ est modélisée par la combinaison d'un retard m/f_s (fréquence d'échantillonnage f_s), une atténuation g et un filtre passe-bas $H_{LP}(z)$:

$$\begin{aligned} H_i(Z) &= 1 \\ H_c(Z) &= g z^{-m} H_{LP}(Z), \qquad H_{LP}(Z) = \frac{1-a}{1-aZ^{-1}} \end{aligned} \tag{5.2}$$

Ce modèle étant établi, les structures envisageables pour inverser le *cross-talk* sont illustrées Figure 2.19 (page 126). Gardner [Gar95] préconise et utilise la structure *shuffler* (schéma du milieu droit). Nous avons quant à nous expérimenté l'*inversion directe* (schéma de droite), qui s'avère satisfaisante.

Pour des haut-parleurs en $\pm 30^\circ$, le paramétrage proposé par Gardner est le suivant: $g \approx 0{,}85$ pour l'atténuation interaurale pleine-bande, $m \approx \tau f_s$ avec $\tau \approx 0{,}2$ ms comme retard interaural, et une fréquence de coupure d'environ 1000 Hz pour le filtre passe-bas. Dans l'interface présentée section 5.3, une boîte de dialogue est dédiée au contrôle de ces paramètres, pour un ajustement temps-réel par l'utilisateur.

5.2.2 Pan-Pot horizontal

Il était incontournable de réaliser, ne serait-ce qu'à titre de comparaison avec l'approche ambisonique, le principe très classique de pan-pot par paire de haut-parleurs (*Pair-Wise Pan-Pot*). C'est le VBAP [Pul97] que nous avons choisi comme stratégie. Sa mise en oeuvre ne réserve pas de surprise et est presque aussi simple que la description de son principe en 2.3.2. Prenons l'exemple de la position θ_a entre les haut-parleurs placés en ϕ_2 et ϕ_3 (Figure 5.2), les gains G_2 et G_3 associés à ces haut-parleurs sont tels que:

$$\begin{aligned} G_1 &= \frac{G'_1}{\sqrt{(G'_1)^2+(G'_2)^2}} \\ G_2 &= \frac{G'_2}{\sqrt{(G'_1)^2+(G'_2)^2}} \end{aligned} \quad \text{où} \quad \begin{aligned} G'_1 &= \frac{\cos\theta_a \sin\phi_2 - \sin\theta_a \cos\phi_2}{\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2} \\ G'_2 &= \frac{-\cos\theta_a \sin\phi_1 + \sin\theta_a \cos\phi_1}{\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2} \end{aligned} \tag{5.3}$$

7. Rappelons que le HRTF sont échantillonnées pour un ensemble discret de directions, et en particulier tous les 5 degrés dans le plan horizontal.

8. Cette méthode est tout à fait analogue au procédé TRADIS de Atal et Schroeder, implémenté par Siebrasse dans les années 70.

En complément, nous signalons seulement à travers la figure 5.2, la façon dont l'interpolation des gains est réalisée lors des déplacements de la source virtuelle, dans notre implémentation.

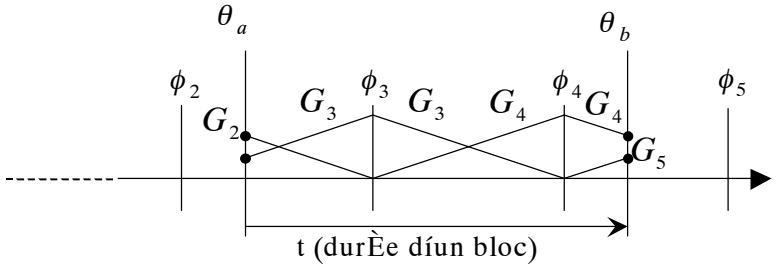

FIG. 5.2 – *Lorsque la source virtuelle a effectué un déplacement (de l'angle θ_a vers l'angle θ_b) depuis la mise à jour précédente (avant le bloc d'échantillons qui vient d'être traité), une interpolation linéaire des gains des haut-parleurs est réalisée sur la durée du nouveau bloc à traiter. Quand θ_a et θ_b ne se trouvent pas dans le même secteur angulaire, voire pas dans des secteurs angulaires adjacents (déplacement brutal), l'interpolation a lieu par sous-blocs, délimités par les angles ϕ_i des haut-parleurs intermédiaires.*

5.2.3 Techniques ambisoniques

Généralités

Les développements en langage C ("première génération") concernant les techniques ambisoniques se réduisent à une implémentation aux ordres 1 et 2 pour une restitution horizontale (ou 2D). Dans cette interface, la structure choisie pour le décodeur est celle qui convient aux cas de décodages dits "réguliers", à savoir: une seule matrice matrice de décodage, dite "basique", et une correction préalable des composantes ambisoniques par des gains ou des filtres pour obtenir les solutions de décodage dérivées (Figure 3.5). La définition (par pseudo-inverse) de la matrice basique et le filtrage correctif (*shelf-filtering*) sont détaillés plus loin.

Comme les autres opérations de spatialisation, le traitement est réalisé par bloc (typiquement 1024 échantillons). Pour les objets sonores individuels, l'encodage est effectué source par source, avec une interpolation linéaire des coefficients d'encodage sur la durée du bloc en cas de changement de position. L'encodage du matériel multi-canal prend en compte les cinq canaux conjointement.

Le *shelf-filtering*

Plusieurs structures de filtres numériques ont été envisagées pour la réalisation du filtrage par palier (*shelf-filtering*). Nous présentons une adaptée au cas d'un étagement en deux paliers, spécifiés par des gains respectivement basse- et haute-fréquence g_{LF} et g_{HF}, avec pour fréquence de transition f_{tr}. Il s'agit d'un filtre du premier ordre de type Regalia/Mitra, déjà utilisée pour diverses tâches au sein du Spatialisateur de l'Ircam.

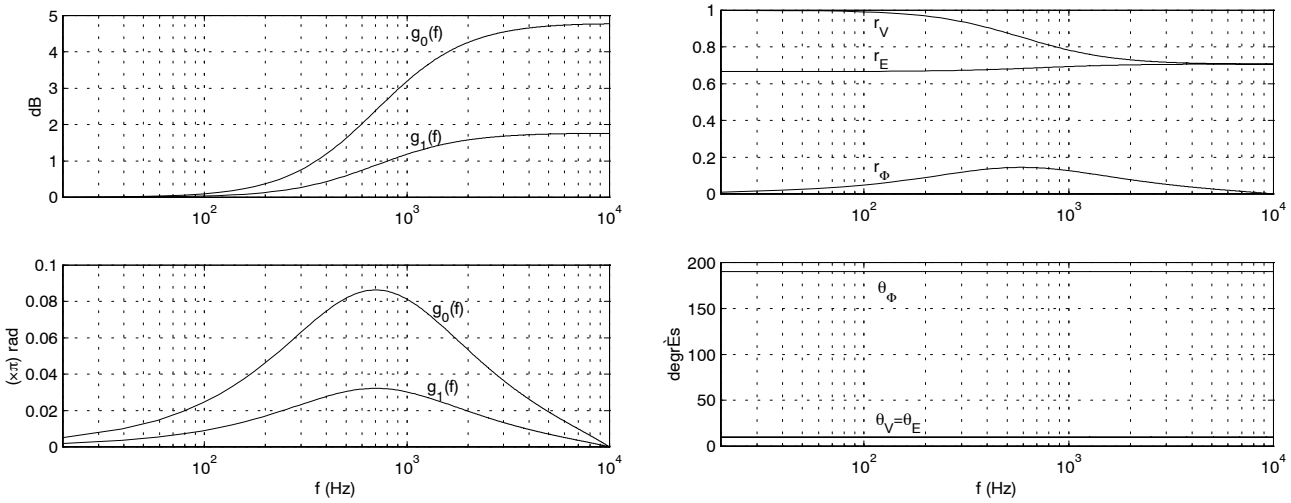

FIG. 5.3 – *Exemple de correction de décodage par* shelf-filtering*: réponses de filtres $g_0(f)$ et $g_1(f)$ et effet sur les vecteurs vélocité (partie réelle de $\vec{V}$, décrite par θ_V et r_V), phasiness (partie imaginaire $\vec{\Phi} = \Im(\vec{V})$ décrite par r_Φ et θ_Φ) et énergie $\vec{E}(r_E, \theta_E)$.*

On note f_s la fréquence d'échantillonnage. Le filtre de paramètres (g_{LF}, g_{HF}, f_{tr}) prend alors la forme:

$$H(Z) = c\frac{1+bZ^{-1}}{1+aZ^{-1}}, \quad \text{avec} \quad \begin{cases} a = \dfrac{r-1}{r+1} & \bar{\omega}_{tr} = 2\pi\frac{f_{tr}}{f_s} \\ b = \dfrac{kr-1}{kr+1} & \text{et} \quad k = \dfrac{g_{LF}}{g_{HF}} \\ c = \dfrac{kr+1}{r+1} g_{HF} & r = \dfrac{\tan(\bar{\omega}_{tr}/2)}{\sqrt{k}} \end{cases} \tag{5.4}$$

A la fréquence de transition, la réponse en dB se situe à mi-chemin entre les deux paliers.

Pour un système d'ordre $M, M+1$ filtres $(g_0(f), \ldots, g_M(f))$ sont définis suivant ce modèle avec la même fréquence d'échantillonnage, pour s'appliquer aux composantes d'ordres m respectifs. Il est indispensable que ces filtres présentent la même réponse de phase sur chacun des paliers pour que le décodage soit valide. La figure 5.3 présente l'exemple d'une correction pour position d'écoute centrée[9] s'appliquant à un système d'ordre 1 et une configuration hexagonale régulière, la fréquence de transition étant fixée à $f_r = 700$ Hz. Un léger écart de phase est observé entre $g_0(f)$ et $g_1(f)$, s'accentuant lors de la transition. Il est responsable de l'émergence d'un terme de *phasiness* $\vec{\Phi}$ (partie imaginaire du vecteur vélocité, équation 1.23 page 23), qui reste marginal. On note que la transition le long de l'axe des fréquences est très douce: le module r_V décroît progressivement à partir de 100 ou 200 Hz pendant que r_E croît pour converger vers une valeur commune en haute-fréquence. Les directions θ_V et θ_E sont quant à elles préservées sur toute la bande de fréquence. Une transition plus rapide pourrait être obtenue en mettant en série deux filtres du même type, de paramètres $(g_{LF}, \sqrt{g_{LF}g_{HF}}, f_{tr})$ et $(1, \sqrt{g_{HF}/g_{LF}}, f_{tr})$. Ce filtre récursif d'ordre 2 reste de coût très raisonnable.

Pour un étagement en trois paliers (gains g_{LF}, g_{MF} et g_{HF}, fréquences de transition f_1 et f_2), il suffirait de mettre deux filtres en série, de paramètres respectifs (g_{LF}, g_{MF}, f_1) et $(1, g_{HF}/g_{MF}, f_2)$.

9. Décodage *basique* et normalisation en amplitude pour les basses-fréquences, décodage *max* r_E et normalisation en énergie pour les haute-fréquences.

Calcul de la pseudo-inverse pour le décodage

Le calcul de la pseudo-inverse d'une matrice $\mathbf{X}$, intervenant dans la définition du décodage (3.63), repose sur sa décomposition en valeurs singulières (SVD: *Singular Value Decomposition*):

$$\mathbf{X} = \mathbf{U}.\mathbf{S}.\mathbf{V}^t, \tag{5.5}$$

où $\mathbf{S}$ est une matrice diagonale de même dimension $m \times n$ que $\mathbf{X}$, et où $\mathbf{U}$ et $\mathbf{V}$ sont des matrices unitaires, *i.e.* telles que: $\mathbf{U}.\mathbf{U}^t = \mathbf{I}_m$ et $\mathbf{V}.\mathbf{V}^t = \mathbf{I}_n$. En notant $\mathbf{S}^{-1}$ la matrice diagonale de dimension $n \times m$ dont les éléments diagonaux sont les inverses de ceux de $\mathbf{S}$, la pseudo-inverse $\mathbf{Y}$ de $\mathbf{X}$ s'écrit:

$$\mathbf{Y} = \text{pinv}(\mathbf{X}) = \mathbf{V}.\mathbf{S}^{-1}.\mathbf{U}^t \tag{5.6}$$

L'implémentation de la SVD a été réalisée en langage C d'après [PTVF92].

Restitution ambisonique en mode binaural

La restitution ambisonique au casque telle nous l'avons conçue repose à la base sur la méthode des haut-parleurs virtuels, comme cela a été présenté en 3.1.3. Pour un coût de traitement minimal et indépendant du dispositif virtuel de restitution, des fonctions transfert des composantes ambisoniques vers les signaux binauraux sont calculées dès le départ (3.30). Dans le cas présent, elles sont notées L_W, R_W, L_X, ..., L_V, R_V. Elles sont calculées sur la base d'un *décodage basique*, l'optimisation du décodage étant assurée par une correction préalable des composantes ambisoniques (Figure 3.5). Les HRTF utilisées sont les mêmes que pour la synthèse binaurale directe (Cf 5.2.1) et l'opération de filtrage est réalisée par le même algorithme de convolution rapide (Figure 5.1).

La plupart du temps, on devrait pouvoir tirer profit de la symétrie – très courante – du dispositif virtuel pour diviser le coût de calcul par deux (Figure 3.6), grâce à la redondance (3.31) des fonctions de transfert. Mais pour laisser la possibilité de tester des configurations non-symétriques à des fins expérimentales, nous avons conservé une structure où sont différenciées les réponses gauche et droite, requérant donc $2K = 4M + 2$ opérations de filtrage (convolution rapide) pour un nombre K de canaux ambisoniques.

5.2.4 Module de réverbération tardive

Approche générale

Une façon de reproduire un effet de réverbération peut simplement consister à mesurer la réponse impulsionnelle d'une salle, puis en effectuer la convolution avec le signal à spatialiser. Cette méthode a le défaut d'être coûteuse[10], et malgré les algorithmes de convolution rapide, on peut se trouver confronté à un compromis difficile entre moindre coût (algorithme de la figure 5.1) et réduction du temps de latence (algorithme décrit dans [Gar96][11]). D'autre part, parce qu'il est important de traduire la décorrélation des signaux perçus aux oreilles en situation naturelle d'écoute, il faudrait multiplier les réponses à convoluer.

Dans le domaine du traitement numérique du signal, il est bien connu que pour obtenir un résultat semblable, il est souvent plus économique de procéder à un filtrage récursif (donnant lieu à une réponse impulsionnelle infinie: RII) qu'à un filtrage convolutif (à réponse impulsionnelle finie: RIF). Suivant une idée similaire et par analogie à un schéma de production des réflexions dans un résonateur (une salle, un conduit),

10. Elle est d'autant plus coûteuse que la réponse impulsionnelle est longue, *i.e.* que la salle a un grand temps de réverbération, ce qui pourtant est souvent apprécié.

11. Cette publication a semble-t-il été précédée d'un dépôt de brevet par la société *Lake*.

les algorithmes de réverbération artificielle consistent à réinjecter dans un réseau de lignes à retard (mémoires tampon pour des signaux) ses propres signaux de sortie en plus des signaux d'entrée, avec des retards qui pourraient correspondre aux trajets aller-retour dans le résonateur afin d'en générer les modes propres et harmoniques[12], et des atténuations (gains ou filtres simples) qui reflètent l'absorption accompagnant les réflexions. La figure 5.4 présente le schéma d'une cellule unique qui peut modéliser l'effet de résonance d'un conduit linéaire. Ce résonateur simple produit cependant un effet de coloration très marqué (modes propres

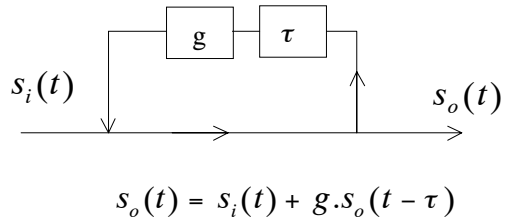

FIG. 5.4 – *Un des schémas possibles pour un résonateur simple: réinjection du signal de sortie avec un retard* τ *et une atténuation (gain g, remplaçable par un filtre passe-bas). Il s'agit d'un filtre en peigne.*

harmoniques ou quasi-harmoniques) qui ne convient pas à notre application. Pour un effet de réverbération convenable, il est nécessaire de faire intervenir plusieurs retards différents, de préférence associés à des lignes différentes, et de procéder à des réinjections mutuelles entre lignes pour un meilleur mélange des modes de résonance et une densification temporelle plus rapide. Pour reprendre l'analogie avec le schéma des réflexions dans une salle, chaque noeud du réseau ainsi constitué (voir par exemple plus loin la figure 5.5) pourrait correspondre à un "réflecteur" (paroi de la salle) qui renvoie et distribue les ondes incidentes à tous les autres réflecteurs [Jot92]. En général, les lignes sont alimentées en entrée par le signal original à spatialisé et des versions retardées, atténuées, voire déjà décorrélées, qui correspondent typiquement au groupe des réflexions précoces.

Propriétés de la réverbération tardive et diffuse

Plutôt que de chercher à imiter un *schéma "géométrique"*, même simplifié, de production des réflexions multiples dans une salle, il est préférable pour un rendu à la fois efficace et acceptable, de s'attacher à ce que les sorties du réverbérateur satisfassent les *propriétés statistiques* de la réverbération diffuse, telles qu'elles ont été rappelées en 1.2.4.

Ces propriétés relèvent des trois aspects suivants:

1. **Propriétés spatiales et directionnelles:** la réverbération diffuse étant caractérisée par un ensemble de réflexions –dense temporellement, mais pas strictement simultanées– venant de toutes les directions (isotropie), le module de réverbération doit fournir *plusieurs sorties décorrélées*, que l'on distribuera, soit directement au niveau des oreilles (présentation binaurale), soit sur plusieurs haut-parleurs ou encore suivant plusieurs directions s'il s'agit d'un encodage ambisonique.
2. **Propriétés temporelles**: le module de réverbération doit assurer une *densité temporelle* suffisante d'impulsions, en réponse à une impulsion originale, et traduire par ailleurs une décroissance temporelle exponentielle avec un contrôle adéquat du temps de réverbération.
3. **Propriétés spectrales et modales**: l'absorption moyenne dépendant en général de la fréquence (plus importante pour les hautes-fréquences), la décroissance exponentielle doit pouvoir traduire un *temps*

12. Ce n'est pas une obligation, et ce n'est pas d'après cette vision des choses que bien des réverbérateurs artificiels sont conçus.

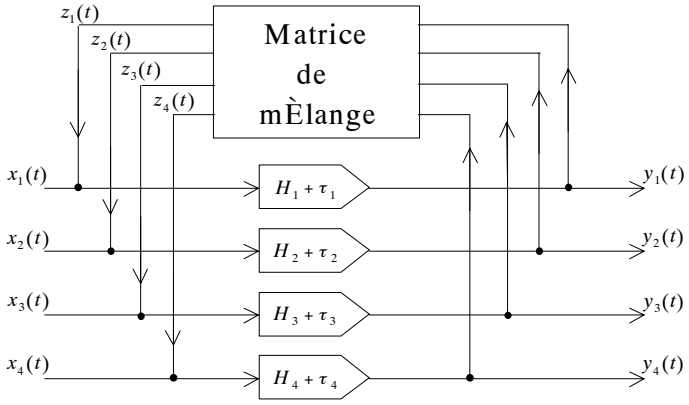

FIG. 5.5 – Feedback Delay Network *(Réseau de Retards Bouclés) d'après [Jot92]: à chaque ligne i (mémoire tampon) du réseau (ici au nombre de quatre pour la clarté du schéma), sont associés un retard τ_i et un filtre $H_i(f)$, ces deux éléments étant rassemblés sous le nom de* retard absorbant.

de réverbération variable suivant la fréquence. Cela se traduit sur le spectre global d'énergie (sur toute la réponse) par une décroissance en fonction de l'énergie. D'autre part, la *densité modale* doit être suffisante pour éviter que l'émergence individuelle de certains modes soit perceptible et induise un effet de coloration non recherché.

Le temps de réverbération –fonction de la fréquence– peut être approximativement estimé d'après les caractéristiques géométriques et physiques (absorption moyenne) de la salle, par les formules de Eyring (1.40) ou de Sabine (1.41). Il se peut que la synthèse artificielle d'effet de salle doive répondre à des attentes esthétiques arbitraires, et non plus à l'objectif d'imiter l'effet d'une salle existante. Dans ce cas, le T_{60}, les caractéristiques énergétiques de la réverbération tardive, ainsi que les autres éléments de la réponse impulsionnelle qui sont reconnus pour avoir un rôle structurant sur le plan de la perception (réflexions précoces et *cluster* [Jot97]), doivent pouvoir être déterminés d'après des attributs subjectifs de type: "présence de la salle", "chaleur", etc... Des relations entre "paramètres perceptifs" et "paramètres énergétiques et temporels" ont été établies à cette fin par des chercheurs de l'Ircam [JKM+93]. Ce type de contrôle est à nouveau évoqué en 5.4.2 à propos de l'incorporation d'une version du *Spatialisateur* de l'Ircam dans une interface 3D.

Le réseau de retards bouclés [Jot92]

Le réseau de retards bouclés (*Feedback Delay Network* [Jot97], Figure 5.5) proposé par Jot [Jot92] a une structure qui permet de répondre de façon très rationnelle aux trois propriétés présentées plus haut, et qui offre un contrôle assez direct des paramètres statistiques. C'est cette structure que nous avons choisie pour réaliser le module de réverbération tardive.

Définition des filtres absorbants. Notons $\sigma(f)$ le taux de décroissance exponentielle associé au temps de réverbération $T_{60}(f)$ pour une fréquence f donnée. La relation entre ces deux quantités s'exprime ainsi:

$$20\log_{10}(e^{-\sigma(f)T_{60}(f)}) = -60 \text{ dB} \qquad \Leftrightarrow \qquad T_{60}(f) = \frac{3\ln 10}{\sigma(f)} \tag{5.7}$$

Rapportée à chaque retard τ_i, l'atténuation requise est réalisée avec un filtre H_i de spectre d'amplitude:

$$|H_i(f)| = \sigma(f)\frac{T_{60}}{\tau_i(f)}, \tag{5.8}$$

en admettant que le retard éventuellement introduit par H_i est minime par rapport à τ_i.

Compromis densité modale / densité temporelle. Disposant d'un nombre plus ou moins limité de lignes –compte-tenu des capacités de calculs disponibles–, on comprend aisément que le choix des retards relève d'un compromis délicat entre densité temporelle et densité modale lorsque l'on a affaire à des temps de réverbération relativement longs: d'un côté, la recherche d'une bonne densité temporelle suggère d'utiliser des retards assez courts; de l'autre côté, le taux d'amortissement requis pour chaque retard est d'autant plus faible ($|H_i|$ d'autant plus élevé) que le temps de réverbération T_{60} est long et que le retard τ_i est court (5.8), ce qui favorise l'émergence des modes et l'effet de coloration. Précisons ce compromis quantitativement:

Pour rappel succinct, d'après [Jot92], la densité modale D_m (en "Hz^{-1}", c'est-à-dire homogène à un temps en secondes) peut être grossièrement estimée comme la somme des retards τ_i:

$$D_m = \sum \tau_i \tag{5.9}$$

La densité temporelle (d'énergie) D_e est quant à elle définie par:

$$D_e = \sum \frac{1}{\tau_i} \tag{5.10}$$

Pour exprimer le compromis entre ces deux densités en fonction du nombre de retards et de lignes, on retient la formule approximative, valable lorsque les retards sont dans des rapports proches de 1:

$$N \simeq= \sqrt{D_m.D_e} \tag{5.11}$$

Choix des retards. Pour éviter l'émergence accidentelle de modes de résonance qui serait due à une coïncidence cyclique de combinaisons de retards, ou encore à des rapports harmoniques trop évidents entre les retards élémentaires τ_i, il est suggéré de baser la définition des retards sur une suite de nombres premiers entre eux (algorithme `delgen`, pour "*delay generator*"). Néanmoins, on constate en pratique que le non-respect de cette recommandation n'est pas si critique.

Matrice de mélange. Le rôle de cette matrice est de réaliser un mélange optimal des signaux avant leur réinjection, ceci afin d'accroître à la fois la densité temporelle et la densité modale au sein de chaque signal, et en outre la *décorrélation* entre les signaux. Elle doit permettre d'éviter une cyclicité trop évidente du passage d'un même signal dans une même ligne. Puisque l'atténuation liée au temps de réverbération n'est contrôlée qu'au niveau des retards absorbants, il est important, d'autre part, que cette matrice *préserve l'énergie* des signaux, c'est-à-dire qu'elle soit *unitaire*. Ces différentes exigences, combinées avec le souci d'un coût de calcul réduit, ont fait porter un intérêt certain [Jot92] aux *matrices de Hadamard*, matrices carrées dont le nombre de lignes ou de colonnes est un puissance de deux, et dont l'opération de multiplication se prête à un *calcul dichotomique rapide* (algorithme "en papillon", à la manière des FFT). Le schéma suivant présente un procédé générique pour les définir:

$$\mathbf{H}_2 = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{H}_4 = \begin{pmatrix} +\mathbf{H}_2 & +\mathbf{H}_2 \\ +\mathbf{H}_2 & -\mathbf{H}_2 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{H}_8 = \begin{pmatrix} +\mathbf{H}_4 & +\mathbf{H}_4 \\ +\mathbf{H}_4 & -\mathbf{H}_4 \end{pmatrix} \quad \Longrightarrow \quad \text{etc...} \tag{5.12}$$

Pour vérifier la propriété d'unitarité, chaque matrice $\mathbf{H}_N$ doit être pondérée par un facteur de normalisation $\sqrt{1/N}$. D'autres familles familles de matrices sont également discutées dans [Jot97] (matrices de *Householder* [Jot92] et *matrices circulantes* [RS97]).

Implémentation réalisée et utilisation simplifiée

L'algorithme tel qu'il vient d'être présenté a été implémenté en langage C, pour les choix de nombre de lignes suivants: 4, 8, et 16. La partie "atténuation" des retards absorbants est modélisée dans chaque cas

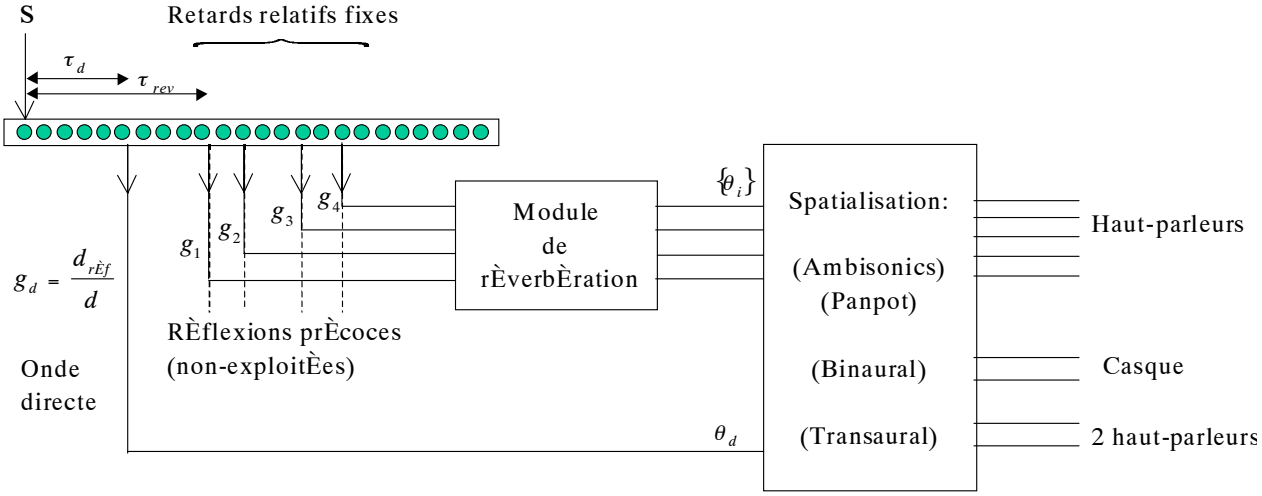

FIG. 5.6 – *Utilisation du module de réverbération au sein d'une structure globale pour la spatialisation de sources sonores avec effet de salle simplifié.*

par des filtres récursifs de différents ordres: cellule d'ordre 2, cellule d'ordre 1, cellule d'ordre 1 purement récursive. Le choix d'un filtre de bas ordre, donc de moindre coût, signifie qu'on modélise moins bien la variabilité de T_{60} en fonction de la fréquence. Le choix de cellules d'ordre 2 permet un contrôle du T_{60} en trois points de l'axe des fréquences: par exemple en très basses-fréquences (0 Hz), à la fréquence de Nyquist, et à une fréquence intermédiaire.

Le coût de calcul pourrait encore être réduit en remplaçant chaque filtre par un simple gain, et en effectuant une égalisation spectrale globale des signaux produits, mais il faut se souvenir que cette méthode approximative ne réalise en toute rigueur qu'un seul T_{60} indépendant de la fréquence. Ces différentes variantes d'implémentation, présentes dans les versions originales du module *Spat* du spatialisateur de l'Ircam [JKM+93], ont depuis été portées en langage C++ par Marc Emerit (CNET Lannion). Nous évoquerons l'utilisation de cette version complète du *Spat* au sein de l'interface 3D réalisée au cours de cette thèse.

L'utilisation et l'intégration de ce module de réverbération ont d'abord été orientées par le souci d'une validation urgente et d'une exploitation rapide. C'est donc un modèle très simplifié de synthèse d'effet de salle que nous présentons ici.

Comme l'indique la figure 5.6, l'effet de salle consiste à ajouter au son direct les sorties du seul module de réverbération tardive. Le *buffer* accordé au son direct et les entrées du module de réverbération puisent dans la même ligne à retards variables (voir 5.2.5), elle-même alimentée par le signal S de la source virtuelle considérée. Le retard τ_d et le gain g_d associés au son direct dépendent de la distance d de l'auditeur à la source virtuelle conformément aux lois de la propagation sphérique: $\tau_d = d/c$ et $g_d = d_{\text{réf}}/d$, où la distance de référence $d_{\text{réf}}$ est typiquement fixée à 1 m. Ce signal est alors soumis à l'une des techniques de pan-pot (ambisonique, VBAP, binaural, ...) en fonction de la direction θ_d de la source virtuelle. Les signaux injectés en entrée du module de réverbération pourraient être exploités de la même façon que le son direct, en tant que réflexions précoces, mais ce n'est pas le cas dans l'implémentation actuelle. Les retards qui leur sont associés sont choisis arbitrairement *fixes par rapport à* τ_d. Leurs gains sont identiques et déterminés soit arbitrairement par l'utilisateur, soit d'après une formule de raccordement énergétique décrite plus loin. Les sorties du réverbérateur sont alors réparties "panoramiquement" par le module de pan-pot.

S'il y a plusieurs sources virtuelles, le même module de réverbération reçoit en entrée les groupes de

signaux venant des différentes lignes à retards.

Paramétrage

A l'aide de l'interface présentée plus loin, l'utilisateur peut déterminer directement les paramètres $T_{60}(f)$ (en deux ou trois fréquences f) et l'atténuation énergétique à l'entrée du module de réverbération, ou bien spécifier des paramètres physiques, tels que:

- Dimensions $L_x \times L_y \times L_z$ de la salle supposées parallélépipédique, ou bien surface S et volume V.
- Absorption moyenne $\bar{\alpha}(f)$ par les parois (en deux ou trois fréquences).
- Prise en compte ou non de l'absorption atmosphérique $\mu(f)$ (d'après [ANS78]).

C'est alors la formule de Eyring (1.40) qui est utilisée pour déterminer le T_{60}. Le raccordement énergétique est calculé automatiquement, d'après la densité d'échos D_e propre au réseau de retards bouclés (5.10) et celle correspondant au modèle de salle parallélépipédique (1.43), ce qui donne un gain $g_{\text{réverb}}$ à l'entrée de chaque ligne:

$$g_{\text{réverb}} = \sqrt{\frac{4\pi c d_{\text{réf}}^2}{D_e V}} \tag{5.13}$$

Le jeu de retards du FDN peut être modifié à la main, ou bien généré automatiquement en donnant le retard le plus court et le retard le plus grand. A cause de l'absence de réflexions précoces et de *cluster*, on est contraint de fixer le plus petit retard assez bas. De manière générale, la simplicité excessive du modèle d'effet de salle présenté ici oblige à ajuster les paramètres à l'oreille, et notamment le raccordement énergétique[13], pour obtenir un effet satisfaisant.

Finalement, quelques jeux de paramètres (jeu de retards compris) ont été prédéfinis pour des effets de salles typiques: “cathédrale", “studio", “salle de bain".

5.2.5 Ligne à retard variable: effet Doppler

L'utilisation d'une ligne à retard variable s'impose pour la modélisation des effets dynamiques de rapprochement et d'éloignement d'une source sonore, bien connus sous le nom d'*effet Doppler*: lors du rapprochement de la source par exemple, le signal reçu par l'oreille subit une contraction temporelle qui se traduit par un déplacement de la hauteur tonale perçue vers les aigus.

La ligne à retard variable que nous avons implémentée est relativement simple, elle est schématisée Figure 5.7. Elle se greffe sur la structure d'un *buffer* circulaire constitué de N blocs de T échantillons. Lorsque le retard τ (ici en nombre d'échantillons) varie entre le début et la fin (retard τ') de la lecture du bloc, $T' = T + \tau - \tau'$ échantillons sont parcourus par le pointeur de lecture pour en retenir $T \neq T'$. Il arrive alors que ce pointeur de lecture se situe entre deux échantillons, auquel cas l'échantillon retenu est calculé par interpolation linéaire des deux. Lorsque le retard τ se stabilise, il est arrondi à un nombre entier d'échantillons afin d'éviter un effet de filtre passe-bas (interpolation figée).

Cette opération ne réalise qu'une approximation assez grossière de la compression ou la dilatation temporelle[14], mais l'expérience montre qu'elle remplit son rôle de façon tout à fait acceptable. Les changements de hauteur tonale sont effectifs, et l'on évite les effets de clic qui apparaîtraient s'il n'y avait pas du tout d'interpolation.

En option, une réallocation dynamique du *buffer* circulaire est possible au besoin, si le nouveau retard spécifié est excessif. Les cas de retournement temporel ($T' < 0$) sont gérés lors de variations très rapides

13. Le gain donné par la formule (5.13) est souvent excessif pour les petits volumes.
14. De meilleures approximations peuvent être obtenue avec des interpolations d'ordres supérieurs, par exemple.

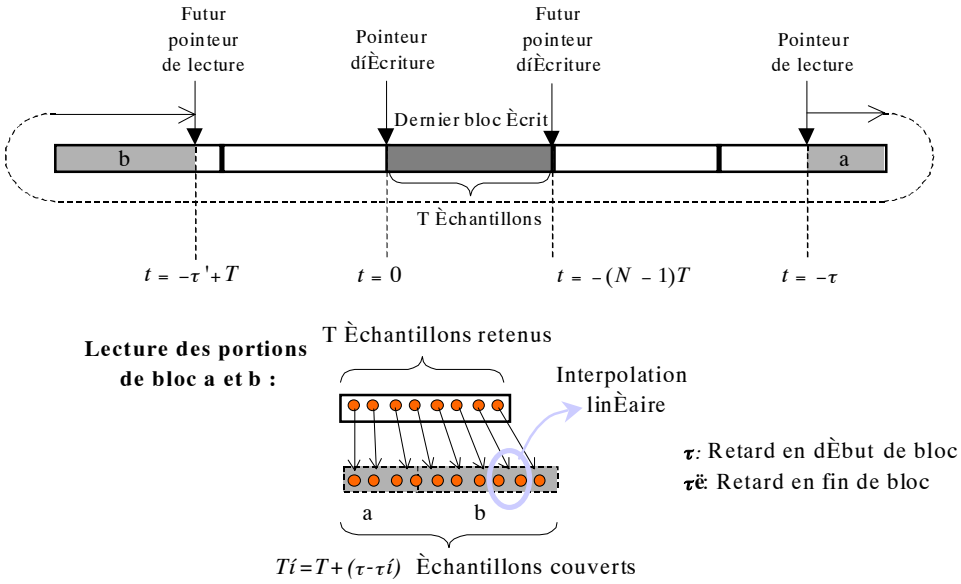

FIG. 5.7 – *Principe d'une ligne à retard variable.*

(supersoniques!). Enfin, plusieurs pointeurs de lecture à retard variable peuvent être greffés sur le même *buffer* circulaire.

5.3 Incorporation dans une interface: un outil de test et de démonstration

5.3.1 Architecture du programme

Généralités

L'interface se présente comme une application *Windows* de type "Boîte de dialogue", développée sous *Visual C++*. A la différence des applications MS-DOS, les saisies de l'utilisateur sont gérées par l'application *via* des messages émis par le système d'exploitation Windows. D'après ces messages ou événements – associés par exemple à un *clic* sur un bouton, le déplacement d'un curseur (*slider*) ou la saisie de caractères dans la boîte de dialogue – se déclenchent des fonctions[15] liées à l'objet "Boîte de Dialogue" en question.

Description globale

Pour simplifier la description (voir également Figure 5.8), l'application est *d'une part* composée d'un ensemble de boîtes de dialogue prenant en compte les actions de l'utilisateur de façon *événementielle*, et comprend *d'autre part* une fonction dédiée aux traitements de spatialisation et aux entrées-sorties audio, exécutée *en parallèle* sous forme de *Thread* (processus). Il s'agit ainsi d'une application *multithread*.

Une boîte de dialogue principale permet d'ouvrir les autres, permet l'exécution ou l'interruption du *thread* (Boutons "*Marche*" et "*Arrêt*"), et aussi la sélection ou la commutation des techniques de spatialisation à appliquer lors du traitement. Les boîtes de dialogue peuvent être regroupées en trois catégories:

– Celles dédiées à la saisie ou spécification des entrées-sorties (fichiers et carte), au paramétrage de la

15. Dans un jargon C++: "des fonctions membres de la classe dont l'objet "Boîte de Dialogue" est une instance.

configuration de restitution, à certaines initialisations (chargement des HRTF pour le filtrage binaural)... Beaucoup sont accessibles sous formes d'onglets. Dans l'implémentation actuelle, leur accès est interdit pendant l'exécution du thread.

– Celles dédiées au contrôle temps-réel des paramètres pour le traitement audio: décodeur ambisonique, réverbérateur, filtrage transaural.
– Celles dédiées au contrôle des sources virtuelles: leur position relative, leur trajectoire, dont la gestion fait l'objet de quelques particularités évoquées en 5.3.2.

La durée de vie du *thread* correspond à l'exécution de la fonction appelée sur commande du bouton "*Marche*" – appelons-là `ProcessAudio3D`. Cette fonction commence par une phase d'allocation mémoire et d'initialisations, se poursuit par une boucle "infinie" où sont effectués les traitements audio par bloc, et se termine par une phase de désallocation (et fermeture de carte et fichiers-son). Le corps de la boucle comprend lui-même trois grandes phases, dont certains aspects sont détaillés plus loin:

– La mise à jour des paramètres de spatialisation (position des sources, choix et paramètres des techniques), à partir de variables modifiées par l'intermédiaire des boîtes de dialogue et rendues accessibles au *thread*[16].
– Le traitement des blocs d'échantillons acquis (blocs de 1024 échantillons, soit 21 ms à 48 kHz): lecture à retard variable, réverbération, encodage/décodage ambisonique et/ou synthèse binaurale et/ou tout autre technique de positionnement 3D. Sont utilisées essentiellement les fonctions C "première génération" évoquées section 5.2.
– Les entrées-sorties audio, carte multi-son et/ou fichiers. L'attente d'une réponse de la requête d'envoi de la part de la carte multi-son a un effet bloquant qui permet de cadencer la boucle, donc les traitements.

Une variable globale `STOP_PROCESS`, positionnée à la valeur 1 sur clic du bouton "*Arrêt*" de la fenêtre principale, sert de *sémaphore* pour sortir de la boucle.

5.3.2 Spécificités et fonctionnalités

Traitement par blocs

La Figure 5.9 détaille l'étape de traitement audio comprise dans la boucle infinie du thread (Figure 5.8). Les *buffers* lus, traités et/ou émis vers la carte-son sont de taille $T = 1024$ échantillons, exceptés les *buffers* circulaires pour les sources monophoniques de taille multiple de T (Figure 5.7). Les buffers de sorties – P_i vers les haut-parleurs, couples (L_b,R_b) et (L_t,R_t) pour les restitutions binaurale et transaurale – font l'objet d'un routage décrit ci-après.

Gestion des entrées-sorties et routage

Les entrées-sorties sur des fichiers-son ne présentent pas de grande particularité. Utilisant la librairie `AFsp`[17], les principaux formats sont pris en compte: **wave, aiff, au, pcm**, etc... La sortie sur fichiers a été abandonnée faute d'usage. Une couche a été ajoutée à la librairie d'origine pour lire les fichiers en boucle, ce qui offre la possibilité d'exploiter de façon répétitive des bruitages (bruits de pas, moteur, téléphone) ou des extraits sonores courts. Signalons à ce sujet une particularité: si la lecture initiale d'un fichier sur disque dur peut avoir un effet ralentissant, sa "relecture virtuelle" est fluide car il est en fait gardé en mémoire vive.

16. Bien que cela ne soit pas très recommandé en programmation objet, c'est essentiellement par des variables globales que se fait l'échange d'informations. Cet "archaïsme" provient des premières ébauches de l'interface et a survécu depuis.
17. `AFsp-V3R2.tar.Z`, *McGill University*, `http://www.TSP.EE.McGill.CA/software.html`

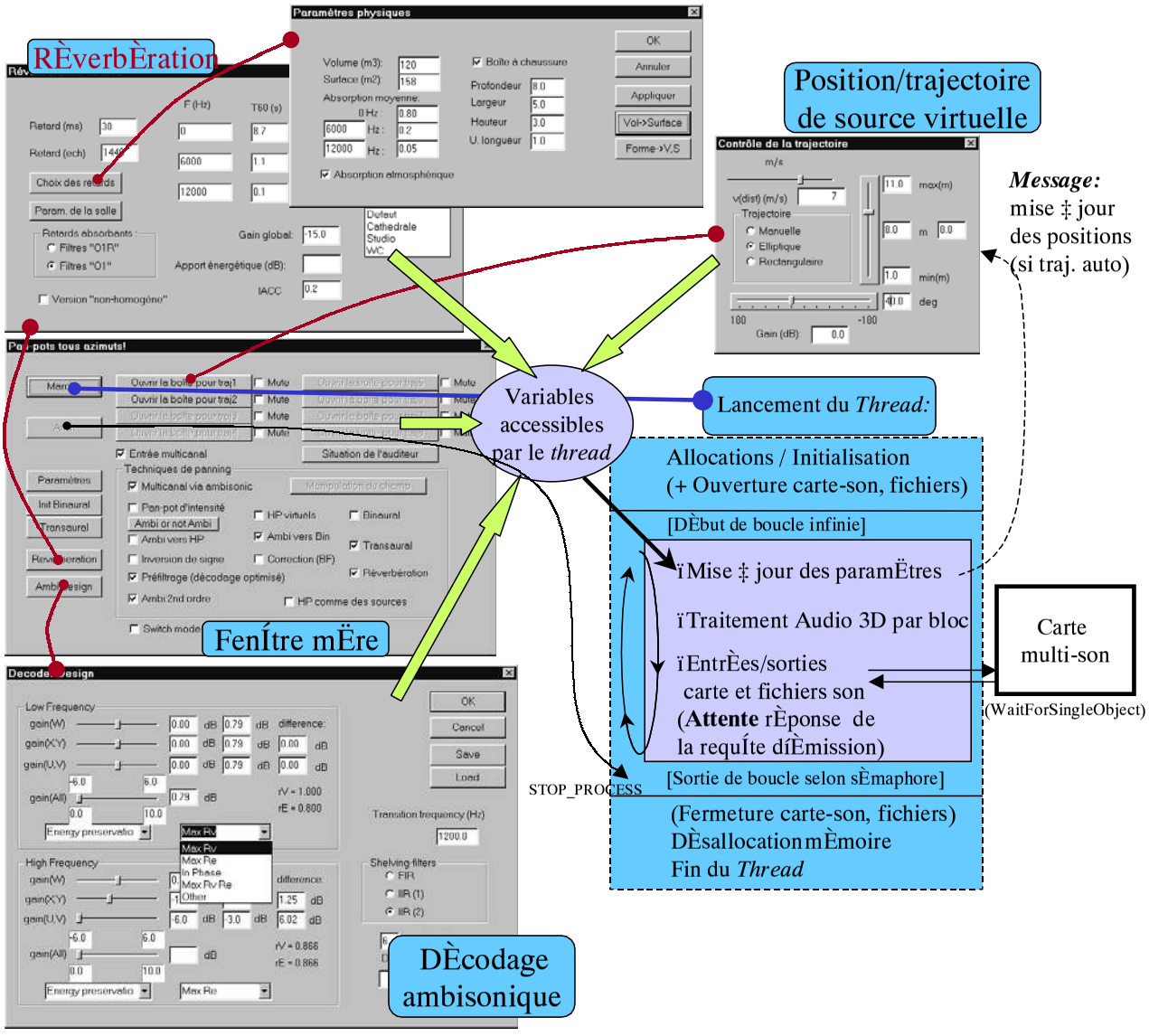

FIG. 5.8 – *Structure schématisée de l'application.*

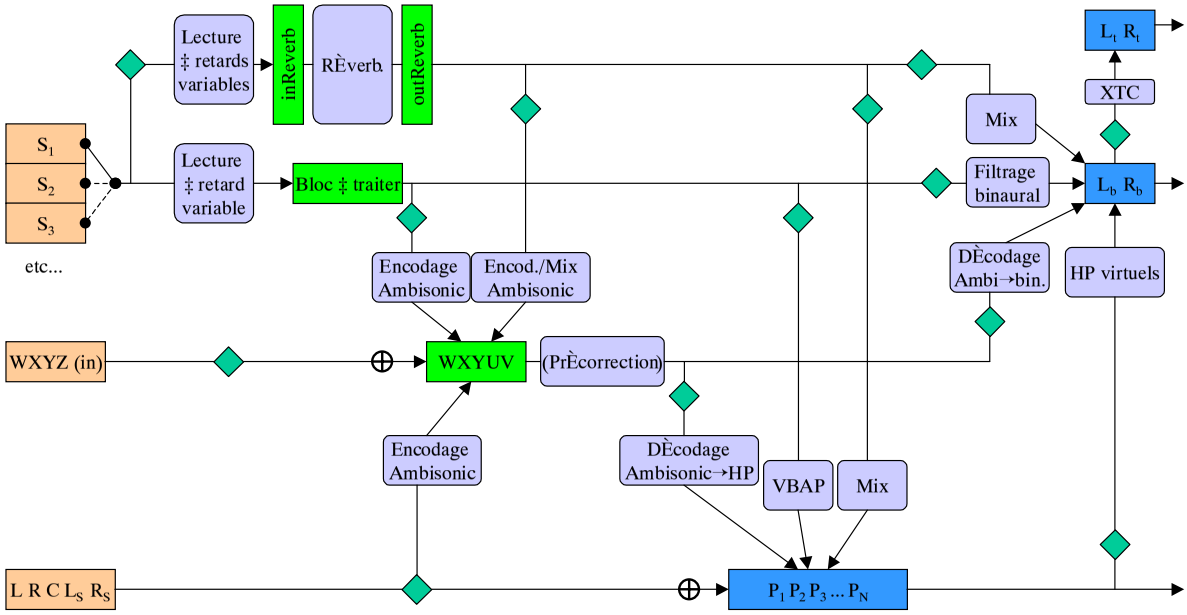

FIG. 5.9 – *Synoptique de la partie "traitement par bloc" du thread (Figure 5.8). Conventions locales: les rectangles désignent des buffers ou des groupes de buffers (circulaires pour les sources S_i); les rectangles aux coins arrondis désignent les opérations de traitement; les losanges indiquent un passage optionnel, selon les sélections faites par l'utilisateur dans la boîte de dialogue principale. Remarque: les sources monophoniques sont d'abord prises en compte individuellement et successivement, en ce qui concerne les opérations de lecture à retard variable (vers un buffer intermédiaire ou les buffers d'entrée du module de réverbération), puis selon sélections, d'encodage ambisonique, de VBAP, et/ou de filtrage binaural.*

La carte multi-son joue quant à elle un rôle déterminant sur le déroulement du programme. L'espace mémoire de la carte accessible par le programme se présente comme un grand *buffer* subdivisé en plusieurs *buffers* correspondant aux différentes voies. Mais c'est *ce même espace mémoire* qui est utilisé aussi bien pour la lecture des échantillons acquis par la carte, que pour écrire les échantillons à faire jouer par la carte. De même que la lecture doit précéder l'écriture, des requêtes de réception et d'émission sont faites à la carte l'une après l'autre, puis la fonction de lecture-écriture attend de la carte un événement qui lui indique que l'espace d'échange est prêt pour une lecture-écriture. C'est par cette attente (fonction bloquante `WaitForSingleObject`) que le *thread* de traitement se synchronise sur la carte. Nous avons choisi de placer ces opérations en fin de boucle, le premier tour de boucle se fait donc "à vide".

Les différents *buffers* d'entrée – *buffers* circulaires pour les sources monophoniques S_i, *buffers* simples pour les sources multi-canal et ambisonique – et de sortie – voies dédiées aux haut-parleurs (HP_i), au binaural (L_b et R_b) et au transaural (L_t et R_t) – font l'objet d'un routage vers les voies de la carte multi-son et la liste des fichiers-son, routage décrit Figure 5.10.

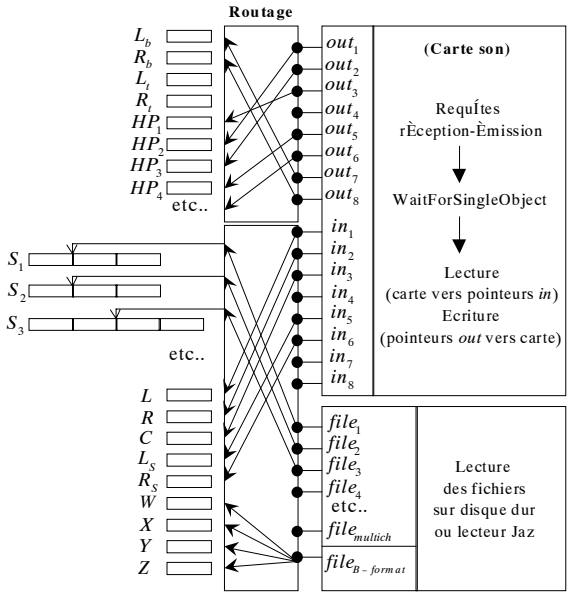

FIG. 5.10 – *Routage des entrées-sorties audio: des pointeurs fournis aux routines de lecture-écriture de la carte son ou de lecture de fichiers sont positionnés sur les différents buffers (voire sous-blocs de buffer circulaire pour les sources monophoniques S_i, cf Figure 5.7), en fonction des paramètres de configuration donnés par l'utilisateur.*

Gestion des positions et trajectoires des sources monophoniques

A *chaque* objet sonore individuel est assignée une boîte de dialogue pour le contrôle de sa position et/ou de sa trajectoire. C'est la position relative (en azimut et en distance) de la source par rapport au point de vue de l'auditeur qui est contrôlée. Ce choix est assez bien adapté au contexte d'évaluation des techniques de positionnement 3D. Pour des applications plus démonstratives ou plus ludiques, on peut préférer l'interface présentée plus loin (section 5.4.2).

Deux modes de contrôle sont proposés (Figure 5.11):

– Un *mode manuel* où l'utilisateur joue sur l'azimut et la distance de la source au moyen de curseurs (ou

sliders).

– Un *mode automatique* où une trajectoire elliptique ou bien rectangulaire est assignée à l'objet sonore. La direction du grand axe de la trajectoire – dont l'auditeur est le centre – est définie par l'azimut initial. Les distances minimale et maximale sont saisies également dans la boîte de dialogue. L'utilisateur peut alors régler la vitesse absolue de la source virtuelle.

Le contrôle du mouvement des sources est un problème délicat car la saisie des actions de l'utilisateur n'est pas naturellement synchronisée avec le *thread* de traitement. L'acquisition de ces informations, qui se fait par messages, peut même être très irrégulière selon la charge de travail du processeur (CPU). Il est bon qu'un *lissage temporel* soit effectué au début de la boucle de traitement d'après les valeurs acquises au cours des tours précédents. Cela est particulièrement important lorsque l'utilisateur manipule la distance, puisque cela se reporte sur l'effet Doppler (changement de hauteur tonale).

En mode automatique, c'est le *thread* lui-même, puisqu'il est cadencé, qui met à jour la position de la source en fonction de la trajectoire spécifiée et de sa vitesse. Il envoie alors un message à la boîte de dialogue en question pour qu'elle remette à jour ses valeurs et ses indicateurs (Figure 5.11). Mentionnons au passage que la gestion des messages émis par le *thread* peut être un facteur ralentissant, et surtout, que le rafraîchissement intempestif de la fenêtre (remise à jour des *sliders*, etc...) s'avère très coûteux. Il vaut donc mieux la fermer!

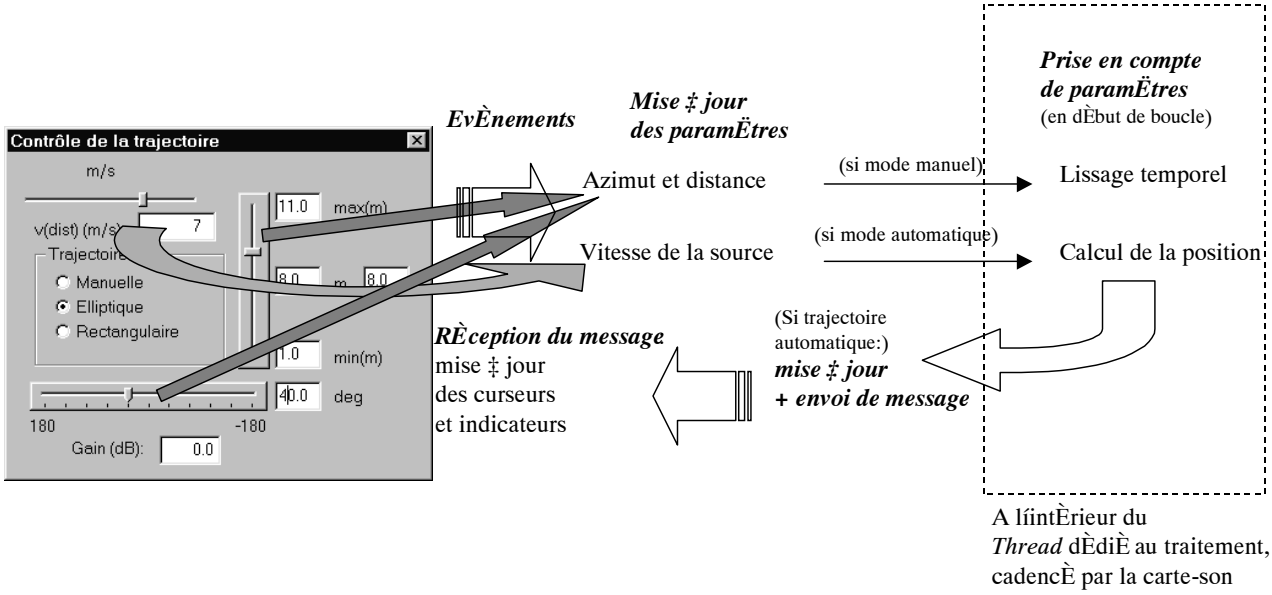

FIG. 5.11 – *Boîte de dialogue pour le contrôle de la position et de la trajectoire d'une source virtuelle, et échange des informations avec le thread dédié au traitement.*

Divers: contrôle des paramètres de quelques modules

La boîte dédiée au **contrôle du décodage ambisonique** est présente Figure 5.8, et visible plus en détails sur la figure 6 de [DRP98] (Annexe B). On y règle les gains correcteurs caractérisant les décodages modifiés

(cf 3.3.2) pour chacune des sous-bandes basse- et haute-fréquence, ainsi que la fréquence de transition. Le choix de l'ordre de l'encodage/décodage (1 ou 2) et de l'application de la correction en une seule ou deux sous-bandes se fait dans la fenêtre principale.

Les **spécifications des entrées-sorties et du routage** sont saisies au sein des onglets d'une fenêtre ouverte grâce au bouton "Paramètres" de la fenêtre principale.

Le **contrôle de la réverbération** (aperçu Figure 5.8) est réparti sur trois fenêtres: une boîte principale (T_{60} en trois fréquences, gain global, effets prédéfinis, ouverture de deux autres boîtes), une boîte pour le contrôle de paramètres physiques simples et une pour le choix des retards du réseau bouclé (non montrée).

Outre sa fonction d'ouverture des autres boîtes et de lancement/interruption des traitements, la **fenêtre mère** permet la sélection/commutation des techniques à appliquer lors du traitement (Figure 5.9) ainsi que des entrées à utiliser (mode *mute* pour les sources monophoniques, l'entrée multi-canal, l'entrée ambisonique). Des précisions sur le mode *switch* (commutation): lorsqu'il est validé, la sélection du mode *binaural direct* entraîne la dé-sélection du mode *Ambi vers bin*, et *vice versa*. Sinon, il est possible de superposer le produit des deux techniques, et – plus instructif! – d'en "écouter la différence" à l'aide du mode *Inversion de signe* applicable aux signaux ambisoniques. Les mêmes possibilités de commutation ou de combinaison sont présentes entre *Ambisonic* et le VBAP[18].

5.3.3 Utilisation: expériences et composition de scène

Configurations du système

Le matériel utilisé a été décrit en 5.1.2. Les configurations de haut-parleurs choisies ont été le plus souvent les suivantes: carrée, rectangulaire et hexagonale (avec deux haut-parleurs frontaux à $\pm 30^{\circ}$ pour laisser la place au PC entre les deux). Le lieu d'écoute étant relativement peu spacieux, le rayon du dispositif était typiquement compris entre 1 m et 1 m 50.

Expérimentation des différentes techniques et effets - Commentaires

Simulation binaurale. Malgré un effet de latéralisation bien respecté, la simulation binaurale n'échappe pas aux artefacts les plus couramment rencontrés, en partie dus à l'usage de HRTF non-individualisées: image intra-crânienne, rejet des images frontales vers le haut ou vers l'arrière. Cependant, on a pu parfois apprécier une nette amélioration de l'extériorisation, dans des conditions d'expériences particulières: ajout d'un effet de salle voisin de celui du lieu d'écoute[19], et surtout, fluctuation azimutale rapide de la source virtuelle[20].

Mode transaural. Le filtrage transaural (en aval du filtrage binaural) se révèle assez efficace, malgré une modélisation simplissime de l'inversion du *cross-talk* (section 5.2.1). Des images peuvent être perçues de façon stable à l'extérieur des haut-parleurs, et ceux-ci sont dématérialisés aux oreilles du sujet convenablement placé. L'impression varie néanmoins selon les sujets: certains perçoivent les sources arrière au bon endroit alors que d'autres les perçoivent repliées vers l'avant. Placé à l'écart de l'axe médian des haut-parleurs, le sujet perçoit un fort effet de coloration et les haut-parleurs sont identifiés comme sources individuelles.

Effet de salle. Le module de réverbération est de composition très succincte, comme nous l'avons déjà signalé, et ne permet pas à lui seul de disposer de la palette d'effets offerte par exemple par le spatialisateur de l'Ircam (le *Spat~*). Cela étant, on obtient des effets corrects en ajustant convenablement les paramètres. L'effet de distance apparente de la source est à l'occasion considérablement amélioré. Parmi les attributs

18. Nommé injustement "pan-pot d'intensité" Figure 5.8.

19. En quelque sorte: effet d'homogénéisation de l'espace sonore virtuel avec le lieu d'écoute réel.

20. Les fluctuations d'ITD et d'ILD perçus sont des facteurs déterminants pour l'impression spatiale [Gri96b], cf 1.3.4

perceptifs qui semblent faire défaut, on note un certain manque de présence de la salle, faute de réflexions précoces.

Pan-pot ambisonique. Un certain nombre d'expériences ont été évoquées en 4.1.3, le lecteur est invité à s'y reporter. Une expérience apparemment anecdotique mérite également d'être rapportée: lors de la restitution au casque d'une source fixe émettant un glissendo des fréquences graves vers les fréquences aiguës, une légère excursion de la direction apparente autour d'une direction moyenne est observée, contrairement à une simulation binaurale directe de la même source. Il serait intéressant de quantifier les variations de la direction en fonction de l'ordre et de la solution de décodage ambisonique employée.

Effet Doppler. Une illustration classique de ce phénomène est celui d'un bruit de moteur (fichier court mis en boucle) passant à assez grande vitesse (par exemple 20 m/s, soit 72 km/h, en trajectoire elliptique) à proximité de l'auditeur (1 m). L'effet de changement de hauteur tonale est tout à fait convaincant.

Mélange de sources, composition de scène

L'interface permet de donner un aperçu sonore d'une scène qui pourrait être composée dans le cadre de la navigation 3D sur Internet, par exemple. C'est aussi l'occasion d'éprouver la capacité de l'approche ambisonique à représenter de façon synthétique une scène sonore composite et à en restituer correctement les informations spatiales sur divers dispositifs, en se libérant du même coup de la contrainte d'un dispositif dédié à un des matériels originaux (notamment le multi-canal).

L'interface peut accueillir en entrée jusqu'à 8 sources monophoniques (fichiers et/ou carte-son). Pour les démonstrations, nous utilisons surtout des fichiers-son relativement courts issus de CD de bruitages. Des bruits de pas – auxquels on assigne une trajectoire rectangulaire –, une sonnerie de téléphone, des voix parlées ou des sons musicaux... tout cela, associé à un effet de salle, peut ainsi composer une scène sonore familière dans laquelle l'auditeur n'a aucune peine à se sentir immergé.

Les enregistrements multi-canal dont nous disposions et qui ont pu être superposés à cette scène sont produits pour l'essentiel par Radio-France (extraits musicaux et théâtre radiophonique). Grâce à l'encodage ambisonique, ces scènes sonores enregistrées sur cinq canaux peuvent être restituées sur quatre ou six haut-parleurs tout en respectant les informations directionnelles originales. Il est vrai que certaines qualités des images sonores, comme la consistance et la couleur, sont dénaturées par ce biais-là. Ces aspects sont discutés dans [DRP98] (annexe B), et en 6.1 dans un contexte où le B-format est envisagé comme intermédiaire de transmission, mais pour une restitution 3:2.

Les enregistrements au B-format dont nous disposons nous ont été gracieusement fournis par Dave Malham (Université de York, Royaume-Uni). Leur exploitation a été très tardive. Dans l'implémentation actuelle, en cas de mélange avec des sources encodées à l'ordre 2, c'est un décodage d'ordre 2 qui est appliqué communément: s'il est optimal pour l'ordre 2, il est donc en général sous-optimal pour la restitution du B-format original. La question d'un décodage mixte adapté à l'ordre 1 comme à l'ordre 2 est abordée en 6.2.1.

Coûts de calcul et capacité d'accueil

Pour une restitution sur haut-parleurs (hors Transaural), le coût de calcul imputable aux procédés de positionnement 3D, que ce soit *Ambisonic* ou le VBAP, est très modéré par rapport à ce qu'exige le module d'effet de salle, ou encore la gestion des trajectoires automatiques des sources monophoniques[21]. Hormis une phase de lecture initiale des fichiers-son sur disque dur, le système peut donc supporter le nombre maximal de sources sonores permis par l'interface tout en assurant la synthèse d'effet de salle.

21. Le coût de rafraîchissement de chaque fenêtre ouverte (Figure 5.11) peut être de l'ordre de 20% du CPU, sur le PC utilisé.

La restitution au casque ou sur deux haut-parleurs (procédé transaural) est systématiquement plus coûteuse, surtout à cause du filtrage binaural. En ne regardant que le problème de positionnement 3D de sources fixes, on peut considérer que le coût de traitement est approximativement proportionnel au nombre N_{TF} de fonctions de transferts utilisées selon les cas, par la synthèse binaurale directe ($N_{\text{TF}} = 2N_{\text{sources}}$), le rendu ambisonique virtuel ($N_{\text{TF}} = 2K = 4M + 2$: coût divisible par 2 par réduction de redondance), ou la combinaison “VBAP + haut-parleurs virtuels" ($N_{\text{TF}} = 2N_{\text{HP}}$: coût également réductible). Comparons ces coûts en imaginant une configuration virtuelle à $N_{\text{HP}} = 6$ haut-parleurs, et une restitution ambisonique d'ordre $M = 2$ (soit $K = 5$). Le procédé “*Ambisonic* virtuel" ($N_{\text{TF}} = 10$) est toujours un peu moins coûteux que le “VBAP virtuel" ($N_{\text{TF}} = 12$), et devient moins coûteux que le “binaural direct" dès que le nombre N_{sources} de sources virtuelles excède 5. C'est ce qu'on a pu vérifier expérimentalement. Ajoutons que dans le cas de sources mobiles, le coût du filtrage binaural individuel double momentanément à chaque fois que le changement de position entraîne une transition entre deux paires de HRTF, soit tous les 5 degrés, ce qui n'apparaît pas avec *Ambisonic* ou le VBAP. Enfin, si l'on tient compte du fait que le dispositif virtuel est symétrique par rapport au plan médian, le coût peut être divisé par 2 pour *Ambisonic*[22] ($N_{\text{TF}} = K = 2M + 1$), qui devient moins coûteux que le binaural direct à partir de $N_{\text{sources}} = 3$ si $M = 2$ et de $N_{\text{sources}} = 2$ si $M = 1$.

5.3.4 Améliorations envisageables en tant qu'outil d'évaluation

L'interface telle qu'elle se présente actuellement a déjà permis de réaliser des tests informels d'écoute. Pour mener à bien des tests de plus grande envergure – notamment dans l'optique d'une validation subjective des techniques ambisoniques –, quelques modifications peuvent être requises. Il s'agit tout d'abord de faciliter et d'assurer la *répétabilité* des séquences, en s'affranchissant au maximum d'un opérateur humain et en masquant de préférence le contrôle des sources, du décodage, etc... Une première solution consiste à stocker les séquences générées sous forme de fichiers-son ou sur cassette Hi8 (pour les écoutes multi-haut-parleurs). Une deuxième possibilité envisageable est l'*automatisation* du traitement, c'est-à-dire des fonctions de traitement, la modification des paramètres, etc., d'après un scénario défini à l'avance.

Parmi les conditions d'écoute traitées dans le cadre de la restitution ambisonique (partie II), le cas d'un auditoire élargi, donc de positions d'écoute excentrées, est celui qui recèle encore le plus d'inconnues quant à la perception directionnelle et au choix du meilleur décodage ambisonique (sections 3.1.3 et 4.2.2). D'après quelques arguments de bon sens, on se doute que la définition de zones critiques ou de périmètres de transition entre telle ou telle solution de décodage, suit une métrique qui n'est ni proportionnelle au rayon du dispositif R_{HP}, ni indépendante. Le projet d'une évaluation expérimentale qui puisse être exploitée pour des dispositifs de différentes envergures risque ainsi de se heurter à des obstacles d'ordre matériel, nécessitant un dispositif expérimental à géométrie variable. Ces contraintes matérielles peuvent disparaître si l'on envisage une simulation binaurale de la restitution sur haut-parleurs en position excentrée. Cela pourrait être réalisé avec l'application présente au prix de quelques adaptations seulement. Il suffirait de rediriger les sorties dédiées aux haut-parleurs (Figure 5.10), issus du pan-pot ambisonique d'une source S_1, vers d'autres buffers circulaires d'entrée ($S_9, \ldots, S_{8+N_{\text{HP}}}$), traités quant à eux par filtrage binaural en leur assignant la position des haut-parleurs par rapport à l'auditeur dans l'espace virtuel de restitution. Mais il faut rester conscient qu'un tel procédé ne traduit pas un certain nombre de caractéristiques naturelles de l'écoute et de la restitution: l'effet des mouvements de la tête, l'effet de la salle virtuelle de restitution, la réponse et la directivité des haut-parleurs, sans parler d'un éventuel masquage inter-individuel.

Ces idées sont reprises en conclusion comme propositions pour la définition d'un outil d'évaluation plus complet, outil dont la réalisation serait facilitée par une conception objet de la programmation.

22. Cf 3.1.3. Cette réduction de coût ne demande qu'une infime correction des fonctions implémentées.

5.4 Développements ultérieurs

5.4.1 Conception “objet": portage en C++ et extensions

Généralités

Une approche objet comme le C++ est reconnue pour apporter une grande modularité et réutilisabilité à la programmation. Pour rappel succinct: une *classe* C++ est définie par un ensemble de variables membres (comme une *structure* en C) et de fonctions membres qui ont un accès naturel aux variables membres. Généralement, les phases d'allocations et d'initialisation sont assurées par une fonction-membre dite “constructeur" exécutée automatiquement à la création de chaque instance de la classe (=objet), un “destructeur" étant appelé pour “nettoyer le terrain" à la fin de la vie de l'objet. Les objets sonores, les signaux associés, les opérations de traitement audio, d'entrée-sortie, peuvent alors être conçus comme des “briques" qu'il suffit d'agencer les unes aux autres, les étapes d'allocation, d'initialisation et d'adéquation des paramètres étant rendues transparentes. Un autre intérêt est de pouvoir bénéficier facilement – et toujours de façon transparente – des fonctionnalités de classes de plus bas niveau, par les principes de dérivation/héritage ou de composition (encapsulation d'objets de classes différentes).

Depuis l'incorporation des développements en C “première génération" au sein de l'interface présentée plus haut, un certain nombre de fonctionnalités ont été portées en C++. Ce portage ressemble bien souvent à une “encapsulation" des fonctions C et des variables traitées au sein des classes, les étapes d'allocation/initialisation et de désallocation étant respectivement reportées dans les constructeurs et le destructeur de la classe. Certaines de ces classes ont pu progressivement se substituer aux fonctions C originales dans l'application de la section 5.3. D'autres fonctionnalités exigent une refonte plus radicale: il est souhaitable, d'une part, de mettre en place des structures communes et réutilisables pour les différents types de traitement audio (notamment pour le positionnement 3D); d'autre part, la volonté d'appliquer les développements relatifs à l'extension d'ambisonique aux ordres supérieurs (partie II), sans limitation *a priori* de l'ordre, nécessite une implémentation beaucoup plus générique que celle évoquée en 5.2.3.

Nous nous contentons ci-après d'énumérer succinctement les classes ou librairies conçues, qu'elles soient achevées et opérationnelles ou seulement ébauchées, à l'état de projet.

Librairies constituées et projets naissants

Les versions C++ des entrées-sorties fichiers (avec la fonction de lecture en boucle) sont regroupées dans une librairie `AudioFileIO.lib`. Quelques éléments de base du traitement audio – buffer circulaire, ligne à retard variable, filtres RII – sont présents dans `AudioProcessing.lib`. Il resterait à compléter cette librairie en portant en C++ les algorithmes de convolution rapide par FFT (utilisés notamment pour le filtrage binaural). Les opérations matricielles dont nous avons besoin pour Ambisonic sont quant à elles déjà disponibles en C++ (`MatrixProcessing.lib`).

Les techniques de positionnement 3D (*Ambisonic*, VBP, binaural et transaural) dont la réécriture en C++ est en projet, utilisent dans l'ensemble les classes évoquées ci-dessus, et justifient par ailleurs la conception de la classe d'utilité commune `CLdSpkCfg`, attenante à la configuration des haut-parleurs. Les classes relatives à *Ambisonic*, actuellement en cours de construction, méritent quelques mots supplémentaires.

Projet: système ambisonique “universel"

L'intention qui à l'origine de ce projet est de pouvoir appliquer l'ensemble des développements et propositions théoriques présentés dans la partie II, liés à l'extension d'*Ambisonic* aux ordres supérieurs (2D,

3D, ou hybride). Un système ambisonique se compose de plusieurs modules, aussi plusieurs classes aux fonctionnalités différentes sont-elles proposées:

- Une classe `CAmbisonicFormat`: où sont spécifiées les conventions d'encodage, le nombre et l'agencement de composantes ambisoniques; elle définit les opérations d'encodage (calcul de vecteur ou de matrice d'encodage[23]); elle assure les opérations de conversion entre différents formats possibles (vecteur de conversion). (Se reporter à la section 3.1.2).
- Une classe `CAmbisonicEncoder` dédiée à l'encodage: pour un nombre donné de sources fixes, la fonction d'encodage est spécifiée par une simple matrice. La gestion des sources en mouvements implique une interpolation des vecteurs d'encodage. Plusieurs encodeurs peuvent être instanciés. Un encodeur peut également être asservi à un décodeur (voir plus bas) pour assurer une égalisation énergétique 3D (cf 3.1.4).
- Classes `CAmbisonicConverter` et `CAmbisonicTransform`, dédiées respectivement à la conversion et aux transformations: ces opérations décrites par des matrices diagonales (conversion) ou souvent creuses (transformation) ne sont pas appliquées systématiquement. Un détail: les transformations usuelles (cf 3.1.5) ne sont pas encore définies de façon générique pour les ordres supérieurs avec tous les degrés de liberté (3D)!
- Une classe `CAmbisonicMaterial`: elle comprend, outre un objet de type `CAmbisonicFormat`, des *buffers* représentant les signaux ambisoniques à considérer à différentes étapes du système (à l'issue de la lecture ou de l'encodage, ou d'une transformation, avant le décodage...).
- Une classe `CAmbisonicDecoder`: elle définit le décodeur ambisonique en fonction de paramètres comme le format d'encodage en vigueur (objet `CAmbisonicFormat`), la configuration de haut-parleurs (objet `CLdSpkCfg`), le choix d'une stratégie de décodage (style de décodage, différenciation éventuelle des critères en plusieurs sous-bandes fréquentielles, structure mono- ou multi-matrice), ou encore à plus haut niveau, des conditions d'écoute spécifiées (Sections 3.1.3 et 3.3). Ainsi sont définis les éléments moteurs du décodage, à savoir des variables membres de type matrice (`CMatrix`) et de type filtre (`CFilterIIR`) principalement.
- Une classe `CBinauralAmbisonicDecoder`: elle définit un décodeur dédié à la restitution binaurale. En plus des spécifications de `CAmbisonicDecoder`, on peut imaginer spécifier un dispositif virtuel de restitution différent du dispositif virtuel de décodage (proposition de la section 4.1.4). Elle contient en outre un ensemble de fonctions de transfert (objets de type `CFFTConvolver`).

Il resterait, à terme, à prévoir la lecture-écriture de fichiers ambisoniques (classe `CAmbisonicFile`): les spécifications sur le format à placer dans l'en-tête des fichiers (*header*) doivent pour cela être clairement définies.

Une grande partie des opérations de traitement s'exprime en terme de matrices, dont certaines peuvent être creuses (*i.e.* contenant beaucoup de zéros), ou diagonales, ou "neutres" (identité). Un traitement global efficace et sans surcoût nécessite une gestion intelligente de ces opérations: il faut savoir les "condenser" lorsque cela est possible (produit de matrices), éventuellement traiter les matrices creuses en sous-blocs, etc... On peut espérer réaliser cela de façon automatique en mettant en place, en "sur-couche", une sorte de gestionnaire-optimiseur des opérations de matriçage (`CAmbisonicManager`).

23. Calcul réalisé de façon récursive à n'importe quel ordre.

Librairie `geo3D`

C'est d'abord dans l'intention de mettre en oeuvre des méthodes géométriques pour la synthèse d'effet de salle – notamment la méthode de sources-miroir[24], au moins pour générer les réflexions précoces – que cette librairie `geo3D` a été constituée. Elle met en jeu les objets élémentaires de la géométrie 2D et 3D (points, vecteurs 2D et 3D, base, référentiel 3D...) et les principales opérations qui peuvent y être associées, dont produit vectoriel, rotations, projections, changement de référentiel...

A défaut d'avoir été exploitée pour la synthèse d'acoustique virtuelle jusqu'à présent, ses fonctionnalités ont pu être directement utilisées pour la définition d'une interface 3D, pour une manipulation et une visualisation des objets sonores plus naturelle que dans l'interface précédente. C'est ce qui est illustré dans la section 5.4.2 suivante.

5.4.2 Une interface pour la manipulation et la visualisation 3D des objets sonores

Fonctionnalités

Les classes de la librairie `geo3D` peuvent s'appliquer aussi bien au monde visuel qu'à l'audio. Un objet de type `CCamera` utilise à bon escient les opérations de projection ("orthographique" ou bien "perspective") pour fournir une représentation visuelle 2D à partir de la spécification d'objets 3D. En assignant à chaque objet sonore et à l'auditeur une forme en plus d'une position et d'une orientation dans l'espace virtuel, ce procédé offre un moyen de visualisation et de manipulation plus intuitif et naturel que l'interface précédente.

Dans l'interface présentée ici – encore développée sous *Visual C++* mais cette fois de type *Single Document Interface* – chaque objet est simplement représenté par une sphère et trois segments décrivant son référentiel local. Un "dallage" au sol et sur deux murs verticaux aide à repérer leur position. Il est possible de sélectionner un objet (dont l'auditeur), le déplacer à la souris ou à l'aide des flèches du clavier, et de changer son orientation. Le point de vue de la caméra est également réglable (zoom, travelling, etc...) et peut aussi être asservi à celui de l'auditeur. Enfin, des boîtes de dialogue peuvent être ouvertes pour chaque objet afin de contrôler quantitativement sa position et son orientation, absolues ou par rapport à l'auditeur, en coordonnées cartésiennes et sphériques. Les figures 5.12 et 5.13 montrent deux modes de projection visuelle: projection orthographique (ou orthogonale), et mode perspective.

Couplage avec la version du *Spat* portée en C++

Pour illustration, cette interface a été couplée avec une version du *Spat~* (le spatialisateur de l'Ircam) portée en C++ par Marc Emerit (CNET Lannion). Dans cette version, un module complet de spatialisation est instancié pour chaque objet sonore. La position de l'objet sonore dans le référentiel de l'auditeur est prise en compte pour le rendu directionnel (azimut et site), mais se reporte aussi, à travers la distance, sur le contrôle des paramètres subjectifs de spatialisation. Ces neuf paramètres subjectifs sont accessibles et contrôlables *via* une boîte de dialogue (Figure 5.14).

Autres possibilités offertes

On peut imaginer exploiter la représentation des référentiels locaux, non plus seulement pour les sources monophoniques, mais aussi pour la *manipulation (rotations) d'un champ ambisonique* préexistant et superposé à la scène sonore.

24. Nous avions implémenté cette méthode en C, mais elle n'a pas été directement portable, à cause de sa lourdeur, dans l'application temps-réel sur PC.

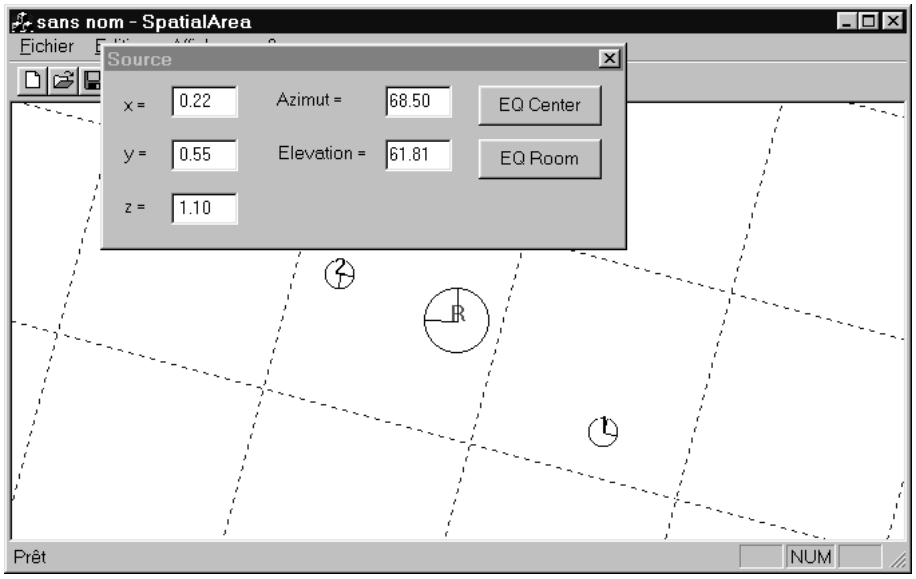

FIG. 5.12 – *Aperçu de l'application* SpatialArea*, en mode* projection orthographique *(vue de dessus), idéal pour la manipulation de sources sonores et l'évolution de l'auditeur (récepteur R) dans un espace 2D, ou parallèlement au plan horizontal. Cela n'exclue pas la prise en compte des sites relatifs des sources: la petite boîte de dialogue indique les coordonnées (cartésiennes et sphériques) de la source 1 dans le référentiel lié à l'auditeur (R).*

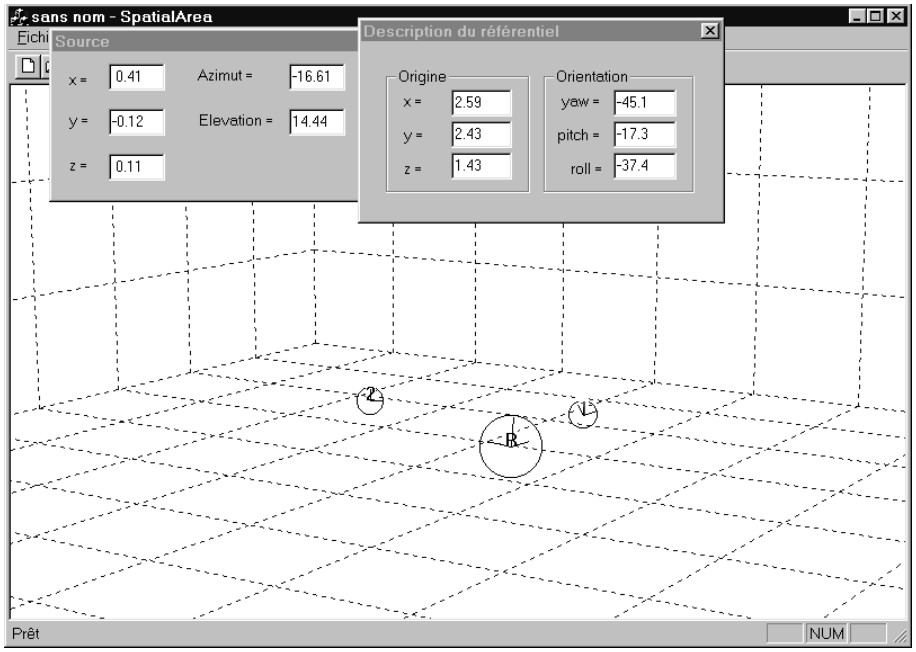

FIG. 5.13 – *Aperçu l'application* SpatialArea*, en mode* projection perspective.

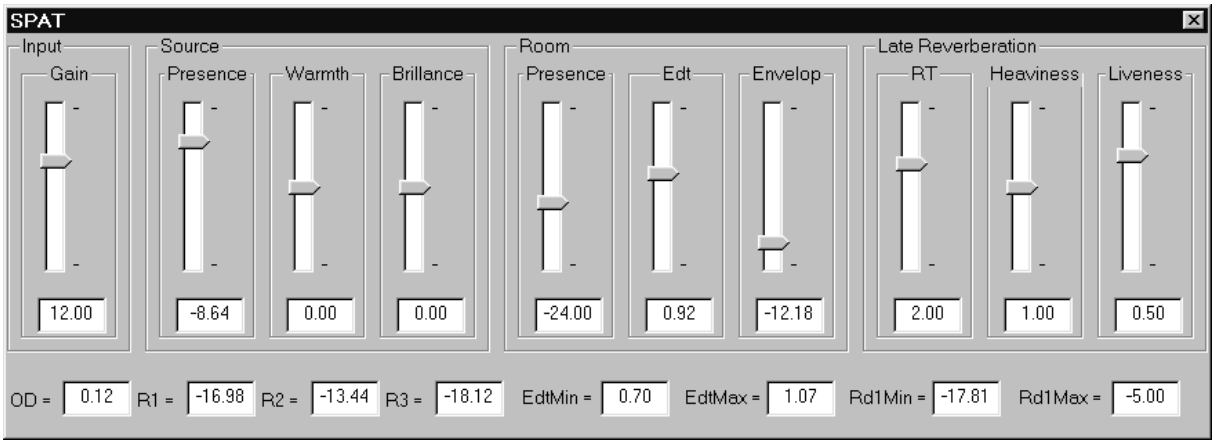

FIG. 5.14 – *Boîte de dialogue (due à Marc Emerit) pour le contrôle des paramètres "perceptuels" du* Spatialisateur. *Une boîte étant associée à chaque source, les déplacements relatifs de la source par rapport au récepteur (distance) dans la scène virtuelle (Figure 5.13) sont automatiquement reportés sur le paramètre* Présence de la source.

Enfin, dans un contexte d'évaluation du rendu sonore au casque – simulation binaurale d'une écoute sur haut-parleurs en position excentrée, encore évoquée en 5.3.4 –, ce type d'interface pourrait être utilisé pour la représentation des haut-parleurs dans l'espace virtuel *de restitution*, au sein duquel l'auditeur pourrait mouvoir son "point de vue".

5.5 Conclusion et propositions

5.5.1 Bilan

Travail présenté

Dans cette section ont été présentés les principaux développements réalisés – en langages C et C++ – au cours de cette thèse pour l'expérimentation et l'écoute. Un premier outil (section 5.3) a permis de donner un aperçu des principales techniques de spatialisation (5.2) – dont les techniques de positionnement 3D évoquées aux chapitres 2 et 3 –, même si pour certaines d'entre elles, l'implémentation est assez simplifiée ou ne représente pas l'état de l'art actuel (binaural, transaural, réverbération). Cette première interface a également servi à expérimenter la composition de scène, le mélange de sources de natures différentes (monophonique, multi-canal, ambisonique), et à démontrer l'intérêt des techniques ambisoniques dans ce contexte. Elle peut être enfin exploitée comme un outil d'évaluation des techniques ambisoniques et a déjà permis de réaliser des tests subjectifs très instructifs bien qu'informels.

Des développements en langage C++, plus génériques et plus modulaires, ont été présentés sous formes de librairies ou de projets, ainsi qu'une interface 3D pour la manipulation et la visualisation des objets sonores.

Poursuite

Dans un contexte d'application, il serait préférable de faire appel à des versions plus performantes des techniques binaurales, transaurales et de génération d'effet de salle, techniques qui n'ont pas été l'objet prin-

cipal de cette thèse. Il est prévu d'achever la réalisation du projet d'implémentation générique des techniques ambisoniques, permettant d'appliquer toutes les propositions évoquées dans la partie II. L'ensemble de ces techniques mériterait enfin d'être démontré en étant appliqué au sein d'une interface 3D: pourquoi pas, dans un premier temps, celle que nous avons développée[25] (section 5.4.2), ou bien des navigateurs MPEG4 ou VRML...

Pour un contexte d'évaluation, nous faisons dans la section 5.5.2 suivante un certain nombre de propositions qui concernent d'une part l'amélioration des outils et des moyens d'évaluation, et d'autre part la méthodologie.

5.5.2 Propositions pour l'évaluation des techniques ambisoniques

Spécifications sur l'outil d'évaluation

Deux modes de restitution sont envisagés pour l'évaluation: la présentation sur haut-parleurs et la présentation au casque basée sur le principe des haut-parleurs virtuels. Parmi les différentes conditions d'écoute que l'on peut rencontrer pour la restitution ambisonique, l'écoute en position excentrée est celle qui met en jeu le plus de paramètres, et nécessiterait donc les expérimentations les plus abondantes. Il faut donc concevoir un outil souple, mettant à profit une extension du principe d'évaluation par haut-parleurs virtuels (simulation binaurale) pour ce type de conditions d'écoute. La restitution au casque étant en effet très abordable d'un point de vue matériel, elle permet d'évaluer des restitutions qui devrait nécessiter un grand nombre de haut-parleurs, une carte son avec autant de sorties, un lieu de restitution adapté[26], et permet de changer très facilement de configuration virtuelle de restitution.

Dans le principe, l'écoute au casque permet d'observer facilement la relation entre les jugements subjectifs et la position d'écoute par rapport au dispositif virtuel, ainsi que de contrôler parfaitement les conditions d'écoute et d'en assurer la reproductibilité. Mais il faut garder à la conscience que ce la méthode introduit un biais considérable par rapport à une écoute naturelle sur haut-parleurs et dans une salle. C'est pourquoi il est important de réduire les prétentions de cette approche à des critères de jugement précis (continuité de la perception directionnelle dans le cas d'une source virtuelle en mouvement, acuité et qualité de localisation/latéralisation, effet de hauteur...), et sur un mode comparatif (la référence pouvant être la simulation binaurale directe d'une source unique).

Suggestions méthodologiques

La possibilité d'une référence – la simulation binaurale directe – lors d'une écoute au casque, offre un moyen de caractériser de façon affinée le *degré de latéralisation* des sources fantômes créées par pan-pot ambisonique. Considérons le rendu ambisonique virtuel d'une source d'incidence donnée, et pour une position d'écoute centrée. Ainsi qu'il a été observé objectivement et subjectivement (cf 4.1), l'incidence perçue est plus proche du plan médian. Par comparaison alternée avec une simulation binaurale directe appliqué au même son, le sujet est invité à ajuster l'incidence (azimut et site) associée à cette deuxième simulation afin qu'elle soit le plus proche possible de la simulation ambisonique dans l'espace. Les résultats pourraient alors être comparés aux courbes d'ITD et d'ILD présentées en 4.1.2, le but étant surtout d'observer l'influence des critères de décodages (en particulier $max\ r_E$).

25. D'autres librairies plus performantes (optimisées) pourraient également être utilisées pour la visualisation 3D: OpenGL, Direct3D, etc...

26. Seul le coût de calcul est susceptible d'être plus important, mais on peut considérer que les ressources CPU des machines actuelles et futures, surtout dans un contexte de laboratoire, ne constituent d'ores et déjà plus un facteur limitant!

L'évaluation en position d'écoute excentrée est rendu beaucoup plus souple par le principe des haut-parleurs virtuels. Elle doit permettre de vérifier certains résultats théoriques et de préciser le domaine d'application des solutions de décodage ambisonique.

Tout d'abord, en plaçant les haut-parleurs virtuels au loin, il devrait être possible de juger la pertinence du *vecteur énergie $\vec{E}$ comme prédicteur de la latéralisation basse-fréquence en position excentrée*, en vérifiant l'amélioration apportée par les solutions *max* r_E à la latéralisation.

D'autres expériences devraient servir à déterminer le décodage optimal à utiliser en fonction de l'étendue de l'auditoire (ou de la position occupée la plus critique) et de la géométrie du dispositif. Les remarques de la section 4.2.2 indiquent qu'il faudrait caractériser, pour une position excentrée donnée, non seulement la précision des images sonores, mais aussi la continuité et l'homogénéité de leurs caractéristiques, en particulier lorsqu'elles se déplacent. Plusieurs expériences peuvent donc se baser sur le cas d'une source en trajectoire circulaire centrée[27]. On peut faire alterner (à chaque tour) deux décodages (notamment *max* r_E et *in-phase*)[28], faire se déplacer lentement l'auditeur sur un rayon, en partant du centre du dispositif (virtuel), en lui demandant à quel moment sa préférence passe d'un décodage à l'autre, selon des critères de précision et/ou de continuité. On peut également tenter de définir, pour une position d'écoute donnée, une interpolation optimale – au goût de l'auditeur – entre deux solutions de décodages.

Et pourquoi pas...

Le faible investissement matériel requis (à part la machine) et la commodité de l'écoute au casque permettraient de multiplier les expériences sur une population importante de sujets: les tests se prêtent facilement à des écoutes simultanées par plusieurs auditeurs dans les mêmes conditions, et peuvent être répartis sur plusieurs laboratoires voire proposés à des particuliers, avec la possibilité d'une collecte automatique des résultats par courrier électronique.

On peut encore imaginer enrichir la reproduction des conditions d'écoute à l'aide d'un système de suivi des mouvements de la tête (*Head-Tracking*) pour tenir compte des rotations de la tête[29].

27. Là encore, la simulation au casque offre la possibilité d'une référence, qu'il faut choisir comme étant l'effet d'une source unique (simulation binaurale directe) parcourant le périmètre des haut-parleurs.

28. Dans un premier temps, on peut les appliquer en pleine-bande.

29. Tant qu'on y est, on peut imaginer enregistrer les mouvements de la tête dans une situation d'écoute donnée, puis les interpréter pour déterminer de quelle façon les mécanismes de latéralisation dynamique rentrent en jeu.

Chapitre 6

Applications et perspectives liées à *Ambisonic*

6.1 Application au codage et à la transmission de matériel multicanal

6.1.1 Motivations et idées de base

La diffusion multi-canal doit son essor à l'enrichissement considérable de l'expérience auditive spatiale qu'elle est susceptible de proposer par rapport à la stéréophonie conventionnelle (Cf 2.3). En contrepartie, l'augmentation du nombre de canaux entraîne un coût de transmission plus élevé dans un contexte de diffusion radiophonique, télévisée, ou sur Internet. Face au facteur limitant qu'est le débit des informations transmises, les techniques de compression audio-numérique – appliquées aux cinq canaux audio séparément – ne sont pas toujours jugées suffisantes. D'autres pistes pour réduire la quantité de données transmises reposent sur l'idée de redondance des informations spatiales perçues. Le constat qu'un auditeur ne possède que *deux oreilles* suggérerait naïvement qu'il suffit de transmettre deux signaux, en faisant appel notamment aux techniques binaurales et transaurales (haut-parleurs virtuels, page 2.5.1). Mais cette approche impose finalement des contraintes d'écoute peu naturelles et restrictives. En poussant le raisonnement un peu plus loin, émerge l'idée d'une représentation minimale des *informations directionnelles du champ acoustique*: l'encodage ambisonique au format B horizontal (trois canaux W, Y, Z) apparaît au premier abord comme un candidat idéal pour réaliser cette idée.

L'encodage ambisonique du matériel multi-canal suit le même principe que dans le contexte de mélange de sources abordé en 5.3.3: chaque voie est encodée comme une source virtuelle ayant comme direction d'incidence celle du haut-parleur associé, dans le dispositif de restitution prévu. On note $\mathbf{C}$ la matrice d'encodage correspondante (équation 3.25, page 158). Il ne reste alors qu'à transmettre les canaux W, X, Y. Contrairement aux expériences du 5.3.3, nous ne nous intéressons ici qu'à une restitution sur le dispositif 3:2 prévu initialement: l'opération de décodage y est donc dédiée et l'on peut faire appel, dans la mesure du possible, aux matrices de décodage de Gerzon [Ger92a] pour assurer la préservation des informations directionnelles au sens des vecteurs vélocité $\vec{V}$ et énergie $\vec{E}$ (cf 2.4.3 et 3.3.4).

Ce principe simple a d'abord été expérimenté en définissant la matrice de décodage par pseudo-inverse, et dans des conditions ne se prêtant pas à des tests très développés, les écoutes comparatives entre le matériel original et la version codée/décodée étant réalisées au studio-son après transfert par le réseau des fichiers-son traités. Cette piste a été ensuite approfondie à l'occasion d'un stage effectué par Louis-Cyrille Trébuchet [Tre97].

6.1.2 Application du principe, résultats et interprétations

Premières expériences

Lors des premières expériences, la disposition 3:2 des haut-parleurs était caractérisée par des angles $\phi_F = 30°$ et $\phi_B = 70°$ (deux paires avant et arrière, plus un haut-parleur frontal, cf Figure 2.6, page 93). Les matériels originaux étaient issus de mixages de sources de nature différentes (voix parlée, musique, bruitage). Malgré le choix sans-doute sous-optimal de la pseudo-inverse de **C** comme matrice de décodage, l'impression retenue est celle d'une différence très peu sensible entre la restitution du matériel original et celle de la version codée/décodée. Mais étant donné le peu d'expériences réalisées dans ces conditions, nous commenterons plutôt celles réalisées par la suite, notamment dans la configuration $(\phi_F = 45°, \phi_B = 50°)$ pour laquelle nous disposons de solutions de décodage optimisées.

Expériences suivantes: résultats

Les expérimentations réalisées avec la configuration $(\phi_F = 45°, \phi_B = 50°)$ mettent en jeu un décodage optimisé en deux sous-bandes suivant les critères de Gerzon sur les vecteurs vélocité $\vec{V}(r_V, \theta_V)$ et énergie $\vec{E}(r_E, \theta_E)$: la matrice appliquée en basse-fréquence (< 700 Hz) vérifie les critères $\theta_V = \theta_E$ et $r_V = 1$, et celle appliquée en haute-fréquence vérifie $\theta_V = \theta_E$ en maximisant r_E (solution dite "*max* r_E"). Outre les solutions proposées par Gerzon [Ger92a], nous disposons d'une seconde version de matrice haute-fréquence définie par Trébuchet [Tre97] (cf 2.4.3 et 3.3.4), dont l'interface *Ambicast* a été utilisée pour ces expérimentations.

Le matériel multi-canal qui a servi d'original est essentiellement constitué d'enregistrements de Radio-France: prises de son musicales, théâtre radiophonique, et prises de son expérimentales utilisant des techniques de microphones non-coïncidents (Figure 2.7, section 2.3.3). La restitution du matériel original a pu être comparée à la version encodée via *Ambisonic* et décodée pour le même dispositif de haut-parleurs ("*codec* ambisonique").

Les tests d'écoute (tests informels) ont été menés dans une salle assez volumineuse (plateau-vidéo du CCETT) mais modérément réverbérante, les haut-parleurs (grosses enceintes Genelec, modèle 10-32A) étant placés sur un cercle de rayon 3,3 m. Les impressions retenues sont les suivantes. Pour l'auditeur placé au centre, la direction des images sonores est préservée, mais leur définition – en azimut comme en profondeur – et leur consistance sont moins bonnes à l'issue du *codec* ambisonique. Leur couleur semble également quelque peu dénaturée. En position d'écoute un peu excentrée, les images présentent en outre une bien moindre stabilité qu'avec la diffusion multi-canal originale. Notons tout de même que la stabilité est sensiblement accrue lorsque l'on utilise les solutions de décodage optimisées (évoquées plus haut), plutôt qu'une solution sous-optimale (comme la pseudo-inverse de **C** appliquée pleine-bande).

Interprétation: nature de l'enregistrement original et diaphonie

A l'issue de cet encodage/décodage, le champ sonore restitué est en principe conforme à l'original... dans les limites de validité de l'approximation ambisonique, c'est-à-dire – à l'échelle de l'auditeur – pour les basses fréquences et pour une position centrée! En dehors de ce domaine, il est devenu d'usage, dans ce document et dans toute la littérature relative à *Ambisonic*, de faire intervenir le vecteur énergie pour caractériser l'effet de localisation auditive. Mais une grande différence démarque l'expérience présente des conditions d'étude qui ont présidé jusqu'ici ([DRP98], Annexe B, section 5): les cinq canaux originaux encodés ne représentent plus, en général, des images sonores indépendantes ni ne diffusent des "événements acoustiques élémentaires" indépendants, mais sont voués à créer *ensemble* une ou des images sonores en étant diffusés par des haut-parleurs *distincts*, faisant intervenir des mécanismes psychoacoustiques différents

selon les techniques de prise de son ou de mixage. Or les cinq voies finalement diffusées à l'issue du codec sont des versions *mélangées* des cinq voies originales.

Les artefacts audibles attendus hors du domaine de reconstruction acoustique se révèlent dépendre assez fortement des techniques utilisées pour la création des images sonores dans le matériel original. Le critère discriminant est sans doute la présence ou non de retards temporels ("ΔT") entre les canaux pour la formation d'une image. Dans le cas où l'image fantôme est créée uniquement par différences d'intensité ("ΔI") entre les voies dans l'enregistrement original, le mélange des voies n'entraîne aucune altération du contenu spectral associé à la source sonore dans chaque canal. Par contre, lorsque le signal associé à la source fantôme est présent avec des retards entre les canaux (prise de son avec des microphones non-coïncidents), le mélange des voies provoque un effet de filtre en peigne donc une altération du contenu spectral par rapport aux signaux d'origine. Par ailleurs la décorrélation inter-canaux propre à ce type de prise de son se trouve dégradée dans une mesure plus importante qu'avec les techniques ΔI. En conséquence, les caractéristiques de l'image sonore et les impressions spatiales plus globales en pâtissent plus.

Pour poursuivre ces commentaires avec une analyse quantitative, il est intéressant d'interpréter ce procédé comme une opération de matriçage/dématriçage passif, avec un avantage cependant sur un procédé comme *Dolby Surround* (cf 2.3.4): le fait qu'un canal supplémentaire est transmis, qui permet de préserver sans ambiguïté l'information directionnelle dans le plan horizontal. Une façon de caractériser la dégradation imputée à cette opération dans un cas général consiste à estimer *la diaphonie* qui en résulte (*interchannel crosstalk*), c'est-à-dire la "quantité de signal" qui est passée d'un canal à un autre.

Quantifier la diaphonie

La "quantité de signal" transmise d'un canal à chaque autre au terme de l'opération apparaît directement dans le produit $\mathbf{D}.\mathbf{C}$ des matrices d'encodage et de décodage (dans une sous-bande fréquentielle donnée si besoin). Dans les tableaux qui suivent, elle est exprimée en dB par rapport à la quantité restante dans le canal d'origine. Les colonnes correspondent aux canaux d'origine, et les lignes aux canaux d'arrivée. Nous comparons ainsi les différents décodage associés à la configuration $(\phi_F = 45^\circ, \phi_B = 50^\circ)$: pseudo-inverse, matrice basse-fréquence (version Gerzon), haute-fréquence (version Gerzon), haute-fréquence (version Trébuchet).

On constate tout d'abord que les matrices (Table 6.1) ne sont pas symétriques, ce qui signifie que la perméabilité entre deux canaux n'est en général pas la même dans un sens ou dans l'autre. Ce qu'il paraît surtout important d'observer, c'est la séparation entre canaux "éloignés" (*i.e.* dont les haut-parleurs associés sont éloignés angulairement) et la séparation latérale, dont dépend la préservation de la décorrélation interaurale et de l'impression spatiale. La stabilité des images et l'étendue du *sweet-spot* en dépendent également.

On note ainsi que les solutions *max* r_E dédiées aux hautes-fréquences assurent la meilleure séparation entre le canal centre C et les canaux arrière L_B et R_B. La version de Gerzon (`HF(G)`) assure une séparation record de 100 dB dans le sens $C \rightarrow L_B, R_B$. Dans l'autre sens, c'est la version de Trébuchet (`HF(T)`) qui est un peu meilleure. L'ensemble des solutions n'offre qu'une séparation modérée entre les canaux frontaux adjacents, surtout dans le sens $C \rightarrow L_F, R_F$. La séparation latérale est quant à elle meilleure à l'avant (entre L_F et R_F) qu'à l'arrière (entre L_B et R_B), où curieusement, les solutions *max* r_E n'apparaissent pas très performantes ($-6{,}5$ dB contre $-9{,}3$ dB pour `LF(G)` et $-10{,}7$ dB pour `pinv`). Habituellement, c'est entre les canaux *surround* L_B et R_B que l'on cherche à préserver la meilleure décorrélation possible. Mais il faut noter que dans la configuration présente, les haut-parleurs arrière ($\phi_B = 50^\circ$) ne sont pas aussi "latéraux" que dans les configurations les plus couramment rencontrées ($60^\circ \leq \phi_B \leq 80^\circ$), à l'inverse des haut-parleurs frontaux ($\phi_F = 45^\circ$ contre bien souvent 30°). Cela signifie que la séparation à l'avant est ici presque aussi importante qu'à l'arrière pour la préservation de la décorrélation interaurale, donc des impressions spatiales.

Pour conclure, les solutions *max* r_E se démarquent le plus nettement des autres au regard de leurs perfor-

pinv	L_B	L_F	C	R_F	R_B
L_B	0.0	-5.2	-17.0	-7.8	-10.7
L_F	-8.5	0.0	-2.0	-18.7	-11.1
C	-21.9	-3.6	0.0	-3.6	-21.9
R_F	-11.1	-18.7	-2.0	0.0	-8.5
R_B	-10.7	-7.8	-17.0	-5.2	0.0

LF(G)	L_B	L_F	C	R_F	R_B
L_B	0.0	-6.9	-13.4	-9.3	-9.3
L_F	-10.2	0.0	1.6	-18.0	-7.6
C	-21.3	-6.1	0.0	-6.1	-21.3
R_F	-7.6	-18.0	1.6	0.0	-10.2
R_B	-9.3	-9.3	-13.4	-6.9	0.0

HF(G)	L_B	L_F	C	R_F	R_B
L_B	0.0	-5.4	-100.3	-15.9	-6.5
L_F	-6.6	0.0	-1.4	-19.2	-13.7
C	-31.8	-4.1	0.0	-4.1	-31.8
R_F	-13.7	-19.2	-1.4	0.0	-6.6
R_B	-6.5	-15.9	-100.3	-5.4	0.0

HF(T)	L_B	L_F	C	R_F	R_B
L_B	0.0	-5.7	-29.1	-15.0	-6.4
L_F	-6.4	0.0	-1.0	-18.6	-14.8
C	-36.7	-4.6	0.0	-4.6	-36.7
R_F	-14.8	-18.6	-1.0	0.0	-6.4
R_B	-6.4	-15.0	-29.1	-5.7	0.0

TAB. 6.1 – *Diaphonie (en dB) entre canaux résultant du codage/décodage ambisonique de matériel multi-canal. Colonnes: canaux d'origine. Lignes: canaux d'arrivée. Avec comme matrice de décodage, de haut en bas et de gauche à droite: pseudo-inverse (pinv), matrice basse-fréquence version Gerzon (LF(G)), matrice haute-fréquence version Gerzon (HF(G)), matrice haute-fréquence version Trébuchet (HF(T)).*

mances de séparation entre les canaux latéraux (L_F, R_F, L_B, R_B) et le canal central C, limitant ainsi un facteur de réduction de la décorrélation interaurale. L'application des autres solutions est de toutes façons réservée à un domaine fréquentiel de reconstruction acoustique où l'effet de la diaphonie n'est pas nuisible.

Conclusions

L'analyse qui vient d'être faite pourrait inviter à travailler sur d'autres matrices d'encodage/décodage qui privilégient la séparation selon les "axes" les plus vitaux... mais on s'éloigne alors du principe ambisonique comme "concept-clé", en risquant d'en perdre les avantages, comme l'assurance de la préservation des informations directionnelles. Il pourrait être intéressant de développer l'idée d'un décodage "actif", à la manière des systèmes 5.2.5 (cf 2.3.4) et avec l'avantage d'un canal intermédiaire supplémentaire qui devrait faciliter les prises de décision du décodeur et améliorer la séparation.

Quoiqu'il en soit, l'idée qu'un enregistrement multi-canal présente des redondances spatiales réductibles de façon transparente par simple encodage ambisonique d'ordre 1, ne s'avère pas juste dans un cas général. Les contre-exemples les plus manifestes se rencontrent avec les enregistrements issus de techniques "ΔT". Les seuls cas où cette idée se trouve vérifiée sont ceux... de matériels issus d'un décodage ambisonique!

6.1.3 Codage ambisonique et compression audio-numérique combinés

Quelques mots sur la compression audio-numérique et le codage psychoacoustique

Les stratégies habituelles de réduction de débit pour la transmission – ou le stockage – de matériel multi-canal de type 5.1 (ou 3:2), reposent sur une compression audio-numérique de chaque canal séparément (codage MPEG2 par exemple). Le codage par sous-bandes consiste à représenter le signal par trames temporelles et par sous-bandes fréquentielles, et de faire des économies d'information en diminuant la quantification (nombre de bits accordé à chaque "échantillon") des sous-bandes où le signal est peu présent. Le codage

psychoacoustique s'appuie en outre sur la connaissance des "défauts" de l'oreille – courbes de masquage, seuils d'audition – pour répartir le bruit de quantification accompagnant la compression dans des régions du spectre où il pourra être masqué par les autres sons, et devenir si possible inaudible [Dur98]. Jusqu'à certains taux de compression[1], selon le codeur utilisé et la nature du signal, le bruit de quantification est en général inaudible: le codage est dit transparent. Au-delà, la dégradation du signal devient perceptible et peut devenir gênante.

Vers un compromis entre dégradation spatiale et dégradation du signal

Les expériences de codage/décodage de matériel multi-canal par un système ambisonique d'ordre 1 montrent une dégradation en termes de précision et de consistance des images sonores, quoique l'information directionnelle reste préservée. Les modifications du champ perçu en termes d'impressions spatiales n'ont pas été clairement discernées ou identifiées jusqu'à présent. Bien que la modification des images sonores puisse s'accompagner d'une altération de la coloration subjective, il n'y a pas à proprement parler de "dégradation du signal" au sens de l'émergence d'un bruit de quantification ou d'une réduction de bande-passante. On ne parlera donc que d'une *dégradation spatiale*.

Le codage des cinq canaux originaux vers trois canaux ambisoniques transmis W, X, Y correspond à une réduction de débit de 3/5. En combinant le codage ambisonique à des techniques de compression audio-numérique, c'est-à-dire en appliquant ces dernières aux canaux intermédiaires W, X, Y, on réalise un compromis entre dégradation du signal et dégradation spatiale, qui peut se révéler plus satisfaisant à l'oreille, *pour un même débit global*, que la compression audio-numérique des cinq canaux originaux (Figure 6.1). Cette idée a fait l'objet d'une expérimentation [Tre97], décrite ci-après.

Expérience réalisée

Pour l'expérience comparative, réalisée au cours d'un stage par Louis-Cyrille Trébuchet [Tre97], deux chaînes de codage-transmission-décodage ont été considérées (Figure 6.1), avec une commutation possible entre les deux[2]. Un codeur et un décodeur MPEG2 ont été utilisés, en mode 3:2 dans un cas (a: MPEG2 seul) et 3:0 dans l'autre (b: Ambisonic+MPEG2). Le décodage ambisonique était celui préconisé par Gerzon (ou celui de Trébuchet) pour la configuration ($\phi_F = 45^\circ$, $\phi_B = 50^\circ$) évoquée plus haut (6.1.2). Afin de rester dans l'esprit d'une transmission où le récepteur pourrait ne pas posséder de décodeur ambisonique, l'encodage ambisonique a été légèrement modifié pour produire un format compatible stéréo (canaux L_t, R_t, O_t), mais l'expérience aurait pu être faite en transmettant directement le format B horizontal (W, X, Y). Le matériel original était constitué d'enregistrements de Radio-France essentiellement basés sur des techniques de microphones non-coïncidents.

Des débits relativement critiques ont été choisis pour rendre l'expérience concluante. Le débit global de 360 kbts/s représente un taux de compression de chaque canal intermédiaire de 10,7 dans le cas (a) et de 6,4 dans le cas (b) (*i.e.* avec le codage ambisonique). Avec 320 kbts/s, le taux de compression est de 12 pour le cas (a) et de 7,2 (b). Notons qu'avec de tels taux de compression, la dégradation perçue du signal est très perceptible, voire très désagréable. Elle est évidemment moins marquée avec la combinaison "Ambisonic+MPEG2".

Une douzaine de personnes ont pu participer à cette expérience – restée néanmoins de nature informelle – qui s'est déroulé dans une salle assez volumineuse (le plateau vidéo du CCETT), les enceintes (grosses en-

1. Actuellement, de l'ordre de 1/5 ou 1/6 avec de bons codeurs.

2. Pour des raisons matérielles, la partie "codage" a été réalisée préalablement, les signaux résultants étant stockés (sur cassettes Hi8) plutôt que transmis.

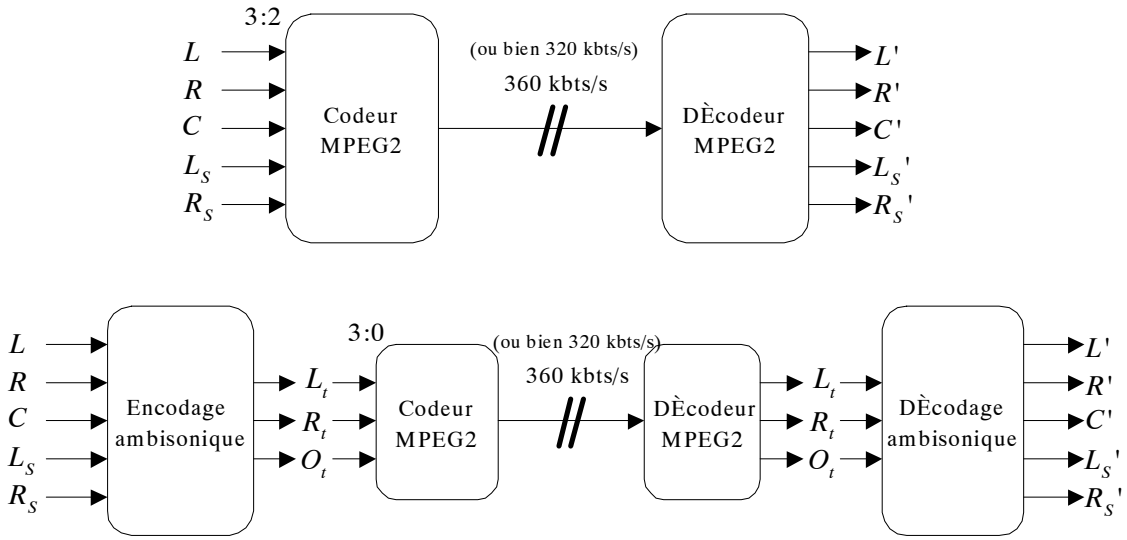

FIG. 6.1 – *Deux schémas de codage/décodage pour la transmission d'un matériel multicanal avec réduction de débit (d'après [Tre97]): (a) par compression audio-numérique (codage MPEG2) des 5 canaux séparés; (b) par encodage ambisonique (modifié) suivi d'une compression MPEG2 des trois canaux intermédiaires L_t, R_t, O_t.*

ceintes Genelec, modèle 10-32A) étant disposées sur un cercle d'un peu plus de 3 m de rayon. Tous les sujets ont pu apprécier la différence entre les résultats des deux chaînes, et ont trouvé le résultat de la combinaison "Ambisonic+MPEG2" le plus agréable – ou le moins désagréable devrait-on dire –, particulièrement avec le débit 320 kbts/s.

Notons qu'avec une réduction de débit trop faible, le codage-décodage (a) aurait pu être transparent tandis qu'une dégradation spatiale aurait subsisté *via* la chaîne (b).

Validité de la combinaison "Ambisonic+MPEG2"

Deux objections quant à la validité de cette méthode concernent les perturbations mutuelles – interactions parasitaires – entre les deux types de codage/décodage. D'une part, on peut s'attendre à ce que la compression indépendante des canaux intermédiaires diminue la cohérence des composantes directionnelles du champ ambisonique et dégrade l'information spatiale. Ce type d'artefact n'a pas été remarqué lors des écoutes, peut-être parce que les autres défauts étaient plus saillants. D'autre part, la validité du codage "psychoacoustique" est mise en défaut, en principe, par le fait que les signaux sur lesquels il est appliqué, sont par la suite recombinés entre eux, avant d'être restitués[3]. Mais dans notre application, le codage est loin d'être transparent et les phénomènes de masquage largement inopérants étant donné le niveau du bruit de quantification.

3. Cette objection devrait aussi s'appliquer, en toute rigueur, au fait de compresser 5 signaux indépendamment sans prise en compte de leur combinaison au niveau des oreilles au moment de la restitution.

Conclusions

A défaut d'expériences plus approfondies, il pourrait être utile de consulter les tables des performances du codeur MPEG2[4] (notes subjectives) en fonction du taux de compression, afin de déterminer dans quelle(s) gamme(s) de débit global la combinaison "Ambisonic+MPEG2" est susceptible d'apporter un avantage suffisamment significatif en termes de qualité de signal pour être préférée au "tout-MPEG2" malgré la dégradation spatiale. Les expériences réalisées pour le moment laissent l'impression que cette stratégie constitue malgré tout une sorte de "solution du pauvre".

Il faut enfin remarquer que les tendances actuelles vont désormais dans une toute autre direction, en tout cas dans un contexte de diffusion *via* des supports solides (DVD, SACD, etc...): avec le G-format, c'est le format ambisonique qui se plie aux contraintes et standards en vigueur (format 5.1), bref, le contraire d'une réduction de quantité d'informations!

6.2 Nouvelles problématiques liées à *Ambisonic* et aux ordres supérieurs

De nouvelles problématiques accompagnent l'extension d'*Ambisonic* aux ordres supérieurs: il s'agit par exemple du décodage conjoint de matériels de résolutions différentes (*i.e.* encodés à des ordres différents), de la définition d'un format mixte, mais aussi de la représentation de l'effet de salle et de l'exploitation de réponses impulsionnelles 3D, ainsi que d'autres aspects abordés plus brièvement.

6.2.1 Décodage mixte

Nous avons déjà souligné l'intérêt pratique que représente l'approche ambisonique pour composer et manipuler des scènes sonores virtuelles à partir de matériels sonores de natures diverses (cf 5.3.3) [DRP98]. Un cas de figure particulier est le mélange d'un matériel ambisonique d'ordre 1 "préconstitué" – un enregistrement au format B, par exemple – avec un matériel d'ordre 2, synthétisé "sur place" par exemple. On décrira les deux jeux de composantes par $(W^{o1}, X^{o1}, Y^{o1}, [Z^{o1}])$ et $(W^{o2}, X^{o2}, Y^{o2}, [Z^{o2}], U^{o2}, V^{o2}, [S^{o2}, T^{o2}, R^{o2}])$, les composantes à caractère vertical (entre crochets) n'étant pas requises pour une restitution 2D. Si l'on souhaite appliquer un décodage optimisé au matériel ce chaque ordre, chaque jeu de composantes requiert son propre décodeur. Il ne suffirait pas, par exemple, de mélanger les composantes d'ordre 1 à celles d'ordre 2 et d'appliquer un décodage optimisé pour l'ordre 2: celui-ci serait sous-optimal – au sens des mêmes critères – pour l'ordre 1. Il est par contre possible, dans certains cas, de *factoriser les opérations de décodage* propres à chaque ordre. Les décodages pour configurations régulières ou semi-régulières (sections 3.3.1, 3.3.2 et 3.3.3) consistent en effet en un matriçage dit "basique" précédé d'une correction des composantes ambisoniques (Figure 3.5). Or, la matrice "basique" pour le décodage d'ordre 2 est compatible avec celle d'ordre 1 dans le sens où cette dernière en constitue un sous-bloc. Une matrice de décodage unique et commune est alors suffisante, le mélange des composantes ambisoniques des deux jeux étant précédé d'une correction par des gains g_m^{o1} et g_m^{o2} propres à chaque jeu (Figure 6.2). Ces gains, éventuellement dépendants de la fréquence, sont déterminés d'après la table 3.10 (page 184) en fonction des critères de décodage.

Pour une configuration régulière de haut-parleurs, la matrice de décodage est basée sur le *principe de projection* (cf 3.3.1). En notant `conv` la convention d'encodage en vigueur et d'après (3.23) et (3.64), cette matrice s'écrit:

$$\mathbf{D}_{\text{proj}}^{(\texttt{conv})} = \frac{1}{N} \left(\mathbf{C}^{(\texttt{conv})} \right)^t . \left[\text{Diag} \left(\underline{\alpha}^{(\texttt{N2D})\,\texttt{conv}} \right) \right]^2 \tag{6.1}$$

4. ... ou d'un autre codeur audio-numérique.

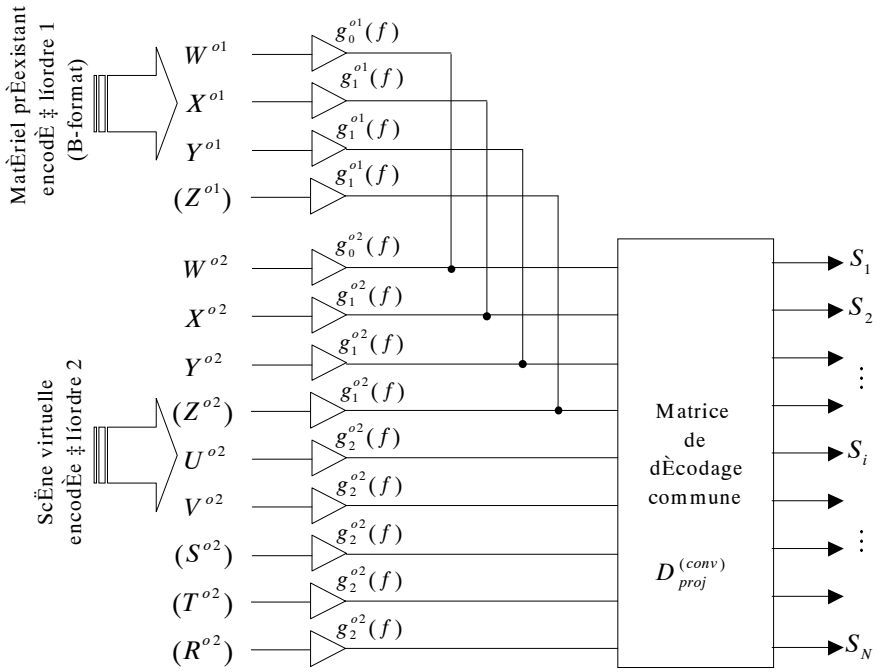

FIG. 6.2 – *Décodage mixte avec matrice de décodage commune aux matériels d'ordres 1 et 2.*

si la configuration est horizontale. Les coefficients de conversions $c_{mm}^{(\mathtt{N2D})\mathtt{conv}}$, constituants du vecteur $\underline{\alpha}^{(\mathtt{N2D})\mathtt{conv}}$, sont décrits table 3.3. Pour une restitution 3D, il faut remplacer la convention `N2D` par la convention `N3D` pour bien appliquer le principe de projection. Avec les configurations semi-régulières[5], la matrice basique (3.92) – définie à l'origine comme une pseudo-inverse (3.63) – n'a plus tout à fait cette forme, mais la compatibilité entre les ordres 1 et 2 subsiste. Enfin, la définition (6.1) peut être adoptée avec des configurations non-régulières[6] (cf 3.3.5 par exemple), notamment pour appliquer le critère *in-phase* étendu (cf 3.3.2).

Ce principe de *décodage mixte* peut être très aisément étendu au mélange avec du matériel ambisonique encodé aux ordres supérieurs à 2.

6.2.2 Format mixte [Mal99c] et son décodage

Le décodage mixte qui vient d'être proposé se présente comme une solution économique dans le cas de mélange de matériels ambisoniques encodés à différents ordres ("représentation mixte"). Mais on observe que le mélange s'effectue après correction des composantes, qui dépend des solutions de décodage, elles-mêmes choisies de préférence en fonction des conditions d'écoute (Figure 3.4). Dans un contexte de transmission, une telle "représentation mixte" de la scène sonore induit donc quelques problèmes. D'un côté, il serait coûteux de transmettre séparément les deux jeux de composantes ambisoniques. D'un autre côté, il n'est pas acceptable de transmettre une représentation ambisonique d'ordre 2 qui soit seulement le mélange des deux jeux de composantes d'origine – avec correction préalable ou non –, puisque cela imposerait des possibilités de décodage figées. Pour résoudre ce dilemme, Malham [Mal99c] a fait la proposition d'un format d'ordre 2 avec dédoublement de la composante W, donc au prix d'un canal transmis supplémentaire W'. Le mélange

5. Cas assez marginaux et peu intéressants pour le décodage d'ordre 2.

6. Dans ce cas, $\mathbf{D}_{\text{proj}}$ ne traduit plus un décodage "basique" au sens d'une reconstruction des composantes ambisoniques, et les critères de colinéarité sur les vecteurs vélocité et énergie ($\vec{u}_E = \vec{u}_V = \vec{u}_S$) ne sont plus assurés.

des deux jeux de composantes est préconisé comme suit:

$$\begin{array}{llll} W & = W^{o1} & W' = W^{o1} + W^{o2} & \\ X & = X^{o1} + X^{o2} & Y = Y^{o1} + Y^{o2} & \text{etc...} \\ U & = U^{o2} & V = V^{o2} & \text{etc...} \end{array} \tag{6.2}$$

A partir des canaux transmis $(W,W',X,Y,[Z],U,V,\ldots)$, il est alors possible de définir un décodage qui soit optimal à la fois pour le matériel originellement encodé à l'ordre 1 et celui encodé à l'ordre 2, à la manière et dans les conditions de la structure de la figure 6.2. Les composantes d'origine W^{o1} et W^{o2} peuvent en effet être extraites sans problème, cependant que les composantes d'ordre 1 sont définitivement mélangées les unes aux autres (X^{o1} avec X^{o2}, etc...) et indissociables, imposant la contrainte: $g_1^{o1}(f) = g_1^{o2}(f) = g_1(f)$. Cela signifie que l'on ne peut contrôler indépendamment que les rapports de gains g_1^{o1}/g_0^{o1}, g_1^{o2}/g_0^{o2} et g_2^{o2}/g_0^{o2}, et non les gains absolus g_m^{oM}: les critères de préservation d'amplitude ou d'énergie ne peuvent donc pas être vérifiés pour les deux matériels ambisoniques originaux à la fois[7]. Nous précisons maintenant la définition et la structure du *décodage mixte optimal*.

Admettons donc que le style de décodage soit choisi – *par exemple* la solution *in-phase* pleine-bande – pour chacun des deux matériels originaux, déterminant ainsi les rapports de gains sus-cités d'après la table 3.10. *Dans l'exemple choisi et pour une restitution 2D*: $g_1^{o1}/g_0^{o1} = 1/2$, $g_1^{o2}/g_0^{o2} = 2/3$ et $g_2^{o2}/g_0^{o2} = 1/6$. Le choix d'un critère de préservation pour le décodage d'ordre 1 permet de déterminer le gain absolu g_0^1, ainsi que les gains[8] $g_1 = g_1^{o1} = g_1^{o2}$, g_0^{o2}, g_2^{o2} par voie de conséquence. Un décodage équivalent à celui de la figure 6.2 peut ainsi être obtenu: sa structure est présentée Figure 6.3. La pondération adéquate des canaux transmis W et W' découle de la relation:

$$g_0^{o1}W^{o1} + g_0^{o2}W^{o2} = (g_0^{o1} - g_0^{o2})W + g_0^{o2}W' \qquad \text{d'après (6.2)} \tag{6.3}$$

Il faut souligner que ce format mixte ne permet un décodage adapté que dans les conditions de restitution énoncées dans la section 6.2.1 précédente. Pour une transposition de ce principe au mélange de matériels d'ordres 1, 2 et 3, par exemple, il faudrait ajouter un canal $W'' = W^{o1} + W^{o2} + W^{o3}$, plus deux ou trois canaux X', Y' (et Z')[9]. Enfin, de tel formats mixtes peuvent évidemment être exploités par un décodeur ambisonique d'ordre 1, en omettant les composantes d'ordre(s) supérieur(s) (U, V, etc...).

6.2.3 *Ambisonic* et effets de salles ou champ réverbéré

Le formalisme ambisonique offre un moyen très intéressant d'appréhender les effets de salle, et plus précisément les propriétés du champ diffus. Les réponses impulsionnelles ambisoniques qui pourraient être mesurées[10] en point donné, dans une salle et pour une source données, constituent en effet une représentation acoustique "réaliste", à la fois temporelle et spatiale de l'effet de salle. Cette représentation est plus souple qu'une paire de réponses impulsionnelles binaurales puisqu'elle n'est pas dédiée à une orientation de tête particulière. Elle est facilement transmissible, peut par la suite être exploitée pour une écoute binaurale ou sur haut-parleurs. Ces aspects sont abordés dans les paragraphes qui suivent. Ils s'accompagnent d'une réflexion sur les propriétés de champ diffus vues à travers ces réponses ambisoniques, et sur la modélisation qui peut en être faite.

7. Dans ces conditions et d'après la table 3.10, l'excès d'énergie du décodage d'ordre 2 par rapport au décodage d'ordre 1 est, selon le type de solution et pour une restitution 2D: 2,22 dB (*basique*), 1,76 dB (*max* r_E) et 1,13 dB (*in-phase*).

8. En toute généralité, ces gains peuvent dépendre de la fréquence. Ce sont alors des filtres (*shelf-filters*).

9. Par exemple, $X' = X^{o1} + X^{o2}$ tandis que $X = X^{o1} + X^{o2} + X^{o3}$.

10. A l'aide d'un microphone ambisonique, s'il existait aux ordres supérieurs.

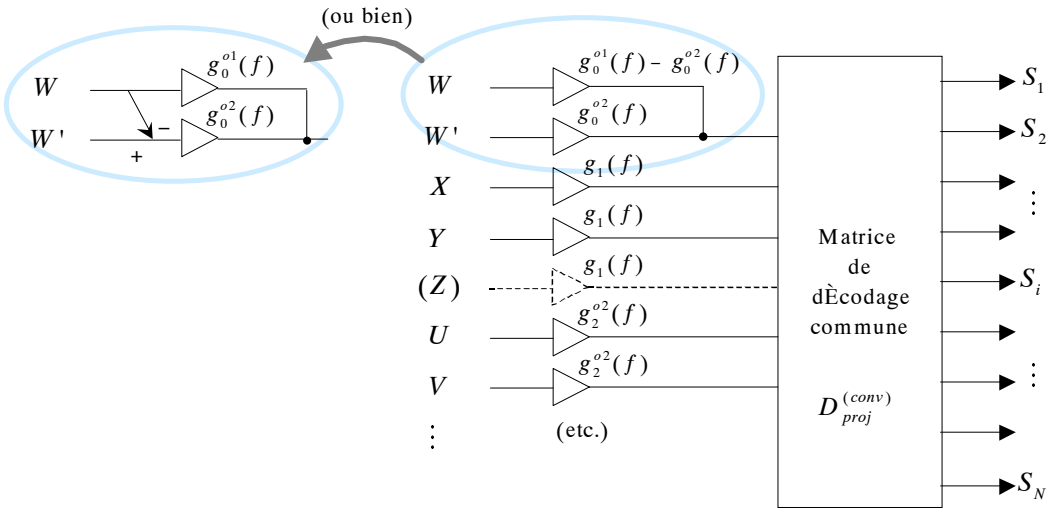

FIG. 6.3 – *Structure exploitant le format mixte (issu d'encodages d'ordres 1 et 2 mélangés) proposé par Malham [Mal99c] pour un décodage optimal, équivalent à celui de la figure 6.2 (à la contrainte* $g_1^{o1}(f) = g_1^{o2}(f) = g_1(f)$ *près).*

En l'absence de microphones pour mesurer de telles réponses, il convient de proposer un moyen de les simuler, par exemple à l'aide d'algorithmes d'acoustique prévisionnelle.

Couplage avec l'acoustique prévisionnelle

Il est d'ores et déjà tout à fait envisageable d'incorporer le principe d'un encodage ambisonique d'ordre élevé à des logiciels d'acoustique prévisionnelle, sans impliquer un véritable surcoût *a priori* (Figure 6.4). Nommons comme antécédent à ce type de couplage, le logiciel *Ramsete*[11] [Far] qui peut fournir des réponses ambisoniques d'ordre 1 (au format B). Le calcul de réponses impulsionnelles ambisoniques d'ordres supérieurs peut ensuite donner lieu, par convolution avec des enregistrements anéchoïques, à la synthèse d'un enregistrement "en salle" convaincant, bien que virtuel.

La modélisation statistique du champ diffus tardif qui apparaît Figure 6.4 est discutée plus loin.

Exploitation des RI3D - Ecoute binaurale

Disposant d'un signal anéchoïque à spatialiser et des réponses impulsionnelles ambisoniques (RI3D), l'étape d'*auralisation* consiste en effectuer le produit de convolution pour calculer le champ ambisonique, puis de le décoder. La figure 6.5 décrit le cas particulier d'une décodage pour une écoute au casque avec *head-tracking*. Cette architecture a l'avantage pouvoir compenser les mouvements de la tête avec un assez faible temps de latence – juste celui du décodage binaural – à condition de pouvoir assurer l'ensemble des calculs.

Si l'on choisit de fixer au départ l'orientation de la tête dans le champ acoustique virtuel, un traitement équivalent (Figure 6.5, droite) peut être réalisé à moindre coût en n'utilisant que deux réponses impulsion-

11. `http://www.ramsete.com`

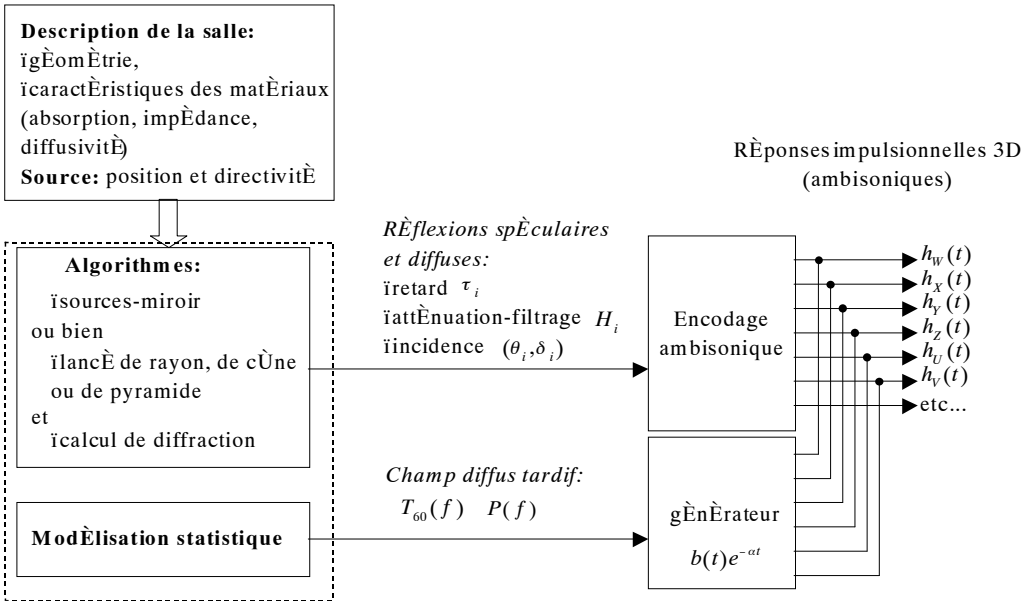

FIG. 6.4 – *Génération de réponses impulsionnelles 3D ambisoniques, par couplage d'algorithmes d'acoustique prévisionnelle avec un encodeur ambisonique et un générateur de bruit blanc pondéré par enveloppe exponentielle.*

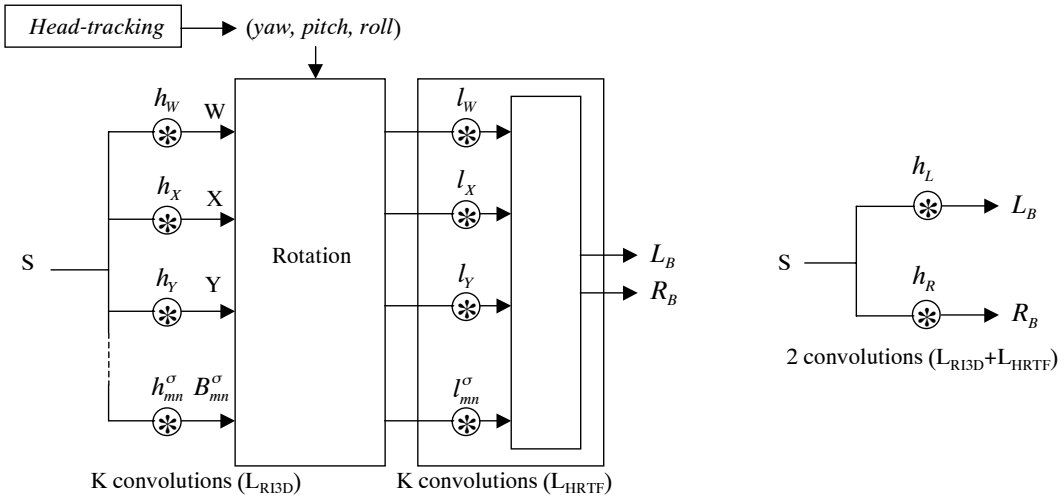

FIG. 6.5 – *A gauche: auralisation du signal anéchoïque S par convolution avec les réponses impulsionnelles ambisoniques 3D (de longueur L_{RI3D}), puis décodage pour présentation binaurale (convolutions par des réponses de longueur L_{HRTF}) (Figure 3.6 pour plus de détails). Une opération intermédiaire de rotation du champ ambisonique permet de tenir compte en temps réel des mouvements de la tête (*head-tracking*). A droite: traitement ne permettant pas de tenir compte des mouvements de la tête.*

nelles $h_L(t)$ et $h_R(t)$:

$$\begin{aligned} h_L &= \sum_{m,n,\sigma} h^{\sigma}_{mn} * l^{\sigma}_{mn} \\ h_R &= \sum_{m,n,\sigma} \sigma h^{\sigma}_{mn} * l^{\sigma}_{mn} \end{aligned} , \tag{6.4}$$

où l'on use d'une notation générique pour désigner les réponses $h_W = h^1_{00}$, $h_X = h^1_{11}$, $h_Y = h^{-1}_{11}$, $h_Z = h^1_{10}$, $h_U = h^1_{22}$, etc...

Au-delà du seul contexte d'acoustique prévisionnelle, ce procédé d'auralisation – qu'il s'agisse d'une restitution binaurale ou sur haut-parleurs – peut être envisagé plus globalement pour nombre d'applications de type "réalité virtuelle". Dans une logique de transmission minimale de données pour des applications sur Internet par exemple, on peut imaginer transmettre séparément des réponses impulsionnelles 3D (ambisoniques) et le signal anéchoïque (éventuellement sous forme de spécifications de son synthétique, cf 6.3.1).

Modélisation statistique du champ diffus

Un module de génération du champ diffus par modélisation statistique est signalé dans la figure 6.4. La modélisation de la réponse tardive diffuse comme un bruit gaussien pondéré temporellement par une exponentielle décroissante [Pol88] a déjà évoquée en 1.2.4. Cette modélisation (équation 1.42, page 31) n'est d'abord proposée que pour une mesure omnidirectionnelle. Nous proposons ici de l'adapter à la partie tardive des réponses impulsionnelles ambisoniques:

$$\begin{aligned} &h'_W(t) = b_W(t)e^{-\alpha(f)t} \\ &h'_X(t) = b_X(t)e^{-\alpha(f)t} \\ &h'_Y(t) = b_Y(t)e^{-\alpha(f)t} \\ &\text{etc...} \end{aligned} \qquad \text{avec} \quad \alpha(f) = \frac{3\ln 10}{T_{60}(f)} \tag{6.5}$$

où nous usons volontairement d'une écriture mathématiquement impropre pour signifier que la décroissance temporelle, de même que l'énergie totale de chaque réponse, sont fonctions de la fréquence: elles dépendent de $T_{60}(f)$ et de $P(f)$. En pratique, la modélisation "*bruit*×*exponentielle*" doit être appliquée par sous-bandes fréquentielles pour générer chaque réponse.

On considère pour le moment que dans l'hypothèse d'un *champ diffus idéal* et *en adoptant implicitement la convention N3D (normalisation 3D)*, les réponses impulsionnelles tardives doivent être de puissances égales et complètement décorrélées. Il doit donc en être de même des bruits $b_W(t)$, $b_X(t)$, etc... La question des propriétés réelles du champ diffus, vues à travers l'encodage ambisonique, est discutée dans la suite.

Champ réverbéré et diffus: qualité de représentation et sensibilité à la restitution

Les impressions spatiales – au sens large (cf 1.3.4), incluant la qualité d'enveloppement – dont est responsable le champ réverbéré dépendent à la fois de la partie précoce et de la partie plus tardive de la réverbération, en plus de la nature du signal émis (sa dynamique, son contenu spectral, etc...). Comme nous l'avons déjà expliqué (en 2.4.4, page 2.4.4, et en 4.3.2), une restitution ambisonique de bas ordre s'accompagne d'un "étalement angulaire" de chaque contribution élémentaire du champ acoustique original (étalement partiellement traduit par l'indice r_E). Plus l'ordre ambisonique M est bas[12], plus il faut s'attendre à une limitation des différences interaurales et de leur fluctuations *dans un domaine haute-fréquence*, ainsi qu'à une réduction de

12. Et rappelons qu'à ordre M égal, la restitution 3D est "moins bonne" dans le plan horizontal que la restitution 2D!

la décorrélation interaurale, soit, sur le plan perceptif, à une dégradation des impressions spatiales. Il devrait être assez simple de vérifier mathématiquement l'impact sur l'IACC, en se basant par exemple sur le modèle d'un champ diffus idéal (voir plus haut).

En d'autres termes, cela signifie que la troncature du champ ambisonique à un ordre M a un effet "passe-bas" sur la qualité du champ diffus restitué, observé dans un domaine haute-fréquence. Cela peut remettre en cause partiellement la mesure objective de l'impression spatiale basée sur la seule observation des composantes latérale Y et omnidirectionnelle W. Néanmoins, les propriétés du champ restitué et perçu (par un auditeur centré) restent conformes à l'original dans un domaine basse-fréquence, lequel – certes – s'élargit avec l'ordre M.

Il resterait à caractériser et modéliser plus finement le champ diffus, vu à travers la représentation ambisonique, dans des cas réels sinon réalistes. Autant la partie précoce de la réverbération peut être simulée de façon fiable – à défaut de microphone ambisonique d'ordre supérieur pour la mesurer –, autant les algorithmes géométriques (Figure 6.4) peuvent introduire des artefacts dans la partie tardive. Il serait néanmoins intéressant de vérifier si la réponse générée tend bien vers les caractéristiques diffuses idéales suggérées plus haut (6.5). Enfin, il faudrait réaliser des tests subjectifs pour quantifier l'effet de la troncature ambisonique sur la qualité perçue du champ diffus, en fonction de l'ordre M.

Conséquences

Les réponses impulsionnelles h^{σ}_{mn} ont d'autant moins d'impact sur le champ basse-fréquence perçu en position d'écoute centrée (ou en simulation binaurale) que leur ordre m est élevé. Puisque l'atténuation du son lors de sa propagation entraîne généralement une décroissance beaucoup plus importance de la réverbération en hautes fréquences qu'en basses fréquences, on peut supposer qu'il est possible de diminuer la longueur des réponses h^{σ}_{mn} d'ordre m élevé sans que cela soit perceptible. On peut ainsi espérer faire des économies de calcul sur les opérations de convolution.

Terminons par une remarque au sujet de la transposition de modélisation de la réponse diffuse tardive (6.5) au cas d'une réverbération par réseau de retards bouclés (cf 5.2.4), réseau dont les sorties sont supposées décorrélées. Par analogie, il devrait être correct d'assigner chaque sortie du réseau FDN à un canal ambisonique différent[13]. Mais il serait bon d'évaluer la pertinence de cette stratégie dans le cas d'une relativement faible densité temporelle d'échos en sortie du réverbérateur.

6.2.4 Travaux et études en perspective

Nous regroupons ici un certain nombre de thèmes relatifs à l'extension d'*Ambisonic* aux ordres supérieurs, susceptibles ou méritant d'être développés à l'avenir.

Prises de son d'ordre supérieur à 1

Bien qu'apparemment couverte, dans le principe, par le brevet [CG77], la réalisation de microphones ambisoniques d'ordres supérieurs à 1 n'a jusqu'ici pas vu le jour. Nous n'en avons nous-mêmes exploré que quelques aspects théoriques dans la section 3.4. Il s'agit cependant d'une prospection qui intéresse vivement le monde de la prise de son, et qui fait partie des défis à relever par les "ambisonicistes". Une thèse est d'ailleurs entamée sur les microphones d'ordre 2 (Philip Cotterell, Université de Reading, UK).

13. Au lieu d'une distribution panoramique, comme c'est le cas dans notre implémentation simplifiée.

Spécifications pour l'extension du format ambisonique aux ordres supérieurs

Il faut d'ores et déjà s'intéresser au stockage et à la transmission de matériel ambisonique d'ordre supérieur. Il est important pour cela d'établir des conventions claires et de dresser la liste des spécifications à introduire dans les entêtes des fichiers-son ambisoniques. En voici quelques propositions:

- Spécifications classiques d'un format audio-numérique: fréquence d'échantillonnage, quantification, répartition des canaux en fichiers multiples ou bien voies entremêlées, etc...
- Spécifications des conventions d'encodage: aspects développés en 3.1.2.
- Ordonnancement des canaux ambisoniques.
- Spécification de canaux supplémentaires (dédoublés pour un format mixte, cf 6.2.2).
- Suivi de la production (historique/"traçabilité"): origine (enregistrement, synthèse), mélange éventuel de plusieurs matériels différents, correction éventuelle de l'encodage pour une égalisation énergétique spatiale (cf 3.1.4), etc...

Il conviendrait d'ajouter à la liste des formats, les formats dérivés traditionnels (UHJ, BHJ, etc...) ou plus récents (G-format) évoqués en 2.4.2.

Décodage actif: détection et émulation des ordres supérieurs

L'idée d'un décodage ambisonique actif n'a été que très peu évoquée au cours de ce document. Cette piste a pourtant fait l'objet d'une réflexion approfondie, et mériterait d'être encore développée. Elle se base sur un principe de détection – à partir du B-format – des événements acoustiques élémentaires (ondes planes) qui peuvent émerger ou prédominer sporadiquement. Pour chaque événement détecté[14], il s'agirait d'isoler le mieux possible le signal associé, puis de reconstituer les composantes ambisoniques d'ordres supérieurs. Même si cette émulation des ordres supérieurs ne peut pas être espérée en permanence dans le cas d'une scène sonore complexe, on peut s'attendre à une amélioration lors des transitoires, qui on le sait, sont d'une importance capitale dans le processus de localisation.

6.3 Contextes d'applications et conclusion

6.3.1 Applications multimédias et liées à Internet

Une des caractéristiques des applications multimédias est le brassage d'une quantité d'objets sonores, visuels et audio-visuels de natures de plus en plus diversifiées. Dans les applications de navigation 3D en particulier, les scènes audio-visuelles visitées deviennent facilement hautement composites: sources sonores spatialisés et objets "solides" 3D, plaquage 2D (matériel audio-visuel, films...) sur des surfaces 3D, ambiance sonore stéréophonique, son multi-canal, etc... (Figure 3.1 page 148 et figure 6.6 pour les aspects audio). Il faut de plus faire face à des ressources et configurations variées au niveau de l'utilisateur, ce qui implique au niveau du rendu sonore, de faire appel à des techniques de spatialisation qui puissent s'adapter au dispositif de restitution, exploiter au mieux les ressources, et être modérément coûteuses au besoin. Enfin, le contexte *Internet* impose des contraintes supplémentaires, notamment la capacité du réseau (limitation du débit), qui pousse à la restriction de la quantité de données transmises du serveur vers le client.

L'approfondissement dont l'approche ambisonique a fait l'objet au cours de cette thèse, a été en partie motivé par le fait qu'elle semblait proposer des solutions intéressantes en réponse à ces enjeux. Nous avons déjà insisté sur la notion de *restitution "à géométrie variable"*, c'est-à-dire adaptable à divers dispositifs de

14. Plusieurs événements peuvent en principe être détecté simultanément, dans la limite du nombre de canaux ambisoniques analysés.

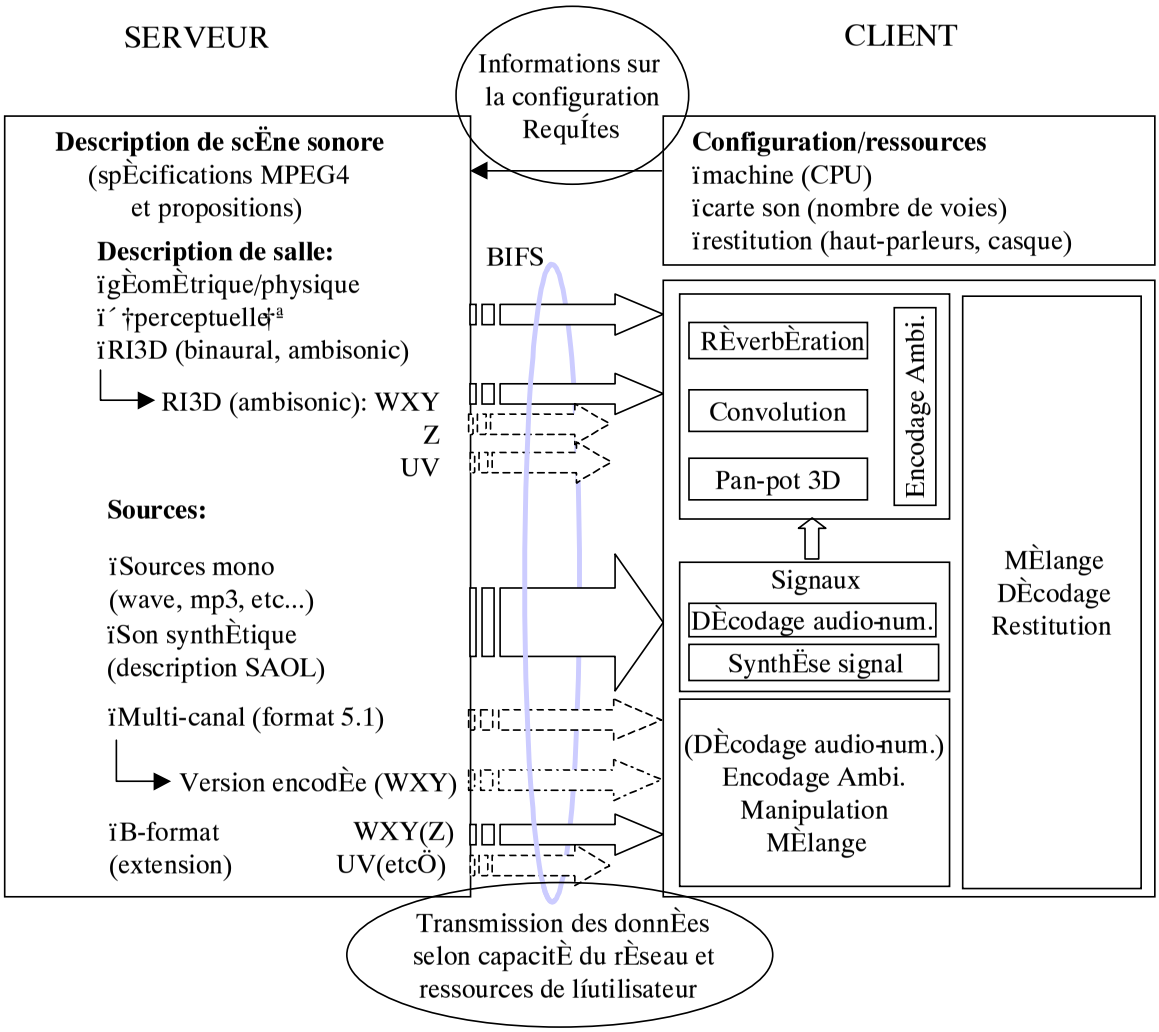

FIG. 6.6 – *Schéma serveur-client simplifié relatif à la partie sonore de la navigation 3D dans les scènes virtuelles sur Internet. Quelques adjonctions aux spécifications de la norme MPEG4 mettent en avant certaines potentialités de l'approche ambisonique.*

restitution et à différentes conditions d'écoute (voir la conclusion du chapitre 3, la section 3.1.1 et presque toute la partie II) [DRP98]. L'idée d'une *représentation compacte* du champ sonore a été appliquée à la représentation de scène sonore composite (Figure 3.1), et envisagée pour la transmission *via* le format B horizontal de matériel à l'origine multi-canal (section 6.1): cette option est signalée dans le schéma de la figure 6.6. Ajoutons au passage que dans une version moins interactive, la spatialisation et le mélange des sources vers le format B pourraient être faits au niveau du serveur et non plus chez le client, afin de limiter – selon les cas – la quantité de données à transmettre.

L'extension de l'approche ambisonique aux ordres supérieurs fait du format ambisonique une représentation spatiale à plusieurs niveaux de résolution. Elle lui donne ainsi une prédisposition naturelle à la *scalabilité*[15], c'est-à-dire la possibilité de n'utiliser – par exemple selon les limites du dispositif de restitution – ou de ne transmettre – dans un contexte de transmission à débit variable – qu'un sous-ensemble des canaux (ceux de plus bas ordre), ce qui revient à tronquer la représentation. Dans la figure 6.6, cela illustré par la transmission optionnelle des canaux U, V (et Z) qui pourraient permettre d'atteindre une meilleure précision de l'imagerie sonore à condition d'avoir suffisamment de haut-parleurs, mais sans lesquels les informations directionnelles sont préservées et l'intelligibilité spatiale de la scène subsiste.

Enfin, *Ambisonic* propose une autre possibilité de représenter de façon concise un champ acoustique: la représentation spatiale et temporelle du champ réverbéré par des réponses impulsionnelles ambisoniques (cf 6.2.3), associée à un signal monophonique et anéchoïque. Ce dernier peut lui-même être représenté et transmis sous une forme minimale, en étant décrit par un ensemble de paramètres de synthèse (langage SAOL[16]), et synthétisé au niveau du client. Cette association est suggérée figure 6.6.

6.3.2 Autres contextes d'application

De par son aptitude à s'adapter à des dispositifs et des situations d'écoute variés, *Ambisonic* est une technique de restitution polyvalente qui est susceptible d'intéresser – et qui intéresse déjà! – de nombreux domaines. En plus des contextes "traditionnels" que sont la diffusion et la création musicales, et la diffusion cinématographique, l'extension vers les ordres supérieurs peut donner à *Ambisonic* un rôle important à jouer dans des contextes requérant une haute-qualité d'image sonore spatiale et/ou dédiés à des auditoires étendus, comme la téléprésence, le cinéma "3D", les applications de réalité virtuelle.

Téléprésence

La téléprésence est une forme avancée de la téléconférence où l'on cherche à rendre transparente l'interface audio-visuelle de communication entre des groupes d'interlocuteurs distants. Cette interface est généralement présentée sous la forme d'un ou plusieurs pans de mur – écran de projection et mur de haut-parleurs étant *a priori* confondus. Les solutions les mieux pressenties pour remplacer les systèmes multi-stéréophoniques jusqu'ici utilisés, et qu'on soupçonne être une source de fatigue pour les sujets, sont les *systèmes holophoniques* [Nic99] dont le *WaveField Synthesis* (WFS) [BVT95] est un bon représentant. Alors qu'un système ambisonique traditionnel d'ordre 1 aurait fait un piètre candidat face à l'exigence du zone d'écoute étendue[17], il est intéressant de reconsidérer dans ce contexte les atouts de systèmes ambisoniques d'ordre(s) élevé(s) qui, comme le WFS, savent tirer profit d'un grand nombre de haut-parleurs. Nous nous contenterons ici de replacer quelques remarques faites à l'occasion de l'étude 4.2.3.

15. Néologisme surtout utilisé dans le domaine du codage d'image.
16. Voir par exemple `http://saol.net`.
17. Et a rapidement été écarté pour cette raison [Nic99].

Une scène visuelle projetée sur un écran, c'est *l'unique point de vue* d'une seule caméra offert à tous, l'information de profondeur restant accessible à chaque spectateur par les effets de perspective et d'ombrage. *Ambisonic* offre lui aussi un "point de vue" (ou plutôt "d'écoute") unique[18], qui peut être extrapolé à un auditoire élargi de différentes façons (cf 4.2.3). Avec un décodage sans correction du champ proche des haut-parleurs et pour un système d'ordre élevé, une source sonore lointaine tend elle-même à être projetée sur l'écran audio-visuel – l'information de profondeur étant préservée grâce au rapport entre le son direct, les réflexions précoces et le champ réverbéré – et peut ainsi coïncider, quel que soit le point de vue du spectateur, avec l'image visuelle associée. Il n'en est plus de même si l'effet de champ proche des haut-parleurs est compensé au niveau du décodage ambisonique, ou encore avec l'holophonie de type WaveField Synthesis, qui permet de "placer" des images sonores *au-delà* du mur de haut-parleurs. Une question se pose alors: «Faut-il, pour des raisons de cohérence audio-visuelle, préférer la restitution ambisonique (première version)? Ou bien est-il préférable, pour des raisons de confort, de restituer le plus fidèlement possible le champ acoustique original, et accepter une distorsion directionnelle avec les images visuelles projetées?». Un autre aspect semble mettre en faveur l'holophonie (WFS): le fait qu'elle assure la même qualité de reconstruction en tout point de la zone d'écoute, ce qui n'est pas le cas avec *Ambisonic* (position centrale privilégiée).

D'un point de vue plus matériel, il faut enfin reconnaître que les dispositifs concentriques habituellement recommandés pour la restitution ambisonique ne s'accordent pas tout à fait avec la géométrie – plutôt plane et frontale – du mur de téléprésence.

Projection audio-visuelle sur un dôme géodésique

La restitution ambisonique ambisonique d'ordres supérieurs telle que nous l'avons décrite en 3.3.5 pour des configurations hémisphériques pourrait être exploitée en association avec certaines formes de cinéma 3D basées sur une projection à l'intérieur d'un dôme géodésique (par exemple, "La Géode"[19] ou le projet de l'ACAT [Ven]). Certaines applications de réalité virtuelle pourraient adopter des structures similaires, quoique de moindre envergure. La notion de projection audio-visuelle coïncidente évoquée plus haut avec la téléprésence garde ici tout son importance.

Productions musicales et cinématographiques

Les dispositifs multi-canal (à cinq, six, sept ou plus de haut-parleurs) pour la diffusion musicale ou le cinéma, devraient pouvoir exploiter les nouvelles techniques de pan-pot ambisonique d'ordre supérieur et encourager leur amélioration dans les cas plus délicats de géométrie non-régulière (cf 3.3.4).

6.3.3 Conclusion

Au cours de ce chapitre, nous avons évoqué – et parfois illustré en détail – quelques unes des nombreuses potentialités offertes par l'approche ambisonique et son extension aux ordres supérieurs, applicables au domaine multimédia, à la navigation 3D sur Internet comme à bien d'autres contextes: réalité virtuelle, création musicale, cinéma, etc...

En complément aux développements théoriques (partie II) et pratiques (5) que nous avons proposés au cours de cette thèse, différents axes de recherche et de développement se dessinent, qui risquent d'accompagner l'essor de cette approche:

– Le développement de microphones ambisoniques d'ordre(s) supérieur(s).

18. Ce point de vue est celui du microphone, bien qu'il n'en existe pas encore pour les ordres supérieurs. On peut aussi adapter le concept de "source notionnelle" [BVT95] grâce au simple principe d'encodage ambisonique.

19. `http://www.cite-sciences.fr/francais/ala_cite/spectacl/geode/global_fs.htm`

– L’établissement de conventions et de spécifications pour un format ambisonique “universel".
– L’optimisation et l’évaluation du décodage ambisonique pour la présentation binaurale.
– L’approfondissement du décodage hybride (faisant intervenir les ordres supérieurs) pour les systèmes multi-canal courants.
– La conception d’un décodage actif (cf 6.2.4).

Conclusion générale

Développements, résultats théoriques et pratiques

Conformément aux objectifs énoncés en introduction, ce travail de thèse a concerné pour une grande partie l'extension de l'approche ambisonique aux ordres supérieurs (partie II), sa mise en oeuvre et son utilisation pour la spatialisation et le mélange de sources (partie III). Au préalable (Partie I), nous avons eu à coeur de justifier et interpréter les théories de la localisation basées sur les vecteurs vélocité $\vec{V}$ (basse-fréquence) et énergie $\vec{E}$ (haute-fréquence), introduites par Gerzon pour l'optimisation "psychoacoustique" de décodage ambisonique, et plus généralement de caractériser les artefacts perceptibles d'une reconstruction imparfaite du champ acoustique. Après avoir montré dans cadre général (chapitre 1) le lien de prédiction entre caractérisation de la propagation – locale avec $\vec{V}$, globale avec $\vec{E}$ – et l'effet supposé de localisation selon la mobilité de la tête, nous avons étendu et complété l'analyse avec les systèmes de restitution (chapitre 2) et mis en évidence la relation entre mode de représentation et/ou dispositif de restitution, et les qualités potentielles de restitution. Cette relation se montre particulièrement intime avec ambisonic, qui offre – d'après des arguments théoriques – le meilleur compromis entre concision et souplesse de la représentation, et conditions d'écoute naturelles.

Nous avons présenté ensuite (chapitre 3) l'extension du formalisme d'encodage et des principes de décodage (2D et 3D) à tout ordre. Ayant précisé la notion d'échantillonnage directionnel et sa propriété de régularité, qui intervient pour les problèmes de décodage et de prise de son, nous avons généralisé les formes de décodage existant à l'ordre 1 en *trois familles de solutions*: "*basic*" qui optimise la reconstruction locale (BF-centrée) – décodage de base dont dérivent les deux autres –, "*max* r_E" qui optimise la propagation globale (HF-excentrée), et "*in-phase*" qui minimise les distorsions directionnelles hors-centre ou à proximité des haut-parleurs. Les matrices de décodage associées sont applicables en pleine-bande ou combinées par sous-bandes de fréquence en vue d'un rendu optimal selon les conditions d'écoute (individuelle/centrée ou collective/excentrée).

Des évaluations objectives de la restitution pour diverses conditions d'écoute, (chapitre 4), étayées par des écoutes informelles, ont mis en évidence l'apport des ordres supérieurs et des solutions optimisées. Nous montrons ainsi d'après des données objectives, que les ordres supérieurs permettent non seulement d'élargir la zone de reconstruction acoustique – proportionnellement à la longueur d'onde –, mais que même en dehors du domaine de reconstruction, ils améliorent la précision et la robustesse des images sonores et préservent mieux les impressions spatiales (séparation latérale). En effet, ce sont aussi les caractéristiques de propagation globale ($\vec{E}$) associées à chaque événement acoustique élémentaire (fronts d'onde originaux) qui sont améliorées.

Parmi d'autres techniques de spatialisation (pan-pot, binaural, transaural et synthèse d'effet de salle), nous avons mis en oeuvre et expérimenté les techniques ambisoniques jusqu'à l'ordre 2 – incluant les modes de restitution binaurale et transaurale –, le tout étant incorporé au sein d'une interface (sur PC) qui permet la manipulation et le mélange de sources monophoniques, multi-canal et ambisoniques (chapitre 5). On a

pu vérifier par exemple qu'en mode binaural, le passage par Ambisonic – même à l'ordre 2 – permettait de manipuler un plus grand nombre de sources que par simulation binaurale directe.

Enfin, d'autres expériences ont illustré le codage/décodage ambisonique (WXY) de matériel multi-canal (codec "5-3-5", cf 6.1), pour une transmission à débit réduit. On en retient, avec un certain nombre d'enregistrements, une assez nette dégradation des qualités spatiales (précision, consistance, stabilité). La combinaison de cet encodage avec la compression audio-numérique MPEG2 sur les canaux WXY intermédiaire, montre cependant un compromis "dégradation spatiale/bruit de quantification" plus acceptable que ce que propose le codage MPEG2 appliqué aux cinq canaux originaux, pour un même débit global *relativement faible* (donc une qualité assez médiocre).

Bilan et perspectives

Si d'une certaine manière, Ambisonic apporte une réponse globale très satisfaisante aux enjeux initiaux, elle n'est pas forcément *la* solution optimale à *chaque* problème, s'agissant par exemple du codec "5-3-5" (cf 6.1) – auquel certains procédés *surround* "5-2-5" (cf 2.3.4) peuvent être préférés – et de la synthèse binaurale performante d'une scène complexe – face aux nouvelles stratégies de type *binaural B-format* (cf 2.5.1). Quoiqu'il en soit, elle constitue une approche véritablement prometteuse, dont les développements intéressent non seulement le domaine multimédia dans son sens le plus large, mais également la production/diffusion musicale et artistique.

En complément des écoutes informelles et des interprétations théoriques approfondies données dans ce document, il resterait à effectuer des tests d'écoute formels pour valider complètement l'approche ambisonique et les théories sous-jacentes. L'outil présenté au chapitre 5 pourrait servir à ce type d'évaluation, moyennant éventuellement quelques adaptations. Ces tests devraient aussi permettre de préciser les conditions d'application optimale des différentes solutions de décodage proposées en fonction de la zone d'écoute.

Il semble également important d'achever le projet d'une implémentation générique (en C++) des différents modules d'un système ambisonique "universel", sans limitation d'ordre, incluant toutes les propositions du chapitre 3 et des sections 6.2.1 et 6.2.2. Dans le même temps, il serait bon de préciser les spécifications d'un format ambisonique universel.

On espère par ailleurs pouvoir améliorer la restitution en mode binaural, et ainsi que le décodage pour les configurations non-régulières.

L'extension d'Ambisonic aux ordres supérieurs fait l'objet d'un intérêt croissant depuis une demi-décennie. Outre la présente thèse et d'autres études parallèles[20], il faut s'attendre à ce que son essor soit bientôt promu par le développement de microphones d'ordres supérieurs, par l'exploitation – hybride (cf 3.3.4) – des ordres supérieurs pour la production multi-canal (même au format 5.1), et par l'enrichissement en nombre de haut-parleurs des dispositifs "courants".

20. Par exemple aux Universités de York (D. Malham) et de Derby, Royaume-Uni.

Bibliographie

[ANS78] "American National Standard Method for the Calculation of the Absorption of Sound by the Atmosphere". ANSI S1.26-1978, American Institute of Physics (for Acoustical Society of America), New York, 1978.

[Bar70] M. Barron. The subjective effects of first reflections in concert halls – The need for lateral reflections. *J. Sound Vib.*, 15(4):475–494, 1970.

[Bau61] Benjamin B. Bauer. "Phasor Analysis of Some Stereophonic Phenomena". *J. Acoust. Soc. Am*, 33(11):1536–1539, November 1961.

[Bau97] Jerry Bauck. A New Loudspeaker Technique for Improved 3D Audio. *presented at the AES 14th International Conference, Seattle*, June 1997.

[BC92] Jerry Bauck and Duane H. Cooper. "Generalized Transaural Stereo". *presented at the 93rd Convention AES, Preprint 3401*, October 1992.

[BCDS89a] M.C. Botte, G. Canévet, L. Demany, and C. Sorin. *Audition binaurale et localisation auditive. Aspects physiques et psychoacoustiques.*, chapter 3, pages 83–122. In Canevet [BCDS89b], 1989.

[BCDS89b] M.C. Botte, G. Canévet, L. Demany, and C. Sorin. *Psychoacoustique et perception auditive*. Audition. INSERM/SFA/CNET, 1989.

[Beg94] Durand R. Begault. *"3-D Sound for Virtual Reality and Multimedia"*. AP Professional, Cambridge, Massachusetts, 1994.

[Ber75] Benjamin Bernfeld. "Simple Equations for Multichannel Stereophonic Sound Localization". *J. Audio Eng. Soc.*, 23(7):553–557, September 1975.

[Bla83] Jens Blauert. *"Spatial Hearing: The Psychophysics of Human Sound Localization"*. MIT Press, Cambridge, Massachusetts, 1983.

[Blu33] A.D. Blumlein. "Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproduction systems". British Patent Specification 394,325, issued June 14, 1933.

[Bru96] J. Bruck. The KFM 360 Surround - A purist approach. *Presented at the 103rd AES Convention, preprint 4637*, 1996.

[Bru98] Michel Bruneau. *Manuel d'acoustique fondamentale*. Hermès, 1998.

[BV95] Jeffery S. Bamford and John Vanderkooy. "Ambisonic Sound for Us". *presented at the 99th Convention AES, preprint 4138*, October 1995.

[BVT95] M. M. Boone, E. N. G. Verheijen, and P. F. Van Tol. Spatial Sound-Field Reproduction by Wave-Field Synthesis. *J. Audio Eng. Soc*, 43(12):1003–1011, 1995.

[Cas94] La perception auditive des sons musicaux*, in* Psychologie de la musique. P.U.F., 1994. Sous la direction de A.Zenatti.

[CB89] Duane H. Cooper and Jerald L. Bauck. Prospects for transaural recording. *J. Audio Eng. Soc.*, 37(1/2):3–19, Jan./Feb. 1989.

[CCI92] "Multi-channel stereophonic sound system with and without accompanying picture". Recommendation 775, CCIR, 1992.

[CG77] Peter Graham Craven and Michael Anthony Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs, August 1977. US Patent#4042779.

[CM82] Lothar Cremer and Helmut A. Müller. *Principles and Applications of Room Acoustics*. Applied Science Publishers, 1982. Translated by Theodore J. Schultz (original version: 1978).

[Con78] Roland Condamines. *"Stéréophonie - Cours de relief sonore théorique et appliqué"*. Masson, 1978. Collection technique et scientifique des télécommunications.

[CS72] D. H. Cooper and T. Shiga. Discrete-Matrix Multichannel Stereo. *J. Audio Eng. Soc.*, 20:346–360, June 1972.

[DK99] Glenn Dickins and Rodney Kennedy. Towards optimal sound field representation. *Presented at the AES 106th Convention, Munich*, 1999.

[DKS94] Bengt-Inge Dalenbäck, Mendel Kleiner, and Peter Svensson. "A Macroscopic View of Diffuse Reflection". *J. Audio Eng. Soc.*, 42(10):793–807, October 1994.

[Dre93] R. Dressler. *"Dolby Pro Logic Decoder Principles of Operation"*. Dolby Laboratories, Inc., 1993. URL: http://www.dolby.com/ht/ds&pl/whtppr.html.

[DRP98] Jérôme Daniel, Jean-Bernard Rault, and Jean-Dominique Polack. Ambisonic encoding of other audio formats for multiple listening conditions. *Presented at the AES 105th Convention, preprint 4795*, September 1998.

[DRP99] Jérôme Daniel, Jean-Bernard Rault, and Jean-Dominique Polack. Acoustic properties and perceptive implications of stereophonic phenomena. *Proceedings of the AES 16th International Conference*, April 1999.

[Dud] Richard O. Duda. A general sound localisation model. URL: http://www-engr.sjsu.edu/ duda/Duda.R.GSLM.html.

[Dur98] Xavier Durot. *Définition d'un modèle psychoacoustique dans le contexte du codage audionumérique à réduction de débit*. PhD thesis, CCETT/ Université Rennes 1, Janvier 1998.

[Ele98] Richard G. Elen. "G+2", A Compatible, Single-Mix DVD Format for Ambisonic Distribution, 1998.

[Far] Angelo Farina. *Angelo Farina Home Page*. URL: http://pcfarina.eng.unipr.it.

[Fra64] N.V. Franssen. *Stéréophonie*. Bibliothèque technique Philips, 1964.

[Fur99] Richard Furse. 3d audio links and information, 1999. URL: http://www.muse.demon.co.uk/3daudio.htm.

[Gar95] William G. Gardner. "Transaural 3-D audio". Technical Report 342, M.I.T Media Laboratory Perceptual Computing Section, July 1995.

[Gar96] William G. Gardner. "Efficient Convolution without Input-Output Delay". *J. Audio Eng. Soc.*, 43(3):127–136, March 1996.

[Gas98] Jean-Dominique Gascuel. *?* PhD thesis, Thèse de Doctorat, 1998.

[Ger73] Michael A. Gerzon. "Periphony: With-height Sound Reproduction". *J. Audio Eng. Soc.*, 21:2–10, January 1973.

[Ger74] Michael A. Gerzon. "Surround Sound Psychoacoustics". *Wireless World*, 80:483–486, December 1974.

[Ger77] Michael A. Gerzon. "Criteria for Evaluating Surround-Sound Systems". *J. Audio Eng. Soc.*, 25:400–408, June 1977.

[Ger85] Michael A. Gerzon. “Ambisonics in Multichannel Broadcasting and Video”. *J. Audio Eng. Soc.*, 33(11):859–871, November 1985.

[Ger92a] Michael A. Gerzon. “Ambisonic Decoder for HDTV”. *presented at the 92nd Convention AES, Preprint 3345*, March 1992.

[Ger92b] Michael A. Gerzon. “General Metatheory of Auditory Localisation”. *presented at the 92nd Convention AES, Preprint 3306 (but mostly written in 1976)*, March 1992.

[Ger92c] Michael A. Gerzon. “Hierarchical System of Surround Sound Transmission for HDTV”. *presented at the 92nd Convention AES, Preprint 3339*, March 1992.

[Ger92d] Michael A. Gerzon. “Panpot Laws for Multispeaker Stereo”. *presented at the 92nd Convention AES, Preprint 3309*, March 1992.

[Ger92e] Michael A. Gerzon. “Psychoacoustic Decoders for Multispeaker Stereo and Surround-Sound”. *presented at the 93rd Convention AES, Preprint 3406*, October 1992.

[Ger92f] Michael A. Gerzon. “The Design of Distance Panpots”. *presented at the 92nd Convention AES, Preprint 3423*, March 1992.

[Gla95] Raplh Glasgal. Ambiophonics: The Synthesis of Concert-Hall Sound Fields in the Home. *Presented at the 99th AES Convention, preprint 4413*, October 1995.

[GM94] Bill Gardner and Keith Martin. “HRTF Measurements of a KEMAR Dummy-Head Microphone”. Technical Report 280, M.I.T Media Laboratory Perceptual Computing Section, May 1994. `http://sound.media.mit.edu/~ kdm/hrtf.html`.

[Gri92] David Griesinger. “IALF - Binaural Measures of Spatial Impression and Running Reverberance”. *presented at the 92nd Convention AES, Preprint 3292*, March 1992.

[Gri96a] David Griesinger. Multichannel Matrix Surround Decoders for Two-Eared Listeners. *Presented at the 101st AES Convention, Preprint 4402*, November 1996.

[Gri96b] David Griesinger. “Spaciousness and Envelopment in Musical Acoustics”. *presented at the 101st Convention AES, Preprint 4401*, November 1996.

[Gri97] David Griesinger. Progress in 5-2-5 Matrix Systems. *Presented at the 103th AES Convention*, 1997.

[HK97] Jyri Huopaniemi and Matti Karjalainen. Review of Digital Filter Design and Implementation Methods for 3-D Sound. *Presented at the AES 102nd Convention, preprint 4461*, March 1997.

[ITU94] “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”. Recommendation BS. 1116, ITU-R, 1994.

[Jes73] M. Jessel. *Acoustique théorique, propagation et holophonie*. Masson, Paris, 1973.

[JKM+93] Jean-Pascal Jullien, Eckhard Kahle, Martine Marin, Olivier Warusfel, Georges Bloch, and Jean-Marc Jot. “Spatializer: A Perceptual Approach”. *presented at the 94th Convention AES, Preprint 3465*, March 1993.

[JLP99] Jean-Marc Jot, Véronique Larcher, and Jean-Marie Pernaux. A comparative study of 3-d audio encoding and rendering techniques. *Proceedings of the AES 16th International Conference*, April 1999.

[JLW95] Jean-Marc Jot, Véronique Larcher, and Olivier Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. *Presented at the 98th AES Convention, preprint 3980*, February 1995.

[Jot92] Jean-Marc Jot. *Etude et réalisation d’un spatialisateur de sons par modèles physiques et perceptifs*. PhD thesis, Ecole Nationale Supérieure des télécommunications, September 1992.

[Jot97] Jean-Marc Jot. Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In *ICMC*, September 1997.

[JWL98] Jean-Marc Jot, Scott Wardle, and Véronique Larcher. Approaches to binaural synthesis. *Presented at the 105th AES Convention, preprint 4861*, September 1998.

[KIH99] Abhijit Kulkarni, S.K. Isabelle, and H.S.Colburn. Sensitivity of human subjects to head-related transfer-function phase spectra. *J.Acoust.Soc.Am*, 105(5):2821–2840, May 1999.

[KNH97] Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. "The "Stereo Dipole" – Binaural Sound Reproduction Using Two Closely Spaced Loudspeakers". *presented at the 102nd Convention AES, Preprint 4463*, March 1997.

[KNKH97] Yuvi Kahana, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada. "Multi-Channel Sound Reproduction Using a Four-Ear Dummy-Head". *presented at the 102nd Convention AES, Preprint 4465*, March 1997.

[Kru96] David Kruglinski. *Atelier Visual C++ (Version 4.0)*. Microsoft Press, 1996.

[Kuh77] George F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J.Acoust.Soc.Am.*, 62(1):157–167, July 1977.

[Kut79] H. Kuttruff. *Room Acoustics*. Applied Science Publishers Ltd, 1979.

[Lip86] Stanley P. Lipshitz. Stereo Microphone Techniques... Are the Purists Wrong? *J. Audio Eng. Soc.*, 34(9):716–744, Sept. 1986.

[LJ97] Véronique Larcher and Jean-Marc Jot. Techniques d'interpolation des filtres audio-numériques: Applications à la reproduction spatiale des sons sur écouteurs. *Présentation au Congrès Français d'Acoustique (S.F.A., TEKNEA 1997).*, Avril 1997.

[LJGW00] Véronique Larcher, Jean-Marc Jot, J. Guyard, and Olivier Warusfel. Study and Comparion of Efficient Methods for 3D Audio Spatialization Based on Linear Decomposition of HRTF Data. *Presented at the 108th AES Convention, preprint 5097*, 2000. February.

[Mak62] Y. Makita. "Localisation directionnelle du son dans un champ sonore stereophonique". *Revue de l'U.E.R., Cahier A - Technique*, (73):102–108, Juin 1962.

[Mal90] David Malham. "A Technique for Low Cost, High Precision, Three Dimensional Sound Diffusion". *Original version given at ICMC Glasgow*, 1990.

[Mal92] David Malham. "Experience with Large Area 3D Ambisonic Sound Systems". *Proceedings of the Institute of Acoustics*, 14-5:209–215, 1992.

[Mal93] Dave G. Malham. *"3-D sound for virtual reality systems using Ambisonic techniques"*. University of York, 1993. URL: http://www.york.ac.uk/inst/mustech/3d_audio/vr93papr.htm.

[Mal95] Dave G. Malham. *"Basic Ambisonics"*. University of York, 1995. URL: http://www.york.ac.uk/inst/mustech/3d_audio/ambsfaq.htm.

[Mal99a] D.G. Malham. Higher order Ambisonics systems for the spatialisation of sound. *Proceedings ICMC99, Beijing*, Octobre 1999.

[Mal99b] D.G. Malham. Homogeneous and non-homogeneous surround sound systems. *Paper given to AES-UK Conference "Audio- the Second Century"*, June 1999.

[Mal99c] D.G. Malham. Second Order Ambisonics - the Furse-Malham Set, Octobre 1999. URL: http://www.york.ac.uk/inst/mustech/3d_audio/secondor.htm.

[Mar96] Martine Marin. *Etude de la localisation en restitution pour la téléconférence de haute qualité*. PhD thesis, Université du Maine, spécialité Acoustique, Octobre 1996.

[Mer62] H. Mertens. Principes d'une étude quantitative de l'ouïe directionnelle en stéréophonie. *L'onde électrique*, (420):172–182, March 1962.

[Mer65] H. Mertens. Directional hearing in stereophony – theory and experimental verification. *E.B.U. Review*, (92):146–158, August 1965.

[MI68] Philip M. Morse and K. Uno Ingard. *Theoretical Acoustics*. McGraw-Hill, 1968.

[NE98] Rozenn Nicol and Marc Emerit. Reproducing 3d-sound for videoconferencing: a comparison between holophony and ambisonic. *In Proc. DAFX98*, November 1998.

[NE99] Rozenn Nicol and Marc Emerit. 3d-sound reproduction over an extensive listening area: A hybrid method derived from holophony and ambisonic. *Proceedings of the AES 16th International Conference*, April 1999.

[Nic99] Rozenn Nicol. *Etude de la restitution du son spatialisé dans une zone étendue: Application à la téléprésence*. PhD thesis, Thèse de Doctorat CNET-Université du Maine, Décembre 1999.

[PBJ98] Jean-Marie Pernaux, Patrick Boussard, and Jean-Marc Jot. Virtual sound source positioning and mixing in 5.1 implementation on the real-time system genesis. *In Proc. DAFX98*, April 1998.

[Pol88] Jean-Dominique Polack. *La transmission de l'énergie sonore dans les salles*. PhD thesis, Université du Maine, Le Mans, 1988. Thèse de Doctorat d'État.

[Pol96a] Mark Poletti. "The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems". *J. Audio Eng. Soc.*, 44(11):948–963, November 1996.

[Pol96b] Mark Poletti. "The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems". *J. Audio Eng. Soc.*, 44(11):948–963, November 1996.

[PP88] J. D. Polack and X. Pelorson. Spatial Impression Evaluation With Omnidirectional Microphones. *Proceedings of The Institute of Acoustics*, 10(2), 1988.

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.

[Pul97] Ville Pulkki. "Virtual Sound Source Positioning Using Vector Base Amplitude Panning". *J. Audio Eng. Soc.*, 45(6):456–466, June 1997.

[RJ93] Karsten Bo Rasmussen and Peter Møller Juhl. "The Effect of Head Shape on Spectral Stereo Theory". *J. Audio Eng. Soc.*, 41(3):135–141, March 1993.

[RS97] Davide Rocchesso and Julius O. Smith. Circulant and Elliptic Feedback Delay Networks for Artificial Reverberation. In *IEEE Transactions on Speech and Audio Processing*, volume 5, January 1997.

[SA63] M. R. Schroeder and B. S. Atal. Computer simulation of sound transmission in rooms. *IEEE Conv. Record*, 7:150–155, 1963.

[SCDH88] Barbara G. Shinn-Cunningham, Nathaniel I. Durlach, and Richard M. Held. Adapting to supernormal auditory localization cues (i-ii). *J.Acoust.Soc.Am.*, 103(6):3656–3676, June 1988.

[Sch62] M. R. Schroeder. Natural sounding artificial reverberation. *J. Audio Eng. Soc.*, 10(3):219–223, 1962.

[Sla] Malcom Slaney. Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work. URL: http://web.interval.com/papers/1998-010/.

[Sou] SoundField Research Ltd. *The SoundField Research Home Page*. URL: http://www.proaudio.co.uk/sndfield.htm.

[SS94] Giuliano Schiffrer and Domenico Stanzial. Energetic properties of acoustic fields. *J. Acoust. Soc. Am*, 96(6), December 1994.

[Str71] A.H. Stroud. *Approximate Calculation of Multiples Integrals*. Prentice-Hall Inc, 1971.

[Sur] Sursound. *Surround Sound Mailing List* (liste de discussion électronique). Inscription: `mailto:majordomo@lists.uoregon.edu`.

[Tre97] Louis-Cyrille Trebuchet. “Etude et mise en oeuvre des techniques ambisoniques pour la spatialisation du son”. Technical report, ENST Paris, CCETT, 1997.

[Val95] Claude Valette. *Mécanique vibratoire appliqué à l'acoustique instrumentale*. Cours du DEA ATIAM, 1995.

[Ven] Kimmo Vennonen. Ambisonics work: A practical system for three-dimensional sound projection. URL (The Australian Centre for the Arts and Technology): http://online.anu.edu.au/ITA/ACAT/Ambisonic/Ambisonicswork.html.

[Wal96] James K. Waller. The Circle Surround 5.2.5 5-Channel Surround System. White Paper, Rocktron Corporation, 1996.

[Wat90] Bill Waterson. *Calvin and Hobbes: Weirdos from another planet!* Universal Press Syndicate, 1990.

[Wei99] Eric W. Weisstein. CRC Concise Encyclodedia of Mathematics, 1996-1999. http://www.treasure-troves.com/.

[WF95] M.J. Walsh and D.J. Furlong. “Improved Spectral Stereo Head Model”. *presented at the 99st Convention AES, Preprint 4128*, October 1995.

[ZF81] Eberhard Zwicker and Richard Feldtkeller. *Psychoacoustique. L'oreille, récepteur d'information.* Collection technique et scientifique des télécommunications. Masson, 1981. Traduit de l'allemand par Christel Sorin.

Annexe A

Formalismes en harmoniques sphériques et cylindriques: résolution de problèmes

A.1 Décompositions cylindrique et sphérique d'un champ

A.1.1 Décomposition en harmoniques cylindriques

Dans un système de coordonnées cylindriques où un point de mesure $\vec{r}$ est décrit par son rayon r, son azimut φ et sa composante verticale z, l'équation de propagation des ondes s'écrit:

$$\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2}{\partial\varphi^2}+\frac{\partial^2}{\partial z^2}+k^2\right]p=0 \tag{A.1}$$

qui admet [Bru98] des solutions à variables séparées de la forme:

$$p=R(r)\Phi(\varphi)Z(z)e^{j\omega t} \tag{A.2}$$

Dans la suite, le terme de dépendance temporelle $e^{j\omega t}$ (implicite) pourra être omis. L'injection de (A.2) dans (A.1) permet de diviser le problème en plusieurs équations, d'où se dégagent [Bru98] les fonctions de la forme:

$$\left\{\begin{array}{lcl} R(r) & = & A_{1m}J_m(k_w r)+A_{2m}N_m(k_w r) \\ \Phi(r) & = & B_{1m}\cos(m\varphi)+B_{2m}\sin(m\varphi) \\ Z(z) & = & A_1\cos(k_z r)+A_2\sin(k_z r) \end{array}\right. , \tag{A.3}$$

avec $k_w^2+k_z^2=k^2=\omega^2/c^2$. J_m et N_m sont les fonctions de Bessel et de Neumann, encore nommées respectivement fonctions de Bessel de la première et de la seconde espèce. Les fonctions de Bessel sont à valeur finie, au contraire des fonctions de Neumann, qui divergent en 0. Une base orthogonale de fonctions angulaires $\{Y_m^{\pm 1}\}$, dites *harmoniques cylindriques*, est définie comme suit:

$$\left\{\begin{array}{lcll} Y_m^1(\varphi) & = & \cos(m\varphi), & m\geq 0 \\ Y_m^{-1}(\varphi) & = & \sin(m\varphi), & m\geq 1 \end{array}\right. \tag{A.4}$$

Ce type de formalisme nous intéressera essentiellement pour une représentation bidimensionnelle du champ (c'est-à-dire restreinte au plan horizontal), en le supposant invariant suivant l'axe vertical z, soit $k_z=0$. Les

harmoniques cylindriques en constitue alors une base de décomposition. Sur tout anneau $\{r_1 \leq r \leq r_2\}$ où le champ ne prend pas de valeur infinie, il peut être décrit par un développement en série de Fourier-Bessel:

$$p(r,\varphi) = \sum_{0\leq m,\sigma=\pm 1} Y_m^\sigma(\varphi)\, j^m \left(A_m^\sigma J_m(kr) + jB_m^\sigma N_m(kr)\right), \tag{A.5}$$

Typiquement, les champs exempts de source sur un disque (ou un cylindre infini) de rayon r_1 sont tels que $B_m = 0$ dans (A.5), cette écriture étant valable à l'intérieur du cylindre. En revanche, dans le cas d'une onde se propageant "vers l'extérieur" du cylindre (donc avec une source à l'intérieur), l'écriture du champ (valable hors du cylindre) impose l'égalité $A_m = -B_m$, faisant apparaître les *fonctions de Hankel divergentes* [Bru98] $H_m^- = J_m - jN_m$, alors que les *fonctions de Hankel convergentes* $H_m^+ = J_m + jN_m$ apparaîtraient dans le cas d'une propagation "vers le centre".

Une base orthonormée d'harmoniques sphériques $\tilde{Y}_m^\sigma$ peut être définie de la manière suivante:

$$\tilde{Y}_m^\sigma(\varphi) = \sqrt{\varepsilon_m} Y_m^\sigma(\varphi), \qquad \sigma = \pm 1 \text{ et } \begin{cases} \varepsilon_0 &= 1 \\ \varepsilon_m &= 2 \quad \text{pour } m \geq 1 \end{cases}, \tag{A.6}$$

de sorte que $\langle \tilde{Y}_m^\sigma | \tilde{Y}_{m'}^{\sigma'} \rangle_{2\pi} = \delta_{mm'}\delta_{\sigma\sigma'}$, où le produit scalaire pour des fonctions angulaires (azimutales) $F(\varphi)$ et $G(\varphi)$ est défini par:

$$\langle F|G\rangle_{2\pi} = \frac{1}{2\pi}\int_0^{2\pi} F(\varphi)G(\varphi)\, d\varphi, \tag{A.7}$$

Cas d'une onde plane

Considérons une onde plane d'incidence $\vec{u}_p$ (azimut φ_p) se propageant parallèlement au plan horizontal:

$$p = Ae^{jk\vec{r}.\vec{u}_p} \tag{A.8}$$

soit encore, en notant $\cos\gamma = \vec{u}_r.\vec{u}_p$ [MI68][Bru98]:

$$\begin{aligned} p(r,\varphi) &= A\sum_{m=0}^{\infty} \varepsilon_m j^m \cos(m\gamma) J_m(kr) \\ &= \sum_{m\geq 0,\sigma=\pm 1} j^m \tilde{Y}_m^\sigma(\varphi)\tilde{Y}_m^\sigma(\varphi_p) J_m(kr) \end{aligned} \tag{A.9}$$

A.1.2 Décomposition en harmoniques sphériques

Expansion radiale et harmoniques sphériques

En utilisant les coordonnées sphériques (Figure A.1) qui sont le rayon r, l'angle polaire[1] θ et l'azimut φ, l'équation de propagation (1.10) prend la forme suivante:

$$\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} + k^2\right] p = 0 \tag{A.10}$$

Elle admet [MI68][Bru98] des solutions à variables séparées de la forme:

$$p = R(r)\Theta(\theta)\Phi(\varphi) \tag{A.11}$$

1. Il faut noter la différence entre l'angle polaire et le site δ utilisé pour le repérage directionnel lié à la tête: ils sont liés par $\theta = \pi/2 - \delta$.

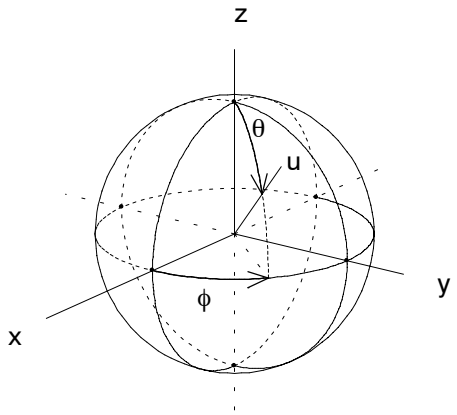

FIG. A.1 – *Système de coordonnées polaires: une incidence $\vec{u}$ est repérée par son azimut ϕ et son angle polaire θ, formé avec l'axe polaire $\vec{z}$. Noter les différences de convention et de notation par rapport à la figure 1.4.*

Suivant un procédé similaire à celui décrit en A.1.1, les fonctions mono-variable (respectivement radiales, polaires et azimutales) obtenues sont:

- les fonctions de Bessel et de Neumann sphériques[2] d'ordre m $j_m(kr)$ et $n_m(kr)$, ou encore les fonctions de Hankel sphériques convergentes $h_m^+ = h_m = j_m + jn_m$ et divergentes $h_m^- = h_m^* = j_m - jn_m$;
- les fonctions polaires (ou longitudinales) $P_{mn}(\cos\theta)$, $0 \geq n \geq m$, qui utilisent les polynômes de Legendre $P_m = P_{m0}$ et associés P_{mn} (A.12);
- les fonctions azimutales de la forme $\cos(n\varphi)$ et $\sin(n\varphi)$.

Aux fonctions (ou polynômes) de Legendre et fonctions de Legendre associées[3]:

$$P_{mn}(\eta) = (-1)^n(1-\eta^2)^{n/2}\frac{d^n}{d\eta^n}P_m(\eta), \quad 0 \leq n \leq m \quad \text{avec} \quad P_m(\eta) = P_{m0}(\eta) = \frac{1}{2^m m!}\frac{d^m}{d\eta^m}(\eta^2-1)^m \tag{A.12}$$

peut être appliquée la semi-normalisation de Schmidt[4]:

$$\tilde{P}_{mn}(\eta) = \sqrt{\varepsilon_n \frac{(m-n)!}{(m+n)!}}P_{mn}(\eta), \qquad \text{où } \varepsilon_0 = 1 \text{ et } \varepsilon_n = 2 \text{ pour } n \geq 1, \tag{A.13}$$

2. Encore appelées fonctions de Bessel sphériques de la première et de la seconde espèce. Leur relation avec les fonctions cylindriques du même nom sont rappelées en annexe A.2.1.

3. Elles peuvent être calculées par des relations de récurrence, rappelées en annexe A.2.2.

4. Ces fonctions sont disponibles parmi les routines matlab, avec et sans l'option "semi-orthonormalisation de Schmidt". Après vérifications, il semble que dans la version non normalisée, le terme $(-1)^n$ (mentionné dans l'aide pour les versions antérieures à la 5.3) ait été omis. Il est réintégré dans la version semi-orthonormalisée, mais l'aide mentionne un facteur $\sqrt{2\frac{(m-n)!}{(m+n)!}}$ (avec nos notations) alors que celui appliqué est bien $\sqrt{\varepsilon_n\frac{(m-n)!}{(m+n)!}}$.

En appariant fonctions polaires et azimutales, on définit enfin les *harmoniques sphériques* $Y^{\sigma}_{mn}(\theta,\varphi)$, de *degré* $m \geq 0$ et d'*ordre*[5] $0 \leq n \leq m$:

$$Y^{\sigma}_{mn}(\theta,\varphi) = \tilde{P}_{mn}(\cos\theta) \times \begin{cases} \cos(n\varphi) & \text{si } \sigma = 1 \\ \sin(n\varphi) & \text{si } \sigma = -1 \text{ et } n \geq 1 \end{cases}, \tag{A.14}$$

Elles sont au nombre de $2m+1$ par degré m et forment elles-mêmes une base orthogonale:

$$\langle Y^{\sigma}_{mn} | Y^{\sigma'}_{m'n'} \rangle_{4\pi} = \frac{1}{2m+1} \delta_{mm'} \delta_{nn'} \delta_{\sigma\sigma'}, \tag{A.15}$$

en faisant usage du produit scalaire défini pour des fonctions sphériques $F(\theta,\varphi)$ et $G(\theta,\varphi)$ par:

$$\langle F|G\rangle_{4\pi} = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} F(\theta,\varphi) G(\theta,\varphi) \sin\theta \, d\theta \, d\varphi, \tag{A.16}$$

Les fonctions $Y^{\sigma}_{mn}(\theta,\varphi)\, j_m(kr)$ et $Y^{\sigma}_{mn}(\theta,\varphi) n_m(kr)$ constituent une base de décomposition pour toute solution de l'équation de propagation (A.10), ce qui permet, de façon similaire au cas cylindrique, d'exprimer le champ par un développement en série de Fourier-Bessel (sphérique):

$$p(r,\theta,\varphi) = \sum_{m,n,\sigma} Y^{\sigma}_{mn}(\theta,\varphi)\, j^m \left(A^{\sigma}_{mn} j_m(kr) + j B^{\sigma}_{mn} n_m(kr) \right), \tag{A.17}$$

Cette expression est valable sur tout anneau sphérique $\{r_1 \leq r \leq r_2\}$, et les mêmes commentaires qu'en A.1.1 peuvent s'appliquer ici.

Il est utile, pour un usage ultérieur, de définir une base orthonormée d'harmoniques sphériques:

$$\tilde{Y}^{\sigma}_{mn}(\theta,\varphi) = \sqrt{2m+1}\, Y^{\sigma}_{mn}(\theta,\varphi) \tag{A.18}$$

telle que:

$$\langle \tilde{Y}^{\sigma}_{mn} | \tilde{Y}^{\sigma'}_{m'n'} \rangle_{4\pi} = \delta_{mm'} \delta_{nn'} \delta_{\sigma\sigma'} \tag{A.19}$$

Les harmoniques Y^1_{m0} ou $\tilde{Y}^1_{m0}$ (d'ordre $n=0$) n'ont pas de dépendance azimutale et sont appelées *harmoniques axiales* (s'agissant de l'axe polaire). Les autres sont les *harmoniques tesserales*. Notons également qu'en fixant l'angle polaire θ à $\pi/2$, les harmoniques Y^{σ}_{mm} ou $\tilde{Y}^{\sigma}_{mm}$ (d'ordre égal au degré) forment une base équivalente à la base des harmoniques cylindriques Y^{σ}_{m} ou $\tilde{Y}^{\sigma}_{m}$, à des facteurs multiplicatifs près qui sont explicités dans 3.1.2.

Les fonctions Y^{σ}_{mn} sont données sous des formes plus explicites Table 3.1 et sont illustrées Figure 3.2.

Calcul des coefficients de la décomposition

Supposons une région exempte de source, délimitée par une sphère de rayon R centrée en $\vec{r} = 0$. Les coefficients $\tilde{A}^{\sigma}_{mn}$ de la décomposition –valable sur cette région–:

$$p(r,\theta,\varphi) = \sum_{m,n,\sigma} \tilde{A}^{\sigma}_{mn} \tilde{Y}^{\sigma}_{mn}(\theta,\varphi)\, j^m\, j_m(kr) \tag{A.20}$$

peuvent être obtenus simplement par projection de la fonction sphérique $p_R(\theta,\varphi) = p(r=R,\theta,\varphi)$ sur la base orthonormée des fonctions harmoniques sphériques $\tilde{Y}^{\sigma}_{mn}$:

$$\langle p_R | \tilde{Y}^{\sigma}_{mn} \rangle_{4\pi} = j^m\, j_m(kR) \tilde{A}^{\sigma}_{mn}, \tag{A.21}$$

sous réserve que $j_m(kR) \neq 0$.

5. Par la suite, lorsque le degré et l'ordre ne seront pas évoqués conjointement, on parlera par abus de langage d'harmoniques d'ordre m plutôt que de degré m

Cas d'une onde plane

Dans le cas une onde plane se propageant selon la direction du vecteur $\vec{u_p}$, ce dernier indiquant la provenance de l'onde, l'expression du champ de pression en un point de mesure $\vec{r} = r\vec{u_r}$ est la suivante:

$$p = Ae^{jk\vec{r}.\vec{u}_p} \tag{A.22}$$

soit encore [MI68][Bru98]:

$$p = A\sum_{m=0}^{\infty}(2m+1)\,j^m P_m(\cos\gamma)\,j_m(kr) \tag{A.23}$$

où $\cos\gamma = \vec{u}_r.\vec{u}_p$. En décrivant les vecteurs unitaires $\vec{u_r}$ et $\vec{u}_p$ par leurs angles polaires et azimutaux respectifs (θ_r,φ_r) et (θ_p,φ_p), il vient:

$$P_m(\cos\gamma) = \sum_{n=0}^{m}\varepsilon_n\frac{(m-n)!}{(m+n)!}P_{mn}(\cos\theta_r)P_{mn}(\cos\theta_p)\cos\left[n(\varphi_r-\varphi_p)\right] \tag{A.24}$$

En adoptant la convention (A.14) comme définition des harmoniques sphériques, et en réincorporant le terme temporel $e^{j\omega t}$, le champ de pression s'écrit:

$$\begin{cases} p &= \sum_{m,n,\sigma} A^{\sigma}_{mn}Y^{\sigma}_{mn}(\theta_r,\varphi_r)\underbrace{(2m+1)\,j^m j_m(kr)e^{j\omega t}}_{\text{terme de propagation}} \\ A^{\sigma}_{mn} &= AY^{\sigma}_{mn}(\theta_p,\varphi_p) \end{cases} \tag{A.25}$$

Si c'est la définition (A.18) qui est choisie:

$$\begin{cases} p &= \sum_{m,n,\sigma} \tilde{A}^{\sigma}_{mn}\tilde{Y}^{\sigma}_{mn}(\theta_r,\varphi_r) \quad \underbrace{j^m j_m(kr)e^{j\omega t}}_{\text{terme de propagation}} \\ \tilde{A}^{\sigma}_{mn} &= A\tilde{Y}^{\sigma}_{mn}(\theta_p,\varphi_p) \end{cases} \tag{A.26}$$

Cas d'une onde sphérique

L'équation 1.14 est reprise ici pour décrire une onde sphérique divergente. En supposant que le lieu$\vec{\rho}$ de la source soit sur le demi-axe polaire $\theta = 0$ ($z > 0$), le champ de pression p, considéré dans la boule $r < \rho$, s'écrit comme la somme:

$$p(\vec{r}) = A\,k\,d\sum_{m=0}^{\infty}(2m+1)(-j)h^{-}_m(k\rho)P_m(\cos\theta)\,j_m(kr) \tag{A.27}$$

Dans le cas d'une incidence quelconque $(\theta_\rho,\varphi_\rho)$, les coefficients A^{σ}_{mn} et $\tilde{A}^{\sigma}_{mn}$ dans les équations (A.25) et (A.26) deviennent:

$$A^{\sigma}_{mn} = AY^{\sigma}_{mn}(\theta_\rho,\varphi_\rho)dkh^{-}_m(k\rho)\,j^{-(m+1)} \tag{A.28}$$

$$\tilde{A}^{\sigma}_{mn} = A\tilde{Y}^{\sigma}_{mn}(\theta_\rho,\varphi_\rho)dk\left(j_m(k\rho) - jn_m(k\rho)\right)j^{-(m+1)} \tag{A.29}$$

Applications

Ce formalisme conduit à la résolution naturelle du problème de la diffraction autour d'une sphère (Cf A.3), et constitue la base indispensable pour l'extension des systèmes ambisoniques aux ordres supérieurs (chapitre 3).

A.2 Rappels sur les fonctions de Bessel, Legendre, etc.

Les définitions et des propriétés utiles (dont relations de récurrence, intégrales, etc.) concernant ces fonctions sont rappelées dans [MI68] et [Wei99], par exemple (et sans-doute Morse et Feshbach, mais je ne l'ai pas). Nous en reportons quelques unes ci-après.

A.2.1 Fonctions de Bessel, de Neumann et de Hankel

Les fonctions de Bessel et de Neumann sphériques j_m et h_m sont définies comme suit:

$$\left\{\begin{array}{lcl} j_n(x) & = & \sqrt{\frac{\pi}{2x}}J_{n+1/2}(x) = (-1)^n x^n \left(\frac{d}{xdx}\right)^n \frac{\sin x}{x} \\ j_0(x) & = & \frac{\sin x}{x} \\ j_1(x) & = & \frac{\sin x}{x^2} - \frac{\cos x}{x} \\ j_2(x) & = & (\frac{3}{x^3} - \frac{1}{x})\sin x - \frac{3}{x^2}\cos x \\ n_n(x) & = & (-1)^{n+1}\sqrt{\frac{\pi}{2x}}J_{-n-1/2}(x) = (-1)^{n+1}x^n \left(\frac{d}{xdx}\right)^n \frac{\cos x}{x} \\ n_0(x) & = & -\frac{\cos x}{x} \\ n_1(x) & = & -\frac{\cos x}{x^2} - \frac{\sin x}{x} \\ n_2(x) & = & -(\frac{3}{x^3} - \frac{1}{x})\cos x - \frac{3}{x^2}\sin x \end{array}\right. \tag{A.30}$$

Les fonctions de Hankel sphériques h_m ont pour expression générique:

$$h_m(\zeta) = j_m(\zeta) + jn_m(\zeta) = \frac{j^{-m}}{j\zeta}\sum_{n=0}^{m}\frac{(m+n)!}{(m-n)!}\left(\frac{j}{2\zeta}\right)^n e^{j\zeta} \tag{A.31}$$

et la tendance asymptotique suivante:

$$h_m(\zeta) \underset{\zeta\to\infty}{\to} (-j)^{m+1}\frac{e^{j\zeta}}{\zeta} \tag{A.32}$$

Les relations de récurrence suivantes, données pour j_m, sont transposables à n_m et à h_m:

$$\frac{2m+1}{\zeta}j_m(\zeta) = j_{m-1}(\zeta) + j_{m+1}(\zeta), \qquad m \geq 1 \tag{A.33}$$

$$(2m+1)\,j'_m(\zeta) = m\,j_{m-1}(\zeta) - (m+1)\,j_{m+1}(\zeta), \qquad m \geq 1 \tag{A.34}$$

Mentionnons aussi l'égalité:

$$h'_m(\zeta)\,j_m(\zeta) - j'_m(\zeta)h_m(\zeta) = \frac{j}{\zeta^2} \tag{A.35}$$

A.2.2 Fonctions de Legendre

A partir des relations de récurrence sur les fonctions de Legendre et fonctions de Legendre associées:

$$\begin{array}{lcl} P_{m+1}(z) & = & P_{m+1,0}(z) = \frac{1}{m+1}\left((2m+1)zP_m(z) - mP_{m-1}(z)\right) \\ P_{m+1,n+1}(z) & = & P_{m-1,n+1}(z) + (2m+1)\sqrt{1-z^2}P_{m,n}(z) \end{array} \tag{A.36}$$

nous pouvons en donner un définition explicite, ici pour les premiers ordres:

$$\begin{array}{lll} P_0^0(z) = 1 & & \\ P_1^0(z) = z & P_1^1(z) = \sqrt{1-z^2} & \\ P_2^0(z) = \frac{1}{2}(3z^2-1) & P_2^1(z) = 3z\sqrt{1-z^2} & P_2^2(z) = 3z(1-z^2) \\ P_3^0(z) = \frac{1}{2}z(5z^2-3) & P_3^1(z) = \frac{1}{2}(5z^2-1)\sqrt{1-z^2} & \\ P_3^2(z) = 15z(1-z^2) & P_3^3(z) = 15(1-z^2)^{3/2} & \\ P_4^0(z) = \frac{1}{8}(35z^4-30z^2+3) & P_4^1(z) = \frac{5}{2}z(7z^2-3)\sqrt{1-z^2} & P_4^2(z) = \frac{15}{2}(7z^2-1)(1-z^2) \\ & P_4^3(z) = 105z(1-z^2)^{3/2} & P_4^4(z) = 105(1-z^2)^2 \end{array} \tag{A.37}$$

La récurrence pour la définition de la dérivée des polynômes de Legendre pourra également se révéler utile:

$$(2m+1)P_m(z) = P'_{m+1}(z) - P'_{m-1}(z) \tag{A.38}$$

A.2.3 Polynômes de Chebychev et fonctions trigonométriques

Les polynômes de Chebychev T_m nous intéressent surtout pour la propriété suivante:

$$T_m(\cos\theta) = \cos(m\theta) \tag{A.39}$$

Nous en retiendrons une définition pratique, d'après la relation de récurrence suivante:

$$\begin{array}{l} T_0(\zeta) = 1 \\ T_1(\zeta) = \zeta \\ T_{m+1}(\zeta) = 2\zeta T_m(\zeta) - T_{m-1}(\zeta) \qquad \text{pour } m \geq 1 \end{array} \tag{A.40}$$

Cette récurrence, combinée avec la récurrence croisée:

$$\sin((m+1)\theta) = \sin(m\theta)\cos\theta + \cos(m\theta)\sin\theta \tag{A.41}$$

permet en outre de définir les valeurs de $\sin(m\theta)$ pour les m élevés.

A.3 Diffraction par une sphère rigide

A.3.1 Cas général

Le problème général posé ici consiste à expliciter le champ résultant de la perturbation par une sphère rigide centrée en $\vec{r} = 0$, à partir de l'expression du champ sur une zone englobant le volume occupé par la sphère, mais en l'absence de celle-ci, et en admettant qu'il ne prenne pas de valeur infinie sur la zone considérée. Le champ de pression résultant est la somme du champ p_l en l'absence de sphère (*champ libre*) et du champ dit *diffracté* p_d:

$$p_t = p_l + p_d \tag{A.42}$$

Les champs de vélocité particulaire (ou acoustique) associés sont $\vec{u}_t$, $\vec{u}_l$ et $\vec{u}_d$, de composantes radiales u_t, u_l et u_d. Les caractéristiques du champ diffracté sont semblables à celles d'un champ rayonné par une sphère [MI68][Bru98]: en particulier, sa vélocité $\vec{u}_d$ se réduit à sa composante radiale u_d, qui s'exprime comme somme pondérée des fonctions de *Hankel* divergentes $h_m^- = j_m - jn_m$ (propagation "vers l'extérieur"). La rigidité de la sphère impose que la composante radiale de la vélocité résultante soit nulle à la surface:

$$u_t(r=a) = \frac{j}{\rho\omega}\left.\frac{\partial p_t}{\partial r}\right|_{r=a} \tag{A.43}$$

En écrivant les champs de pression, puis les champs de vélocité, comme sommes d'harmoniques sphériques:

$$p_l(r,\theta,\varphi) = \sum_{m=0}^{\infty} A_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\varphi)\, j^m\, j_m(kr) \tag{A.44}$$

$$p_d(r,\theta,\varphi) = \sum_{m=0}^{\infty} D_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\varphi)\, j^m h_m^-(kr) \tag{A.45}$$

La contrainte (A.43) entraîne naturellement:

$$p_t = \sum_{m=0}^{\infty} A_{mn}^{\sigma} j^m \left(j_m(kr) - \frac{j_m'(ka)}{h_m^{-\prime}(ka)} h_m^-(kr) \right) Y_{mn}^{\sigma}(\theta,\varphi) \tag{A.46}$$

et en utilisant (A.35), l'écriture du champ de pression à la surface de la sphère se simplifie comme suit:

$$p_t(r=a,\theta,\varphi) = \sum_{m=0}^{\infty} \frac{j^{m-1}}{(ka)^2 h_m^{-\prime}(ka)} A_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\varphi) \tag{A.47}$$

Il est courant de rencontrer dans la littérature ([MI68], par exemple) une autre écriture du champ résultant (A.46), en introduisant notamment les amplitudes (réelles) $B_m(ka)$ et angles de phase $\delta_m(ka)$, définies par $jB_m(\zeta)e^{j\delta_m(\zeta)} = h_m'(\zeta)$. En remarquant que $j_m'(\zeta) = \Re(h_m'(\zeta))$, l'équation (A.46) devient[6]:

$$p_t = \sum_{m=0}^{\infty} A_{mn}^{\sigma} j^m \left(j_m(kr) + j\sin\delta_m(ka) e^{j\delta_m(ka)} h_m^-(kr) \right) Y_{mn}^{\sigma}(\theta,\varphi) \tag{A.48}$$

6. Attention: on ne trouve pas exactement cette écriture dans [MI68] parce que l'écriture complexe de la dépendance temporelle qui y est adoptée est $e^{-j\omega t}$ et non $e^{+j\omega t}$, comme nous en avons choisi la convention!

A.3.2 Cas d'une onde plane

En adoptant la convention (A.14) et en utilisant (A.25), le champ total résultant de la diffraction d'une onde plane d'incidence (θ_S,φ_S) (vecteur $\vec{u}_S$) sur la sphère, s'écrit:

$$\begin{aligned} p_t(r,\theta,\varphi) &= S\sum_{m=0}^{\infty} Y_{mn}^{\sigma}(\theta_S,\varphi_S)\, j^m \left(j_m(kr) - \frac{j_m'(ka)}{h_m^{-\prime}(ka)} h_m^-(kr) \right) Y_{mn}^{\sigma}(\theta,\varphi) \\ p_t(\vec{r}) &= S\sum_{m=0}^{\infty} (2m+1)\, j^m \left(j_m(kr) - \frac{j_m'(ka)}{h_m^{-\prime}(ka)} h_m^-(kr) \right) P_m(\vec{u}_r.\vec{u}_S) \end{aligned} \tag{A.49}$$

où S est l'amplitude de l'onde pour le nombre d'onde k, et $\vec{u}_r = \vec{r}/r$.

A la surface de la sphère:

$$\begin{aligned} p_t(r=a,\theta,\varphi) &= S\sum_{m=0}^{\infty} \frac{j^{m-1}}{(ka)^2 h_m^{-\prime}(ka)} Y_{mn}^{\sigma}(\theta_S,\varphi_S) Y_{mn}^{\sigma}(\theta,\varphi) \\ p_t(r=a,\vec{u}_r) &= S\sum_{m=0}^{\infty} \frac{j^{m-1}}{(ka)^2 h_m^{-\prime}(ka)} (2m+1) P_m(\vec{u}_r.\vec{u}_S) \end{aligned} \tag{A.50}$$

A.3.3 Modélisation simplifiée d'une tête: HRTF et indices de localisation

HRTF

Le champ de pression à la surface de la sphère est exprimé jusqu'ici en fonction de l'angle polaire par rapport l'incidence de l'onde. La réponse fréquentielle à une onde plane d'incidence $\vec{u}_S$, mesurée au point $a\vec{u}_r$ de la sphère, est donnée par:

$$H(\mu,f) = \sum_{m=0}^{\infty} \frac{j^{m-1}}{(ka)^2 h_m^{-\prime}(ka)} (2m+1) P_m(\mu), \tag{A.51}$$

où $f = \frac{k}{2\pi c}$ et $\mu = \cos\theta = \vec{u}_S.\vec{u}_r$.

Dans un repère où l'axe interaural est dirigé suivant $\vec{y} = \vec{u}_r$, les réponses binaurales d'une tête sphérique sont alors données par:

$$\begin{aligned} H_l(\vec{u}_S,f) &= H(\vec{u}_S.\vec{y},f) \\ H_r(\vec{u}_S,f) &= H(-\vec{u}_S.\vec{y},f) \end{aligned} \tag{A.52}$$

Différences interaurales et variations

Le calcul des différences interaurales ITD et ILD fait appel aux formules données en 1.4.2. Il est intéressant de faire apparaître l'expression approximative du retard de phase, valable en basse fréquence. On se reportera pour cela à [Kuh77].

On s'intéresse également aux variations des indices par légères rotation de la tête. Il faut pour cela s'appuyer sur l'expression:

$$\partial_\theta H(\cos\theta) = \frac{\partial H(\cos\theta)}{\partial\theta} = -\sum_{m=0}^{\infty} \frac{j^{m-1}}{(ka)^2 h_m^{-\prime}(ka)} (2m+1) P_m'(\cos\theta)\sin\theta \tag{A.53}$$

En écrivant H sous la forme générale $H = |H| e^{j\varphi_H}$, il est facile de montrer que:

$$\frac{\partial_\theta H}{H} = \frac{1}{2}\partial_\theta \ln|H|^2 + j\partial_\theta\varphi_H \tag{A.54}$$

En appliquant cette relation aux réponses gauche H_l et droite H_r (attention au signe!), et en extrayant de chaque quantité la partie imaginaire, on obtient rapidement la variation du retard interaural de phase par rotation de la tête:

$$\frac{\partial \mathrm{ITD}_{\mathrm{phase}}}{\partial \theta} = \frac{1}{2\pi f} \Im \left(\frac{\partial_\theta H_l}{H_l} - \frac{\partial_\theta H_r}{H_r} \right) \tag{A.55}$$

A.4 Optimisation généralisée du décodage ambisonique

A.4.1 Caractéristiques des décodages modifiés

Nous reportons d'abord les résultats qui concernent les décodages 2D pour les configurations régulières non-minimales. Les formules montrées dans [DRP98] (Annexe B) sont données ici dans le cas plus général de coefficients correcteurs g_m complexes. La norme r_E du vecteur énergie a pour valeur:

$$r_E^{2D} = \frac{2\sum_{m=1}^{M} \Re(g_{m-1}g_m^*)}{|g_0|^2 + 2\sum_{m=0}^{M} |g_m|^2} \tag{A.56}$$

et l'énergie réduite associée $\mathcal{E}$ (3.81) vaut:

$$\mathcal{E}^{2D} = \left(|g_0|^2 + 2\sum_{m=1}^{M} |g_m|^2 \right) / |g_0|^2 \tag{A.57}$$

Attelons-nous maintenant au cas des décodages 3D, toujours dans le cadre de configurations régulières –pour l'échantillonnage de la base d'harmoniques sphériques– non-minimales. Il est supposé ici que les configurations considérées satisfont des propriétés de régularité au second degré (voir 3.2.3) jusqu'à l'ordre M considéré. Dans ce cas, le gain associé à chaque haut-parleur, résultant des étapes de codage et de décodage d'une onde plane d'amplitude A, s'écrit:

$$\begin{aligned} G_k &= \frac{A}{N}\sum_{m=0}^{M}(2m+1)P_m(\vec{u}_p.\vec{u}_k) \\ G_k &= \frac{A}{N}\sum_{m=0}^{M}(2m+1)P_m(\cos\theta_k\cos\theta_p - \sin\theta_k\sin\theta_p\cos(\varphi_k-\varphi_p)) \end{aligned} \tag{A.58}$$

Un décodage modifié va mettre en jeu des gains associés à chaque ordre de la décomposition:

$$G_k = \frac{A}{N}\sum_{m=0}^{M} g_m(2m+1)P_m(\vec{u}_p.\vec{u}_k) \tag{A.59}$$

Le vecteur énergie s'écrit:

$$\vec{E} = \frac{\sum_{k=1}^{N} |G_k|^2 \vec{u}_k}{\sum_{k=1}^{N} |G_k|^2} = r_E\vec{u}_E \tag{A.60}$$

Afin de simplifier l'écriture, mais sans restreindre la généralité des résultats, l'onde plane est représentée comme se propageant selon l'axe polaire. On admettra sans démonstration que:

$$r_E = \frac{\sum_{k=1}^{N} |G_k|^2 \cos\theta_k}{\sum_{k=1}^{N} |G_k|^2} \tag{A.61}$$

En faisant appel aux résultats sur les intégrales de produits de fonctions de Legendre [MI68], les expressions suivantes sont obtenues:

$$\sum_{n=1}^{N} |G_k|^2 = \frac{A^2}{N} \sum_{m=0}^{M} (2m+1)|g_m|^2, \tag{A.62}$$

soit une énergie réduite associée $\mathcal{E}^{3D} = \sum_{m=0}^{M}(2m+1)|g_m/g_0|^2$

$$\begin{aligned} \sum_{n=1}^{N} |G_k|^2 \cos\theta_k &= \frac{2A^2}{N} \sum_{m=0}^{M} m\Re(g_{m-1}g_m^*) \qquad \text{(Cas général: } g_m \text{ complexe)} \\ &= \frac{2A^2}{N} \sum_{m=0}^{M} m g_{m-1} g_m \qquad (g_m \text{ réel}) \end{aligned} \tag{A.63}$$

$$r_E^{3D} = \frac{2\sum_{m=1}^{M} m\Re(g_{m-1}g_m^*)}{\sum_{m=0}^{M}(2m+1)|g_m|^2} \quad \text{ou} \quad r_E^{3D} = \frac{2\sum_{m=1}^{M} m g_{m-1} g_m}{\sum_{m=0}^{M}(2m+1)g_m^2} \quad \text{si les } g_m \text{ sont réels.} \tag{A.64}$$

A.4.2 Optimisation *"max r_E"*

L'optimisation du décodage ambisonique "hautes-fréquences", basée sur la maximisation en module $\vec{r}_E$ du vecteur énergie, avait été généralisée aux ordres supérieurs pour des configurations horizontales régulières dans [DRP98] (Annexe B). La méthode est appliquée ci-après au cas d'une représentation et d'une restitution du champ en trois dimensions, c'est-à-dire à une décomposition en harmoniques sphériques, et non plus cylindriques.

Maximiser r_E d'après (A.64) revient à annuler ses dérivées par rapport aux gains g_n, ce qui donne la relation:

$$\frac{\partial r_E}{\partial g_m} = 0 \Rightarrow (2m+1) r_E g_m = (m+1) g_{m+1} + m g_{m-1}, \tag{A.65}$$

où $m = 0,1,\ldots,M$, en ayant introduit les paramètres $g_{-1} = 0$ et $g_{M+1} = 0$. Cette relation récurrente est à rapprocher d'une équation vérifiée par les fonctions de Legendre [MI68]:

$$(2m+1)\eta P_m(\eta) = (m+1)P_{m+1}(\eta) + mP_{m-1}(\eta) \tag{A.66}$$

Ainsi, le jeu de paramètres g_m maximisant r_E peut être déterminé *à une constante multiplicative près* en posant $g_m = P_m(\eta)$ et $\eta = r_E$, ce qui fixe en particulier les valeurs $g_0 = 1$ et $g_1 = r_E$, et en général:

$$g_m = P_m(r_E), \quad \text{pour} \quad m = 0,1,\ldots,M \tag{A.67}$$

r_E étant défini comme la plus grande racine de P_{M+1}:

$$P_{M+1}(r_E) = g_{M+1} = 0 \tag{A.68}$$

A.4.3 Solutions *in-phase* généralisées ou "à loi de pan-pot monotone"

Il s'agit plus précisément des solutions de décodage pour configurations régulières qui assurent une décroissance des gains associés aux haut-parleurs en fonction de leur éloignement de la source virtuelle. Ces solutions doivent leur nom "*in-phase*" au fait que les haut-parleurs sont alors alimentés en phase ($\mathcal{G}(\theta) \geq 0$),

propriété qui coïncide avec celle de décroissance pour le décodage d'ordre 1 tel que Malham l'a introduit [Mal92]. Pour les systèmes d'ordres supérieurs, il existe une infinité de familles de solutions vérifiant la stricte propriété *in-phase* ($\mathcal{G}(\theta) \geq 0$), dont on peut isoler des cas ayant des propriétés complémentaires intéressantes comme l'optimisation des critères vélocité ou énergie (voir en A.4.4 pour les solutions à l'ordre 2), ou bien la propriété de décroissance (monotonie), que nous développons maintenant.

Cas d'une restitution 2D (plan horizontal)

Dans le cas de configurations régulières 2D (haut-parleurs disposés aux sommets d'un polygone régulier), la loi de pan-pot ou la prise de son équivalente utilise une fonction de directivité de la forme:

$$\mathcal{G}(\theta) = g_0 + 2\sum_{m=1}^{M} g_m \cos(m\theta) \tag{A.69}$$

Il s'agit alors d'assurer la décroissance de $\mathcal{G}(\theta)$ sur l'intervalle $[0\,\pi]$, avec condition de nullité en $\theta = \pi$, ce qui est réalisé en imposant les conditions:

$$\frac{d^{2n}\mathcal{G}}{d\theta^{2n}}(\theta = \pi) = 0, \qquad \text{pour } n = 0,\ldots,M-1, \tag{A.70}$$

où:

$$\frac{d^{2n}\mathcal{G}}{d\theta^{2n}}(\theta) = 2\sum_{m=1}^{M} g_m (-1)^n m^{2n} \cos(m\theta) \qquad \text{quand } n > 1 \tag{A.71}$$

En posant $g'_m = g_m/g_0$, vérifier (A.70) revient alors à résoudre le système d'équation:

$$\mathbf{M}.\begin{pmatrix} g'_1 \\ g'_2 \\ \vdots \\ g'_M \end{pmatrix} = \begin{pmatrix} -1/2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{A.72}$$

où la matrice $\mathbf{M}$ a pour éléments les coefficients $m_{ij} = (-1)^j j^{2i-2}$ (matrice de Vandermonde modifiée), les indices lignes i et colonnes j allant de 1 à M. Les solutions du système ont la forme:

$$g'^{\,2D}_m = \frac{(M!)^2}{(M+m)!(M-n)!} = \frac{C_{2M}^{M-m}}{C_{2M}^{M}}, \tag{A.73}$$

où l'on fait usage des coefficients binomiaux (ou combinaisons) $C_n^k = \frac{n!}{(n-k)!k!}$.

La table A.1 donne les coefficients g'_m trouvés pour les premiers ordres et le coefficient g_0 pour une normalisation en énergie dans le cas d'une restitution à N haut-parleurs.

Cas 3D

Pour une restitution 3D sur une configuration régulière (au sens défini en 3.2.3), la même méthode s'applique, mais sur le diagramme polaire suivant:

$$\mathcal{G}(\theta) = \sum_{m=0}^{M} (2m+1) g_m^{3D} P_m(\cos\theta), \tag{A.74}$$

M	1	2	3	4	5	6
$g_0/\sqrt{N}$	$\sqrt{3/2}$	$\sqrt{18/35}$	$\sqrt{100/231}$	...	...	...
g'_1	1/2	2/3	3/4	4/5	5/6	6/7
g'_2	-	1/6	3/10	2/5	10/21	15/28
g'_3	-	-	1/20	4/35	5/28	5/21
g'_4	-	-	-	1/70	5/126	1/14
g'_5	-	-	-	-	1/252	1/77
g'_6	-	-	-	-	-	1/924

TAB. A.1 – *Paramètres des solutions "in-phase" 2D généralisées pour les premiers ordres*

M	1	2	3	4	5	6
$g_0/\sqrt{N}$	...	...	...	...	...	...
g'_1	1/3	1/2	3/5	2/3	5/7	3/4
g'_2	-	1/10	1/5	2/7	5/14	5/12
g'_3	-	-	1/35	1/14	5/42	1/6
g'_4	-	-	-	1/126	1/42	1/22
g'_5	-	-	-	-	1/462	1/132
g'_6	-	-	-	-	-	1/1716

TAB. A.2 – *Paramètres des solutions "in-phase" 3D généralisées pour les premiers ordres*

où P_m désigne le polynôme de Legendre d'ordre m. Chaque terme $\cos(m\theta)$ dans l'expression (A.69) peut s'exprimer comme un polynôme en $\cos\theta$: $\cos(m\theta) = T_m(\cos\theta)$, où T_m est le polynôme de Chebychev[7] de degré m. Les deux expressions (A.69) et (A.74) s'écrivent donc comme des polynômes de degré M, et l'association des termes de même degré permet de déduire les nouveaux coefficients g_m^{3D}, à un facteur g_0^{3D} près (qui n'est pas égal à 1 en appliquant cette méthode). Les coefficients obtenus sont de la forme:

$$g'^{\,3D}_m = \frac{M!(M+1)!}{(M+m+1)!(M-n)!} = \frac{C_{M+1}^{m+1}}{C_{M+m+1}^{m+1}} \tag{A.75}$$

La table A.2 donne les paramètres réduits g'_m obtenus, et le coefficient $g_0 = \sqrt{\frac{N(2M+1)}{(M+1)^2}}$ requis pour une normalisation en énergie (à compléter?), l'énergie réduite $\mathcal{E}$ valant $\frac{(M+1)^2}{2M+1}$.

A.4.4 Combinaisons de critères

L'optimisation en posant le critère "*in-phase*" (fonction $\mathcal{G}(\theta)$ positive) comme contrainte associée au critère "max r_V" ou bien "max r_E", a été réalisée pour les systèmes horizontaux d'ordre 2. A titre indicatif, nous donnons Table A.3 les caractéristiques de ces solutions parmi d'autres du second ordre. Puisqu'elles semblent sous-optimales pour toutes les conditions d'écoute, nous ne les présentons que pour éviter une perte de temps au lecteur curieux de les connaître.

Puisqu'un enjeu important semble être de minimiser le gain associé aux haut-parleurs dans la direction opposée à la source, il semblerait plus judicieux de conserver le critère $\mathcal{G}(\pi) = 0$. Associé à cette contrainte,

7. Ces polynômes sont définis par la relation de récurrence $T_{m+1}(\zeta) = 2\zeta T_m(\zeta) - T_{m-1}(\zeta)$ pour $-1 \leq \zeta \leq 1$ avec $T_0 = 1$ et $T_1(\zeta) = \zeta$.

Ordre	Critère	g_1/g_0	g_2/g_0	r_V	r_E
2	“Basique"	1	1	1	0.8
2	“In-phase" (monotone)	2/3	1/6	0.667	0.8
2	“In-phase" (max r_V)	$1/\sqrt{2}$	1/4	0.707	0.832
2	“In-phase" (max r_E)	0.6795	0.31916 (*)	0.679	0.843
2	“$r_V = 1$, max r_E"	1	$\sqrt{5/2}-1$	1	0.860
2	“max r_E"	$\sqrt{3}/2$	1/2	0.866	0.866

TAB. A.3 – *Caractéristiques des solutions de décodage ambisonique 2D du second ordre, incluant les solutions à critère combiné avec la contrainte “*in-phase*". (*): il s'agit de la racine réelle du polynôme* $74X^3 - 42X^2 + 9X - 1 = 0$.

le critère de maximisation du coefficient r_E donne naissance à une autre famille de solutions de décodage, qui pourrait se révéler d'un certain intérêt. En revanche, cette famille de solutions ne vérifie pas la propriété “*in-phase*" selon laquelle la fonction $G(\theta)$ est positive, et n'en mérite plus la dénomination. L'étude de cette famille n'est pas approfondie ici.

Annexe B

AES 105th Convention (San Francisco, sept. 98) [DRP98] (Version corrigée)

Annexe C

AES 16th International Conference (Rovaniemi, avril 98) [DRP99] (Version corrigée)