Localisation auditive en contexte de synthèse binaurale non-individuelle

Hélène Bahu

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Traitement du signal

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Hélène BAHU

Pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Localisation auditive en contexte de synthèse binaurale non-individuelle

soutenue le 14 décembre 2016

devant le jury composé de :

M. Gérard Assayag	Directeur de thèse
M. Olivier Warusfel	Encadrant
M. Etienne Parizet	Rapporteur
M. Mathieu Paquier	Rapporteur
M. Bruno Gas	Examinateur
Mme Rozenn Nicol	Examinatrice

Remerciements

Je tiens à remercier mon équipe avec laquelle j'ai travaillé à l'IRCAM. Merci tout d'abord à Olivier qui m'a énormément appris pendant ces trois années très enrichissantes. Merci à Thibaut et Markus pour leur connaissance et leur savoir faire, et pour m'avoir aidée à entreprendre les travaux de recherche tout au long de ma thèse. Je tiens à remercier particulièrement Mounira mais également Adrien pour leur soutien dans les moments difficiles. Merci aussi à Marine pour son aide précieuse. Merci pour la sympathie et la bienveillance de toutes les personnes qui m'ont entourée à l'IRCAM. Travailler dans cet établissement a été une réelle chance. La mixité des savoirs et des disciplines y est unique. Merci à tous les membres du projet BiLi pour cette expérience professionnelle passionnante. Merci au jury pour avoir participé à l'évaluation de ma thèse.
Un grand merci à ma famille, mes parents, mes frères et ma soeur, qui m'ont toujours encouragée et sans qui je ne serais pas arrivée jusqu'ici. Merci à mes amis. Enfin, merci à Maxime avec qui j'ai partagé l'expérience de la thèse... L'aventure continue.

Résumé

L'écoute binaurale sur casque reste à ce jour la seule technique autorisant la restitution exacte aux oreilles d'un auditeur de l'ensemble des indices acoustiques responsables de la localisation auditive. Elle repose sur une technique d'échantillonnage de la fonction de directivité de la tête. Pour chaque direction, le couple de fonctions de transfert relevées (Head Related Transfer Functions) consigne un ensemble d'indices acoustiques résultant des phénomènes de diffraction subis par l'onde entre la source et chacun des conduits auditifs. A l'étape de synthèse, la diffusion sur casque d'un signal monophonique préalablement filtré par ce jeu de HRTFs se traduit par la perception d'une source spatialisée dans la direction correspondante. La dépendance spatiale et spectrale des indices encodés dans les HRTFs résulte de la morphologie de l'individu et en constitue une signature acoustique aussi singulière qu'une empreinte digitale. Ainsi, une synthèse effectuée avec des HRTFs non-individuelles se traduit en général par des défauts de localisation et/ou de timbre.

Afin de juger de la qualité du rendu spatial des sources sonores spatialisées, les tests de localisation sont les tests les plus conventionnels. Ils permettent de vérifier que les sources virtuelles sont localisées de manière conforme à la situation d'écoute réelle. Cependant, l'identification des directions perçues par l'auditeur implique l'utilisation d'une méthode de report qui introduit inévitablement une erreur liée à l'éventuel biais et à la dispersion avec lesquels l'auditeur reporte son jugement. Une première contribution de la thèse concerne la comparaison de trois méthodes de pointage pour le report de la localisation auditive. Pour ce faire, un test de localisation de sources sonores réelles distribuées en 3 dimensions a été mis en œuvre. Les mérites d'une méthode, dite de pointage proximal, n'impliquant que le mouvement de la main et peu commentée dans la littérature, sont discutés en comparaison des méthodes plus répandues de pointage avec la tête ou bras tendu qui sollicitent le corps entier.

L'acquisition individuelle des HRTFs souffre de nombreuses contraintes pratiques qui en limitent l'utilisation à grande échelle. Des méthodes d'individualisation ont été développées afin de s'affranchir de la mesure en chambre anéchoïque et s'appuient pour certaines sur l'exploitation de larges bases de données de HRTFs. Une base de données de HRTFs à haute résolution spatiale mesurée sur une cinquantaine d'individus constitue la seconde contribution de la thèse.

Parmi les méthodes d'individualisation, nombreuses sont celles qui font appel à un calcul de similarité entre HRTFs, en particulier pour leur classification. Ces outils de classification, qui tentent de dégager des sous-groupes de HRTFs, s'appuient sur la définition d'une métrique opérant notamment sur la représentation spectrale des HRTFs (amplitude et phase). La définition d'une métrique spectrale forme également la base des modèles de prédiction de la localisation auditive. Une contribution majeure de la thèse est le développement d'un modèle de localisation auditive visant à prédire les directions perçues de sources virtuelles synthétisées avec des HRTFs non-individuelles. Les paramètres du modèle, dont la métrique spectrale, sont analysés et optimisés de sorte à s'approcher au mieux des observations réalisées lors d'un test de localisation auditive. L'objectif ultime de ce travail est d'évaluer si le modèle peut être utilisé dans le cadre de la sélection d'un jeu de HRTFs pour un nouvel individu à partir de l'observation des directions pointées lors de l'écoute de sources virtuelles synthétisées avec des HRTFs non-individuelles.

Summary

With the spread of headphone listening, binaural technology appears as the most appropriate solution to democratize the access to spatialized audio contents. Binaural synthesis of virtual sound sources is based on the use of filters called HRTFs (Head Related Transfer Functions), which provide the listener with accurate localization cues. These cues are however highly listener-dependent and the use of non-individual HRTFs may lead to localization and timbre artefacts.

Individual acquisition of HRTFs requires a complex measurement setup installed in an anechoic chamber which is incompatible with large scale deployment. Therefore, individualization methods have been devised in order to offer alternatives to this individual measurement. They are often based on the exploitation of large HRTFs databases. To this end, a new HRTFs database with high spatial resolution has been created.

The development of a model that predicts the perceived direction of a virtual source synthesized with non-individual HRTFs is the core of the thesis work. The choice of the metric used for quantifying the similarity between HRTFs receives a particular attention. The ultimate goal is to evaluate how such a model can be used to select automatically the optimal HRTFs set for an individual, from the observation of his responses in a localization test of virtual sound sources synthesized with non-individual HRTFs. The implementation of such a test implies the use of a reporting method, which may introduce some bias in the responses. This thesis includes a comparative study of 3 reporting methods and the recommendation of a method more suitable in the context of binaural listening through headphones.

Table des matières

Liste des abréviations

Abréviations	Définition
ACP	Analyse en Composantes Principales
ANOVA	ANalysis Of VariAnce
CI	Confidence Interval
DCT	Discrete Cosine Transform
DTF	Directional Transfer Function
EC	Equalization-Cancellation
ERB	Equivalent Rectangular Bandwidth
GELP	God's Eye Localization Pointing
GT	Gammatone
GUI	Graphical User Interface
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
ILD	Interaural Level Difference
ITD	Interaural Time Difference
JND	Just Noticeable Differences
MAA	Minimum Audible Angle
MFCC	Mel-Frequency Cepstral Coefficient
MSE	Mean Square Error
ROC	Receiving Operating Characteristic
SI	Similarity Indices
SNR	Signal-to-Noise Ratio
SOFA	Spatially Oriented Format for Acoustics
VAS	Virtual Auditory Space

Introduction

Depuis l'apparition des systèmes de captation et de reproduction sonore, toute scène sonore peut être jouée indépendamment du moment et de l'espace où elle a été enregistrée. Avec l'intensification de la demande de l'industrie audiovisuelle et des applications de réalité virtuelle, les techniques de spatialisation sonore (enregistrement, synthèse et reproduction) connaissent un essor important. Elles cherchent à reproduire le plus fidèlement possible les caractéristiques du champ sonore, en particulier en termes de sensations auditives spatiales. Contrairement à la stéréophonie, les techniques de spatialisation 3D favorisent l'immersion de l'auditeur en sollicitant l'ensemble de la sphère auditive. Pour garantir cette capacité d'immersion et une restitution fidèle des sources dans l'ensemble de l'espace, les techniques de reproduction sur haut-parleurs nécessitent un très grand nombre d'enceintes, de l'ordre de quelques dizaines pour des techniques comme HOA (*Higher Order Ambisonics*) à plusieurs centaines pour la WFS (*Wave Field Synthesis*), pourtant restreinte en pratique à la restitution du plan horizontal. La technique binaurale se distingue de ces méthodes par l'extraordinaire simplicité du dispositif de restitution, un simple casque audio standard, à travers lequel elle restitue les informations spatiales directement aux oreilles de l'auditeur. Avec la généralisation de l'écoute au casque en mobilité, cette technique se présente ainsi comme une solution privilégiée pour démocratiser l'accès à des contenus sonores spatialisés. Cette thèse s'est inscrite dans le cadre du projet collaboratif BiLi, *Binaural Listening*[1], visant le développement et le déploiement de solutions pour la diffusion de contenus sonores 3D pouvant être écoutés en mode binaural sur casque.

La technique binaurale consiste à capter, reproduire ou synthétiser, le plus directement possible, les informations acoustiques reçues à l'entrée des conduits auditifs lors d'une écoute en situation naturelle. Elle vise ainsi à offrir à l'auditeur l'illusion de la présence de sources sonores précisément positionnées dans l'espace qui l'entoure. Pour reproduire virtuellement cet espace (*Virtual Auditory Space*, VAS), une première méthode, la captation en binaural natif, consiste à enregistrer directement la scène sonore en plaçant des microphones miniatures à l'entrée des canaux auditifs d'un auditeur ou d'une tête artificielle. Au moment de la restitution, l'usage du casque permet de garantir la confidentialité de la transmission des signaux captés à l'oreille droite et à l'oreille gauche vers leur oreille respective, évitant ainsi de sur-ajouter une information spatiale dont la reproduction sur haut-parleurs peut plus difficilement s'affranchir. Une seconde méthode, la synthèse binaurale, consiste à utiliser un jeu de filtres qui décrivent la fonction de directivité de la tête selon un ensemble de directions couvrant l'ensemble de la sphère auditive. Cette technique permet alors de simuler de manière flexible la direction des différentes sources sonores composant l'espace et tire tout son avantage dans le cadre d'applications interactives et de réalité virtuelle, lorsque ces filtres sont modifiés en temps réel pour simuler une source mobile ou le mouvement propre de l'auditeur au sein de la scène sonore.

Ces filtres, dénommés HRTFs (*Head Related Transfer Functions*), encodent les transformations acoustiques que subit l'onde sonore entre une source et les canaux auditifs d'un auditeur. Ces transformations sont liées aux phénomènes de diffraction et de réflexion de l'onde sur le corps de l'auditeur (tête, épaules, pavillons d'oreilles, etc.) et fournissent au système auditif des indices de localisation. Ces indices de localisation, temporels et spectraux, monauraux et interauraux, dépendent de la direction d'incidence de la source mais également de la morphologie de l'auditeur, ce qui leur confère un caractère fortement individuel. En d'autres termes, la relation entre indices de localisation et direction d'incidence est propre à chaque individu. Notre faculté à construire une représentation mentale de l'espace sonore à partir d'indices acoustiques parfois subtils contenus dans les signaux binauraux résulte d'un processus d'apprentissage multisensoriel. Celui-ci peut s'appuyer par exemple sur la congruence visuo-auditive pour la localisation d'objets frontaux ou sur une congruence entre variations des indices acoustiques et indices proprioceptifs liés à nos propres mouvements. Au moment de la synthèse d'un contenu binaural, utiliser les HRTFs relatives à un

1. http ://www.bili-project.org/

individu différent de l'auditeur risque de rompre cette relation acquise et entretenue quotidiennement par le système auditif de l'auditeur et de provoquer des artefacts perceptifs : défauts de localisation ou de timbre. La nécessité d'individualiser les filtres binauraux a été démontrée dans des études préalables [WAKW93, MSJH96] afin d'assurer la reproduction fidèle des indices de localisation. Cependant, la mesure des HRTFs requiert un processus long et complexe reposant sur un équipement important installé dans une chambre anéchoïque. Cette contrainte freine le déploiement grand public de la technique binaurale. Pour cette raison, de nombreux travaux de recherche sont menés dans l'optique de trouver d'autres moyens que la mesure en conditions contrôlées (environnement anéchoïque, signaux d'excitation large bande, échantillonnage dense de l'espace, tête immobile, etc.) pour acquérir les HRTFs individuelles d'un individu.

A partir de l'observation que les informations contenues dans les HRTFs sont intimement liées aux caractéristiques morphologiques de l'auditeur, on peut supposer que des traits de ressemblance morphologique entre individus se traduiront également par des signatures acoustiques entretenant certaines caractéristiques similaires. On peut dès lors imaginer exploiter des banques de données de HRTFs et sélectionner parmi celles-ci un jeu de HRTFs qui conviennent à un nouvel individu ne disposant pas de ses mesures.

Evaluer la similarité entre HRTFs de différents individus requiert la définition d'une métrique. Etant donnée la forte variabilité inter-individuelle de la composante spectrale des HRTFs ainsi que l'importance de son rôle dans la localisation auditive, il apparaît clairement qu'une telle métrique doit prendre en considération cette composante. La recherche d'une métrique permettant à la fois de traduire le rôle des composantes spectrales dans la localisation auditive et leur caractère individuel reste une question active dans la littérature. Elle fait l'objet d'un intérêt particulier dans le cadre de ce travail de recherche.

Outre son potentiel pour les applications audio, la technique binaurale se présente également comme un outil précieux pour l'étude des mécanismes de la localisation auditive. En effet, elle offre à l'expérimentateur la possibilité de manipuler les indices délivrés aux oreilles de l'auditeur (modifications [Mid99b], suppression d'indices dans certaines bandes fréquentielles [LB02], etc.) et d'en observer les effets sur les directions perçues à travers des tests de localisation auditive. Certaines études sur la localisation auditive ont ainsi pu mettre en évidence que, lorsque les indices spectraux sont partiels ou altérés, les directions perçues sont davantage guidées par une ressemblance spectrale entre le stimulus et les HRTFs de l'auditeur plutôt que par la direction de la source sonore. Les HRTFs de l'auditeur contiennent en effet les indices acoustiques à partir desquels l'auditeur se base pour localiser. A partir de ces observations, des études ont émis l'hypothèse que les directions perçues pouvaient être prédites par une comparaison objective entre les indices acoustiques délivrés par le stimulus et les HRTFs de l'auditeur [Mid92, LB02, BML14]. Cette hypothèse constitue la base du développement de modèles de prédiction de la localisation auditive. Ici encore, la définition d'une métrique spectrale capable d'évaluer la similarité objective entre indices spectraux apparaît essentielle.

Etant donné la difficulté d'acquisition des filtres individuels, la synthèse binaurale est, pour une majeure partie des individus, réalisée avec des filtres de spatialisation non-individuels. Pour cette raison, la condition d'écoute en synthèse binaurale non-individuelle mérite une attention particulière. Le travail de thèse est motivé par un désir de mieux comprendre comment la localisation auditive est gérée par le système auditif dans cette situation d'écoute. Pour ce faire, un modèle de localisation auditive probabiliste a été développé. Il reprend l'hypothèse des modèles antérieurs et étend le principe à la localisation auditive en synthèse binaurale non-individuelle. Contrairement à une approche physiologique de la modélisation des mécanismes de traitement du système auditif, l'approche adoptée ici combine des approches psycho-physique et de traitement du signal. Elle tente d'extraire des signaux acoustiques reçus aux oreilles de l'auditeur, l'information permettant d'expliquer les directions pointées par celui-ci lors d'un test de localisation de sources sonores synthétisées avec des HRTFs non-individuelles. La prédiction de la localisation se base sur la connaissance conjointe des informations acoustiques délivrées au casque et de celles contenues dans les HRTFs de l'auditeur. Elle est réalisée à partir de la comparaison de ces informations. L'optimisation des paramètres du modèle est l'occasion d'évaluer la pertinence de différentes métriques spectrales proposées dans la littérature. De plus, ce travail est motivé par la recherche de solutions d'individualisation des filtres HRTFs par une approche de sélection guidée d'un jeu de HRTFs dans une base de données. L'idée est qu'à partir de l'observation et de la modélisation des réponses d'un individu relevées dans un test de localisation de sources sonores virtuelles, nous puissions sélectionner un jeu de HRTFs optimal, ou du moins lui proposer un ensemble restreint de jeux de HRTFs susceptibles de lui convenir. La mise en œuvre de ce modèle requiert à la fois les données expérimentales de localisation et les HRTFs individuelles des participants. La première étape réside dans l'acquisition d'une base de données de HRTFs d'une cinquantaine de sujets. La

seconde étape consiste à déterminer un protocole pour collecter les données expérimentales de localisation. La méthode de report des directions perçues mérite une attention particulière étant donné qu'elle introduit une part d'erreur dans les jugements collectés. Nous réaliserons en amont une comparaison de différentes méthodes de report afin de guider le choix de celle qui sera utilisée dans le test de localisation.

Les trois premiers chapitres fournissent principalement la base théorique et l'état de l'art des travaux sur la localisation auditive, la synthèse binaurale et les méthodes d'individualisation. Tout d'abord, le chapitre 1 présente le point de départ de la technique binaurale à savoir les mécanismes qui régissent la localisation auditive humaine avec deux oreilles et en trois dimensions. Il permet de prendre conscience de l'importance du caractère individuel des indices acoustiques et permet d'appréhender les défauts liés à l'utilisation de filtres de spatialisation non-individuels à la restitution. Le chapitre 2 expose le principe de technique binaurale, les facteurs qui entrent en jeu dans la qualité du rendu spatial ainsi que les différentes méthodes de visualisation de l'information contenue dans les HRTFs. De plus, elle présente les éléments techniques de l'acquisition de la base de données de HRTFs du projet BiLi, de la captation à l'égalisation des HRTFs. Enfin, le chapitre 3 offre une vue d'ensemble des méthodes alternatives à l'acquisition des HRTFs en chambre anéchoïque, parmi lesquelles s'inscrivent les méthodes de sélection guidée.

Les chapitres 4 et 5 permettent d'aborder les points sensibles qui touchent au développement et à l'optimisation du modèle de prédiction de la localisation auditive étudié aux chapitres 6 et 7. Une étude comparative de trois méthodes de report est présentée au chapitre 4. Les mérites de la méthode proximale, peu commentée dans la littérature, sont discutés comparativement à des méthodes de pointage plus répandues comme le pointage avec la tête ou bras tendu. Cette étude a fait l'objet d'un article publié dans la revue *Acta acustica united with Acustica* en 2016. La problématique de la métrique spectrale est ensuite abordée en détails au chapitre 5. Après avoir répertorié les principales métriques utilisées dans la littérature, nous proposons une première comparaison d'un sous-ensemble d'entre elles.

Le chapitre 6 introduit le modèle de prédiction de la localisation auditive proposé, en soulignant les éléments tirés des précédents modèles de la littérature ainsi que les points qui l'en distinguent. Les différentes étapes et différents paramètres qui mènent à l'obtention d'une distribution de probabilité de réponse dans l'espace, ainsi que la méthode d'évaluation des résultats, y sont exposés en détails. Le chapitre 7 présente le test de localisation de sources virtuelles synthétisées avec des HRTFs individuelles et non-individuelles mené en parallèle de la mise en place du modèle. Il discute des paramètres permettant d'optimiser la prédiction des directions pointées et offre une comparaison directe des résultats avec la littérature. Enfin, nous évaluons le modèle sous forme d'une mise en oeuvre préfigurant une procédure de sélection guidée dans laquelle on recherche le jeu de HRTFs susceptible de convenir à un individu sur la seule base des observations de ces réponses à un test de localisation réalisé en condition non-individuelle.

Chapitre 1

Localisation auditive spatiale

La simulation d'une source spatialisée au casque requiert tout d'abord la compréhension du mécanisme de localisation auditive binaurale et tridimensionnelle. La connaissance de l'ensemble des indices acoustiques qui confèrent au système auditif l'information spatiale nécessaire à la localisation d'une source sonore est le point de départ de la synthèse binaurale. Elle permet également d'appréhender les défauts de spatialisation, lorsque par exemple les indices de localisation sont non-individuels, et d'envisager la modélisation du processus de localisation auditive. La représentation perceptive de l'information spectrale contenue dans les HRTFs, tenant compte des aspects physiologiques du système auditif, définit la base d'une métrique de similarité spectrale à pertinence perceptive.

Pour une source sonore située dans le plan médian de l'auditeur (plan X-Z sur la figure 1.1), les caractéristiques acoustiques reçues aux deux oreilles sont semblables, étant donné le caractère quasi-symétrique de la tête. Mais lorsque la source est décalée par rapport au plan médian, le son parvient aux oreilles gauche et droite avec un décalage temporel et un niveau sonore différent engendré par l'obstacle de la tête. Ces différences interaurales de temps et d'intensité (*Interaural Time Differences*, ITD et *Interaural Level Differences*, ILD, respectivement) augmentent avec la latéralisation de la source et varient selon la circonférence de la tête. La propagation de l'onde sonore est perturbée par le profil morphologique de l'auditeur avant de parvenir à l'entrée de ses canaux auditifs. Ces perturbations sont à l'origine de modifications spectrales en hautes fréquences (zones d'amplification ou d'atténuation marquées) qui sont intimement liées aux caractéristiques morphologiques de l'auditeur. Ces transformations dépendent de la direction d'incidence et varient plus particulièrement avec l'élévation. Ainsi, les différences interaurales et les transformations spectrales sont deux caractéristiques acoustiques qui dépendent de la direction d'incidence du son par rapport à l'auditeur. Le mécanisme auditif les utilise pour identifier la position de la source sonore dans l'espace. Elles constituent donc des indices acoustiques de localisation.

Ces différents indices sont complémentaires pour la localisation tridimensionnelle. Dans un premier temps, les indices interauraux fournissent principalement des informations sur la position latérale de la source. Cependant, à une valeur d'indice interaural donnée correspond un ensemble de directions distribuées en première approximation sur une enveloppe conique ayant l'axe interaural pour axe de symétrie. Cette seule valeur interaurale ne permet donc pas de discriminer ces différentes directions et on parle de "cône de confusion". Par ailleurs, les indices spectraux permettent de déterminer l'élévation et de lever les ambiguïtés liées aux cônes de confusion.

Le rôle de chacun de ces indices varie en fonction de la fréquence. Le contenu fréquentiel initial de la source détermine donc quelles informations spatiales sont délivrées au système auditif pour localiser. Pour un intervalle fréquentiel restreint, des erreurs de localisation peuvent alors apparaître. Par exemple, si la source est dépourvue de hautes fréquences supérieures à 4 kHz, les indices spectraux majeurs sont absents. En condition d'écoute statique, on observe alors des erreurs de localisation en élévation ainsi que des ambiguïtés avant-arrière. En condition d'écoute dynamique, les résultats sont moins dramatiques. En effet, les mouvements de la tête engendrent une évolution simultanée des indices interauraux et des indices proprioceptifs qui forment les indices dynamiques de localisation. Ils permettent la discrimination avant-arrière et l'identification de l'élévation.

Les différences interaurales de temps et d'intensité dépendent de la largeur de la tête et les transformations spectrales sont principalement déterminées par la taille et la forme des pavillons d'oreille. Les indices de localisation sont donc individuels. A travers l'expérience multi-sensorielle, chaque auditeur a appris à les reconnaître et à en extraire les informations spatiales, se construi-

sant ainsi une carte auditive spatiale reliant indices et directions d'incidence. Cependant, lorsque les indices sont altérés ou étrangers à l'auditeur (e.g. en synthèse binaurale avec HRTFs non-individuelles), les performances de localisation sont dégradées. Par ailleurs, nous présentons une capacité d'adaptation : une re-calibration de notre carte auditive spatiale peut être obtenue suite à l'apprentissage de nouveaux indices de localisation.

L'analyse des indices acoustiques et de leurs variations en fonction de la direction d'incidence est rendue possible par l'observation des signaux captés au moyen de deux microphones miniatures placés à l'entrée des canaux auditifs gauche et droit de l'auditeur. A partir d'un tel enregistrement, il est tout d'abord possible d'extraire les différences de temps d'arrivée et les différences de niveau entre les deux oreilles (ITD et ILD, respectivement). Le contrôle de la position spatiale de la source sonore permet d'observer l'évolution des indices avec la direction d'incidence de l'onde sonore. Moyennant la connaissance de la réponse en fréquence de la source sonore, une opération de filtrage inverse conduit à l'extraction des filtres gauche et droit représentatifs de toutes les transformations spectrales subies par l'onde sonore pendant son trajet entre la source et les canaux auditifs. Dans le cas idéal d'un enregistrement effectué en milieu anéchoïque, celui-ci est dénué d'effet de salle, c'est-à-dire que seul le son direct est capté et on s'affranchit alors des réflexions parvenant sous plusieurs incidences qui apparaissent en milieu réverbérant. Cette procédure correspond à l'acquisition de ce que l'on appelle les fonctions de transfert relatives à la tête, ou *Head Related Transfer Functions* (HRTFs). En pratique, ces HRTFs sont obtenues par la transformée de Fourier des HRIRs (*Head Related Impulse Responses*), issues de la mesure. Elles encodent les indices acoustiques et permettent ainsi une analyse en fonction de la direction de mesure et des individus mesurés.

Les HRTFs sont utilisées pour la spatialisation sonore au casque audio (technique binaurale). Une source virtuelle spatialisée dans une direction donnée peut être synthétisée par le filtrage du signal de la source par une paire de HRTFs gauche-droite mesurée à la direction correspondante. Les HRTFs constituent non seulement un outil puissant pour la spatialisation sonore mais également un outil de représentation précieux pour l'étude des diverses caractéristiques des indices acoustiques, en offrant à l'expérimentateur un contrôle précis des indices acoustiques délivrés à l'auditeur. Par exemple, certaines bandes fréquentielles peuvent être supprimées afin d'identifier celles qui sont déterminantes pour la localisation ou encore différents degrés de lissage spectral peuvent être appliqués pour estimer quelles approximations sont acceptables pour le système auditif.

Ce chapitre introduit tout d'abord les principaux indices acoustiques en jeu dans la localisation sonore 3D (direction et distance). Puis, il présente les caractéristiques liées à la capacité du système auditif à localiser, à discriminer les fréquences et les directions de l'espace ainsi qu'à s'adapter à de nouveaux indices de localisation. La connaissance précise de ces caractéristiques est indispensable si l'on veut pouvoir reproduire pour n'importe quel individu une situation de localisation auditive en environnement virtuel.

Systèmes de coordonnées

Avant toute chose, il est nécessaire de définir les outils de représentation spatiale associés à la description de la position des sources sonores dans le repère de l'auditeur, ainsi que la terminologie associée.

L'axe matérialisant la droite passant par les deux oreilles est appelé axe interaural. Les plans horizontal et vertical qui coïncident avec cette droite sont respectivement appelés le plan horizontal et le plan frontal. Le plan frontal sépare les hémi-champs avant et arrière. Le centre de la tête est généralement défini par le point de l'axe interaural situé à équidistance des deux oreilles. Basé sur le système de référence de l'anatomie, le plan médian est le plan vertical qui sépare les hémi-espaces gauche et droit. Plus généralement, les plans sagittaux désignent les plans parallèles au plan médian. Lorsque la source sonore est latéralisée, on appellera "oreille ispilatérale" l'oreille en face de la source et "oreille controlatérale", l'oreille opposée à la source, située dans l'ombre de la tête.

Pour décrire la position de la source sonore relative à l'auditeur, différents systèmes de coordonnées ont été utilisés dans les études sur la localisation auditive. Le système de coordonnées sphériques est le plus largement répandu. Il utilise les angles d'azimut θ et d'élévation ϕ (anglicisme du terme "site", moins répandu), pouvant être assimilés à la longitude et à la latitude, respectivement. Ce système, aussi référencé par *single-pole coordinate system*, peut être visualisé

à gauche de la figure 1.1. Morimoto et Aokata [MA84] ont introduit le système de coordonnées latéral-polaire, jugé plus adapté à la représentation de la localisation sonore. L'angle latéral Θ est défini (en référence à Wallach (1940) [Wal40]) par la longueur horizontale d'arc entre le point et le plan médian. L'angle polaire Φ correspond à l'angle d'élévation à l'intérieur d'un cercle d'angle latéral constant, c'est-à-dire un plan sagittal, et est déterminé par les indices spectraux. Un tel système de coordonnées peut être visualisé à droite de la figure 1.1. Le système de coordonnées double-pôle, notamment utilisé par Middlebrooks [Mid92], est une combinaison de ces deux systèmes dans le sens où il partage la définition de l'élévation en terme d'angle polaire Φ au sein d'un plan sagittal et la définition d'azimut θ du système sphérique.

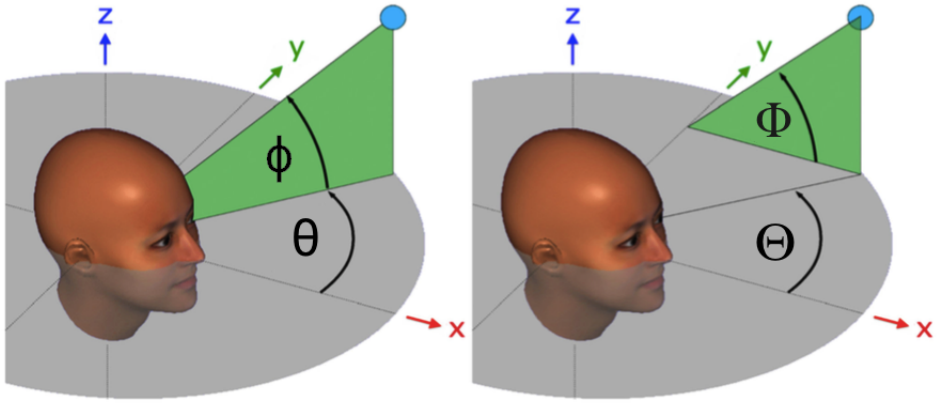


FIGURE 1.1 – Systèmes de coordonnées sphérique et latéral-polaire (source : [ZM14]).

Par la suite, les angles d'azimut θ et latéral Θ seront compris entre 0° (direction frontale) et 180° (direction arrière sur le plan médian) avec des valeurs positives à droite et négatives à gauche. Les angles d'élévation ϕ et polaire Φ seront compris entre 0° (plan horizontal) et 90° (pôles) avec des valeurs positives au-dessus du plan horizontal et négatives en-dessous.

1.1 Indices de localisation auditive

Le processus de localisation auditive consiste à identifier la position de la source sonore dans 3 dimensions : la distance, l'azimut, et l'élévation. Différents indices, relatifs à chacune de ces dimensions, sont à l'origine de notre capacité à localiser une source sonore dans l'espace. Ils sont liés aux phénomènes acoustiques intervenant sur le trajet de l'onde sonore entre la source et les canaux auditifs de l'auditeur.

1.1.1 Indices interauraux

Les indices interauraux (ou binauraux) se réfèrent aux différences de temps et d'intensité de l'onde sonore aux oreilles gauche et droite. Ils sont majoritairement responsables de la localisation latérale mais ne permettent pas à l'auditeur d'identifier une direction spatiale unique à cause des ambiguïtés présentes sur les cônes de confusion.

1.1.1.1 Différences interaurales de temps et d'intensité

Lord Rayleigh fut un pionnier dans l'étude et la compréhension de la localisation auditive binaurale. Partant de l'observation que l'onde sonore (de direction d'incidence en dehors du plan médian) arrive avec un certain retard à l'oreille controlatérale et avec une intensité moindre dus à l'obstacle de la tête, il mit en évidence les différences dans les signaux atteignant l'oreille droite et gauche. Sa théorie, introduite en 1907, et fondée sur la localisation de sons de fréquence pure (*pure tone*), affirme que la localisation auditive gauche-droite réside dans ces différences interaurales de temps et d'intensité (*Interaural Time Difference*, ITD et *Interaural Level Difference*, ILD, respectivement). Cette théorie dite "Duplex" constitue le fondement de la compréhension de la localisation auditive binaurale latérale. Plus récemment, Macpherson et Middlebrooks [MM02a] ont étudié plus spécifiquement le rôle relatif des indices d'ITD et d'ILD pour la localisation de stimuli large bande, filtrés passe-bas et passe-haut.

Les différences interaurales de temps et d'intensité sont proches de zéro dans le plan médian étant donné que le trajet de l'onde sonore jusqu'aux deux oreilles est équivalent, et augmentent en valeur absolue au fur et à mesure que la source sonore se latéralise, jusqu'à atteindre un maximum à azimut $\theta = \pm 90°$ et élévation 0°. La figure 1.2 permet de visualiser ces différences sur un exemple d'HRIRs gauche et droite mesurées à $\theta, \phi = (90°, 0°)$. La variation des indices interauraux avec l'azimut est non-linéaire. Pour un même pas d'azimut, les changements d'ITD et d'ILD sont plus importants autour du plan médian que sur les côtés.

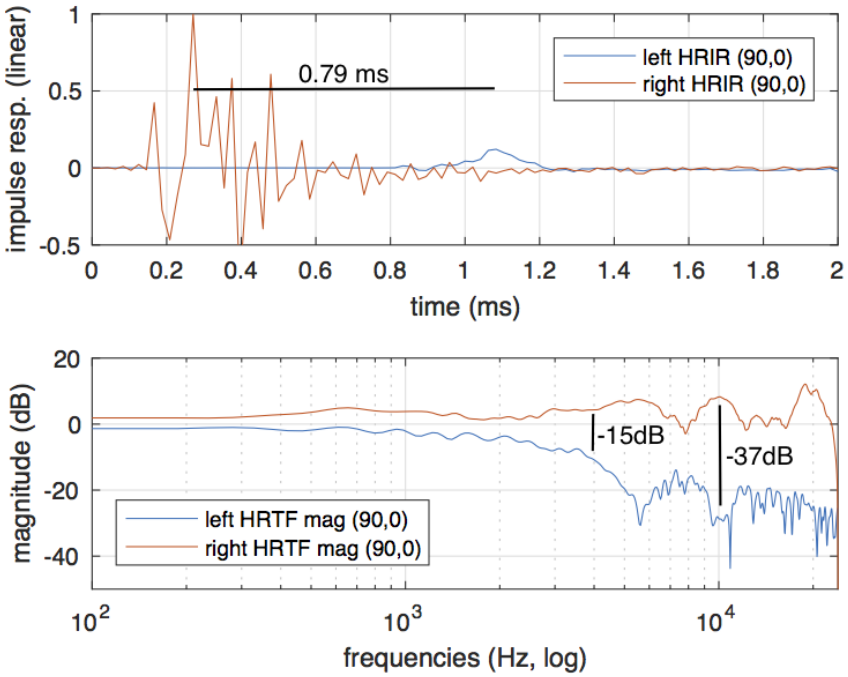


FIGURE 1.2 – Exemple d'HRIRs et de spectre d'amplitude de HRTFs (obtenues par transformée de Fourier des HRIRs) gauche et droite à la direction $(\theta, \phi) = (90°, 0°)$. L'ITD et l'ILD (à 4 kHz et 10 kHz) relatifs à cette mesure sont donnés sur la figure).

Les différences interaurales d'intensité sont liées au fait que l'onde sonore est diffractée par la tête et atteint donc l'oreille controlatérale avec une intensité moindre qu'à l'oreille ipsilatérale (où l'intensité y est amplifiée dans certains cas). Plus les longueurs d'onde sont courtes, plus l'obstacle de la tête est important. Pour des longueurs d'onde supérieures au diamètre de la tête, c'est-à-dire pour des fréquences inférieures à 500 Hz, ces différences sont négligeables. De cette façon, les hautes fréquences sont d'autant plus atténuées que les basses fréquences ce qui se traduit par des différences d'intensité plus grandes en hautes fréquences et qui varient plus rapidement avec l'azimut. Ainsi, l'ILD maximale est d'environ 20 dB à 4 kHz et de 35 dB à 10 kHz à $\theta = 90°$ [MMG89] (voir figure 1.2).

L'ILD varie également avec la distance. Pour des sources situées à moins de 50 cm du centre de la tête, l'ILD présente une forte dépendance avec la distance [BR99]. Cela est dû au fait que l'amplitude de la HRTF augmente à l'oreille ispilatérale et décroît à l'oreille controlatérale avec le rapprochement de la source. Cette augmentation d'ILD est davantage marquée en basses fréquences et pour des sources latéralisées (l'ILD reste nul sur le plan médian). Les valeurs peuvent atteindre jusqu'à 20 dB à des distances de 12cm [BR99]. Des valeurs d'ILDs élevées en basses fréquences caractérisent la région proximale, étant donné le contraste avec les valeurs faibles associées au champ lointain (à des distances de plus d'un mètre). L'ILD constitue donc un véritable indice de distance pour l'auditeur.

Les différences interaurales de temps d'arrivée entre les deux oreilles dépendent de la différence de chemin acoustique entre les deux oreilles et sont donc dépendantes des dimensions de la tête de l'auditeur. L'ITD maximale moyennée sur plusieurs individus de la base de données CIPIC [ADTA01] est de 646 μsec avec une déviation standard de 33 μsec. Dans [Mid99a], l'ITD maximale

évaluée sur une population de 33 individus varie de 657 μsec à 792 μsec avec une moyenne de 709 μsec et une déviation standard de 32 μsec. Etant donné que le seuil de discrimination (*Just Noticeable Difference*) de l'ITD est d'environ 10 μsec [Bla97], ces différences inter-individuelles de l'ITD maximale sont significatives, ce qui s'explique par une forte corrélation avec la taille de la tête. Lorsqu'un individu localise une source virtuelle synthétisée à partir de HRTFs correspondant à une tête plus petite, on observe une sous-latéralisation de la source sonore [Mid99b].

Pour une source lointaine située dans le plan horizontal et pour un modèle de tête sphérique, Woodworth (1938) propose un modèle géométrique de calcul de l'ITD (Δt) :

$$\Delta t = \frac{\Delta d}{c} = \frac{r \cdot (\theta + \sin\theta)}{c} \tag{1.1}$$

où Δd représente la différence de chemin acoustique entre les deux oreilles, r le rayon de la tête, c la célérité de l'onde sonore et $0 \leq \theta \leq \frac{\pi}{2}$. Cependant, ce modèle est approximatif étant donné que la tête n'est pas exactement sphérique.

Le rôle de l'ITD et de l'ILD dans la localisation sonore dépend de la fréquence de l'onde sonore. Les différences interaurales de temps constituent un indice prédominant pour la localisation de signaux à basse fréquence, ou plus généralement contenant des basses fréquences, inférieures à 1.5 kHz [WK92, Bla97, MM02a]. Entre 1.5 kHz et 4 kHz, l'ILD et l'ITD participent tous les deux à la localisation latérale. Au-delà de $3-4$ kHz (pour des sons purs haute fréquence ou pour des sources filtrées passe-haut), les jugements de localisation dans la dimension gauche-droite semblent être déterminés en majeure partie par l'indice d'ILD avec une faible contribution de l'ITD. On voit donc que ces deux indices ne sont pas totalement décorrélés. Cependant, on note un rôle plus important de l'ITD qui prédomine sur l'ILD lorsque toutes les fréquences sont disponibles ou lorsque les indices sont en contradiction [WK92]. Ces considérations sont relatives au cas de sources éloignées par rapport à l'auditeur. Dans une situation de champ proche, c'est-à-dire lorsque la source est à une distance de moins d'un mètre de l'auditeur, l'ILD devient très important en basses fréquences alors que l'ITD est inchangé.

1.1.1.2 Cônes de confusion

Les différences interaurales de temps et d'intensité sont ambigües en élévation à cause du caractère quasi-symétrique de la tête et des pavillons. Cette ambiguïté spatiale a été définie pour la première fois en 1939 par Wallach [Wal40]. Elle se représente par des contours sur lesquels les indices d'ITD et d'ILD sont constants et s'apparentent à des cônes dont le sommet est localisé au centre de la tête et dont l'axe coincide avec l'axe interaural. Ils sont appelés "cônes de confusion" [OP84] en référence aux ambiguïtés de localisation en élévation sur ces contours. Dans la littérature, on distingue souvent les confusions avant-arrière et haut-bas [WAKW93, Bro95].

Pour un modèle de tête sphérique et dont l'axe interaural est un diamètre de la sphère, ces lignes d'iso-ITD et d'iso-ILD sont circulaires, représentées par des plans sagittaux, i.e. centrées sur une angle latéral fixe. Cependant, la tête possède une géométrie plus complexe qui a pour conséquence de rendre ces lignes légèrement dépendantes de l'angle latéral, comme on peut le voir dans la figure 1.3 tirée de l'article de [WK99].

Les indices interauraux sont donc insuffisants pour déterminer la direction de la source sonore sur un cône de confusion. Afin de lever ces ambiguïtés, l'auditeur utilise les indices spectraux et les mouvements de la tête.

1.1.2 Indices spectraux

Durant son trajet entre la source et l'entrée des canaux auditifs, l'onde sonore subit des transformations spectrales dues aux phénomènes de diffraction et de réflexion qui interviennent au contact des épaules, de la tête, des pavillons, etc. Elles sont dépendantes de la morphologie de l'auditeur et spécifiques à la position de la source dans l'espace. Les multiples réflexions de l'onde sur le pavillon ont pour conséquence de créer des interférences constructives et destructrices en hautes fréquences et se traduisent par des pics et des creux dans le spectre de la source reçu aux tympans. Ce filtrage acoustique de l'onde sonore constitue des indices acoustiques appelés indices monauraux et véhiculent au système auditif des informations sur la direction de provenance du son. Des études ont montré qu'ils constituent des indices prédominants pour la localisation en élévation et pour la discrimination avant-arrière [HW74b, Bro95]. A l'inverse, ils contribuent peu à la localisation

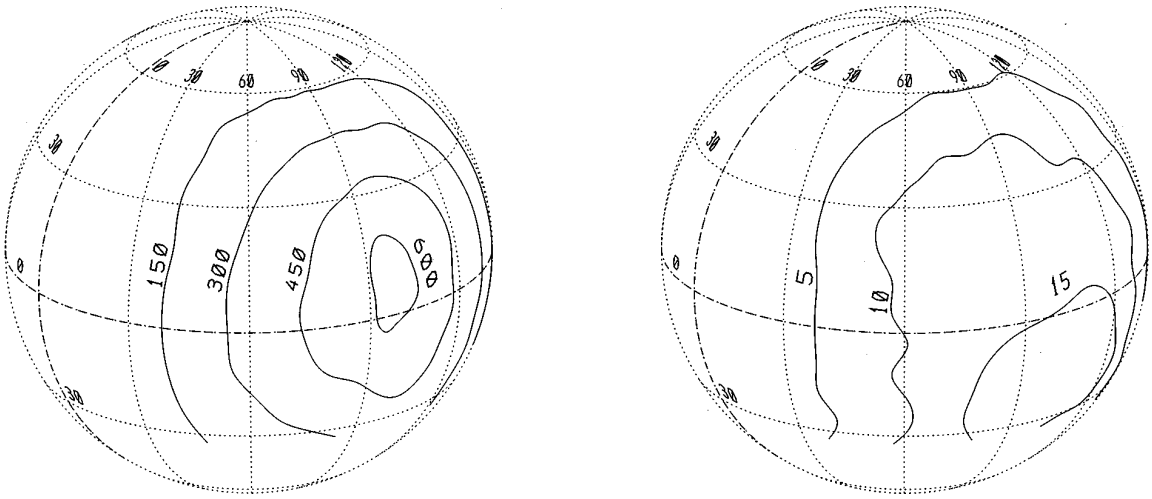


FIGURE 1.3 – Contours d'iso-ITD à gauche, en μsec, et d'iso-ILD à droite, en dB, d'un même individu positionné face à la direction frontale (0°, 0°) indiquée par l'intersection des arcs pointillés matérialisant le plan horizontal et le plan médian (source : [WK99]).

latérale [MM02a]. Ces indices interviennent particulièrement en hautes fréquences étant donné la taille des détails morphologiques qui en sont à l'origine. En effet, les pavillons d'oreille jouent un rôle sur les fréquences supérieures ou égales à $5 - 6$ kHz. Bronkhorst [Bro95] observe par exemple une réduction des erreurs de localisation de 13° à 9% et une réduction des confusions avant-arrière de 41% à 21% lorsque la fréquence de coupure passe de 7 kHz à 15 kHz (sources réelles). Langendijk et Bronkhorst [LB02] montrent que la contribution des indices spectraux de fréquences inférieures à 4 kHz dans le processus de localisation est négligeable.

La dépendance des caractéristiques spectrales hautes fréquences en fonction de la direction fait de ces indices monauraux des indices de localisation. Par exemple, l'onde sonore n'est pas réfléchie par les mêmes éléments du pavillon selon sa direction d'incidence. Par conséquent, les transformations spectrales sont différentes, apparaissent décalées en fréquence, ou sont plus ou moins marquées, comme on peut le voir figure 1.4 tirée du livre de Xie [XDNXB13].

Plusieurs études se sont intéressées à l'identification des caractéristiques spectrales variables selon l'angle d'incidence par l'étude du spectre d'amplitude des HRTFs. Hebrank et Wright [HW74b] ont par exemple observé le déplacement d'un zéro vers les basses fréquences (entre 5 et 11 kHz) à mesure que la source se déplace du haut vers le bas dans le plan médian frontal. Cette observation a également été mentionnée par Butler et Belendiuk [BB77], ainsi que Langendijk et Bronkhorst [LB02] qui suggèrent que cet indice serait responsable de la discrimination haut-bas sur un cône de confusion. Ces auteurs mettent également en évidence l'existence d'un pic spectral dans la région $[8 - 16]$ kHz présent à l'avant mais absent ou atténué à l'arrière, et qui jouerait par conséquent un rôle pour la discrimination avant-arrière. Toutefois, ils ont montré que la suppression de certains indices spectraux d'intérêt, tels que ceux mis en évidence pour la discrimination haut-bas ou avant-arrière, n'affectait pas la localisation si le reste du spectre était conservé. Cette observation suggère qu'il existe une multitude d'indices spectraux distribués sur toute la bande spectrale qui participent à la localisation en élévation (voir aussi Zhang et Hartmann [ZH10]). Au contraire, une réduction de la largeur de bande fréquentielle dégrade de façon significative la localisation en élévation. Des études ont montré que la localisation de stimuli à bande fréquentielle étroite est davantage guidée par la fréquence centrale de la bande fréquentielle disponible plutôt que par la position spatiale effective de la source [Mid92]. Ce phénomène est associé à la notion de "*covert peak*" et intimement lié aux "bandes directionnelles" discutées par Blauert [Bla97]. Par ailleurs, la mise en évidence des caractéristiques spectrales pertinentes pour la localisation auditive est rendue difficile à cause des larges différences qui existent entre les individus. En effet, les variations inter-individuelles de taille et de forme du pavillon d'oreille qui font obstacle aux fréquences supérieures ou égales à 4 kHz sont à l'origine de différences spectrales dans le filtrage du pavillon observé chez chaque individu. On observe par exemple des variations au niveau de la position fréquentielle, de l'amplitude et de la largeur des pics et des creux [LB02] mais aussi des décalages spatiaux [Gui09]. De nombreuses études travaillent à la mise en correspondance des différences spectrales et morphologiques. Etant donné la complexité et la singularité des profils spectraux, il

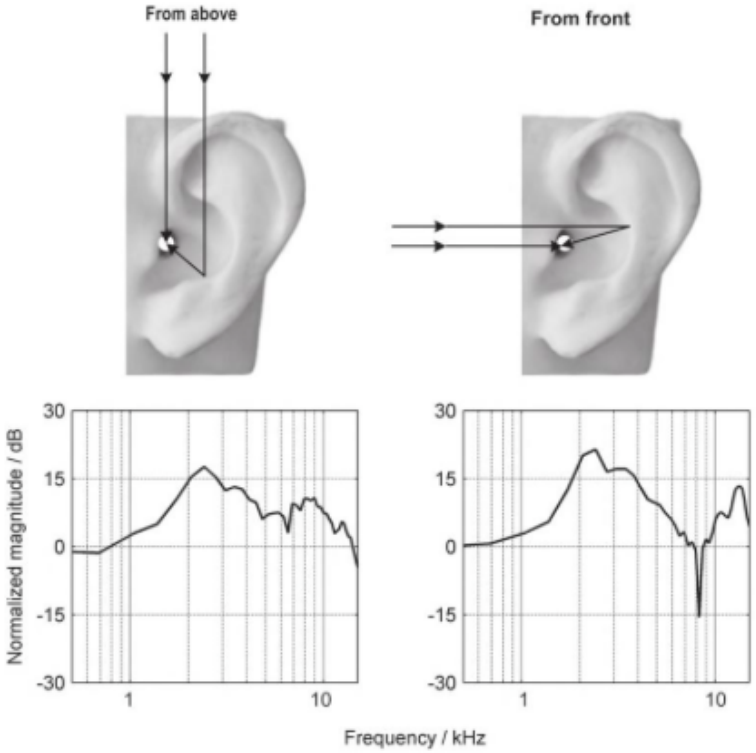


FIGURE 1.4 – Indices spectraux dépendants de la direction d'incidence de l'onde sonore sur le pavillon d'oreille (Figure 1.12 du livre de Xie [XDNXB13]).

apparaît cependant difficile de résumer les différences inter-individuelles à de simples décalages ou facteurs d'échelle.

Nous avons vu que les indices spectraux sont déterminants pour la localisation sur un cône de confusion. Cependant, le système auditif n'est capable d'interpréter ces indices que s'ils lui sont propres. En effet, sa faculté à localiser les sources sonores résulte de processus d'apprentissage et de calibration spécifique à ces indices acoustiques. Des erreurs de localisation, en particulier de type confusions avant-arrière et haut-bas, apparaissent lorsque ceux-ci ne sont pas individuels [WAKW93]. Middlebrooks [Mid99b] a mis en évidence que la distribution fréquentielle de ces caractéristiques individuelles (pics et creux) est plus importante que leur forme détaillée. En effet, le système auditif possède une certaine résolution spectrale et n'est pas sensible à tous les détails spectraux contenus dans les HRTFs (en particulier hautes fréquences). Plusieurs études ont montré qu'une réduction des détails spectraux par lissage spectral n'affectait pas la localisation sonore à condition qu'elle soit réalisée sur la base d'une échelle perceptive [KC98, HZK99].

Les indices spectraux sont des indices principalement monauraux. Hebrank et Wright [HW74a] ont montré que des individus entraînés sont capables de localiser une source sur le plan médian avec seulement une oreille. Au contraire, d'autres études ont montré que les deux oreilles sont nécessaires pour la localisation en élévation et la discrimination avant-arrière [ZH10] même si l'oreille controlatérale joue un rôle moins important [JCCvS04]. Morimoto [Mor01] a par ailleurs mis en évidence que lorsque la source est à angle latéral supérieur ou égal à 60° l'oreille controlatérale reçoit un signal trop faible pour contribuer à la localisation en élévation. Certains travaux sur la localisation sonore suggèrent également que le spectre interaural (i.e. la différence de niveau gauche-droite en fonction de la fréquence) joue un rôle pour la localisation en élévation [SBCD75] sans que celui-ci ne semble prépondérant [MN82].

1.1.3 Indices dynamiques

Les indices interauraux de temps et d'intensité qui composent la théorie Duplex permettent au système auditif de déterminer la position latérale de la source mais sont insuffisants pour localiser une source sonore dans une unique position de l'espace. Cela est dû aux cônes de confusion,

formés par les lignes d'iso-ITD et d'iso-ILD, engendrés par le caractère symétrique de la tête. Les indices spectraux permettent de lever cette ambiguïté. Cependant, ces indices sont présents en hautes fréquences, et on peut se demander comment procède le système auditif pour localiser une source en élévation lorsque celle-ci ne contient que des basses fréquences. C'est dans ces conditions qu'interviennent principalement les indices dynamiques.

L'hypothèse selon laquelle les mouvements de la tête jouent un rôle dans la localisation sur un cône de confusion a tout d'abord été proposée par Wallach [Wal40] puis validée par Perrett et Noble [PN97a, PN97b]. Pour comprendre leur rôle dans la discrimination avant-arrière, prenons un exemple. Lorsque la tête bouge de gauche à droite, l'ITD et l'ILD changent. Si la source est localisée à droite et que les indices diminuent avec ce mouvement, cela signifie que la source est positionnée à l'avant. Wallach montre que ce phénomène n'est pas directement lié aux mouvements de la tête mais plutôt aux changements d'ITD et d'ILD engendrés par le mouvement, i.e. aux indices dynamiques d'ITD et d'ILD : la même amélioration est observée lorsque la source est en mouvement et que la tête reste fixe. L'auditeur doit cependant être à l'origine de ces mouvements, en contrôlant par exemple la position de la source à l'aide d'un clavier [WK99]. A partir de l'observation que plus la source s'éloigne verticalement du plan horizontal, plus les indices interauraux changent lentement, Wallach suppose que ces indices jouent également un rôle dans l'identification de l'élévation.

Les études menées par la suite sur la compréhension des processus de localisation en élévation et de discrimination avant-arrière se sont plutôt focalisées sur les indices spectraux. Ceux-ci sont en effet prédominants lorsque la source contient des hautes fréquences (supérieures à 4 kHz environ). Cependant, lorsqu'elle en est dépourvue mais qu'elle contient des basses fréquences ($<$2 kHz), seuls les mouvements de la tête permettent de détecter l'élévation [PN97b]. Pour une source large bande, les mouvements de la tête peuvent aider à préciser la localisation en élévation et surtout de réduire les confusions avant-arrière qui apparaissent en condition d'écoute statique même si les indices spectraux sont disponibles (21% [Bro95], 19% [WAKW93], 6% [WK89]).

1.1.4 Perception de la distance

Lorsqu'une source sonore s'éloigne ou se rapproche de l'auditeur, on note différents changements acoustiques dans le signal reçu aux tympans qui constituent pour l'auditeur des indices de distance [Zah02]. Premièrement, l'intensité de l'onde sonore reçue aux oreilles diminue avec la distance (atténuation géométrique). Plus précisément, elle décroît de 6 dB à chaque doublement de distance. L'intensité est un indice de distance relatif, i.e. nécessitant la pré-connaissance de la puissance émise par la source. En effet, l'intensité perçue est à la fois liée à la puissance acoustique de la source et à sa distance par rapport à l'auditeur. Deuxièmement, en milieu réverbérant, le rapport d'énergie entre le champ direct et le champ réverbéré diminue avec la distance. Le rapport son direct et champ réverbéré est un indice de distance absolu, i.e. ne nécessitant pas la connaissance de la source. Enfin, certains phénomènes spectraux sont à l'origine d'indices de distance. Pour des distances importantes, l'atténuation de l'intensité acoustique est plus marquée en hautes fréquences étant donné les propriétés d'absorption de l'air. Cette dissipation d'énergie acoustique s'ajoute à l'atténuation géométrique. De plus, lorsqu'on se place en champ proche (distances inférieures à un mètre) et pour des sources latéralisées, l'ILD augmente plus rapidement en basses fréquences qu'en hautes fréquences à mesure que la source se rapproche [BR99]. Les différences interaurales d'intensité élevées en basses fréquences sont caractéristiques du champ proche.

1.2 Caractéristiques du système auditif

Le système auditif est capable de localiser une source sonore dans l'espace 3D avec une certaine précision, même en l'absence d'information visuelle. La précision de localisation varie avec la position de la source et selon les individus. Cette capacité a été acquise par l'expérience multisensorielle (visuelle, auditive et proprioceptive), et est donc intimement liée aux indices acoustiques propres à l'auditeur. Cependant, le système auditif présente une faculté à s'adapter à de nouveaux indices, notamment spectraux, à travers une phase d'apprentissage. De plus, il extrait les indices spectraux avec une certaine résolution spectrale.

1.2.1 Résolution spatiale

La résolution spatiale avec laquelle les humains peuvent localiser une source sonore dans l'espace est dénommée par l'angle minimal perceptible (ou *Minimum Audible Angle*, MAA) et est liée aux différences minimales perceptibles (ou *Just Noticeable Difference*, JND) des indices de localisation. Elle dépend à la fois de l'amplitude de variation des indices de localisation avec l'angle et de la capacité de l'auditeur à distinguer ces différences.

Le MAA varie dans l'espace et avec le contenu spectral de la source sonore. Dans la dimension horizontale, il est minimal à la direction frontale et augmente avec la latéralisation de la source. Pour un signal sinusoïdal de fréquence 1 kHz, Mills [Mil58] observe un minimum de 1° près du plan médian et une augmentation avec l'azimut jusqu'à un maximum de 10° sur les côtés. Dans ces conditions (dimension latérale, source à basse fréquence), l'angle minimal audible est relié aux différences d'ITD minimales perceptibles (ou JND) qui se trouvent autour de $10\mu s$ (d'après Blauert [Bla97]). Une première hypothèse concernant l'augmentation du MAA avec l'azimut suppose que la sensibilité du système auditif est plus grande aux ITDs faibles par rapport aux ITDs plus importants [DC77]. Une autre hypothèse, validée récemment par Smith et Price [SP14], réfute la précédente et montre que cette capacité est liée au fait que la quantité de variation d'ITD est plus importante proche du plan médian que sur les côtés pour une différence d'angle identique. Dans la dimension verticale, notre résolution spatiale en élévation est plus limitée que la résolution en azimut et diminue depuis la direction frontale jusqu'aux directions arrière et élevées (voir figure 2.5 du livre [Bla97]). Les valeurs de MAA en élévation sont très variables en fonction de la fréquence et du type de source. Différentes études, mentionnées dans [Bla97], indiquent une imprécision verticale à la direction frontale qui varie entre 4° pour un stimulus de bruit blanc, 9° pour une voix familière et 17° pour une voix non-familière.

Enfin, le système auditif perçoit la distance des sources sonores avec une précision bien moindre que la précision avec laquelle il détermine la direction. Sa précision à identifier la distance d'une source sonore varie avec la distance (voir figure 2.8 du livre [Bla97]) : entre $\approx$ 30cm à 1 mètre et $\approx$ 50cm autour de 5 mètres. De plus, Zahorik [Zah02] note une tendance à sous-estimer la distance de sources lointaines (>1.6m) et à sur-estimer les sources proches (<1.6m).

1.2.2 Plasticité

La localisation auditive est basée sur la connaissance des relations entre les indices acoustiques et l'information spatiale. L'apprentissage de ces relations relève de l'interaction entre plusieurs modalités sensorielles (vision, audition, proprioception) [AMS08]. Selon Poincaré, la perception de l'espace ne peut se développer qu'en présence de mouvements générés par l'individu, combinés à une sensation proprioceptive. L'exploration de l'espace s'effectue grâce à l'interaction entre l'individu et l'environnement i.e. à travers l'expérience de conséquences sensorielles engendrées par des actions motrices volontaires (e.g. des changements acoustiques induit par une rotation de la tête). Comme le dit Poincaré dans "La valeur de la science" (1905) : "Localiser un objet, cela veut dire simplement se représenter les mouvements qu'il faudrait faire pour l'atteindre".

A travers l'expérience multisensorielle, le système auditif a donc acquis une capacité à décoder les indices physiques de l'onde sonore (binauraux et monauraux) afin d'en extraire l'information spatiale. Pour localiser une source sonore en condition statique, on peut donc supposer qu'il analyse l'information acoustique et effectue une comparaison avec les indices, spécifiques à chaque direction d'incidence, qu'il a appris à interpréter avec l'expérience. Partant de l'observation que ces indices acoustiques changent avec l'évolution anatomique du corps humain au cours de l'existence, le système auditif présente une capacité d'adaptation. Autrement dit, il renouvelle en continu sa carte auditive spatiale qui lui permet de localiser les sons.

Plusieurs études ont montré la capacité du système auditif à s'adapter à de nouveaux indices spectraux. Une première méthode consiste à insérer des moules au niveau des pavillons d'oreille [HOR99, CB13, CBK14]. Ces études mettent en évidence une re-calibration de la carte auditive spatiale relative aux nouveaux indices spectraux après une période de quelques semaines (en situation réelle d'apprentissage sensorimoteur). L'amélioration des performances de localisation est observée pour toutes les directions, indépendamment de la zone visuelle [CB13]. De plus, la carte auditive spatiale relative aux indices originaux est conservée à l'issue de cette phase d'adaptation (i.e. quand on retire les moules) [HOR99]. Un seconde méthode consiste à utiliser un environnement virtuel dans lequel des HRTFs non-individuelles sont présentées au casque [ZBS$^+$06, MCD$^+$12, PK12]. La localisation y est immédiatement altérée par rapport à un envrionnement virtuel avec HRTFs individuelles. Cependant, ces études suggèrent que l'adaptation aux HRTFs non-individuelles est rapide : une réduction notable des confusions avant-arrière est observée à l'issue de 2 sessions d'en-

trainement de 12 [PK12] ou 30 minutes [ZBS+06] comprenant un retour audio-moteur, proprioceptif ou audio-visuel. Cependant, la durée nécessaire pour atteindre une re-calibration complète de carte auditive spatiale au travers d'expériences de réalité virtuelle n'est pas claire et semble dépendre des sujets [HOR99, ZBS+06].

1.2.3 Différences inter-individuelles de localisation

Plusieurs études ont révélé des différences inter-individuelles dans les performances de localisation, en particulier en élévation [WAKW93, WK89] et dans le pourcentage de confusions [WK99]. Deux hypothèses principales ont été formulées : (1) la différence provient du fait que les indices spectraux sont plus ou moins marqués selon les individus et qu'ils présentent une dépendance spatiale plus ou moins prononcée (facteur acoustique) ou (2) qu'il existe une différence dans nos capacités d'analyse des indices spectraux afin d'inférer la direction correspondante (facteur perceptif). Middlebrooks [Mid99b] et Wightman et Kistler [WK99] ont proposé que ces différences pouvaient également être liées à la qualité des HRTFs mesurées. Cependant, cette hypothèse a été réfutée par Langendijk et Bronkhorst qui observent que les sujets qui présentent de faibles performances de localisation en condition d'écoute virtuelle avec leur propres HRTFs présentent également des difficultés à localiser des sources réelles [LB02].

Afin d'examiner la première hypothèse, Wenzel et al. [WWKF88] et Middlebrooks [Mid99b] ont comparé les performances de localisation d'un "bon localisateur" à l'écoute des HRTFs d'un "mauvais localisateur" et inversement. Ces deux études observent une réduction des performances de localisation dans le cas où un "bon localisateur" utilise les HRTFs du "mauvais localisateur". Cette observation va dans le sens de l'hypothèse du facteur de qualité des indices spectraux. Cependant, et contrairement à Middlebrooks, Wenzel et al. n'observent pas d'amélioration des performances du "mauvais localisateur" à l'écoute des HRTFs du "bon localisateur". On peut donc supposer que la dégradation observée pour le "bon localisateur" à l'écoute des HRTFs du "mauvais localisateur" soit plutôt liée au fait que les indices spectraux ne sont pas individuels. De plus, si les différences inter-individuelles de localisation provenaient uniquement des différences entre "qualité" des indices spectraux, alors deux sujets auraient les même performances de localisation à l'écoute d'une même paire d'HRTF. Or ce n'est pas le cas, les individus réagissent différemment à l'écoute d'une même paire d'HRTF non-individuelle (voir par exemple [WAKW93]).

Middlebrooks [Mid99a, Mid99b] a mis en évidence une corrélation entre les performances de localisation à l'écoute d'une HRTF non-individuelle et l'amplitude des différences spectrales entre cette HRTF et celle de l'auditeur à la direction correspondante (suivant une certaine métrique spectrale). En effet, leurs résultats révèlent qu'une réduction des différences inter-spectrales permet une amélioration significative des performances de localisation. A l'opposé, les résultats d'Andéol et al. [ASG15] n'ont révélé aucune corrélation significative entre l'amplitude des différences inter-individuelles (au sens de la métrique de Middlebrooks) et les performances de localisation en conditions non-individuelles. Ces auteurs ont également étudié la relation entre la saillance des indices spectraux et les performances de localisation individuelles, mais aucun lien n'a été mis en évidence (voir également [AMS13]). Les résultats des études d'Andéol et al. vont à l'encontre de l'hypothèse du facteur acoustique. Au contraire, ils soulignent une plus forte influence des paramètres non-acoustiques, tels que la sensibilité aux modulations spectrales, sur les performances de localisation. De plus, ils montrent que les performances de localisation d'un auditeur peuvent être significativement améliorées par un processus d'entrainement perceptif à l'extraction des indices spectraux. Les résultats du modèle de Majdak et al. [MBL14] renforcent également la seconde hypothèse en montrant que les différences inter-individuelles semblent davantage être liées au niveau de calibration de l'auditeur par rapport à ses propres HRTFs plutôt qu'à une propriété spectrale intrinsèque à ses HRTFs. Langendijk et Bronkhorst soutiennent également cette hypothèse [LB02].

Enfin, dans le cadre d'une écoute au casque, il semblerait que le niveau d'expertise des sujets ait un rôle important dans les performances de localisation de sources virtuelles. Tout d'abord, les études ayant utilisé des auditeurs naïfs ne montrent pas de différence significative de taux de confusions entre des sources virtuelles synthétisées avec des HRTFs individuelles ou non-individuelles [Bro95, BWA01] contrairement à d'autres études (e.g. [WAKW93]). La différence serait masquée par un taux de confusions particulièrement élevé en condition individuelle dans le cas d'auditeurs naïfs. Middlebrooks [Mid99b] remarque plus globalement une précision de localisation plus faible chez ses auditeurs non expérimentés. Les résultats d'Andéol et al., basés sur une écoute au casque, pourraient également s'expliquer par le niveau d'expertise des sujets à l'écoute binaurale, qui augmente avec l'entrainement intensif à la localisation de sources virtuelles accompagné d'un retour visuel. Cela dit, cela ne remet pas en cause le fait que l'hypothèse non-acoustique soit la plus vraisemblable.

1.2.4 Modélisation des échelles perceptives

La compréhension de la réponse du système auditif aux sons est très importante dans le cadre du développement de dispositifs de restitution sonore. La connaissance des caractéristiques acoustiques pertinentes pour le système auditif permet d'appréhender le niveau de précision avec lequel reproduire de manière optimale un contenu sonore. De plus, elle permet d'estimer objectivement quelles seraient les distorsions perceptibles d'un signal sonore retransmis aux oreilles d'un auditeur à travers un système de restitution.

Premièrement, la sensation auditive d'amplitude varie avec le niveau de pression acoustique et la fréquence. Des échelles perceptives ont été construites pour déterminer le niveau sonore perçu à partir d'une mesure acoustique. Deuxièmement, l'oreille a une sensibilité fréquentielle limitée et variable avec la fréquence et le niveau sonore. Les effets de filtrages de la cochlée ont été étudiés et plusieurs modèles ont été proposés.

1.2.4.1 Echelle perceptive d'intensité

L'unité des décibels permet de quantifier le niveau sonore d'une façon plus proche de la sensibilité de l'oreille. Un décibel représente à peu près le seuil de discrimination humaine de deux sons de pressions acoustiques différentes (soit la plus petite différence de niveau perceptible) et 0 dB, le seuil d'audition. Le niveau sonore exprimé en décibels est relié logarithmiquement au rapport des puissances entre la pression acoustique mesurée p_{eff} (valeur efficace) et la pression acoustique de référence p_{ref} correspondant au seuil d'audition :

$$L_p = 20 \cdot \log_{10} \frac{p_{eff}}{p_{ref}} \tag{1.2}$$

où L_p est le niveau de pression sonore exprimé en décibels (ou plus précisément dB*SPL* pour *Sound Pressure Level*) et $p_{ref} = 20\mu$ Pascals.

Pour un niveau de pression sonore L_p donné, la sensation perçue du niveau varie en fonction du signal et notamment de sa fréquence (un son aigu est perçu plus fort qu'un son grave). La mesure de la sensation d'intensité sonore correspond à la sonie, exprimée en phones. Les courbes isosoniques sont les courbes d'égale intensité sonore perçue (de sonie identique) en fonction des fréquences. Ces profils montrent une sensibilité maximale autour de $2 - 4$ kHz et varient avec le niveau.

La courbe de pondération A permet d'interpréter le niveau sonore perçu d'un son mesuré par bande d'octave ou de tiers d'octave. Une valeur de niveau sonore en dBA correspond à la somme des valeurs efficaces mesurées dans chaque bande. Pour déterminer le niveau sonore perçu d'un son complexe (composé de plusieurs fréquences), l'intervalle fréquentiel audible est généralement divisé en bandes fréquentielles d'octave [Ols72] ou en bandes critiques.

1.2.4.2 Echelle perceptive fréquentielle

Selon l'organisation tonotopique de la cochlée, chaque endroit de la membrane basilaire répond de manière privilégiée à une fréquence donnée. Le traitement des différentes composantes fréquentielles d'un son complexe (composé de plus d'une fréquence) opère dans les différentes régions de la membrane basilaire. Elle peut être modélisée par un banc de filtres passe-bande, qui se chevauchent et sont centrés sur une fréquence caractéristique, que l'on appelle filtres auditifs. La largeur et la forme de chacun de ces filtres imitent la résolution fréquentielle du système auditif : la largeur des filtres augmente avec la fréquence étant donné que la résolution fréquentielle diminue. De plus, les fréquences centrales de chaque bande fréquentielle sont espacées de façon non régulière, approximativement logarithmique. Deux modèles permettent de modéliser le phénomène de résolution non-uniforme selon la fréquence.

La largeur des filtres auditifs peut être approximée par la largeur des bandes critiques, comme introduit par Fletcher dans les années 1940. Ce terme fait référence à de nombreux phénomènes perceptifs liés aux mécanismes cochléaires comme les effets de masquage, la sonie de sons complexes, ou la discrimination fréquentielle [Zwi61]. Par exemple, l'effet de masquage d'un son pur par une bande de bruit croît avec la largeur de la bande B (i.e. augmentation du seuil de détection du son pur) jusqu'à la largeur de bande critique $B = B_c$ à partir de laquelle le seuil n'augmente plus avec la largeur de bande. Les largeurs de bande B_c augmentent avec la fréquence du son pur masqué, ce qui signifie que la discrimination fréquentielle est meilleure en basses fréquences. La largeur de bande critique B_c est à peu près constante pour les fréquences inférieures à 500 Hz ($\Delta f = 100$ Hz) et augmente de façon approximativement proportionnelle à la fréquence au-delà de

500 Hz ($\Delta f = 0.2 \cdot f$). L'échelle des Barks, introduite par Zwicker en 1961 [Zwi61], est l'échelle des fréquences proportionnelle à la largeur des bandes critiques [FZ06] :

$$\Delta f_{CB} = 25 + 75 \left(1 + 1.4 \left(\frac{f_c}{1000} \right)^2 \right)^{0.69} \tag{1.3}$$

Δf_{CB} est la largeur de bande critique associée à la fréquence centrale du filtre f_c, exprimée en Hertz. Une distance de 1 sur cette échelle correspond à une bande critique (l'échelle comprend un total de 24 bandes critiques) et correspond de très près à 100 mels [Zwi61].

L'échelle MEL (définie par Stevens et al. en 1937) est une échelle qui modélise l'interprétation de la hauteur d'un son par l'oreille humaine. Le MEL m est reliée à la fréquence f suivant l'équation :

$$m\left(f\right) = 2595 \log \left(1 + \frac{f}{700} \right) \tag{1.4}$$

Les espacements entre les points de cette échelle correspondent à la même différence de hauteur perçue. Celle-ci est reliée proportionnellement à la fréquence en dessous de 1 kHz puis augmente logarithmiquement avec la fréquence.

Un autre modèle de largeur de bande appelé *Equivalent Rectangular Bandwidth* (ERB) a été introduit par Glasberg et Moore en 1990 [GM90] :

$$\Delta f_{ERB} = 24.7 \left(0,00437 f_c + 1 \right) \tag{1.5}$$

Un exemple de banc de filtres basé sur ce modèle peut être visualisé en figure 1.5. Les filtres ERB modélisent avec plus de précision les phénomènes auditifs observés [MG96]. Par exemple, contrairement à l'échelle des Barks, l'échelle ERB expliquent mieux comment les contours d'égale intensité perçue changent avec le niveau.

Le filtre gammatone a été souvent utilisé pour modéliser la forme des filtres auditifs. Sa réponse impulsionnelle et sa fonction de transfert sont données par la formule [BK01] :

$$\begin{cases} h_G(t) = t^{n-1} e^{-2\pi b_w t} \cos\left(2\pi f_c t + \psi\right) \\ H_G(f, f_c) = \left(\frac{1}{1 + j(f - f_c)/b_w} \right)^n \qquad \text{for } \psi = 0 \end{cases} \tag{1.6}$$

où $t \geq 0$, n correspond à l'ordre du filtre, b_w la largeur de bande, f_c la fréquence centrale du filtre et ψ la phase initiale (ici, $\psi = 0$). Patterson et al. [PRH+92] ont montré que le filtre gammatone d'ordre 4 modélise de façon précise les filtres auditifs, par conséquent $n = 4$. Plus récemment, il a été montré que le banc de filtres gamma-chirp offrent une approximation plus exacte de la réponse fréquentielle du système auditif en ajoutant une correction dépendante du niveau sonore à la réponse fréquentielle du filtre gammatone classique.

La précision avec laquelle le système auditif encode l'information spectrale est très réduite par rapport au degré de précision contenu dans les HRTFs mesurées [CP94] et ne correspond pas à une échelle linéaire (contrairement à l'échantillonnage fréquentiel sous-jacent à la transformée de Fourier). A partir de l'hypothèse que le spectre d'amplitude des HRTFs ne nécessite pas de plus grande résolution que celle de la cochlée, un banc de filtres auditifs peut être utilisé pour simplifier l'information spectrale contenue dans les HRTFs. L'application d'un banc de filtres modélisant le traitement de la cochlée sur les HRTFs permet de lisser le spectre d'amplitude sans générer d'artefacts audibles [BK01]. Cette propriété présente un avantage au niveau de la réduction de données (limitation de la mémoire requise et du coût de calcul). C'est ce que nous étudierons plus précisément dans le chapitre suivant en section 2.3.1.3 concernant la modélisation des HRTFs.

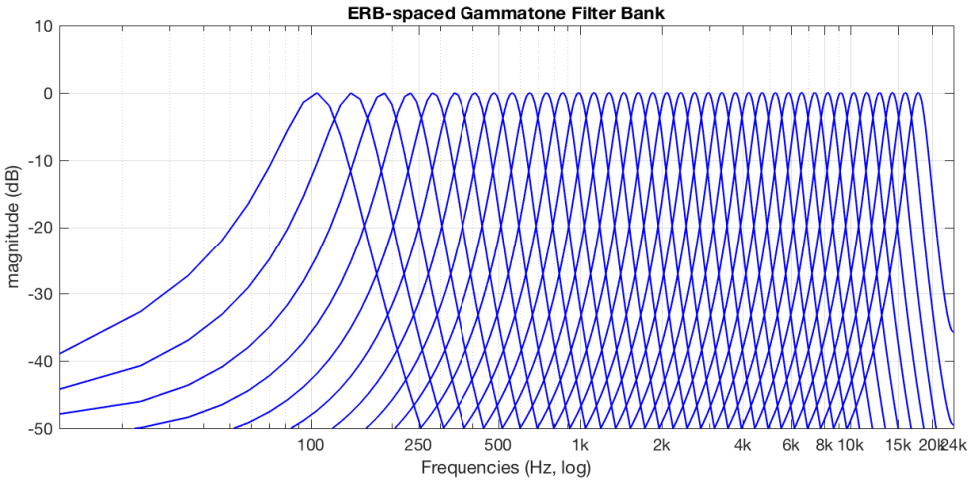


FIGURE 1.5 – Banc de 38 filtres gammatone de largeur de bande ERB et de fréquences centrales espacées d'un ERB entre 100 Hz et 18 kHz.

Conclusion

Le système auditif humain localise les sources sonores par l'analyse de différents indices de localisation résultants de phénomènes acoustiques dépendants de la direction d'incidence de l'onde sonore et de la distance de la source par rapport à l'auditeur. Il existe deux grandes classes d'indices relatifs à la direction d'incidence : les indices interauraux de temps et d'intensité (ITD et ILD) responsables de la localisation latérale, et les indices spectraux indispensables pour la localisation en élévation et la discrimination avant-arrière. Les indices interauraux dynamiques, engendrés par les mouvements de la tête, jouent également un rôle dans l'extraction de l'élévation et la résolution des ambiguïtés avant-arrière. Ces indices interviennent dans différentes bandes de fréquences et sont intimement liés à la morphologie de l'individu.

Le système auditif est donc capable d'interpréter l'information spatiale intrinsèque aux signaux acoustiques reçus aux tympans. Cette capacité a été acquise avec l'expérience multi-sensorielle (vision, audition, proprioception) par l'apprentissage d'une carte auditive spatiale basée sur les indices individuels. Nous avons vu que des différences inter-individuelles peuvent exister dans le niveau de calibration de l'auditeur par rapport à ces indices. Aussi, le système auditif présente une capacité d'adaptation à d'autres indices spectraux, moyennant une phase d'apprentissage passive ou active.

La compréhension du mécanisme de localisation auditive est essentielle pour mener à bien les objectifs fixés dans la thèse. Dans un premier temps, elle permet de prendre conscience de l'ensemble des éléments à considérer pour optimiser la synthèse binaurale au casque et d'analyser l'origine des défauts de spatialisation observés (1) lorsque les filtres utilisés lors de la synthèse sont non-individuels, (2) lorsque la synthèse est statique (mouvements de tête absents), (3) lorsque l'effet de salle est négligé. Dans un deuxième temps, elle est indispensable si l'on souhaite modéliser le mécanisme de localisation auditive et prédire les directions perçues à partir de la connaissance de l'information acoustique reçue à l'entrée des canaux auditifs gauche et droit. Bien que les nombreuses études ayant tenté d'identifier les indices acoustiques de localisation s'accordent à dire que la localisation auditive est en grande partie basée sur le spectre de l'onde sonore arrivant aux canaux auditifs de l'auditeur, il n'existe pas de consensus sur les caractéristiques spectrales pertinentes pour la localisation auditive. De plus, les processus d'extraction et de reconnaissance des indices spectraux restent encore mal connus. Ils constituent la base des modèles de prédiction de la localisation en élévation.

Des études psychoacoustiques ont mis en évidence que le niveau sonore et la hauteur d'un son perçu sont reliés de façon non linéaire à l'intensité acoustique et à la fréquence. De plus, le système auditif présente une résolution fréquentielle limitée et n'utilise qu'une partie de l'information spectrale acquise par la mesure acoustique des HRTFs. Cela suggère que la complexité des HRTFs peut être réduite. En effet, plusieurs études ont montré que la localisation est robuste à un lissage de la structure fine du spectre des HRTFs, notamment par l'application d'un banc de filtres auditifs conçu à partir de considérations psychoacoustiques.

La connaissance de ces aspects physiologiques permet d'appréhender avec quelle précision nous devons reproduire l'information notamment spectrale, afin de garantir que la localisation des sources virtuelles soit conforme à la situation d'écoute réelle. Elle représente un premier pas vers la définition d'une méthode d'estimation objective des distorsions perceptibles (métrique). Dans le cadre de modèles de localisation auditive, la compréhension des traitements physiologiques permet de s'approcher au mieux des processus réels.

La technique binaurale a pour but de reproduire les mêmes percepts qu'en situation d'écoute réelle. Ce chapitre a permis d'identifier les indices acoustiques nécessaires à la localisation auditive en trois dimensions. L'objectif du chapitre suivant est de définir la méthodologie qui permet de les acquérir et de les reproduire de manière optimale.

Chapitre 2

La technique binaurale

Ce chapitre introduit l'ensemble des éléments qui composent la synthèse binaurale, de la captation des fonctions de transfert à l'ensemble des facteurs déterminants la qualité du rendu. L'obtention des filtres pour une large population d'individus est indispensable à la mise en place de tests d'écoute, à l'étude des variations inter-individuelles et au développement de méthodes d'individualisation par la sélection guidée. Les méthodes de représentation plus concise de l'information individuelle et inter-individuelle permettent de simplifier les données, en ne conservant que l'information pertinente pour le système auditif, mais aussi de visualiser les dimensions intrinsèques de ces données à haute dimension.

L'apparition et la généralisation des dispositifs de réalité virtuelle ont créé un besoin de développer des solutions de rendu sonore 3D. Les techniques de spatialisation sonore ont permis de répondre à cette demande en intégrant la composante spatiale à la reproduction d'une scène sonore.Elles offrent un espace sonore immersif où des sources virtuelles peuvent être précisément localisées tout autour de l'auditeur. Contrairement aux systèmes de rendu multi haut-parleurs (HOA, WFS, etc.), la technique binaurale utilise simplement une restitution sur casque audio standard et se présente ainsi comme un outil privilégié pour démocratiser l'audio immersif. L'information spatiale y est délivrée au travers des deux canaux auditifs via l'utilisation des HRTFs. Un contenu sonore binaural, constitué de deux signaux gauche et droit, peut être capté directement en plaçant un microphone miniature à l'entrée de chacun des canaux auditifs d'un individu ou d'une tête artificielle, ou bien synthétisé par filtrage d'une source avec des HRTFs individuelles ou non-individuelles. Le principe de la synthèse binaurale est illustré figure 2.1. La procédure de synthèse peut concerner un son monophonique, stéréophonique ou multi-canal. En effet, la technique binaurale peut décoder les formats audio multi-canaux par la virtualisation des haut-parleurs. Par sa flexibilité et sa simplicité de mise en œuvre, la technique binaurale apparaît comme un outil privilégié pour les applications de réalité virtuelle, le rendu sonore de téléconférences ou de jeux vidéos. L'évolution de ces systèmes de spatialisation sonore requièrent des méthodologies d'évaluation de la qualité du rendu plus développées, qui prennent notamment en compte les attributs perceptifs liés à la perception spatiale. Ces méthodes sont indispensables pour identifier les facteurs en jeu et ainsi optimiser la qualité du rendu spatial.

L'acoustique virtuelle au casque est basée sur l'hypothèse qu'en reproduisant aux oreilles de l'auditeur exactement les mêmes stimulations acoustiques qu'en présence de sources réelles alors les sensations perçues par l'auditeur seront identiques à la situation d'écoute réelle. Pour obtenir un tel résultat, la synthèse binaurale doit remplir différentes conditions. Premièrement, la reproduction de la position spatiale exacte des sources suppose l'utilisation des indices de localisation propres à l'auditeur lors de la synthèse, i.e. les indices spécifiques à sa morphologie, tel qu'exposé au paragraphe 1.1.2. Deuxièmement, la position spatiale des sources sonores virtuelles doit rester constante relativement à l'orientation de la tête de l'auditeur, ce qui sous-entend la synthèse dynamique des sources en temps réel en parallèle de l'utilisation d'un système de repérage spatial de la tête. Troisièmement, la synthèse binaurale ne doit pas se contenter de simuler la présence de la source mais également l'effet de salle engendré naturellement par la propagation des ondes sonores émises par la source dans le milieu. Quatrièmement, les chaînes de captation et de reproduction doivent être transparentes, i.e. que les effets de filtrage des transducteurs sur les signaux enregistrés et transmis doivent être éliminés pour assurer la fidélité de la reproduction. Lorsque ces conditions ne sont pas respectées, des études ont observé différents artefacts tels qu'une augmentation des confusions avant-arrière, une diminution de la précision de la localisation en élévation et un manque d'externalisation (i.e. les sources sont perçues proches voire à l'intérieur de la tête) [BWA01]. Des

travaux ont été menés afin d'identifier les éléments influant sur chacune de ces caractéristiques liées au rendu binaural.

La synthèse binaurale repose sur l'utilisation des filtres (HRTFs) consignant l'ensemble des indices acoustiques de localisation relatifs à chaque individu et chaque direction. L'acquisition des HRTFs nécessite un système de mesure particulier, installé en chambre anéchoïque. Plusieurs bases de données publiques existent dans la littérature. Un des premiers objectifs de cette thèse concerne la mesure d'une base de données de HRTFs sur un échantillonnage sphérique fin et déterminé à partir de considérations mathématiques et pratiques. Pour assurer la transparence de l'ensemble des équipements utilisés, de la captation des HRTFs à la reproduction au casque audio, des méthodes d'égalisation ont été proposées dans la littérature. Nous présentons leur mise en œuvre dans le cadre de la base de données mesurée.

Enfin, les HRTFs sont obtenues sous forme de filtres linéaires qui peuvent être modélisés de plusieurs façons. Des études psychoacoustiques ont montré qu'une réduction de l'information spectrale d'amplitude et de phase est possible étant donnée la résolution limitée du système auditif. La réduction d'information peut être menée par l'intermédiaire de la décomposition des données sur une base de fonctions spatiales ou spectrales, pré-définies ou déterminées statistiquement, en limitant le nombre de coefficients utilisés pour la recomposition. Les méthodes de décomposition et de réduction de dimension sont des outils qui permettent de mettre en évidence les caractéristiques intrinsèques de données complexes, telles que les HRTFs, responsables de la variance présente dans les données.

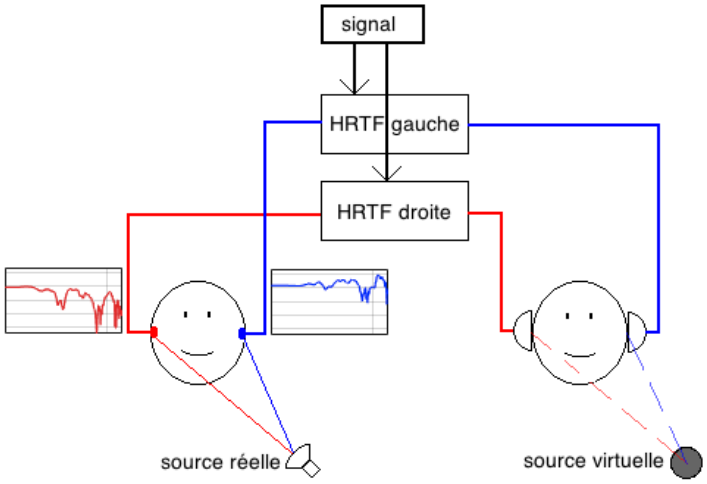


FIGURE 2.1 – Principe de la synthèse binaurale : de la captation de HRTFs à la reproduction d'une source virtuelle au casque audio.

Ce chapitre présente dans un premier temps les dimensions perceptives et attributs associés à la reproduction d'un contenu sonore spatialisé ainsi que les méthodes mises en œuvre dans la littérature pour évaluer l'ensemble de ces caractéristiques. Dans un deuxième temps, nous exposons les éléments qui contribuent à la fidélité de la reproduction des informations directionnelles et d'éloignement des sources virtuelles synthétisées par la technique binaurale ainsi qu'aux sensations d'immersion et de réalisme. Puis, nous présentons les bases de données existantes et les spécificités de la base de données mesurée dans le cadre de cette thèse. L'égalisation de l'ensemble de la chaîne de captation et de reproduction (casque audio) est étudiée en détails. Enfin, les modes de représentation des informations spatiale, spectrale et temporelle des HRTFs ainsi obtenues sont présentées en lien avec les méthodes de décomposition sur des fonctions de base. Une introduction aux méthodes de réduction de dimension appliquées aux bases de données de HRTFs ainsi que les avantages qu'elles présentent en termes d'analyse de ce type de données complexes est donnée en fin de chapitre.

2.1 Fidélité du rendu et facteurs en jeu

Cette section présente les dimensions perceptives qui définissent la qualité d'expérience binaurale et les méthodes qui permettent de les évaluer. Nous parlerons ensuite de l'importance des différents éléments qui composent la synthèse tels que l'individualisation des filtres binauraux, le rôle de l'effet de salle et de la synthèse dynamique, ainsi que les dimensions perceptives en jeu.

2.1.1 Dimensions perceptives et méthodes d'évaluation

L'évolution des techniques de spatialisation sonore soulève un besoin de développer des méthodes d'évaluation de la qualité du rendu. Pour pouvoir comprendre et optimiser la qualité du rendu spatial d'une production binaurale ou multicanale, l'évaluation ne doit pas se limiter à la qualité globale mais à chacune des dimensions perceptives utilisées consciemment ou non par les auditeurs pour juger de la qualité perçue.

Dimensions perceptives De nombreuses études ont tenté d'identifier les dimensions perceptives sous-jacentes à la qualité sonore. Malgré la difficulté liée à l'interprétation sémantique des sensations perceptives, deux grandes classes d'attributs perceptifs ont été mises en évidence : le timbre et l'espace [LBCP10]. Premièrement, bien que les définitions soient diverses, le timbre d'un son est souvent relié à ses caractéristiques spectro-temporelles et dépend des caractéristiques acoustiques de l'effet de salle. La fidélité de la reproduction des caractéristiques timbrales d'une source sonore participe à la sensation de naturel. Deuxièmement, les caractéristiques spatiales sont divisées entre les aspects de localisation et l'impression d'espace. Tout d'abord, la fidélité de la reproduction de la position spatiale des sources (définie par la direction et la distance par rapport à l'auditeur) est une caractéristique d'importance dans le cadre des techniques de spatialisation sonore. C'est pourquoi des tests de localisation sont très souvent employés pour valider les méthodes de spatialisation, en particulier les méthodes d'individualisation des HRTFs [See03, KP12]. L'analyse des performances de localisation permet d'identifier les distorsions spatiales de l'espace virtuel. Par ailleurs, contrairement aux systèmes de reproduction sur haut-parleurs, la perception de distance ou plus précisément l'externalisation d'une source sonore synthétisée au casque (i.e. la localisation à une certaine distance de la tête) est une caractéristique cruciale en synthèse binaurale. En effet, le manque d'externalisation (i.e. la localisation dans la tête ou intra-crânienne) est un artefact commun rencontré avec cette technique. Dans le cadre de la reproduction de scènes sonores complexes, la lisibilité, c'est-à-dire la capacité à distinguer les différentes sources, est un également attribut qui peut être utilisé pour évaluer la qualité du rendu spatial. Enfin, l'impression d'espace est liée aux sensations d'immersion, d'enveloppement, de profondeur et de largeur. Ces sensations sont intimement liées aux caractéristiques acoustiques d'un effet de salle comme le temps de réverbération, le temps d'arrivée et la distribution spatiale des premières réflexions. Le naturel et le réalisme d'une scène sonore sont également des attributs pertinents pour l'évaluation d'environnements de réalité virtuelle et s'expriment notamment par des jugements de préférence [BR00].

Méthodes d'évaluation Parmi les méthodes d'évaluation de la qualité du rendu, il existe tout d'abord les méthodes d'évaluation directe, où le stimulus est noté par le participant sur une échelle étiquetée par un attribut défini en amont par l'expérimentateur, et peut être présenté conjointement avec une référence explicite (voir la méthode MUSHRA [SK12, FMB14]). Les méthodes indirectes consistent à extraire des informations sur la qualité du rendu à partir des performances d'un sujet à réaliser une tâche. Le test de localisation en est un exemple : l'auditeur indique la direction, et éventuellement la distance, perçue de la source sonore et l'expérimentateur déduit la fidélité du rendu spatial par l'analyse des erreurs de localisation. Cela dit, il existe un biais dans les réponses relatif à la méthode de report, comme on le verra au chapitre 4.

La question du stimulus est essentielle et dépend de l'application. Dans le cadre de tests de localisation, les stimuli sont généralement des bruits large bande afin de délivrer à l'auditeur l'ensemble des indices de localisation. Parmi les types de stimuli utilisés dans la littérature, on peut mentionner les stimuli de bruit ou de son pur (impulsifs ou continus), les stimuli de voix (de femme ou d'homme, familière ou non), un extrait sonore de scène complexe ou d'un contenu musical, etc. De plus, dans le cadre des méthodes de spatialisation, le caractère spatio-temporel du stimulus est important. Il peut concerner des trajectoires horizontales et verticales (plan médian) [Iwa06, KP12, AK15] ou un évènement ponctuel. Dans tous les cas, l'expérimentateur doit s'assurer que le stimulus n'est pas trop long pour limiter la durée du test et l'effet de fatigue du participant.

Pour conclure, malgré la diversité des attributs liés à la qualité du rendu spatial d'un contenu sonore spatialisé, la fidélité du rendu de la position spatiale des sources virtuelles est primordial. Par conséquent, les tests de localisation occupent une place importante dans les méthodes d'évaluation du rendu.

2.1.2 Nécessité d'individualiser

Synthèse binaurale avec HRTFs individuelles La synthèse binaurale repose sur l'utilisation des fonctions de transfert relatives à la tête qui délivrent les indices de localisation nécessaires à l'auditeur pour localiser une source virtuelle au casque. La synthèse binaurale au casque utilisant les HRTFs individuelles de l'auditeur offre globalement la même précision de localisation qu'à l'écoute de sources réelles en champ libre [WK89, MSJH96, LB00, MMS01]. Parmi les quelques défauts observés, on note dans un premier temps une augmentation des confusions avant-arrière [WK89, Bro95, KC05]. Cela est probablement lié au fait qu'en condition d'écoute réelle, de très petits mouvements de la tête peuvent être réalisés par l'auditeur même si il reçoit l'instruction de les éviter [WK99]. Ces mouvement donnent lieu à des indices de localisation dynamiques, qui sont absents en situation de synthèse binaurale statique, et jouent un rôle important dans la résolution de confusions avant-arrière (c.f. section 1.1.3). Dans un deuxième temps, Bronkhorst et al. [Bro95] indiquent une précision légèrement dégradée en élévation dans le cas de sources virtuelles avec une déviation standard des réponses dans la dimension verticale qui augmente de 8° à 13°. Cette observation est soutenue par Wightman et Kistler [WK89] qui observent une corrélation en élévation qui décroît de 0.9 à 0.83 dans le cas de sources virtuelles. Les auteurs de ces deux études suggèrent un lien avec l'existence de distorsions dans les HRTFs mesurées.

Synthèse binaurale avec HRTFs non-individuelles En pratique, les HRTFs individuelles de l'auditeur ne sont pas toujours disponibles et les HRTFs mesurées sur un autre individu ou une tête artificielle sont alors utilisées. L'écoute binaurale non-individualisée se traduit par une dégradation globale de la localisation des sources virtuelles due à l'utilisation d'indices de localisation non-individuels (i.e. étrangers à l'auditeur). Par rapport au cas individuel, on note une augmentation significative des confusions avant-arrière [WAKW93, MSJH96, ZBS$^+$06] et une baisse des performances en élévation [WAKW93, Bro95]. Cela s'explique par le rôle important des indices spectraux dans la localisation verticale et la discrimination avant-arrière, ainsi que leur forte corrélation avec les détails morphologiques des pavillons d'oreille, qui varient considérablement entre les individus. De plus, certaines études soulignent une accentuation de la localisation intra-crânienne [KC05] mais cette observation n'a été confirmée ni par Begault et al. [BWA01] ni par Møller et al. [MSJH96] (études basées sur des stimuli de voix). Ces deux études ont par ailleurs noté une tendance à surestimer l'élévation d'une source virtuelle située dans le plan horizontal avec l'utilisation de HRTFs non-individuelles [WAKW93,BWA01]. Du fait de l'utilisation d'indices interauraux non-individuels, des erreurs latérales peuvent également être observées si les différences de circonférence de tête sont importantes entre l'individu sur lequel les HRTFs ont été mesurées et l'auditeur [Mid99b].

Bien que la localisation des sources virtuelles se dégrade avec l'utilisation de HRTFs non-individuelles, le système auditif présente une capacité d'adaptation à de nouveaux indices acoustiques (c.f. section 1.2.2). En effet, des études ont montré que l'apprentissage d'un jeu de HRTFs non-individuelles (avec un retour auditif, visuel, proprioceptif) permettrait d'obtenir des performances de localisation similaires à celles l'écoute en condition individuelle (voir par exemple [ZBS$^+$06, CBK14]).

Niveau d'expertise Le niveau d'expertise des auditeurs semble jouer un rôle dans le taux de confusion observé. En effet, bien que des différences de taux de confusions avant-arrière d'un facteur deux soient observées entre des sources virtuelles synthétisées avec les HRTFs individuelles de l'auditeur et des sources réelles, les valeurs varient en fonction des études : Wightman et Kistler [WK89] observent une augmentation de 6% à 11% alors que Bronkhorst et al. [Bro95] notent des taux qui varient de 21% à 41% pour des auditeurs naïfs. Aussi, les différences de taux de confusions entre HRTFs individuelles et non-individuelles sont très réduites dans le cas d'auditeurs non-expérimentés [Bro95]. Notons que la méthode de calcul du taux de confusions peut également expliquer les disparités entre les études.

2.1.3 Réverbération

Dans les situations naturelles, le champ sonore est généralement composé du son direct et du champ réverbéré constitué des multiples réflexions du son sur les parois et sur les différents

obstacles de l'environnement. La durée de réverbération est une caractéristique intrinsèque de l'environnement et dépend de son volume et de la quantité d'absorption présente. Le rapport d'énergie entre champ direct et champ réverbéré va agir sur la sensation d'espace et plus spécifiquement sur la sensation de distance apparente de la source sonore. Ce rapport diminue avec la distance de la source sonore par rapport à l'auditeur.

Le mécanisme de localisation en milieu réverbérant est régi par l'effet de précédence. Quand le son est suivi par un écho suffisamment précoce, les deux sons fusionnent en un percept unique dont la localisation est dominée par le son direct. Au delà de 50 ms, l'auditeur percevra cependant la présence d'un écho. Bien que dominée par le son direct, la perception spatiale pourra être affectée par la réverbération, notamment sous forme d'un flou de localisation ou de largeur apparente de la source.

L'effet de salle peut être intégré à la synthèse binaurale par l'utilisation de fonctions de transfert mesurées en milieu réverbérant : on parle alors de BRIRs (*Binaural Room Impulse Responses*). Dans le cas où les fonctions de transfert ont été mesurées en chambre anéchoïque, l'effet de salle peut être ajouté en amont du filtrage par les HRTFs, en convoluant le signal à spatialiser avec une réponse impusionnelle de salle mesurée ou synthétisée. Contrairement aux BRIRs, les HRIRs offrent une flexibilité sur le choix des caractéristiques acoustiques de l'effet de salle.

L'ajout d'un effet de salle dans un contenu binaural donne lieu à une meilleure externalisation des sources, en particulier pour les sources frontales. Dans une étude sur des stimuli de voix, Begault et Wenzel [BWA01] observent une augmentation de 40% à 72% du taux de sources externalisées grâce à la simulation d'un effet de salle et montrent que la présence des premières réflexions jusqu'à 80 ms suffit à améliorer significativement l'externalisation. Ces auteurs ajoutent que la réverbération permet de préciser la localisation en azimut (baisse globale de 5° des erreurs non signées en azimut pour des sources sonores situées sur le plan horizontal). Pour une écoute en condition non-individuelle, Völk et al. [VHF08] soulignent cependant que l'effet n'est significatif que si les fonctions de transfert utilisées proviennent d'une tête humaine et non d'un mannequin.

La corrélation entre réverbération et externalisation s'explique par le fait que la distance perçue d'une source sonore est intimement liée au rapport entre le champ direct et le champ réverbéré. Ainsi, si l'énergie du champ réverbéré est nulle, les sources sonores sont perçues proches voire dans la tête.

Wenzel et al. [WAKW93] suggèrent également que l'effet de salle réduirait le taux de confusions avant-arrière en renforçant les différences timbrales entre sources avant et arrière.

2.1.4 Synthèse binaurale dynamique

En situation réelle, les mouvements de la tête et du corps sont omniprésents et sont constitués de mouvements de rotation et de translation dans l'espace. Ces mouvements modifient la position relative de la source sonore par rapport aux deux oreilles. La congruence entre la variation des indices acoustiques et la variation des indices proprioceptifs forme les indices de localisation dynamiques. Comme présenté en section 1.1.3, ils permettent la résolution des ambiguïtés avant-arrière et précisent la localisation verticale [Wal40, PN97b, WK99]. Dans le cadre de la simulation d'un environnement sonore au casque, il apparaît nécessaire de prendre en compte ces effets en traitant de façon dynamique (mise à jour en temps réel) la spatialisation des sources sonores autour de l'auditeur en fonction de la position et de l'orientation de sa tête. Cela nécessite l'interpolation des HRTFs mesurées et l'utilisation d'un système de suivi des mouvements.

Par rapport à la synthèse binaurale statique, la spatialisation binaurale dynamique permet tout d'abord une réduction des ambiguïtés avant-arrière [WK99, BWA01, PRD14]. Begault et al. [BWA01] observent une réduction de 59% à 28% du taux de confusions avant-arrière. De plus, elle améliorerait l'impression de réalisme [XDNXB13]. Durlach et al. [DRP+92] suggèrent également que les sources virtuelles sont mieux externalisées. Cependant, cette dernière hypothèse n'est pas soutenue par Wightman et Kistler [WK99] et Begault et Wenzel [BWA01].

Dans le cadre de stimuli de voix, une fois le suivi de la tête inclus, Begault et al. [BWA01] n'ont plus observé d'impact significatif de l'individualisation des HRTFs sur la précision de localisation et du degré de réverbération sur le niveau de réalisme. Cette étude suggère donc que la spatialisation binaurale dynamique permettrait de palier en grande partie le problème de l'individualisation et de la simulation de la réverbération. Paukner [PRD14] soutient ce résultat dans une étude où il compare l'impact du suivi de la tête sur la localisation de stimuli de voix ou de bruit, filtrés par des HRTFs individuelles ou non. Ici aussi, les différences de performances entre différents jeux d'HRTFs (individuels ou non) ne sont plus significatives lorsque le suivi de la tête est inclus.

2.2 De la mesure au VAS

La technique binaurale repose sur l'utilisation de fonctions de transfert d'oreille, spécifiques à l'auditeur. Dans le cas idéal, celles-ci sont mesurées en environnement anéchoïque selon un grand nombre de directions autour du sujet qui doit rester immobile pendant toute la durée d'acquisition. Le choix de la grille d'échantillonnage sphérique est le résultat d'un compromis entre la résolution spatiale perceptive et les contraintes de temps, qui dépendent du dispositif. Si celui-ci ne comprend qu'un seul haut-parleur alors la durée des mesures augmente considérablement par rapport à un système multi haut-parleurs (possibilité de mesurer plusieurs directions en même temps). La définition de la grille de mesure peut également chercher à répondre à des contraintes "mathématiques" comme par exemple la décomposition en harmoniques sphériques. C'est le cas notamment de la grille d'échantillonnage adoptée pour l'acquisition de la base de données de HRTFs dans le cadre du projet BiLi. La mesure de cette base de données fait l'objet de la première contribution principale de cette thèse. Après en avoir présenté les principales caractéristiques en comparaison avec les bases de données existantes, nous nous intéresserons au traitement post-mesures qui consiste principalement à compenser la chaîne d'acquisition (réponses des haut-parleurs, microphones, etc.).

2.2.1 Bases de données de HRTFs

Il existe plusieurs bases de données d'HRIRs publiques qui se distinguent par la méthode et le système d'acquisition (généralement propres au laboratoire), les positions de mesure (distribution des directions et distance), l'environnement de mesure (HRIRs mesurées en chambre anéchoïque ou BRIRs mesurées en milieu réverbérant), la fréquence d'échantillonnage, la méthode d'égalisation, etc. Généralement, les données à la fois brutes et égalisées sont disponibles. Parmi les bases de données d'HRIRs publiques, on peut mentioner 3 bases de données d'HRIRs mesurées avec la méthode du conduit auditif bloqué (i.e. micros placés juste à l'entrée du canal auditif grâce à des moules individuels) : CIPIC [ADTA01](UC Davis, 45 individus, 1250 directions, égalisation champ libre), Listen [1] (IRCAM, 187 positions, 51 sujets, égalisation champ diffus) ou encore la base de données ARI [2] (120 individus, 1550 directions, milieu semi-anéchoïque, égalisation champ diffus).

Ces 3 bases de données sont disponibles en format SOFA (*Spatially Oriented Format for Acoustics*, [MIC+13]). Ce format, récemment standardisé par le comité de standardisation de l'AES sous le nom d'AES69-2015, permet de rassembler les principales bases de données publiques dans un format standardisé qui donne accès aux mesures mais également aux méta données relatives aux mesures.

2.2.2 Mesure de la base de données BiLi

Une base de HRTFs à haute résolution spatiale a été mesurée à l'IRCAM et à Orange Labs dans le cadre du projet BiLi. Ces mesures permettent d'acquérir une référence individuelle pour un ensemble de participants du projet dans l'optique de mener différents tests de validation perceptive.

Ce travail a fait l'objet d'un article de conférence [CBNW14] présenté en août 2014 au Forum Acusticum à Cracovie. Cet article est reproduit en Annexe A et concerne uniquement le système d'acquisition de l'IRCAM. Nous en présentons ici un résumé en insistant sur les considérations qui ont conduit au choix de la grille d'échantillonnage spatial. Enfin, nous discutons des différences principales entre les mesures de l'IRCAM et d'Orange. En effet, dans le travail qui suit, nous aurons l'occasion de réunir les mesures de HRTFs réalisées dans les deux laboratoires. Il est donc nécessaire d'avoir conscience des éléments qui les distinguent.

2.2.2.1 Équipement et acquisition audio

Ces mesures ont été réalisées dans la chambre anéchoïque de l'IRCAM (voir photo en figure 2.2(a)). Le sujet était assis sur un siège fixé sur une table tournante (B&K 9640) permettant d'atteindre les valeurs d'azimut cibles et un bras articulé, supportant quatre haut-parleurs, permettait de contrôler l'élévation. Deux lasers, fixés de part et d'autre dans l'axe de rotation du bras, matérialisent l'axe interaural et permettent d'ajuster la position de la tête du sujet au début de chaque séance de manière à ce que les conduits auditifs (axe interaural) se trouvent en coïncidence avec ces faisceaux.

Les haut-parleurs sont de marque ELAC et sont placés à une distance d'environ 2 mètres du centre de la chambre anéchoïque (axe de rotation de la table tournante). Les cellules microphoniques

1. http ://recherche.ircam.fr/equipes/salles/listen/
2. https ://www.kfs.oeaw.ac.at/

Knowles (de type FG26107 C34) sont insérées dans des moules de conduits auditifs préalablement réalisés de manière individuelle chez un audio-prothésiste. Ces moules ont d'abord été découpés afin de placer la capsule microphonique juste à l'entrée du canal auditif puis ont été vernis et percés.

La technique de mesure adoptée est celle dite du "sinus glissant" [Far00] balayant les fréquences de 0 à 48 kHz de façon logarithmique sur une durée de 682 ms. La durée du sinus glissant est choisie suffisamment longue pour garantir un bon rapport signal-à-bruit et suffisamment courte pour éviter les risques de mouvements incontrôlés du sujet et ainsi satisfaire l'hypothèse d'invariance temporelle du système. La fréquence d'échantillonnage est de 96 kHz. Une fois la réponse impulsionnelle obtenue par déconvolution du signal d'excitation, une estimation du SNR ("Signal-to-Noise Ratio") est réalisée de telle sorte à répéter la mesure si celle-ci est considérée comme défectueuse (typiquement, lorsque SNR$<$30 dB). Les SNR obtenus au cours de la campagne sont de l'ordre de 72 dB pour le côté ipsilatéral et de l'ordre de 54 dB pour le côté controlatéral.

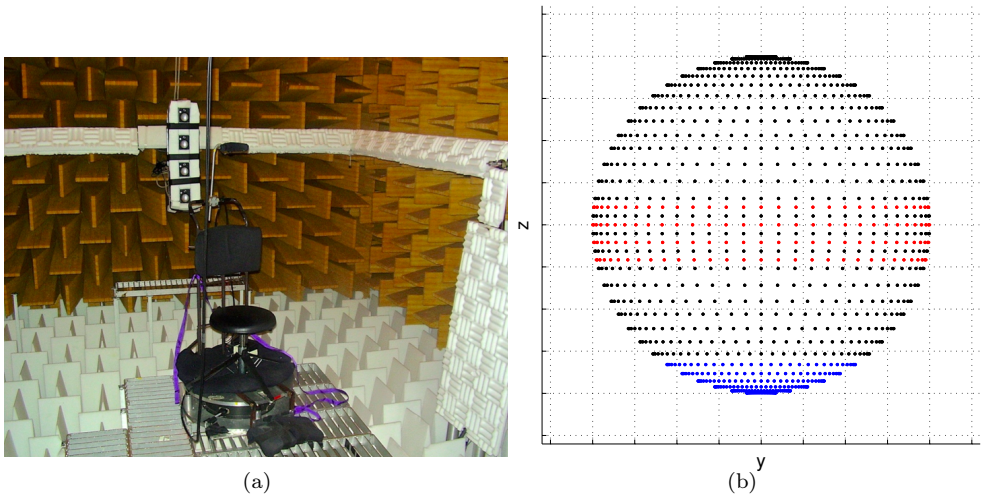
Au cours des mesures, la position et l'orientation de la tête sont relevées par un système de suivi de position OptiTrack composé de 6 caméras. Le système repère une cible constituée de 5 marqueurs rétro-réfléchissants placés sur un casque ajusté sur la tête du sujet. Cela permet de vérifier, pendant les mesures, que le sujet ne s'est pas trop écarté de la position idéale et offre si nécessaire la possibilité de corriger la direction de mesure *a posteriori*.

2.2.2.2 Grille de mesure

La grille de mesure a été choisie selon des critères mathématiques, liés à la volonté de procéder ultérieurement à une décomposition des HRTFs sur la base des harmoniques sphériques, et des critères pratiques, liés à la résolution angulaire des éléments mécaniques mobiles de la chambre anéchoïque (table et bras).

La grille d'échantillonnage sphérique retenue est de type gaussien, d'ordre 29 et composée théoriquement de 1800 points. Le choix de l'ordre 29 est dicté par les conditions de repliement spatial qui répondent au critère suivant : $kr < N$, avec r rayon de la tête, k le nombre d'onde et N l'ordre de décomposition [Raf05]. En d'autres termes, pour r de l'ordre de 10 cm et pour respecter le critère de repliement du spectre spatial sur l'ensemble du spectre audible ($f \approx 16$ kHz), l'ordre de décomposition doit être de l'ordre de 29. Sur le plan théorique, les principaux avantages des grilles gaussiennes sont d'offrir la possibilité de mener une décomposition en harmoniques sphériques à l'ordre N à partir d'un nombre de points relativement faible. Les grilles gaussiennes requièrent en effet $2(N+1)^2$ points d'échantillonnage, tandis que les grilles dites équiangulaires (pas constant en azimut et en élévation) nécessitent $4(N+1)^2$ points d'échantillonnage [Raf05]. D'autres grilles présentent l'avantage de ne requérir que $(N+1)^2$ points. Cependant, contrairement aux grilles gaussiennes, elles n'ont pas l'avantage de présenter un pas angulaire régulier. En effet, pour l'ordre 29, la grille gaussienne repose sur un pas azimutal régulier et égal exactement à 6°. Ce nombre entier est clairement avantageux dans la mesure où la résolution angulaire de la table tournante est de 1°. En revanche, le pas angulaire en élévation n'est pas tout à fait constant. Il varie cependant faiblement entre 5.85° et 5.90°. La grille obtenue en pratique correspond donc à la grille théorique dont les coordonnées des points en élévation ont été approximées en tenant compte des contraintes pratiques. Etant donné que le pas angulaire en élévation est voisin de 6°, les haut-parleurs ont été installés verticalement tous les 6°, comme on peut le voir sur la figure 2.3, et les positions angulaires du bras ont été choisies de sorte à minimiser l'écart angulaire entre les élévations théoriques et les élévations effectives des haut-parleurs. Des contraintes pratiques ne nous permettant pas d'atteindre les élévations proches du pôle inférieur de la sphère (*polar gap*), l'élévation la plus basse mesurée est donc de $-50.5°$. Par ailleurs, à l'ordre 29, la grille gaussienne ne possède pas de points dans le plan horizontal (élévation 0°). L'absence de points à 0° en élévation a été jugée critique car elle impose d'effectuer un travail d'interpolation pour accéder au plan horizontal, naturellement important pour juger de la qualité d'un jeu d'HRTFs. Nous avons décidé d'ajouter cette élévation à la grille de mesure. Etant donné les positions des haut-parleurs, une fois le bras pivotant placé à 0°, cela nous permet de mesurer par la même occasion les élévations $-12°$, $-6°$ et $+6°$. La grille obtenue est présentée en figure 2.2(b) et comporte donc 1680 directions de mesure.

Au total, 54 sujets ont été mesurés dont 42 hommes et 12 femmes ainsi que 3 têtes artificielles (deux têtes artificielles Neumann dont celle de la NASA [ABK15], une tête artificielle Brüel&Kjær avec et sans pavillons). Cette base a fait l'objet d'un post-traitement qui concerne principalement la compensation de la chaîne de mesure (c.f. 2.2.3.2). Elle est également disponible en format SOFA.



(a) (b)

FIGURE 2.2 – (a) Photographie de la chambre anechoïque de l'IRCAM. Une chaise est placée sur une table tournante au centre de la chambre anéchoïque et 4 haut-parleurs sont disposés verticalement sur le bras mécanique permettant d'atteindre les différentes élévations cibles. (b) Illustration de la grille de mesure. Les points noirs correspondent à l'échantillonnage gaussien pour l'ordre 29. Les points bleus en font partie mais représentent les élévations du "vide polaire" qu'il n'est pas possible d'atteindre en pratique. Le points rouges sont les points de mesure rajoutés autour du plan horizontal.

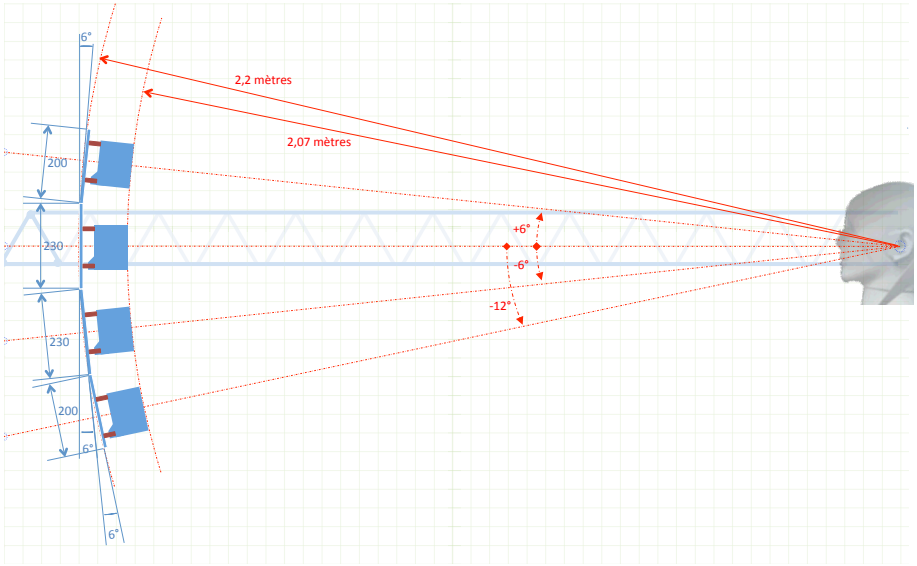


FIGURE 2.3 – Schéma de montage des 4 haut-parleurs sur le bras articulé de la chambre anéchoïque.

2.2.2.3 Mesures complémentaires à Orange Labs

D'autres mesures de HRTFs basées sur la même grille d'échantillonnage sphérique ont été réalisées à Orange Labs (Lannion), partenaire du projet BiLi. Une base de données morphologique a également été créée pour les participants aux mesures de HRTFs (les détails sont donnés dans l'article [ROEK15]).

Le système de mesure d'Orange Labs présente l'avantage de pouvoir acquérir les HRTFs sur un même nombre de points mais en réduisant par 5 la durée de mesure. Cela est rendu possible par l'adoption de la méthode MESM (*Multiple Exponential Sweep Method* [**?**]) en parallèle de l'utilisation de 31 haut-parleurs disposés sur deux arcs. Les haut-parleurs émettent des sinus glissants qui se chevauchent à des temps décalés de 0.1 seconde, soit quasi-simultanément, ce qui permet d'acquérir toutes les élévations aux azimuts matérialisés par les deux arcs en une seule fois. La

rapidité d'acquisition permise par cette technique présente l'avantage de réduire les mouvements du sujet durant la mesure (hypothèse d'invariance temporelle mieux respectée) et surtout de rendre la procédure moins pénible pour le sujet. Plus détails techniques sont donnés dans l'article [FR15].

Les haut-parleurs peuvent être disposés sur les arcs avec davantage de précision que nous le permet notre bras articulé. Ainsi, les élévations mesurées sont plus proches de la grille gaussienne théorique. Elles ne diffèrent cependant des élévations mesurées à l'IRCAM que de 0.35° au maximum. La dernière élévation mesurée sur la calotte inférieure de la sphère est située à $-56.05°$ contre $-50.5°$ à l'IRCAM, soit une élévation supplémentaire au niveau du vide polaire. Notons que des mesures ont été réalisées au-delà mais qu'elles présentent une difficulté d'égalisation à cause des réflexions sur la table tournante et les jambes du sujet. De plus, les mesures additionnelles autour du plan horizontal ont été réduites à la mesure du plan horizontal (une élévation additionnelle seulement, à 0°). Les azimuts de la grille de mesure d'Orange Labs correspondent exactement à ceux de la grille de mesure de l'IRCAM. La grille d'Orange Labs comprend un total de 1560 points de mesure, dont 1500 directions sont communes avec le système de mesure de l'IRCAM.

Les haut-parleurs sont compensés en amont grâce à l'application d'un filtre égalisateur au moment de l'excitation, ce qui permet d'obtenir un meilleur SNR. Plusieurs méthodes d'égalisation ont été testées dont une correspond de très près à celle qui a été réalisée à l'IRCAM. Par conséquent, les deux bases de données peuvent facilement être rassemblées pour créer une base de données de HRTFs plus large et donc plus représentative de la population. L'autre avantage concerne le fait que 10 sujets ont été mesurés à la fois à l'IRCAM et à Orange Labs. Nous utiliserons conjointement les mesures de HRTFs des deux laboratoires au chapitre 5.

2.2.3 Égalisation

L'égalisation du spectre d'amplitude des HRTFs constitue la principale étape de post-traitement des mesures. Elle vise à retirer la composante liée à la chaîne de mesure (microphones, haut-parleur). Nous présentons son principe avant d'en détailler sa mise en pratique pour les HRTFs de la base BiLi-IRCAM. De plus, nous abordons la problématique de l'égalisation de la réponse en fréquence du casque audio (ou HpTF, *Headphone to ear canal Transfer Function*). Celle-ci peut être réalisée de manière couplée ou découplée de l'égalisation des équipements utilisés à la captation. Elle est indispensable pour la synthèse binaurale étant donné que la réponse en fréquence d'un casque possède des détails spectraux dont l'ordre de grandeur est comparable à ceux des HRTFs [KC00]. Négliger l'effet du casque dégraderait l'information spectrale et par conséquent, la localisation auditive de sources virtuelles en termes de discrimination avant-arrière et haut-bas, [WK05, KC05] mais aussi d'externalisation [KC05].

2.2.3.1 Méthodes d'égalisation

Il existe deux catégories d'égalisation : l'égalisation couplée et l'égalisation découplée [Lar01].

Égalisation couplée L'égalisation couplée consiste à déconvoluer la chaîne électroacoustique. Elle est menée en deux étapes : (1) l'égalisation de la réponse en fréquence du haut-parleur, par une mesure de la source en champ libre avec un microphone de référence (à réponse fréquentielle plate) placé à l'endroit de la tête du sujet lorsque celle-ci est absente ; (2) l'égalisation des fonctions de transfert du casque et des microphones par une mesure de casque réalisée directement après les mesures de HRTFs, de manière à s'assurer que la réponse en fréquence et le placement des microphones soient identiques au moment des mesures de HRTFs.

La difficulté réside dans le fait de devoir enchaîner directement sur la mesure de casque après la session de mesures de HRTFs, qui dans certains cas est déjà très longue. De plus, étant donné la mauvaise répétabilité de cette mesure [KC00], il est nécessaire de procéder à plusieurs mesures après replacement du casque sur les oreilles du sujet. Enfin, cette méthode d'égalisation suppose de déterminer en amont le casque audio qui sera utilisé pour la reproduction.

Égalisation découplée L'égalisation découplée consiste à égaliser les fonctions de transfert ainsi que le casque par rapport à un champ sonore de référence, supposé être reproduit de manière stable. Cette méthode offre une grande flexibilité en permettant de traiter indépendamment l'égalisation de la chaîne de mesure et celle de la chaîne de restitution. Le champ sonore de référence peut être le champ libre ou le champ diffus. Le champ libre est constitué d'une onde plane provenant d'une incidence donnée. Le champ diffus est composé d'ondes planes décorrélées provenant d'incidences

distribuées uniformément autour du récepteur (e.g. la réverbération tardive d'une salle). L'égalisation champ libre des HRTFs utilise la HRTF d'une direction donnée, le plus souvent la HRTF frontale. L'égalisation champ diffus des HRTFs repose sur la HRTF en champ diffus, i.e. le filtrage effectué par les éléments morphologiques de l'auditeur (torse, tête, pavillons) sur le champ sonore diffus. Une méthode d'estimation de la HRTF en champ diffus consiste à effectuer une moyenne spatiale des spectres d'amplitude des HRTFs mesurées dans toutes les directions en chambre anéchoïque [Lar01]. Chaque HRTF est alors pondérée par le poids associé à la portion d'angle solide délimitée par la surface sphérique entourant le point de mesure (plus le point est isolé par rapport aux autres, plus sa surface sera grande et plus le poids associé à la mesure sera important). L'estimation du spectre d'amplitude en champ diffus mag_{DF} s'écrit :

$$mag_{DF} = \sqrt{\sum_{i=1}^{M} (mag_i^2 \cdot w_i)} \tag{2.1}$$

avec i l'indice de la direction de mesure ($i = 1 \cdots M$), w_i le poids associé au point de mesure i et mag_i le spectre d'amplitude de la HRTF mesurée à la direction i. Le spectre d'amplitude diffus évalué en pratique n'est qu'une approximation d'un champ diffus idéal, puisque les mesures réalisées ne couvre pas l'ensemble de la sphère (vide polaire).

Les HRTFs sont généralement la combinaison d'une composante dépendante de la direction (*Directional Transfer Functions*, DTF), et d'une composante indépendante de la direction (*Common Transfer Function*, CTF). L'égalisation champ diffus permet d'annuler la composante indépendante de la direction. Les HRTFs ainsi égalisées sont (parfois) désignées par DTFs (*Directional Transfer Functions*).

Les synthèses réalisées par des HRTFs égalisées champ libre ou champ diffus sont compatibles respectivement avec une restitution sur des casques égalisés champ libre ou champ diffus par le fabriquant industriel. Cela permet d'éviter d'importants artefacts de timbre à la reproduction.

Larcher [LJV98, Lar01] préconise l'égalisation champ diffus car elle permet de réduire les différences inter-individuelles en retirant les caractéristiques individuelles indépendantes de la direction. En particulier, on observe une réduction notable des différences inter-individus en dessous de 5 kHz après égalisation champ diffus.

2.2.3.2 Égalisation champ diffus des HRTFs de la base BiLi

Concernant les HRTFs de la base de données acquise ici, nous n'avons pas la possibilité d'appliquer l'égalisation couplée car les mesures de casque n'ont pas été réalisées systématiquement après la session de mesures, pour des raisons pratiques. Les HRTFs de la base de données sont donc égalisées suivant la méthode de l'égalisation découplée. Chaque jeu de HRTFs est compensé par la moyenne spatiale de l'ensemble des mesures et on obtient ainsi une base de données de DTFs. Les principales étapes de post-traitement des HRTFs de la base BiLi-IRCAM sont présentées ici. Elles concernent : (1) l'estimation de filtres d'égalisation relatifs aux haut-parleurs ; (2) le fenêtrage temporel des HRIRs ; (3) l'égalisation champs diffus des HRTFs.

Les mesures ont été effectuées avec 4 haut-parleurs (HP) différents dont les réponses en fréquence peuvent varier. La première étape de traitement consiste à annuler les différences inter haut-parleurs, de sorte à se ramener à un transducteur de référence, ici choisi (arbitrairement) comme le HP n°4. L'annulation des différences inter haut-parleurs peut se faire sur la base d'une mesure en champ libre, ou en champ diffus. Etant donné la configuration pratique des transducteurs (montage "non-symétrique" sur le bras articulé, c.f. figure 2.3), de faibles interactions (acoustiques et/ou vibratoires) du haut-parleur avec son environnement proche sont possibles. Nous avons donc privilégié ici une égalisation par une mesure de référence en champ diffus. La mesure des réponses réalisée par un microphone de référence (Brüel et Kjær) supposé idéal. L'estimation de la réponse en champ diffus de chaque haut-parleur peut être visualisée en figure 2.2.3.2.

Des filtres de compensation sont calculés par le rapport inverse du spectre d'amplitude diffus de chaque HP sur le HP de référence (HP4). Afin d'éviter des pics proéminents dans les filtres inverses, une procédure de régularisation (compression de la dynamique) est mise en oeuvre pour les hautes fréquences (au-delà de 15 kHz). Les filtres de compensation appliqués aux haut-parleurs 1, 2 et 3 sont illustrés dans la figure 2.5. On note que les variations spectrales inter-HPs sont de l'ordre de ± 2 dB. Enfin, on repasse dans le domaine temporel via une reconstruction à phase minimale du spectre complexe et on tronque la réponse impulsionnelle des filtres à 512 échantillons.

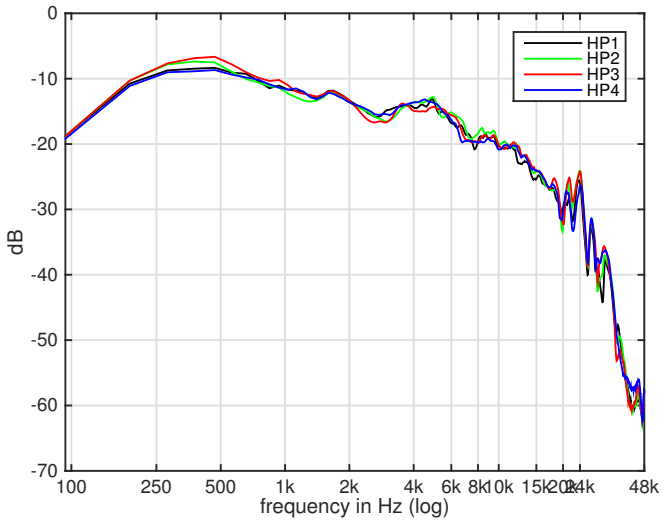


FIGURE 2.4 – Champs diffus estimés pour chacun des 4 haut-parleurs de mesure.

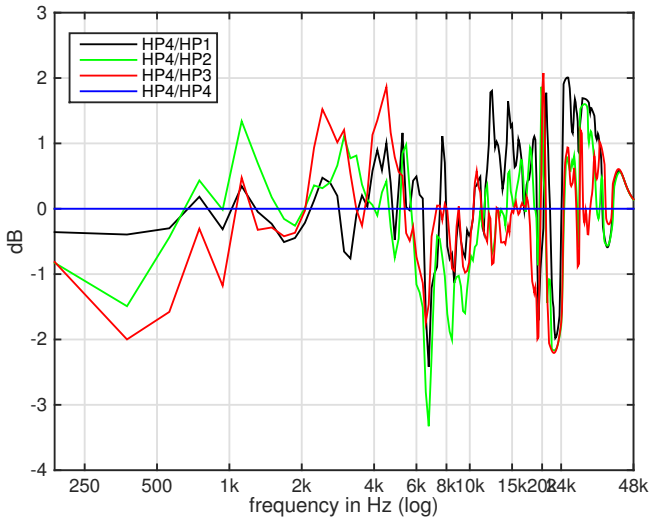


FIGURE 2.5 – Filtres de compensation de chaque haut-parleur par rapport au haut-parleur de référence (HP4).

Les HRIRs mesurées sont fenêtrés temporellement de sorte à ne conserver que la partie utile du signal, soit 2048 échantillons (environ 20 s à 96 kHz). Elles sont ensuite convoluées par la réponse du filtre de compensation (décrit plus haut) relatif au haut-parleur utilisé pour la direction considérée.

La dernière étape concerne l'égalisation champ-diffus des spectres d'amplitude des HRTFs. Par définition, celle-ci est réalisée individuellement pour chaque jeu d'HRTF mesuré. En effet, il s'agit ici de retirer la composante individuelle indépendante de la direction.

Pour procéder au traitement, les HRTFs sont décomposées en un spectre d'amplitude et un spectre de phase. Le champ diffus des HRTFs est estimé selon l'équation 2.1 à partir d'une moyenne énergétique des spectres d'amplitude (dans les $M = 1680$ directions) dont les poids w_i sont proportionnels à la surface de Voronoi S_i associée à chaque point de mesure (soit l'aire sphérique couverte par chaque point de mesure) Enfin, le spectre complexe est obtenu en recomposant spectre de phase et spectre d'amplitude égalisé. Un exemple de HRTF brute et égalisée est donné en figure 2.6 pour la direction frontale ainsi que la réponse en fréquence de la HRTF en champ diffus utilisée pour la compensation. On note que le gain moyen de la HRTF égalisée est proche de zéro, étant donné que le gain global de toutes les HRTFs a été compensé par le processus d'égalisation.

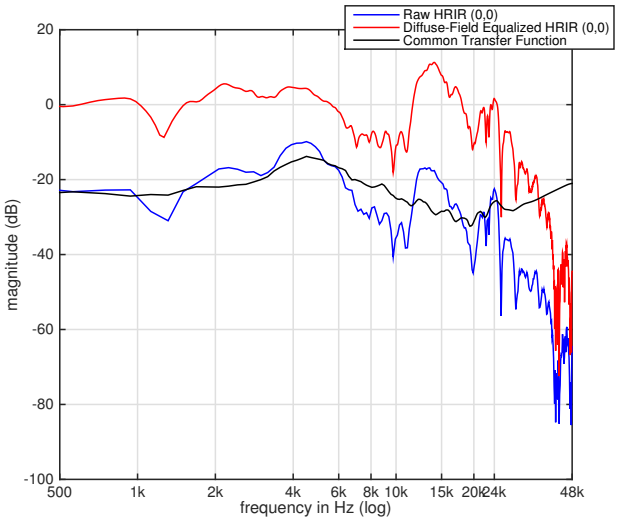


FIGURE 2.6 – Exemple de spectre d'amplitude d'une HRTF gauche mesurée en direction frontale, avant (bleu) et après (rouge) égalisation par le filtre de compensation champ diffus (inverse de la courbe noire).

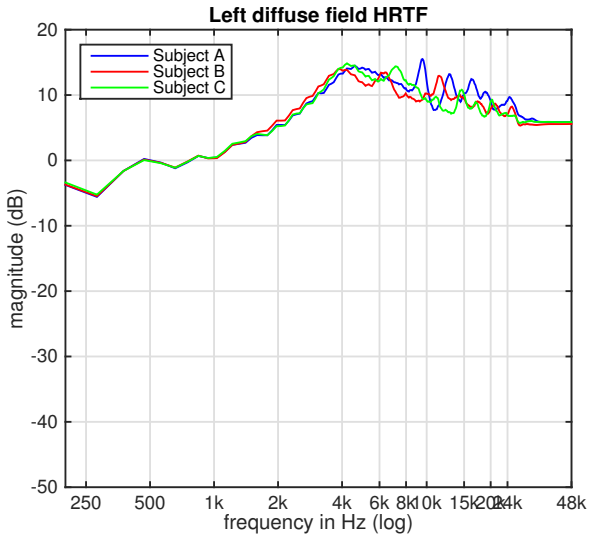
2.2.3.3 Mise en pratique de l'égalisation découplée

Les DTFs, obtenues par l'égalisation champ diffus des HRTFs mesurées, contiennent uniquement la composante dépendante de la direction. Suivant la théorie de la méthode découplée, l'utilisation d'un casque audio égalisé champ diffus permettrait de rendre aux DTFs leur composante indépendante de la direction au moment de la restitution. En effet, un casque égalisé champ diffus par rapport à une tête artificielle de référence, signifie que lorsqu'il émet un champ diffus, le signal reçu aux oreilles de cette tête artificielle a le même spectre que celui reçu lorsque celle-ci est placée dans un champ diffus. En théorie, la HRTF en champ diffus propre à cette tête "standard" conviendrait à restituer la composante indépendante de la direction pour n'importe quel individu [LJV98]. Cela sous-entend que le champ diffus mesuré à l'entrée des canaux auditifs aurait des propriétés similaires à tous les individus. Nous étudions ici le décalage entre la mise en pratique de cette méthode et la théorie.

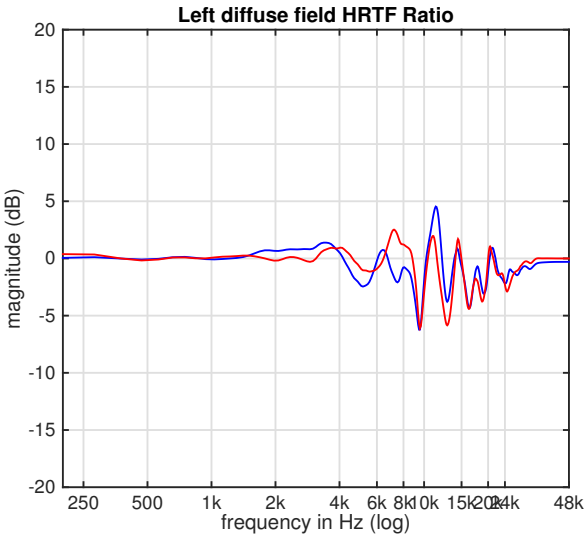
La figure 2.7(a) présente la HRTF en champ diffus estimée à partir des jeux de HRTFs de 3 sujets (A, B et C) de la base de données BiLi, dont l'effet du haut-parleur de référence a été compensé. La seule composante de la chaîne de mesure résiduelle concerne l'effet du microphone, que nous n'avons pas systématiquement mesuré pour chaque individu. Cependant, les capsules miniatures Knowles utilisées possèdent une réponse en fréquence relativement plate, comme l'illustre la figure 2.8. La figure 2.7(b) présente le rapport des estimations de la HRTF en champ diffus des sujets B et C sur la HRTF en champ diffus du sujet A. La variation inter-individuelle est de l'ordre de ± 5 dB. Cela permet de confirmer que le champ diffus mesuré à l'entrée des canaux auditifs de différents individus est relativement similaire.

La figure 2.7(c) présente la réponse en fréquence d'un casque audio mesurée sur les sujets A, B et C. Etant donné la faible répétabilité selon le positionnement du casque sur les oreilles [KC00], les réponses en fréquence pour chacun des individus ont été obtenues à partir de la moyenne de 5 mesures réalisées après replacement du casque sur les oreilles. Les mesures ont été effectuées avec la méthode du conduit auditif bloqué et la réponse en fréquence des capsules microphoniques Knowles utilisées n'a pas été retirée des courbes présentées figure 2.7(c). Le casque utilisé ici est de type Sennheiser HD 650, a été égalisé par rapport au champ diffus par le fabriquant et est circumaural, c'est-à-dire qu'il entoure l'oreille. La figure 2.7(d) nous permet d'identifier que les réponses en fréquence du casque contiennent des variations inter-individuelles de l'ordre de ± 18 dB au-dessus de 6 kHz. Cette observation est en accord avec Pralong et Carlile [PC96]. Ces auteurs ont observé une variabilité inter-sujets importante au-dessus de 6 kHz et dont la déviation standard atteint 17 dB à 9 kHz (étude sur 10 sujets avec un casque circumaural Sennheiser HD 250 égalisé champ diffus). En effet, à partir de 6 kHz, le pavillon d'oreille, dont les détails morphologiques varient considérablement entre les individus, a une influence sur le signal émis par le transducteur du casque, en particulier lorsque le casque est circumaural car il capte les détails morphologiques du pavillon. La présence du pavillon est à l'origine de colorations spectrales au moment de la restitution qui varient selon la morphologie de l'auditeur. Plusieurs études ont mis en évidence la nécessité de réaliser une égalisation du casque de manière individuelle [MHJS95, PC96, WK05].

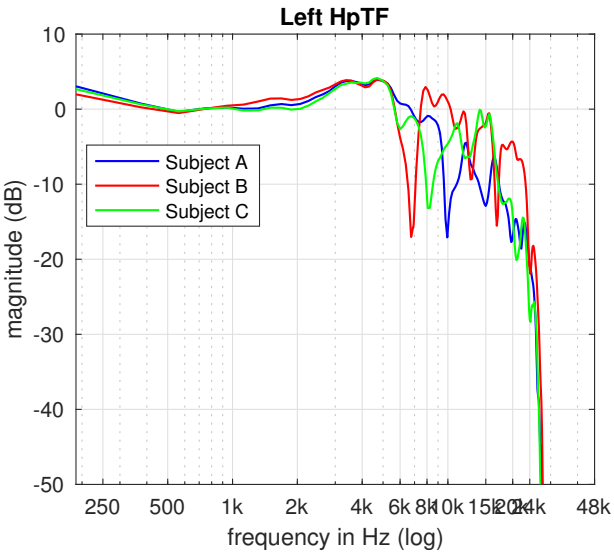
En comparaison à la figure 2.7(b), on note que les variations inter-individuelles sont plus importantes sur la réponse en fréquence du casque que sur la HRTF en champ diffus. La composante champ diffus retransmise au moment de la restitution varie selon l'auditeur. Enfin, la figure 2.9 permet de visualiser la différence entre ce que prévoit la théorie de l'égalisation découplée et ce que l'on obtient en pratique. En théorie, la composante champ diffus des HRTFs (i.e. la composante indépendante de la direction) serait retransmise au moment de la restitution. En pratique, l'hypothèse n'est pas totalement vérifiée étant donné que le casque a une influence qui dépend des détails du pavillon.



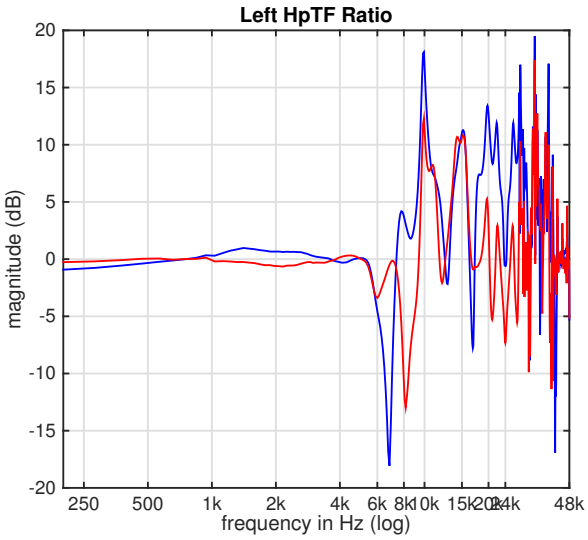
(a) HRTF en champ diffus



(b) Variation inter-individuelle des HRTFs en champ diffus



(c) HpTF



(d) Variation inter-individuelle HpTFs

FIGURE 2.7 – (a) HRTF en champ diffus estimée à partir de 3 jeux de HRTFs issus des individus A, B et C de la base de données BiLi. (b) Rapport en fréquence des HRTFs en champ diffus vis-à-vis d'un sujet de référence. (c) Réponse en fréquence moyenne du casque HD 650 mesurée sur les individus A, B et C. (d) Rapport des réponses en fréquence de casque vis-à-vis d'un sujet de référence.

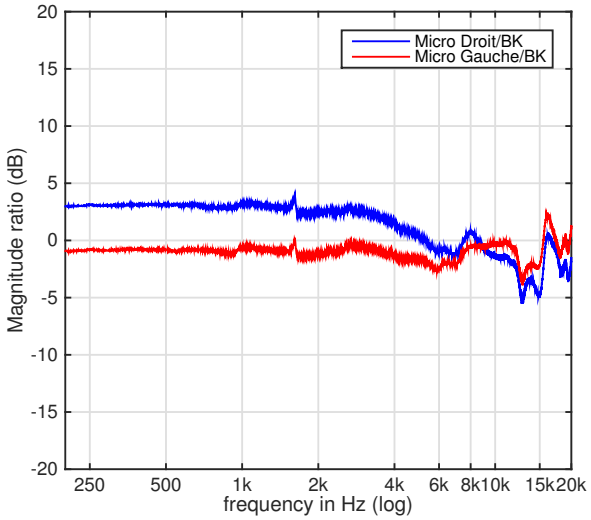


FIGURE 2.8 – Rapport de la réponse en fréquence d'une paire de capsules Knowles sur la réponse en fréquence d'un microphone Brüel et Kjær à réponse fréquentielle plate.

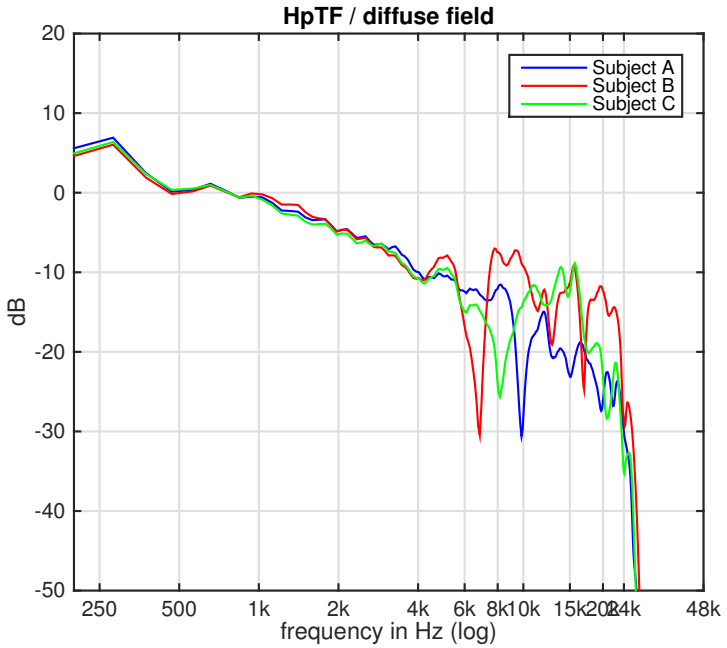


FIGURE 2.9 – Rapport de la réponse en fréquence du casque sur la HRTF en champ diffus pour chacun des 3 sujets.

2.3 Représentation des HRTFs

Les fonctions de directivité de la tête d'un individu peuvent être représentées en fonction du temps (HRIRs), en fonction de la fréquence (HRTFs), ou encore spatialement pour une fréquence donnée.

2.3.1 Visualisation locale d'un jeu de HRTFs

Les HRTFs contiennent à la fois les transformations d'amplitude et de phase subies par l'onde sonore entre la source et le tympan. La représentation fréquentielle se prête à la séparation de ces informations d'amplitude et de phase qui peuvent alors être modélisées et analysées séparément. Dans l'optique de décrire les HRTFs de façon concise, et pour limiter le coût de calcul lié à l'implémentation de filtres, différents modèles relatifs à ces deux composantes existent (en particulier pour la composante spectrale). Basés sur des considérations psychophysiques, ils ont l'avantage de réduire la dimensionnalité des données sans créer d'artefacts audibles. Cette section présente les méthodes de traitement du signal qui permettent de simplifier la représentation des HRTFs en ne gardant que l'information pertinente pour le système auditif.

2.3.1.1 Modélisation composante à phase minimale et retard

Etant donné que les HRTFs peuvent être considérées comme un filtre causal et stable, elles peuvent être décomposées en un filtre à phase minimale et un filtre passe-tout [Lar01] :

$$\begin{aligned} HRTF &= mag \cdot exp(j \cdot \psi) \\ &= mag \cdot \exp(j \cdot mph) \cdot \exp(j \cdot eph) \\ &= H_{min} \cdot H_{exc} \end{aligned} \tag{2.2}$$

La phase originale ψ de la HRTF est décomposée en les phases d'une composante à phase minimale mph et d'une composante d'excès de phase eph. La composante à phase minimale contient les informations spectrales et la composante à excès de phase, les informations temporelles. Dans le cas des HRTFs, l'excès de phase présente généralement un caractère quasi-linéaire jusqu'aux environs de $8-10\,\text{kHz}$ (voir figure 2.10). Par conséquent, il est d'usage de le représenter en première approximation par un retard pur. Une estimation de ce retard consiste, par exemple, à mesurer la pente du spectre d'excès de phase dans cette zone fréquentielle. La composante à phase minimale des HRTFs représente les informations spectrales. Elle peut être implémentée par un filtre RIF (à Réponse Impulsionnelle Finie) ou RII (à Réponse Impulsionnelle Infinie). L'implémentation RII est centrée sur la reproduction des résonances du spectre d'amplitude (pics et creux) et repose sur une modélisation pôles-zéros. Elle permet de modéliser les HRTFs avec moins de coefficients que l'implémentation RIF.

Cette représentation des HRTFs en filtre à phase minimale et retard pur a été validée par des tests perceptifs. Kistler et Wightman [KW92] ont montré que les performances de localisation obtenues avec des HRTFs modélisées étaient identiques à celles obtenues avec des HRTFs mesurées. Kulkarni et Colburn [KIC99] ont montré que l'information haute fréquence de la phase n'était pas utilisée par le système auditif et que, par conséquent, la suppression de ces informations n'entraînait aucun artefact audible. La séparation des informations relatives à la composante à phase minimale et à la composante d'excès de phase peut être visualisée figure 2.10.

2.3.1.2 Estimation du retard interaural

Comme on vient de le voir, le spectre de phase de la composante à excès de phase des HRTFs est généralement approximé par un retard pur. La différence des retards monauraux à l'oreille gauche et droite constitue alors l'indice d'ITD, prépondérant en basses fréquences. Il existe plusieurs méthodes d'estimation de l'ITD à partir des mesures :

— la différence de temps d'arrivée de l'onde à gauche et à droite (méthode de détection de seuil, appelée "Threshold"). Les temps d'arrivée sont définis par le point de la réponse impulsionnelle atteignant un certain pourcentage de la valeur maximale de l'amplitude de la réponse (le choix du seuil n'est pas imposé).

— la différence entre les retards purs à gauche et à droite (référencée par la suite par "DELMOD"). Basée sur l'hypothèse de linéarité de phase, l'estimation du retard pur correspond à la pente de la droite de régression linéaire sur le spectre de phase de la composante à excès de phase dans un certain intervalle fréquentiel (généralement basses fréquences).

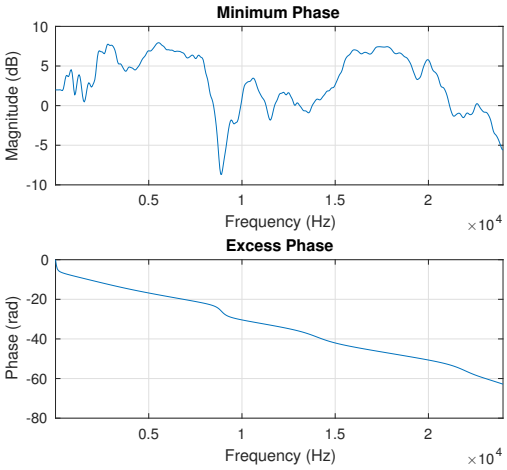


FIGURE 2.10 – (1) Spectre d'amplitude en dB d'une HRTF mesurée à (θ, ϕ)=(54°, 0°) (2) Excès de phase, approximativement linéaire jusqu'à 8.5kHz.

— le maximum de l'inter-corrélation des réponses impulsionnelles droite et gauche. L'estimation peut alors être réalisée sur les réponses impulsionnelles brutes ("MaxIACCr") ou sur l'enveloppe des réponses impulsionnelles ("MaxIACC").

— le centroïde de l'inter-corrélation des réponses impulsionnelles droite et gauche calculé sur les HRIRs brutes ("CenIACCr") ou leurs enveloppes ("CenIACC").

Les méthodes "MaxIACCr", "MaxIACC" et "Threshold" souffrent cependant d'un manque de robustesse dû au mauvais rapport signal-à-bruit des réponses impulsionnelles contralatérales. Plus de détails sur les mérites de chacune de ces méthodes peuvent être trouvés dans l'article [KN14].

2.3.1.3 Modélisation du spectre d'amplitude

Le spectre d'amplitude des HRTFs est défini pour toutes les fréquences, à une direction donnée :

$$mag\left(f, \theta_i, \phi_i\right) = \left|H\left(f, \theta_i, \phi_i\right)\right| \tag{2.3}$$

Afin de limiter le coût de l'implémentation des HRTFs, plusieurs méthodes visent à réduire la complexité des spectres d'amplitude en ne gardant que l'information pertinente pour le système auditif.

Premièrement, le spectre d'amplitude peut être lissé spectralement en appliquant un banc de filtres tenant compte de la résolution fréquentielle limitée du système auditif, comme présenté section 1.2.4. Le spectre lissé résultant $\tilde{H}$ est défini par le spectre d'amplitude moyenné par bande de fréquence b :

$$\tilde{H}(b, \theta, \phi) = 20 \cdot \log_{10} \sqrt{\frac{\sum_{k \in f_b} \left[H(f_k, \theta, \phi)\right]^2}{N_b}} \tag{2.4}$$

où $f_b = f_{l_b} \cdots f_{h_b}$ est le vecteur des N_b fréquences considérées dans la bande fréquentielle b de fréquence minimale f_{l_b} et maximale f_{h_b}. Si la représentation est basée sur un modèle perceptif, alors elle ne crée pas d'artefacts audibles. Cette représentation permet de réduire le nombre d'échantillons fréquentiels à B coefficients, où B est le nombre de bandes du banc de filtres (très variable mais généralement $B < 50$).

Deuxièmement, les HRTFs peuvent être approximées par un modèle RII à ordre faible. Par exemple, Blommer et Wakefield [BW97] ont montré qu'un modèle pôle-zéro d'ordre 60 est équivalent à un filtre RIF d'ordre 2048.

Troisièmement, les HRTFs peuvent être décomposées sur des fonctions cosinusoïdales (transformée en cosinus discret, DCT). Cette méthode a notamment été utilisée par Kulkarni et Colburn [KC98] dans une étude sur la discrimination de sources réelles et virtuelles, synthétisées avec des HRTFs lissées par troncature des coefficients de la décomposition DCT. Les auteurs montrent que la discrimination des sources virtuelles apparait lorsque seuls 8 coefficients sont utilisés. Une reconstruction à partir de 16 coefficients seulement ne génère pas d'artefacts audibles.

2.3.2 Visualisation spatiale

De manière complémentaire à la représentation locale de HRTFs, on peut s'intéresser à la distribution spatiale des spectres d'amplitude pour une fréquence donnée f_k, i.e. sous forme de *Spatial Frequency Response Surfaces* (SFRS) [Gui09] :

$$\text{SFRS}(f_k) = \{DTF\,(f_k, \theta, \phi)\} \tag{2.5}$$

Les SFRS, définies sur la sphère, peuvent être décomposées sur la base des harmoniques sphériques de sorte à être plus facilement manipulées (interpolation spatiale, rotation, etc.) et visualisées, notamment grâce à une représentation spatiale continue.

Décomposition en harmonique sphériques Comme toute fonction F définie sur une sphère de rayon r_0 et de carré intégrable, les SFRS peuvent être décomposées sur une base d'harmoniques sphériques $Y_n^m(\theta, \phi)$ s'écrit [PNW+12] :

$$F(r_0, \theta, \phi, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k, r_0) Y_n^m(\theta, \phi) \tag{2.6}$$

où

$$k = \frac{2\pi f}{c} \tag{2.7}$$

est le nombre d'onde, f la fréquence et c la vitesse du son.

Les coefficients a_{nm} sont les coefficients associés aux harmoniques sphériques et sont obtenus par la transformée de Fourier spatiale :

$$a_{nm}(k, r_0) = \int_{\Omega \in S^2} F(r_0, \theta, \phi, k) Y_n^m(\theta, \phi)^* d\Omega \tag{2.8}$$

En pratique, les HRTFs sont mesurées sur un nombre réduit de points sur la sphère $(\theta_i, \phi_i), i = 1 \cdots M$, les SFRS sont donc définies sur un échantillonnage discret. L'échantillonnage spatial limite l'ordre des harmoniques sphériques représentables à un ordre fini, de la même façon que l'échantillonnage temporel limite la fréquence maximale représentable. Cela implique que le spectre spatial a_{nm} est tronqué à un ordre maximal N. De plus, l'échantillonnage régulier de la sphère n'étant pas possible, il convient de pondérer chaque point d'échantillonnage par un facteur w_i qui dépend du type de pondération (par exemple, la pondération de Voronoi qui consiste à pondérer chaque point par la surface sphérique entourant le point). Ainsi, l'échantillonnage de la sphère consiste à approximer l'intégrale 2.8 de la façon suivante :

$$a_{nm} = \sum_i w_i \cdot SFRS(\theta_i, \phi_i) \cdot Y_n^m(\theta_i, \phi_i)^* \tag{2.9}$$

La discrétisation de la surface de la sphère engendre donc la limitation de l'ordre de la décomposition du champ sonore à l'ordre N. Cet ordre fini a pour effet de limiter la fréquence représentable à $kr \leq N$ à cause de l'aliasing spatial [Raf05]. La décomposition n'est donc possible que jusqu'à la fréquence de coupure $f_c = \frac{Nc}{2\pi r}$.

De plus, le manque de mesures sur la calotte inférieure de la sphère, dû à des contraintes pratiques, et les éventuelles perturbations présentes dans les données mesurées, posent certains problèmes au niveau de la décomposition en harmoniques sphériques. Des solutions de régularisation ont alors été proposées pour répondre à ce problème, comme la régularisation de Tikhonov [PNW+12] ou la méthode d'Ahrens [**?**]. Cette méthode consiste à mener tout d'abord une décomposition à un ordre faible, peu sensible à l'absence de données dans une portion de l'espace, à compléter cette calotte inférieure avec ces données estimées (sur une grille complète) et enfin, à recommencer la décomposition à un ordre plus élevé.

Limiter l'ordre de décomposition sur les harmoniques sphériques permet d'obtenir un lissage spatial. De plus, la décomposition en harmoniques sphériques peut être utilisée pour l'interpolation des HRTFs. Premièrement, l'interpolation des données entre les points de mesure est particulièrement nécessaire pour assurer la continuité du rendu spatial en synthèse binaurale dynamique. Deuxièmement, parallèlement aux HRTFs mesurées, il existe des données de repérage spatial de la tête du sujet pendant les mesures qui offrent la possibilité de récupérer les directions effectivement mesurées. La grille d'échantillonnage obtenue par cette correction risque cependant d'être déformée de manière variable selon les sujets. Cependant, un travail d'interpolation permettrait de ré-interpoler les données sur la grille de mesure commune.

2.3.3 Décomposition des HRTFs sur une base

La décomposition en harmoniques sphériques, présentée dans la section précédente, est un cas particulier de la décomposition des HRTFs en une combinaison linéaire de fonctions spectrales et spatiales, qui peut s'exprimer ainsi [Lar01] :

$$HRTF(\theta, \phi, f) = \sum_{i=1}^{N} C_i(\theta, \phi) L_i(f) \tag{2.10}$$

Afin de décrire les HRTFs sous cette forme, il faut tout d'abord définir une base, qui peut être imposée ou déterminée statistiquement et formée de fonctions spectrales $L_i(f)$ ou spatiales $C_i(\theta, \phi)$. Puis, les coefficients de la décomposition des HRTFs sur cette base doivent être déterminés. Ils transmettent alors l'information individuelle.

Les harmoniques sphériques définissent une base des fonctions spatiales $C_i(\theta, \phi)$ génériques et communes à tout individu et les coefficients de la décomposition (pondérations en fonction de la fréquence) contiennent une information individuelle. Les fonctions de base sur lesquelles sont décomposées les HRTFs peuvent aussi être définies par des fonctions spectrales $L_i(f)$. L'analyse en composantes principales (ACP) permet par exemple d'obtenir une base de fonctions spectrales déterminées statistiquement [KW92]. Si l'analyse en composantes principales a été menée sur une base de données de HRTFs comprenant un grand nombre de sujets, alors cette base de fonctions spectrales revêt un caractère générique tandis que les fonctions spatiales sont individuelles.

Contrairement à la décomposition en harmoniques sphériques, l'ACP ordonne les fonctions de base par ordre décroissant d'importance et fournit ainsi une représentation plus compacte des données. Ainsi, en limitant l'ordre de la décomposition à un nombre réduit de fonctions de base, i.e. en éliminant les axes où la variance est réduite, on obtient une représentation optimisée des HRTFs. De la même façon, les fonctions spatiales pourraient être obtenues statistiquement. Elles seraient alors ordonnées par ordre d'importance et pourraient offrir une représentation plus compacte des données. A l'inverse, on pourrait chercher à imposer les fonctions spectrales (e.g. filtres passe-bande).

2.4 Réduction de dimensionnalité

Les méthodes de réduction de dimension sont des méthodes d'apprentissage non supervisées qui permettent de représenter de manière plus compacte des données complexes à haute dimension. Le problème peut être formulé ainsi : considérons une base de données de N vecteurs $\vec{X_i}$, de dimension D, l'objectif est de déterminer les N vecteurs $\vec{Y_i}$ de dimension d, avec $d << D$, représentant les données dans l'espace à dimension réduite. Ces méthodes permettent, dans un premier temps, de réduire l'information et d'obtenir des données plus appropriées à la classification, la régression linéaire, l'interpolation, etc. Dans un deuxième temps, elles permettent d'éliminer les caractéristiques redondantes et d'extraire un sous-ensemble de caractéristiques pertinentes expliquant la variabilité des données : elles offrent ainsi une meilleure visualisation et interprétation des données.

Prenons par exemple le cas d'un ensemble d'images représentant un visage dont la pose et l'expression varient [RS00]. Même si cette variabilité est facile à observer, les données sont à haute dimension étant donné le nombre de pixels. Les composantes de l'espace de représentation révélé par un algorithme de réduction de dimension, tel que l'algorithme LLE, sont corrélées aux caractéristiques non-linéaires et intrinsèques de l'image, à savoir la pose et l'expression du visage.

Les bases de données de HRTFs constituent elles aussi un cas particulier de données à haute dimension. En effet, elles possèdent une dimension $S \times M \times N$ où S est le nombre de sujets, M est le nombre de directions mesurées pour chaque sujet et N le nombre de points fréquentiels. En appliquant les techniques de réduction de données à une base de données de HRTFs, on peut espérer révéler des caractéristiques correspondant à l'information spatiale (i.e. les directions de mesure) ou à la variabilité inter-sujets et mettre ainsi en évidence des groupes d'individus. La classification d'individus est en effet un problème bien connu des laboratoires de recherche sur l'individualisation des HRTFs. Nous verrons en effet au chapitre 3 comment ils peuvent aider la sélection d'un jeu de HRTFs non-individuel dans une base de données pour un nouvel individu n'ayant pas été mesuré.

2.4.1 Méthodes linéaires et non-linéaires, état de l'art

Il existe des algorithmes linéaires et non-linéaires de réduction de dimension. Parmi les méthodes linéaires, on peut mentionner la méthode du "MultiDimensionnal Scaling" (MDS) ou l'ACP, qui

est la plus utilisée. Cette dernière consiste à déterminer un espace vectoriel linéaire dont les axes sont orthogonaux et maximisent la variance présente dans les données. De façon similaire, l'analyse ICA (Independent Component Analysis) permet d'obtenir un espace de représentation linéaire mais dont les composantes principales sont indépendantes entre elles.

Dans le cas où les données sont bruitées ou bien présentent une distribution non-linéaire et/ou des cas isolés (*outliers*), les méthodes non-linéaires sont mieux adaptées [KMKK08, RDSD10]. Parmi ces méthodes, on peut mentionner les algorithmes Isomap (Isometric mapping) [TdSL00, GMGF14], LLE (Locally Linear Embedding) [RS00, DR05, SR03] ou Laplacian Eigenmaps (carte propres Laplaciennes) [BN03, Ber14].

Nous présentons ici brièvement le principe des méthodes non-linéaires principalement utilisées dans la littérature. Elle se distinguent en deux groupes : les méthodes globales (Isomap), ou locales (LLE et Laplacian eigenmaps).

— Isomap est une généralisation non-linéaire de l'analyse MDS. Elle cherche à préserver la structure globale de l'espace à haute dimension à travers la conservation des distances géodésiques entre toutes les paires de points. La première étape consiste à identifier les plus proches voisins de chaque point et à construire un graphe G (de distances euclidiennes) où chaque point est lié à ses plus proches voisins par des arrêtes pondérées. Puis, elle crée une matrice de distances géodésiques entre toutes les paires de points en calculant le chemin le plus court entre chaque paire de points sur le graphe G. Enfin, elle construit un espace à dimension réduite en utilisant la méthode MDS afin de préserver ces distances.

— Roweis et Saul [RS00] a montré l'avantage de la méthode LLE par rapport à la méthode Isomap. Elle se distingue par le fait qu'elle apprend une structure globale qui préserve uniquement la géométrie locale. La méthode caractérise la géométrie de chaque voisinage par les coefficients linéaires qui reconstruisent chaque point x_i à partir de ses plus proches voisins x_j, sous l'hypothèse que chaque point et son voisinage peuvent être modélisés comme localement linéaires (métrique localement euclidienne). Les coefficients de reconstruction optimaux W_{ij} sont trouvés par la méthode des moindres carrés (minimisation de l'erreur de reconstruction) sous les contraintes suivantes : la contrainte de dispersion, qui consiste à dire que la pondération associée à un point qui n'est pas voisin est égal à zéro ; et la contrainte d'invariance, qui impose que la somme des poids soit égale à 1. De plus, ils répondent à des règles de symétrie : ces pondérations sont invariantes par rotation, translation et changement d'échelle [SR03]. La dernière étape de l'algorithme consiste à construire l'espace à dimension réduite d, uniquement à partir des poids W_{ij} : chaque vecteur $\vec{Y_i}$ est reconstruit à partir de la combinaison linéaire de ses K voisins, utilisant les mêmes poids que ceux calculés pour le vecteur $\vec{X_i}$ correspondant, en minimisant l'erreur de reconstruction.

— La méthode "Laplacian Eigenmaps" préserve également le voisinage local. Le graphe G connectant les points voisins entre eux par des arrêtes pondérées est utilisé pour construire l'espace à dimension réduite. La méthode utilise la décomposition en valeurs propres de la matrice laplacienne normalisée associée au graphe de voisinage G pour déterminer $\vec{Y_i}$. Voir [BN03] pour plus de détails.

2.4.2 Applications aux HRTFs, état de l'art

Kapralos et al. [KMKK08] ont utilisé les méthodes de réduction pour palier le coût du filtrage des sources par les HRTFs à haute dimension lors de la synthèse binaurale. Sur la base d'une évaluation des performances de localisation obtenues avec des HRTFs réduites, ils montrent l'avantage des méthodes LLE et Isomap par rapport à l'ACP.

Les méthodes LLE et Isomap se sont aussi montrées efficaces pour représenter les informations spatiales des HRTFs, à savoir les directions de mesures. D'après [DR05], la caractéristique la plus pertinente apprise par LLE sur une base de données de HRTFs mesurées sur le plan médian est l'information d'élévation, ce qui paraît cohérent avec l'information perceptive. Ainsi, ils montrent que l'algorithme de réduction permet d'obtenir une cartographie des HRTFs dont la métrique est pertinente perceptivement : les distances entre HRTFs dans l'espace de représentation sont proportionnelles à la distance entre les directions de mesure. De plus, ils proposent une méthode d'interpolation basée sur la représentation des HRTFs dans l'espace à dimension réduite qui permet de reconstruire une HRTF à haute dimension dans n'importe quelle direction de l'espace.

De la même façon, [GMGF14] montrent que l'espace réduit, estimé par la méthode Isomap sur une base de données contenant des HRTFs uniquement dans le plan horizontal, représentent la variabilité des HRTFs à travers l'azimut de la source. Dans cette étude, la réduction des données constitue un pré-traitement avant l'étape de régression entre paramètres anthropométriques et

HRTFs réduites, le but étant de prédire les HRTFs d'un nouvel individu à partir de ses paramètres anthropométriques.

Aytekin et al. [AMS08] ont appliqué la méthode LTSA sur des HRTFs humaines et de chauve-souris mesurées à différentes directions. Ils obtiennent eux aussi un espace de représentation dont la topologie respecte l'espace auditif i.e. où les HRTFs mesurées à des directions proches dans l'espace auditif sont proches dans l'espace à dimension réduite.

2.4.3 Réduction des données de la base BiLi

A titre d'exemple, nous avons appliqué l'analyse en composantes principales ainsi que les méthodes Isomap et Laplacian Eigenmaps sur la base de données BiLi pour les directions dans le plan horizontal.

Les données HRTFs de la base BiLi ont préalablement été égalisées en champ diffus de sorte à éliminer les composantes spectrales indépendantes de la direction, notamment liées au système de mesure (microphones, haut-parleurs, etc.). Les spectres d'amplitude des HRTFs sont ensuite lissés par filtrage ERB. Le banc de filtres ERB (semblable à celui présenté figure 1.5) est calculé entre $f = 150\,\text{Hz}$ et $f = 18000\,\text{Hz}$, conduisant à 36 canaux fréquentiels. L'implémentation des différentes méthodes de réduction est réalisée en utilisant la *toolbox* Matlab fournie par Van der Maaten et al. [vdMPvdH08].

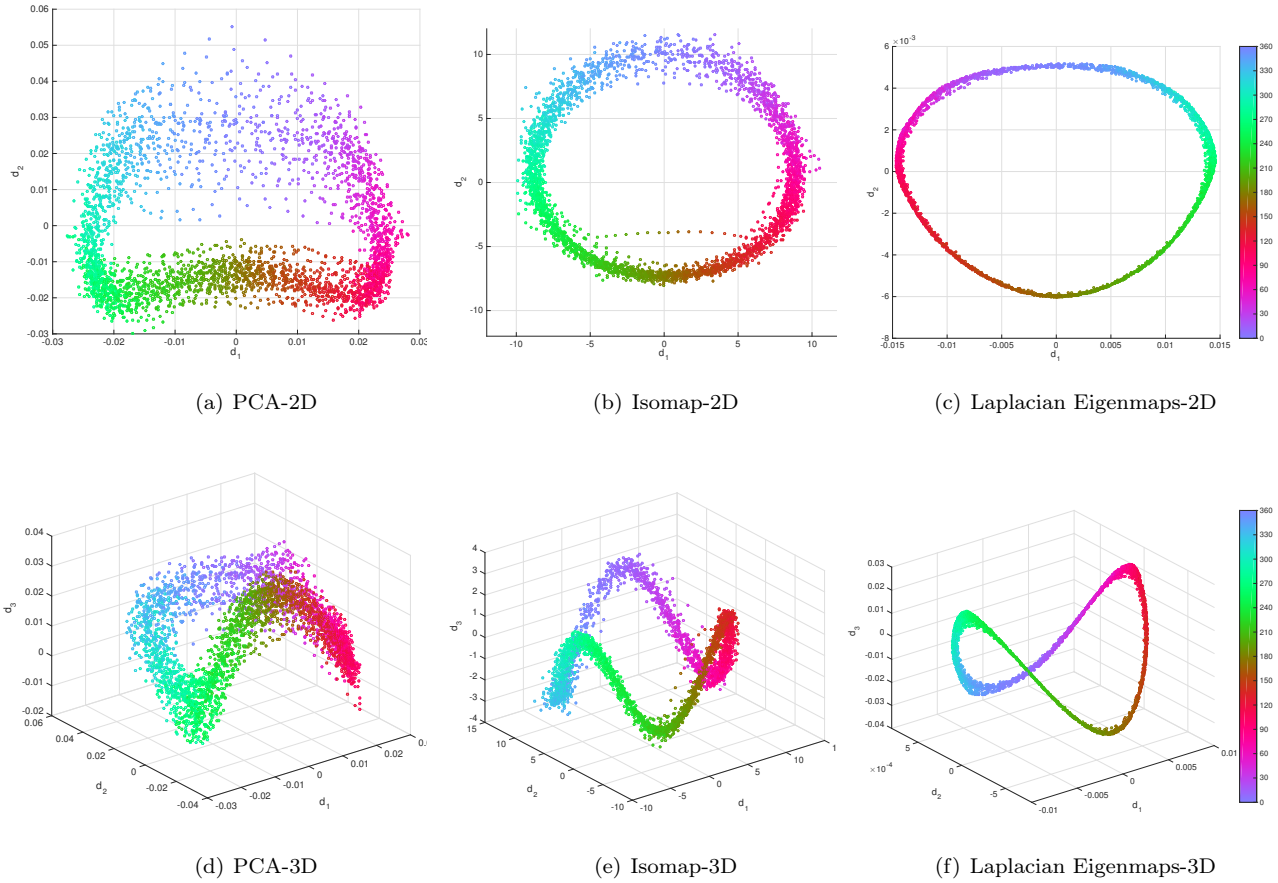
Les résultats de la réduction en 2 dimensions sur la concaténation des HRTFs gauche et droite (représentation par 36×2 canaux fréquentiels) sont donnés en figure 2.11(a), 2.11(b) et 2.11(c) et ceux de la réduction en 3 dimensions sont donnés en figure 2.11(d), 2.11(e) et 2.11(f). Le code couleur associé représente les valeurs d'azimut des points de mesure. Chaque azimut est représenté par 54 points de la même couleur, un pour chaque sujet.

On observe sur ces figures que les algorithmes non-linéaires produisent une distribution 2D des points de forme quasi-circulaire alors que l'algorithme linéaire produit une distribution plus dispersée. De plus, on observe que, dans le cas de l'ACP, les points autour de l'azimut 0° se rapprochent des points à 180° ce qui montre que l'algorithme linéaire est très affecté par la ressemblance avant-arrière. Au contraire, les autres méthodes discriminent mieux les points à l'avant des points à l'arrière. La principale différence de l'ACP réside dans la volonté de préserver les distances globales entre tous les points tandis que les méthodes Laplacian Eigenmaps et Isomap tendent à respecter surtout les relations de voisinage. De cette façon, les méthodes non-linéaires discriminent avec une meilleure efficacité les directions physiquement distinctes (e.g. avant-arrière), bien que celles-ci puissent présenter des similitudes spectrales. Ainsi, elles conduisent à un espace de représentation plus robuste aux confusions avant-arrière et aux différences inter-sujets, qui met en évidence la variabilité des HRTFs à travers les directions de mesure du plan horizontal.

Notons par ailleurs que la "ceinture" de points obtenue en figure 2.11(c) est plus fine que celle obtenue en figure 2.11(b). Cela peut s'expliquer par le fait qu'Isomap est une méthode non-linéaire globale : bien qu'elle repose sur l'hypothèse selon laquelle seules les distances entre paires de points voisins sont connues, elle préserve tout de même les distances géodésiques entre tous les points de l'espace. Au contraire, Laplacian Eigenmaps ne préserve que la géométrie locale.

On remarque également sur la figure 2.11(b) qu'un ensemble de points se détachent de la distribution circulaire. Cet ensemble de points correspond aux données d'un sujet en particulier. La distribution obtenue avec Laplacian Eigenmaps ne présente pas ce cas isolé. Représenter les données de façon extrêmement dense pourrait alors faire l'objet d'une perte d'information si nous nous intéressons aux différences inter-individuelles.

Nous avons également réalisé cette étude sur la totalité des données, i.e. en ajoutant les données mesurées en élévation. Les premières observations obtenues à ce sujet, non présentées ici, valident l'idée que les méthodes non-linéaires permettent une meilleure discrimination des données selon la direction de la mesure. Par exemple, les points de l'espace correspondant à des HRTFs mesurées pour des élévations négatives se distinguent bien de ceux correspondant à des élévations positives lorsque la méthode Laplacian Eigenmaps est utilisée. Au contraire, l'ACP donne un nuage de point compact.



(a) PCA-2D (b) Isomap-2D (c) Laplacian Eigenmaps-2D

(d) PCA-3D (e) Isomap-3D (f) Laplacian Eigenmaps-3D

FIGURE 2.11 – Réduction à $d = 2$ (a, b, c) et $d = 3$ (d, e, f) de la base de données de HRTFs BiLi pour les données dans le plan horizontal uniquement avec les méthodes ACP (a et d) ; Isomap (b et e) ; et Laplacian Eigenmaps (c et f). L'échelle de couleur réfère à l'angle d'azimut de la HRTF.

Conclusion

Ce chapitre a permis de répertorier l'ensemble des éléments de captation, de traitement, d'analyse et de modélisation des HRTFs ainsi que les attributs perceptifs et les facteurs en lien avec l'optimisation du rendu d'un contenu sonore spatialisé en binaural. Bien que les dimensions perceptives sous-jacentes à la qualité d'expérience soient nombreuses, la fidélité du rendu de la position spatiale des sources sonores dans l'espace sonore virtuel reste primordiale pour la plupart des applications ayant recours à la technique binaurale. C'est pour cela que les tests de localisation occupent une place particulièrement importante dans l'évaluation de la synthèse binaurale.

L'individualisation des HRTFs est un facteur important pour la qualité du rendu spatial, en particulier en condition d'écoute statique. En effet, l'utilisation de HRTFs non-individuelles à l'auditeur se traduit par une augmentation des confusions avant-arrière, des erreurs de localisation en élévation mais aussi latérales et une baisse de l'externalisation des sources. Cette dernière caractéristique est également déterminée par l'apport de la synthèse dynamique (suivi de la tête en temps réel) et la simulation d'un effet de salle, qui jouent un rôle sur le réalisme d'une reproduction binaurale. Etant donné la complexité du processus de mesure acoustique des HRTFs d'un individu, l'individualisation des HRTFs pour n'importe quel individu se présente donc comme une problématique majeure. De nombreux travaux ont été menés pour tenter de développer des alternatives à la mesure acoustique. Elles se basent sur des mesures morphologiques ou des tests d'écoute, qui permettent notamment de guider la sélection d'une HRTF non-individuelle dans une large base de données. Le chapitre qui suit expose les différentes méthodes d'individualisation et les tentatives de classification des individus d'une base de données afin de réduire la complexité de la tâche de sélection d'une HRTF non-individuelle pour un nouvel individu.

Enfin, nous avons vu que les méthodes de réduction de dimension permettaient davantage de mettre en évidence la dimension spatiale des HRTFs plutôt que leur caractère inter-individuel. Ce résultat prouve qu'il existe une forte continuité locale dans l'information spatiale contenue dans les fonctions de transfert. Cette caractéristique est essentielle pour autoriser le processus d'apprentissage sensori-moteur de la localisation auditive. Cependant, dans une situation n'invoquant pas la mobilité, on peut se demander comment procède le système auditif pour extraire l'information spatiale à partir d'indices acoustiques statiques et étrangers à la cartographie qu'il s'est construite au fil de l'expérience auditive. La suite de cette thèse s'intéressera à l'identification d'une métrique de comparaison des indices acoustiques permettant d'expliquer les directions perçues par le système auditif dans de telles conditions.

Ce chapitre a exposé l'ensemble des procédés liés à l'obtention et à la modélisation des filtres utilisés en synthèse binaurale. Le caractère individuel de ces filtres et la complexité de leur acquisition acoustique freinent l'exploitation à grande échelle de cette technique de spatialisation sonore. Le chapitre suivant fournit une vue d'ensemble des méthodes d'individualisation développées dans l'objectif de s'affranchir de l'acquisition acoustique individuelle. Nous verrons comment la base de données peut être exploitée à des fins d'individualisation.

Chapitre 3

Individualisation des HRTFs

En pratique, les filtres individuels ne sont pas disponibles pour tout individu souhaitant accéder à des contenus sonores binauralisés. Ce chapitre présente les méthodes d'acquisition des HRTFs alternatives à la mesure acoustique en environnement anéchoïque. Certaines tirent avantage de la corrélation entre les indices acoustiques et certaines caractéristiques morphologiques, d'autres reposent sur des jugements subjectifs. Des méthodes de classification permettent de faciliter les tâches de sélection d'un jeu de HRTFs non-individuel dans une base de données en mettant en évidence des directions ou individus représentatifs.

Par son principe d'analyse-synthèse par échantillonnage, la technique binaurale offre *a priori* toutes les garanties d'une reproduction avec un haut degré de réalisme. En échantillonnant la fonction de directivité de la tête, il est possible de rapporter aux oreilles de l'auditeur l'ensemble des indices responsables de la perception auditive spatiale. Cette capacité à restituer simplement l'ensemble du champ auditif (sphère auditive complète), est primordiale pour la qualité d'immersion [**?**]. La qualité de reproduction spatiale, i.e. permettant de se situer par rapport aux objets ou événements sonores environnant, est également déterminante pour favoriser l'impression de présence dans le monde virtuel. Ce principe direct, ainsi que la simplicité du dispositif de restitution nécessaire (un casque audio standard), ont fait de la synthèse binaurale l'approche privilégiée pour traiter la composante sonore d'un environnement de réalité virtuelle.

La démocratisation des téléphones mobiles et tablettes a considérablement accru l'usage du casque qui devient le véhicule principal pour l'accès au contenus audio. Cette évolution et la simplicité du dispositif d'écoute permettent aussi de proposer au grand public des expériences immersives. Cependant, contrairement aux casques de vision 3D qui ne nécessitent qu'un simple paramètre d'ajustement par le réglage d'écartement des yeux, l'adaptation individuelle de la synthèse binaurale ne se résume pas au simple ajustement d'un "écart" interaural. Comme on l'a vu précédemment, la fonction de directivité de la tête, échantillonnée par les HRTFs, est une signature propre de l'individu. Par principe, il est donc important d'utiliser les HRTFs caractérisées individuellement pour garantir à l'auditeur une reproduction fidèle des indices de localisation qui lui sont propres.

Bien que des paramètres comme le suivi de la tête ou la synthèse de la réverbération rentrent en jeu dans la qualité globale de réalisme (c.f. paragraphe 2.1), ne pas respecter ces caractéristiques individuelles c'est prendre le risque d'introduire des incongruences ou différentes distorsions spatiales (positions absolues ou relative des objets sonores non conformes, inversions avant-arrière, perception intra-crânienne).

Parmi les méthodes d'acquisition des HRTFs d'un individu, la mesure acoustique en environnement anéchoïque ou le calcul numérique à partir de mesures morphologiques par scan sont les méthodes les plus précises. Cependant, ces procédures sont complexes et ne peuvent être envisagées pour une application grand public de la synthèse binaurale.

Une première alternative consiste alors à utiliser des HRTFs "universelles" provenant par exemple d'un modèle générique de tête. La plupart des bases de données de HRTFs comprennent celles mesurées sur une ou plusieurs têtes artificielles. Par ailleurs, certains auteurs ont mis en avant l'existence de jeux de HRTFs convenant à une majorité d'individus. Ces résultats sont issus des tests d'écoute dans lesquels les performances de localisation [WAKW93, MJHS96] ou les jugements qualitatifs [KP12] ont été collectés sur un ensemble de jeux de HRTFs. Cependant, ces études se basent sur des jugements établis sur des populations faibles. Il est également courant d'observer des sujets dont les jugements se démarquent notablement de la tendance globale. Au vu de la forte

variabilité inter-individuelle des indices spectraux, il semble illusoire de penser pouvoir obtenir un rendu spatial adéquat pour tous les individus à partir d'un unique jeu de HRTFs.

Une seconde alternative, consiste à exploiter diverses informations, objectives ou subjectives, relevées ou fournies par l'individu, de sorte à lui proposer un rendu spatial optimal, ou du moins amélioré par rapport à l'utilisation d'une HRTF générique ou non-individuelle aléatoirement sélectionnée. Ces méthodes reposent sur l'utilisation des bases de données de HRTFs existantes mesurées en conditions contrôlées, supposées être représentatives d'une large population d'individus. En partant de l'hypothèse qu'il existe des traits de similitude entre individus, il semble pertinent de rechercher et sélectionner au sein d'une base, la ou les HRTFs qui approchent un rendu satisfaisant sinon optimal pour chaque individu. Cette recherche ou sélection peut s'opérer (1) sur la base d'une proximité morphologique ou évaluée sur le signal (proximité spectrale ou des fonctions de directivité spatiale) ; (2) à l'issue d'expériences d'écoute où le participant est amené à juger le rendu offert par plusieurs jeux de HRTFs ; (3) à partir de mesures acoustiques simplifiées réalisées sur l'individu. De sorte à rendre la tâche moins fastidieuse, des études se sont intéressées à la mise en évidence de HRTFs ou de morphologies archétypiques par l'intermédiaire d'un processus de classification de HRTFs permettant ainsi de réduire le nombre de HRTFs à tester.

3.1 Acquisition

En premier lieu, les méthodes d'acquisition des HRTFs d'un individu par la mesure, qu'elles soient acoustiques ou morphologiques, sont les plus précises car elles permettent d'obtenir les informations individuelles sur une plage fréquentielle large et un échantillonnage spatial fin. Nous verrons qu'elles requièrent un protocole bien défini.

3.1.1 Mesures acoustiques

3.1.1.1 Mesure en chambre anéchoïque sur un échantillonnage dense

Traditionnellement, les HRTFs sont acquises sur une grille sphérique dense. L'ensemble de la procédure d'acquisition et d'égalisation est détaillée en section 2.2.

Les mesures en conditions anéchoïques permettent de s'affranchir de l'effet de salle et laissent ainsi toute liberté lors de la synthèse. Mais elles peuvent également être réalisées en environnement non-anéchoïque. Les réponses mesurées sont alors appelées BRIRs (*Binaural Room Impulse Responses*).

Malgré la précision qu'offrent ce type de systèmes, la mesure acoustique de HRTFs est complexe et fastidieuse, autant pour l'expérimentateur que pour le sujet qui est mesuré.

3.1.1.2 Reconstruction à partir d'un nombre réduit de mesures représentatives

La méthode d'individualisation présentée ici propose d'acquérir les HRTFs selon un nombre réduit de directions et de générer les HRTFs aux positions intermédiaires. Cependant, comme dans tout problème d'interpolation, la performance de la technique diminue à mesure que les données de départ se raréfient. Cette méthode pose donc la question du nombre et des directions où les HRTFs doivent être mesurées pour pouvoir estimer avec une précision suffisante les HRTFs aux autres directions. Pour cela, des méthodes de classification sont généralement utilisées pour mettre en évidence les directions représentatives. L'interpolation spatiale et la prédiction à l'aide de méthodes statistiques sont les principales méthodes de reconstruction utilisées dans la littérature pour reconstruire les HRTFs aux directions manquantes.

Certaines études se sont intéressées au nombre minimal de directions de mesure nécessaires à la reconstruction de l'ensemble des HRTFs par interpolation spatiale, sans que cela ne génère d'artefacts audibles. Carlile et al. [CJvR00] ont par exemple montré que 150 mesures seulement sont nécessaires. Cependant, ces auteurs ont systématiquement utilisé une distribution uniforme des points sur la sphère et n'identifient pas les directions représentatives.

A partir de l'observation que la résolution spatiale du système auditif varie dans l'espace, Fahn et al. [CSL03] utilisent une méthode de classification afin d'identifier les HRTFs qui permettent une minimisation de l'erreur de reconstruction par interpolation spatiale. Les auteurs mettent en évidence un ensemble de 12 et de 181 directions pour le plan horizontal et pour la sphère complète, respectivement. L'étude ne porte cependant que sur le jeu de HRTFs d'une tête artificielle. Rien n'indique que l'ensemble des directions représentatives identifiées soient généralisables à toute une population. Lee et Lee [LpL11] ont par ailleurs montré que le nombre et les positions des HRTFs

nécessaires à une reconstruction complète sur le plan horizontal varient en fonction du jeu de HRTFs.

Lemaire et al. [LCB+05] (voir aussi [Bus06] pour plus de détails) ont utilisé des méthodes statistiques (les réseaux de neurones) à la fois pour la sélection des directions représentatives et la prédiction des HRTFs aux directions manquantes. Les directions représentatives sont obtenues par classification d'un jeu de HRTFs par les cartes de Kohonen sur la base de leur similarité spectrale. L'évaluation de la classification obtenue est réalisée à partir de l'erreur de reconstruction entre les HRTFs estimées à partir des HRTFs représentatives et les HRTFs de référence (issues de la mesure). L'étude montre que, pour un individu donné, l'algorithme de classification par cartes de Kohonen optimise la sélection des HRTFs représentatives par rapport à une sélection géométrique (uniforme) des directions sur la sphère. Cependant, il s'avère que cette sélection de directions ne peut se généraliser à tous les sujets. Lorsque la classification est généralisée à un ensemble de sujets, le recours à une distribution uniforme redevient supérieure à la méthode statistique.

Basée sur la représentation spatiale des HRTFs sous forme d'SFRS, une méthode singulière a été proposée dans la thèse de Guillon [Gui09]. Elle consiste dans un premier temps à classifier une large base de données d'SFRS à différentes fréquences et mesurées sur différents individus. La similarité, basée sur l'inter-corrélation spatiale des fonctions définies sur la sphère, est évaluée à une rotation près et à un décalage fréquentiel près. Plusieurs SFRS archétypiques sont ainsi mises en évidence. Dans un deuxième temps, les SFRS incomplètes, i.e mesurées sur un échantillonnage homogène mais grossier, sont décomposées en harmoniques sphériques et comparées aux SFRS archétypiques par l'intermédiaire d'un processus de reconnaissance de formes. Les SFRS identifiées comme proches sont alors utilisées pour remplacer les SFRS mesurées grossièrement. Les résultats de tests d'écoute en synthèse binaurale dynamique suggèrent que seules 45 à 65 mesures sont nécessaires pour obtenir une spatialisation équivalente aux HRTFs mesurées sur un échantillonnage fin. Cependant, l'auteur ne met pas en évidence de positions de mesures optimales et la méthode est limitée par la décomposition en harmoniques sphériques qui impose que l'échantillonnage spatial soit homogène.

L'étude de Guillon [Gui09] a mis en évidence que des ressemblances dans les fonctions spatiales de différents sujets apparaissent à une rotation près. Aussi, Lemaire et al. [LCB+05] ont découvert qu'un sous-ensemble de directions représentatives n'était pas généralisable à tous les individus. Ces deux observations suggèrent qu'il n'existe pas de directions représentatives universelles.

3.1.2 Mesure morphologique et modèle numérique

Une méthode alternative à la mesure acoustique est le calcul numérique à partir d'une représentation géométrique de la tête de l'auditeur, qui permet de s'affranchir du temps et du bruit de mesure. Elle offre ainsi des possibilités telles que la mesure à différentes distances et l'acquisition sur un échantillonnage plus fin.

Les méthodes des éléments finis (*Finite Element Method*, FEM) et des éléments aux frontières (*Boundary Element Method*, BEM) sont deux méthodes qui permettent d'acquérir numériquement les HRTFs [Kat98, Kah00]. Elles reposent sur la simulation des transformations subies par l'onde sonore lors de sa propagation entre une source et l'entrée des canaux auditifs.

La méthode requiert le maillage numérique 3D de la morphologie de l'auditeur. Celui-ci doit être suffisamment fin pour pouvoir calculer les HRTFs jusqu'à une fréquence élevée. En effet, la fréquence maximale atteinte par le calcul est déterminée par la finesse du maillage. La principale limite de cette technique réside alors dans le coût de calcul qui augmente avec cette précision. Afin de limiter ce coût, le maillage peut être affiné uniquement aux régions nécessitant une fine résolution, comme les pavillons d'oreille (dont les détails morphologiques sont les principaux responsables des caractéristiques spectrales des HRTFs et varient fortement entre les individus) et laissé plus grossier ailleurs. Il est également possible d'imaginer partir d'un maillage "générique" que l'on adapte en entrant les paramètres individuels d'une nouvelle morphologie.

L'acquisition du scan 3D de la tête et des oreilles requiert cependant la mise en place d'un processus complexe et l'utilisation d'équipements particuliers. De plus, le calcul numérique ne prend pas en compte le torse qui est pourtant à l'origine d'indices de localisation en élévation dans l'intervalle $[700 - 1000]$ Hz.

3.2 Adaptation

Les méthodes d'adaptation d'une HRTF non-individuelle à un nouvel individu se basent sur des observations au niveau signal ayant mis en évidence des similarités entre jeux de HRTFs à une

fréquence donnée moyennant un facteur d'échelle sur la dimension fréquentielle, ou à une direction donnée, par rotation des fonctions spatiales. Une disparité spatiale homogène sur l'ensemble du spectre peut s'expliquer par une simple différence de placement du sujet lors des mesures. En revanche lorsque celle-ci concerne une zone fréquentielle particulière, elle est plus vraisemblablement liée à une caractéristique morphologique. Des méthodes d'adaptation de HRTFs non-individuelles à partir de mesures morphologiques ont été proposées.

Ces mêmes observations peuvent être faites sur le paramètre de retard interaural et conduisent à la proposition de méthode d'ajustement individualisé de l'ITD en tirant parti de l'implémentation séparée de la composante à phase minimale et du retard pur (c.f. section 2.3.1.1).

3.2.1 Composante spectrale

La méthode du morphisme fréquentiel (*frequency scaling*) [Mid99a] se base sur l'observation qu'une différence inter-individuelle de taille entre pavillons d'oreille aurait pour impact de décaler en fréquence certaines caractéristiques spectrales (e.g. résonances et zéros) présentes dans les fonctions de transfert. La méthode consiste à appliquer un facteur d'échelle fréquentiel aux spectres des filtres non-individuels en vue d'une adaptation à un nouvel auditeur. Les facteurs d'échelle peuvent être estimés par l'intermédiaire d'un test d'écoute ou à partir de mesures anthropométriques (largeur de tête et taille des pavillons d'oreille) [MMO00]. L'étude montre une amélioration en termes de localisation des sources virtuelles entre les HRTFs non-individuelles brutes et celles ayant subi un morphisme fréquentiel [Mid99b].

De façon complémentaire, Guillon [Gui09] et Maki et Furukawa [MF05] montrent qu'une différence entre l'orientation des pavillons de deux individus peut se résumer par une rotation du système de coordonnées. La méthode d'adaptation proposée consiste à appliquer à la fois des opérations de décalage par rotation et de morphisme fréquentiel sur une HRTF prototypique préalablement identifiée comme proche (voir section 3.1.1.2).

Des méthodes similaires consistent à ajuster itérativement l'amplitude du spectre non-individuel sur certaines bandes de fréquence [TG98,RYW00], ou les poids de la décomposition ACP [HPP08], de sorte à optimiser le rendu spatial.

3.2.2 Retard interaural

Bien que plusieurs méthodes de sélection/estimation de HRTFs ont été validées à travers des tests de localisation, des erreurs liées au fait que l'ITD ne soit pas individualisé ont été observées (erreurs principalement latérales [See03]). En effet, ces études se sont principalement focalisées sur l'individualisation des indices spectraux.

Une première approche d'individualisation de l'ITD consiste à adapter le rayon d'un modèle sphérique d'ITD à une mesure de la taille de la tête, par l'intermédiaire d'une méthode de régression multi-linéaire [AAD01]. Duda et al. [DAA99] ont proposé un modèle plus proche de la réalité, tenant compte du caractère non-sphérique de la tête et de la position non centrée des oreilles. Cependant ces modélisations restent très simplifiées.

Une autre approche consiste à décomposer les ITDs d'un ensemble de jeux de HRTFs, représentés sous forme de fonctions spatiales, en composantes principales. A l'aide de méthodes de régression linéaire, des études (e.g. [HG10, AAK12]) ont mis en évidence la corrélation entre les premières composantes de la décomposition et des paramètres morphologiques comme la taille de la tête. L'estimation des poids de la décomposition ACP associés à l'ITD pour un nouvel individu est donc rendue possible à partir de la mesure individuelle de seulement quelques paramètres morphologiques.

3.3 Sélection guidée à partir de données objectives

3.3.1 Paramètres morphologiques

A partir de l'hypothèse selon laquelle les HRTFs sont intimement liées à certains paramètres morphologiques, comme mis en évidence par exemple par Middlebrooks [Mid99a], on peut imaginer sélectionner une HRTF non-individuelle sur la base d'une proximité morphologique. Cela suppose de posséder une base de données contenant à la fois les HRTFs et les paramètres morphologiques des individus ce qui est le cas, par exemple, de la base CIPIC.

Zotkin et al. [ZDD02] ont mis en place une sélection personnalisée d'un jeu de HRTFs non-individuel issu de la base CIPIC. La sélection est réalisée à partir d'une comparaison de 7 mesures anthropométriques réalisées sur l'individu pour lequel on cherche à sélectionner une HRTF et

celles de la base de données de HRTFs. Bien que les HRTFs ne peuvent se résumer à 7 paramètres, cette méthode de sélection a l'avantage d'être rapide et simple à mettre en œuvre (les mesures anthropométriques sont réalisée à partir d'une photographie de l'oreille).

Aussi, les méthodes statistiques ont été utilisées pour apprendre les relations de haut niveau entre paramètres morphologiques et HRTFs mesurées. Parmi les méthodes utilisées, on trouve les modèles de régression multi-linéaire (*Multiple Linear Regression*, MLR) [JLL+00, HZZ+06, HG10] ou de réseaux de neurones [HZMW08,LH13,GMGF14]. Une étape de réduction de la dimensionalité des données est souvent appliquée en amont comme la décomposition ACP [JLL+00, HZZ+06, HG10] ou des méthodes non-linéaires telles que la méthode Isomap [GMGF14]. Les tests perceptifs associés à ces études montrent qu'une estimation personnalisée des HRTFs permet d'obtenir de meilleurs résultats qu'à l'écoute d'une HRTF non-individuelle.

3.3.2 Enregistrements binauraux

Une alternative à la mesure acoustique "idéale" (i.e. environnement anéchoïque, contrôle des directions de mesure, source large bande) repose sur la sélection guidée à partir d'informations acoustiques captées sur un individu et obtenues dans des conditions non contrôlées. Dans le cas d'enregistrements binauraux réalisés dans des conditions naturelles (e.g. scènes extérieures), la position et le contenu spectral des sources sont inconnus. L'une des difficultés majeures réside dans le fait qu'*a priori* on ne peut accéder qu'aux informations interaurales. L'objectif est alors d'identifier dans une base de données les HRTFs correspondant aux mieux aux indices estimés sur ces mesures "à l'aveugle".

L'étude de Maazaoui et Warusfel [Maa16] explore cette approche en se basant sur le modèle de localisation auditive d'égalisation-annulation [Bas03]. A partir d'enregistrements binauraux, les retards de phase ou d'enveloppe ainsi que le gain interaural sont estimés à tout instant tels qu'ils minimisent l'erreur résiduelle entre les signaux gauche et droit. Cette erreur résiduelle caractérise l'information binaurale issue de l'enregistrement en condition non-anéchoïque. Le couple de HRTFs de la base de données permettant de minimiser la différence d'erreur résiduelle avec le signal enregistré est la plus vraisemblable. Les résultats de cas simulés avec une source fixe ou en mouvement sont prometteurs [Maa16], mais la méthode n'a pas encore été testée dans le cadre de la sélection d'un individu.

3.4 Sélection guidée sur la base de relevés subjectifs

Les méthodes d'évaluation de la qualité du rendu d'un contenu sonore binauralisé (c.f. section 2.1.1) ont été utilisé dans le cadre de travaux d'individualisation pour sélectionner le jeu de HRTFs non-individuel permettant d'optimiser la qualité du rendu par rapport à un ensemble de HRTFs testé. Les tests d'écoute mis en œuvre peuvent être précédés d'une sélection objective d'un nombre réduit de HRTFs pour l'auditeur (par exemple à partir de quelques mesures acoustiques [AR14]) permettant de guider *a priori* la sélection. Ils peuvent aussi être facilités par la mise en évidence d'individus typiques d'un groupe.

3.4.1 Sélection guidée à partir de tests d'écoute

Seeber et Fastl [See03] proposent de guider la sélection à travers deux expériences d'écoute successives. Dans la première expérience, les participants doivent sélectionner 5 HRTFs (sur les 12 présentées) comme offrant la meilleure impression spatiale frontale à l'écoute de trajectoires horizontales. La seconde étape consiste à sélectionner une seule HRTF en termes de régularité d'une trajectoire horizontale, d'externalisation et de taux de confusions. De manière similaire, Iwaya et al. [Iwa06] proposent la sélection d'une HRTF guidée par des jugements qualitatifs suivant une comparaison par paires dans laquelle 32 HRTFs sont testées. Ces deux méthodes se montrent efficaces (15 à 20 minutes) et permettent d'offrir de bonnes performances de localisation avec le jeu de HRTFs sélectionné. Cependant, ces deux études ne se basent que sur l'évaluation de trajectoires horizontales. Or, Andreopoulou et Katz [AK16] ont montré que l'information à la fois horizontale et verticale était nécessaire pour obtenir un jugement complet des HRTFs, étant donné les différences dans les jugements des participants vis-à-vis de chacune d'entre elles.

Katz et al. [KP12] ont réalisé un test d'écoute où les 45 participants doivent juger de trajectoires horizontales et verticales de 46 HRTFs non-individuelles (en utilisant les termes *bad/ok/excellent*). Les résultats montrent que les HRTFs sélectionnées sur la base de ce simple test d'écoute permettent d'obtenir de bonnes performances de localisation.

Bien que la sélection par jugements qualitatifs semble être capable d'identifier une HRTF convenable pour un nouvel individu, la tâche parait fastidieuse étant donné la faible consistence des jugements qualitatifs, en particulier pour les participants non-experts [SK12]. Alternativement, les tests de localisation sont plus intuitifs et permettent de collecter des jugements objectifs sur la qualité du rendu spatial des sources virtuelles synthétisées avec des HRTFs non-individuelles, par l'analyse des erreurs de localisation.

3.4.2 Réduction de la base de données

La taille des bases de données de HRTFs existantes rendent la tâche de sélection d'un jeu d'HRTF dans une base de données fastidieuse.

Recherche de sujets représentatifs A partir de l'hypothèse selon laquelle il existerait des groupes d'individus proches, une étape de classification des sujets permet de réduire la taille de la base de données en ne gardant que les individus représentatifs des groupes. La classification est alors réalisée à partir d'une proximité signal entre les HRTFs des individus.

Shimada et al. [SHH94] ont réalisé la classification de HRTFs mesurées en environnement non-anéchoïque à l'aide de l'algorithme LBG (*Linde-Buzo-Gray*). Ils mettent en évidence une bonne externalisation des sources virtuelles avec les représentants des 8 classes identifiées. Cependant, la classification est réalisée direction par direction.

D'autres études ont examiné la classification de HRTFs en seulement deux directions (avant et arrière) à l'aide des algorithmes de regroupement hiérarchique (*Hierarchical Agglomerative Clustering*, HAC [RBA+10]) et des K-moyennes (*k-means*, [TBK12]). Les résultats montrent qu'une sélection basée sur l'un des représentants des classes identifiées (5 sur 62 HRTFs [TBK12] et 6 sur 196 [RBA+10]) permet une amélioration notable de la localisation, en particulier en termes de réduction des confusions avant-arrière et d'externalisation, par rapport à l'écoute avec de HRTFs génériques ou sélectionnées aléatoirement.

La classification par regroupement hiérarchique a également été utilisée par Xie et al. Dans une première étude [XZZ13], les auteurs s'intéressent la sélection de HRTFs représentatives à partir des 3 directions frontales associées au système de reproduction 5.1 puis généralisent leur méthode à toutes les directions [XZH15]. Ils montrent pouvoir mettre en évidence 7 groupes d'individus parmi les 52 utilisés pour la classification de HRTFs. Par ailleurs, ils observent que les représentants de chaque classe présentent une proximité morphologique avec les individus de la classe qu'ils représentent. Les résultats montrent qu'une sélection subjective parmi les 7 représentants permet d'obtenir des performances de localisation améliorées par rapport à une HRTF générique.

Une autre méthode de sélection de HRTFs représentatives pourrait se baser sur des jugements subjectifs. Katz et al. [KP12] ont par exemple identifié 7 HRTFs convenant à la majorité de leurs auditeurs. Cette sélection se base sur les jugements de qualité de 45 participants à l'écoute de 46 HRTFs présentées suivant des trajectoires circulaires dans le plan horizontal et dans le plan vertical. Seeber et Fastl [See03] ont également mis en évidence une préférence des leurs 46 participants pour 3 des 12 jeux de HRTFs testés. Une étude de classification permettrait d'identifier s'il s'agit de HRTFs représentatives d'un point de vue signal.

Recherche de directions représentatives Les tests d'écoute peuvent également être facilités par la mise en évidence de directions discriminantes, i.e. porteuses d'information individuelle.

Andreopoulou et al. [ARB13] ont utilisé un algorithme de classification linéaire (LDA) afin de mettre en évidence un sous-ensemble de directions capables de discriminer les sujets d'une base de données. L'algorithme permet ainsi de réduire de 64% les données. A partir des HRTFs mesurées uniquement aux directions discriminantes sur un nouvel individu, les auteurs proposent une pré-sélection guidée d'un jeu de HRTFs non-individuel dans une large base de données en terme d'une proximité au niveau signal à ces directions. Les résultats de tests d'écoute montrent que les HRTFs non-individuelles proches, au sens de ce critère signal, obtiennent de meilleurs jugements de préférence que des HRTFs plus éloignées dans le classement de similarité objective.

Conclusion

Les méthodes basées sur les tests d'écoute cherchent à identifier le jeu de HRTFs qui optimise le rendu spatial des sources virtuelles pour un auditeur à travers l'observation de ses réponses. La tâche peut cependant s'avérer assez fastidieuse étant donné qu'elle nécessiterait idéalement de tester l'ensemble des HRTFs d'une base de données exhaustive afin de converger vers le jeu d'HRTF optimal. A l'inverse, on peut imaginer analyser les réponses du sujet à l'écoute de HRTFs non-individuelles quelconques, de manière à identifier le jeu de HRTFs qui caractériserait objectivement le comportement des réponses. Pour ce faire, il est possible d'envisager l'utilisation d'un modèle de prédiction des directions perçues qui testerait une à une les HRTFs de la base de données en terme de leur capacité à expliquer les directions perçues. En principe, le jeu de HRTFs identifié comme optimal véhiculerait des information spatiales proches de celles qui définissent la carte audio-spatiale de l'auditeur, sur laquelle il se base pour localiser. Cette méthode sera testée au chapitre 7.

Parmi les méthodes d'individualisation présentées dans ce chapitre, plusieurs d'entre elles reposent sur une mesure de similarité objective entre HRTFs : la classification d'un jeu de HRTFs, pour l'identification de directions représentatives, ou de l'ensemble des jeux de HRTFs d'une base de données, pour la mise en évidence d'individus représentatifs. Pour garantir la qualité des résultats, il faut s'assurer que la métrique de similarité reflète des caractéristiques perceptivement pertinentes.

Les méthodes présentées dans ce chapitre apportent des solutions concrètes au problème d'individualisation. Cependant, il est difficile de comparer le mérite de chacune de ces méthodes étant donné la variabilité dans leurs protocoles d'évaluation du rendu final. En effet, même si la plupart cherchent à quantifier les performances de localisation de la HRTF obtenue, les tests de localisation mis en œuvre diffèrent de manière significative, que ce soit dans les positions évaluées ou dans la méthode de report utilisée. Cela souligne la nécessité d'établir un protocole expérimental universel qui rendrait la comparaison des études plus évidente. Concernant les tests de localisation, il s'agit de trouver à la fois un ensemble de directions représentatives de la qualité du rendu sonore spatialisé mais aussi une méthode de report qui soit précise et facile à mettre en œuvre.

Intuitivement, il apparait que les directions caractéristiques de la fidélité du rendu spatial qu'un jeu de HRTFs puisse offrir correspondent aux directions auxquelles il se distingue le plus d'un autre jeu de HRTFs. Les études qui se sont penchées sur cette question (voir section 3.1.1.2) ont cependant démontré une difficulté à identifier de telles directions caractéristiques. De plus, ces études s'intéressent davantage aux directions où les HRTFs sont représentatives des HRTFs aux autres directions (au sein d'un même jeu) plutôt que des directions qui discriminent le mieux les individus. La littérature à ce sujet semble insuffisante.

Le deuxième point important pour la mise en parallèle de résultats d'études de localisation concerne la méthode de report de la direction perçue. Or, celle-ci introduit inévitablement une part d'erreur dans les jugements collectés qui ne peut être mise en évidence que par une comparaison directe de plusieurs méthodes. Cela fait l'objet du prochain chapitre, dans lequel un état de l'art des méthodes de report est donné ainsi qu'une étude comparative de 3 méthodes de pointage d'intérêt.

Les approches proposées afin d'acquérir un jeu d'HRTF optimal pour un individu donné sont nombreuses. Cependant, le mérite des méthodes ayant recours à des mesures de similarité objective entre HRTFs reposent directement sur la pertinence de la métrique utilisée. Aussi, l'évaluation de la qualité du rendu offert par le jeu de HRTFs obtenu requiert notamment la mise en place de tests de localisation auditive de sources virtuelles. La question de la méthode de report des jugements utilisée est essentielle, et est adressée au chapitre suivant.

–

Le premier chapitre a permis de comprendre que la localisation auditive se base sur des indices acoustiques individuels et que leur interprétation résulte d'un processus d'apprentissage. On peut dès lors se demander comment fonctionne la localisation auditive en présence d'indices acoustiques non-individuels, étrangers à l'auditeur, n'ayant pas fait l'objet d'un apprentissage. A cause de la complexité de la procédure d'acquisition des filtres individuels, la situation d'écoute binaurale non-individuelle est aujourd'hui majoritaire. Comme développé davantage au chapitre 6, certaines études ont mis en évidence que, dans le cas où les indices acoustiques délivrés à l'auditeur sont altérés ou partiels, la localisation est guidée par une similarité du point de vue acoustique entre le signal perçu et les indices appris par l'auditeur et sur lesquels s'est construite sa carte auditive spatiale. A partir de cette observation, il semblerait possible d'inférer la direction perçue par l'analyse comparative des indices acoustiques reçus et appris. Cette hypothèse forme le fondement des modèles de prédiction de la localisation auditive développés dans la littérature, tels que ceux présentés au chapitre 6. Nous étendons l'hypothèse à la localisation auditive en présence d'indices non-individuels dans un modèle introduit au chapitre 6 et développé au chapitre 7.

Comme présenté au chapitre 1, les indices spectraux occupent une place importante dans la localisation auditive binaurale et sont la cause première des distorsions spatiales en synthèse binaurale non-individuelle. Par conséquent, une attention particulière doit être portée, lors de la mise en place du modèle de prédiction, à la comparaison des indices spectraux. La définition d'une mesure de la similarité en termes d'information spectrale véhiculée par les HRTFs n'est pas triviale. Pourtant, nous avons observé au cours des chapitres 2 et 3 qu'elle constitue la base de nombreux concepts comme la classification de HRTFs, la sélection guidée basée sur une similarité du point de vue signal ou encore l'évaluation de l'ampleur des distorsions perceptibles générées par interpolation ou modélisation des HRTFs. Le chapitre 5 offre un aperçu de la diversité des métriques spectrales de la littérature, utilisées comme mesure de similarité ou comme critère d'erreur. Le développement d'un modèle de prédiction de la localisation est l'occasion de comparer plusieurs métriques spectrales de la littérature en termes de faculté à extraire les indices spectraux pertinents pour la localisation auditive. Pour ce faire, les résultats de prédiction offerts par chacune des métriques seront évalués vis-à-vis de données de localisation réelles.

Un test de localisation de sources virtuelles synthétisées avec différents filtres non-individuels sera mené en parallèle du modèle de prédiction. Cela pose la question de la méthode de report. Celle-ci introduit une part d'imprécision dans les jugements et il est donc nécessaire de traiter cette composante avec attention. Une étude comparative de trois méthodes de pointage (chapitre 4) servira à préciser le protocole expérimental du test de localisation de sources virtuelles au chapitre 7. La prédiction de localisation requiert les fonctions de transfert individuelles des auditeurs pour lesquels elle est menée. La campagne de mesures de HRTFs présentée au chapitre 2 ainsi que la maîtrise de l'ensemble des éléments techniques qui conduisent à la synthèse des sources virtuelles au casque nous permet d'envisager ce travail.

Enfin, le chapitre 3 a exposé un éventail de méthodes d'individualisation des HRTFs. A travers les tests d'écoute, nous avons vu que des jugements subjectifs peuvent guider la sélection d'un jeu de HRTFs non-individuel. Suivant ce raisonnement, nous proposons une méthode qui, par l'observation des directions pointées par un individu lors d'un test de localisation, cherche à identifier le jeu de filtres de la base de données qui caractérise le mieux sa perception spatiale auditive.

–

Chapitre 4

Comparaison des méthodes pour le report de la localisation auditive

La méthode de report est une question essentielle lors de la mise en place de tests de localisation auditive afin de garantir la pertinence des jugements. Ce chapitre dresse tout d'abord un aperçu de l'ensemble des méthodes de report de la localisation auditive utilisées dans la littérature, avant de présenter la mise en pratique de trois méthodes de pointage au sein d'un test de localisation sur sources réelles. Il expose les résultats d'une étude comparative de ces trois méthodes, dont les mérites sont analysés selon plusieurs critères d'intérêt.

Afin d'évaluer la fidélité du rendu binaural par rapport à une écoute en conditions réelles, plusieurs dimensions perceptives doivent être prises en compte : par exemple, la fidélité du rendu du timbre, l'impression globale d'espace et la localisation des sources (c.f. section2.1). Néanmoins, les tests de localisation restent les plus conventionnels compte tenu de l'importance de la reproduction de la localisation pour garantir l'adhésion de l'auditeur à la scène auditive qui lui est présentée (c.f. discussion sur l'immersion et la présence des sources en introduction du chapitre 3). Ils permettent de vérifier si les sources virtuelles synthétisées avec des HRTFs sont précisément localisées. L'évaluation des performances de localisation auditive consiste à relever le biais et la dispersion des jugements de la localisation perçue de la source par rapport à sa position effective. Cette mesure ne dépend pas seulement des performances de la méthode de reproduction et de la capacité du sujet à localiser des sources sonores, mais également de la précision et de la répétabilité avec lesquelles il reporte son jugement. Elle est donc intimement liée à la méthode de report utilisée.

Les méthodes de report employées dans la littérature sont très variables, rendant ainsi les résultats des tests de localisation difficilement comparables. On compte deux catégories principales de méthodes, à savoir les méthodes exocentriques qui se distinguent des méthodes égocentriques où le sujet est au centre du dispositif de report. Certaines études ont mis en évidence que les méthodes égocentriques sont plus précises que les méthodes exocentriques. Néanmoins, il n'existe à ce jour aucune méthode standard ayant révélé des avantages marqués en comparaison à d'autres méthodes de report.

L'objectif de cette étude est d'identifier une méthode de report de la localisation auditive qui offre des jugements précis, c'est-à-dire où les directions indiquées reflètent au plus près les directions perçues, avec une bonne répétabilité. Pour cela, il faut que cette méthode soit intuitive pour les participants. Aussi, nous cherchons une méthode qui permette de collecter un maximum de jugements dans un temps limité. Trois méthode égocentriques sont ici comparées dans un test de localisation de sources réelles (haut-parleurs), de sorte à s'affranchir pour l'instant des problèmes de fidélités de reproduction binaurale. Cette étude a fait l'objet d'un article soumis fin février 2015 à Acta Acustica, et reproduit en Annexe B. Ce chapitre présente l'analyse et les résultats de cette étude, en résumant les parties de description du protocole pour lesquelles plus de détails peuvent être trouvés dans l'article.

4.1 Introduction et état de l'art

De nombreuses méthodes de report de la localisation auditive ont été décrites et appliquées dans la littérature. Avec la méthode verbale (ou *Absolute judgment technique*) [WK89, WAKW93], le participant exprime les valeurs numériques d'azimut et d'élévation de la direction perçue. Bien que cette méthode soit fréquemment utilisée, la tâche s'avère peu intuitive pour les participants non expérimentés [Eva98]. Plusieurs études ont utilisé une méthode de report graphique sur une interface 2D ou 3D représentant une vue schématique de l'environnement de l'auditeur [PEN03, Lar01, BWA01, SWA+14]. La méthode GELP (*God's Eye Localization Pointing*) a été également utilisée quelques fois [GGE+95, DPSB00]. Elle utilise une sphère rigide matérialisant l'espace auditif sur laquelle le sujet indique la direction perçue avec un

stylet. Elle présente l'avantage d'un temps de réponse plus court que les méthodes de pointage manuelles ou avec la tête [GGE+95, DPSB00]. Ces dernières, plus répandues, consistent à demander au participant de se tourner vers la source perçue et de l'indiquer, bras tendu, avec le doigt ou avec un objet tenu dans la main [DPSB00, OP84, Grö02, PEN03, MGL10, GML10, PK12, PRD14], ou encore avec la tête [MM90, Bro95, CLH97, MMS01, BCJvS05, MGL10, AMB13]. Il existe aussi des méthodes de pointage indirectes utilisant une manette rotative (type *joystick*) [LB02, BDPW13] ou une *trackball* [See97] permettant de contrôler un pointeur acoustique (source virtuelle [LB02, BDPW13] ou haut-parleur [PH05] aligné avec le son cible) ou un pointeur visuel, tel qu'un laser [See97]. Ces méthodes ont l'avantage de réduire la complexité des interactions entre les différentes modalités sensorielles (i.e. visuelle, auditive et vestibulaire). Cependant, l'utilisation d'un pointeur visuel restreint le report au champ de vision et la mise en place d'un pointeur acoustique est peu adaptée aux tests de localisation de sources réelles distribuées en trois dimensions.

Le retour visuel peut être utilisé en parallèle du pointage manuel ou avec la tête, en plaçant un laser sur le front [MMS01, FMSZ10, TZV13, AMB13, WRS12] ou sur le pointeur tenu dans la main [TZV13, PRD14], matérialisant ainsi la direction pointée. L'apport d'un retour visuel permet d'améliorer les performances de localisation [MGL10, TZV13]. Tabry et al. [TZV13] ont par ailleurs montré que la précision relative au pointage avec la tête était plus affectée par l'absence d'indices visuels que la méthode de pointage manuel. Un dernier type de méthode de report consiste à indiquer la direction perçue avec les yeux (voir par exemple [HOR99]). Cependant, l'étude est alors restreinte aux sources situées dans le champ de vision.

Certaines études [DPSB00, PEN03, GGE+95] ont montré l'avantage, en termes de précision, des méthodes égocentriques (i.e. où le sujet est au centre du dispositif de report) par rapport aux méthodes verbales et aux méthodes exocentriques (i.e. de report sur une interface 2D ou 3D dont le centre ne coïncide pas avec la tête du sujet). Pour ces dernières, des erreurs systématiques révèlent la difficulté du participant à transposer sa représentation de l'espace sur le dispositif de report [DPSB00, PEN03]. Aussi, les méthodes qui ne nécessitent pas de mouvements du corps présentent l'avantage d'un temps de réponse court et permettent donc de collecter plus de réponses en un temps donné [GGE+95].

Au vu de ces observations, notre étude ne considère que des méthodes de pointage égocentriques. Trois méthodes de pointage ont été évaluées. Tout d'abord, les méthodes de pointage avec la tête ("head pointing") ou avec un pistolet en plastique tenu en main ("gun pointing"). Dans une étude comparative, Majdak et al. [MGL10] montrent que les performances obtenues sont globalement similaires bien que la méthode de pointage manuel donne de meilleurs résultats pour les sources très élevées. Frank et al. ont par ailleurs mis en avant le caractère intuitif et la fiabilité de la méthode de pointage au pistolet [FMSZ10]. Nous avons comparé ces méthodes à une autre méthode de pointage manuel qui, cette fois-ci, n'implique pas les mouvements du corps. La tâche consiste à indiquer la direction perçue en positionnant un stylet, tenu au bout des doigts, dans la région proximale de la tête ("proximal pointing"). Cette méthode est proche de celle appliquée dans l'étude de Djelani et al. [DPSB00] mais s'en distingue toutefois en autorisant indifféremment l'usage des deux mains. On espère ainsi remédier aux erreurs observées lors du pointage de sources situées dans l'hémi-espace opposé à la main utilisée pour pointer [DPSB00]. De plus, contrairement à l'étude de Djelani et al., le sujet reste immobile pendant l'émission du son ce qui évite la contribution des indices de localisation dynamiques et autorise une comparaison plus rigoureuse avec les deux autres méthodes étudiées. Pour finir, notre étude implique la localisation d'un plus grand nombre de sources et utilise des sources réelles (émises par des haut-parleurs) et non virtuelles (délivrées au casque).

Les avantages escomptés pour la méthode proximale sont : (1) de permettre au participant d'indiquer n'importe quelle direction de l'espace en 3 dimensions en lui laissant le choix de la main à utiliser pour pointer ; (2) d'obtenir un temps de réponse rapide étant donné que les mouvements du corps ne sont pas impliqués ; (3) de pouvoir être utilisée dans l'étude de la localisation de sources réelles comme celle de sources virtuelles émises au casque, souvent perçues proches de la tête ; (4) de collecter simultanément des jugements sur la distance perçue dans le cas de systèmes de reproduction binauraux. On peut craindre cependant certaines difficultés motrices pour atteindre les sources situées à l'arrière de la tête. On peut également craindre des différences de performance entre les sources pointées avec la main dominante ou non-dominante.

4.2 Test de localisation

L'étude présentée ici concerne un test sur sources réelles comparant les performances de report de localisation entre trois groupes de participants effectuant chacun le même test mais avec des dispositifs de reports différents.

4.2.1 Participants

39 participants, dont 15 femmes et 24 hommes et dont l'âge varie entre 21 et 55 ans, ont passé l'expérience. 19 d'entre eux sont experts du domaine de l'audio mais seulement 4 avaient déjà participé à une expérience de localisation auditive. Les participants ont été répartis en 3 groupes de 13 personnes (8 hommes et 5 femmes), affectés chacun à une condition expérimentale (H : "Head pointing", G : "Gun pointing", ou P : "Proximal pointing"). Ils n'avaient aucune connaissance sur la position des sources mais étaient informés que le son pouvait provenir de n'importe quelle direction de l'espace en trois dimensions.

4.2.2 Dispositif expérimental

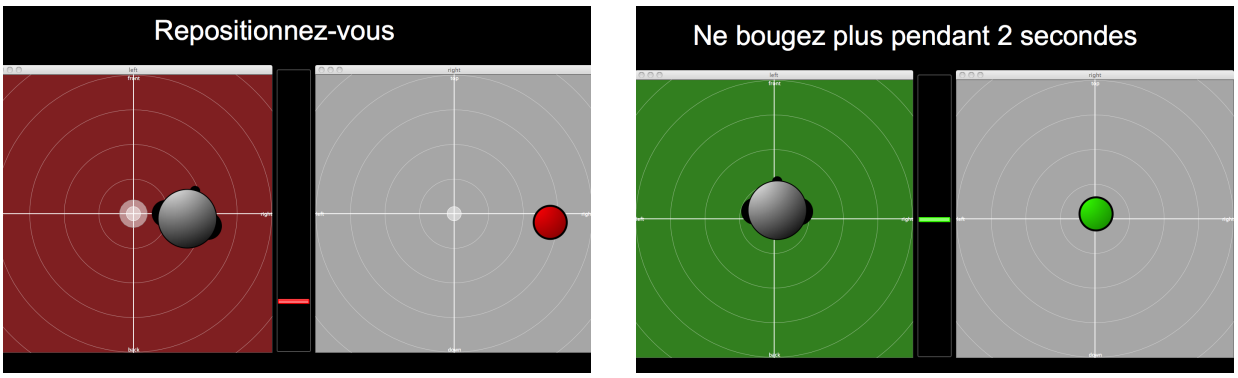
Le test s'est déroulé dans un studio traité en absorbant et équipé de 24 haut-parleurs coaxiaux distribués sur une demi-sphère. Les coordonnées sphériques de ces haut-parleurs sont décrite dans le tableau 4.1. Le participant était assis au centre de l'hémisphère de haut-parleurs sur une chaise rotative et ajustable en hauteur. Un système de suivi de position permettait de relever en temps réel la position et l'orientation de la tête du participant ainsi que la position de ses mains pour les méthodes G et P. L'expérience se déroulait dans le noir de sorte que le participant ne puisse jamais apercevoir la position des HPs et d'éviter toute interaction multisensorielle. Cependant, une interface s'allumait entre chaque stimulus de sorte à guider le participant pour se repositionner parfaitement au centre du dispositif et orienter sa tête selon la direction de référence, i.e. 0° en azimut et 0° en élévation. L'interface peut être visualisée en figure 4.1.

El (°)	Az (°)								
-5	-160	-120	-80	-40	0	40	80	120	160
26	-20	-60	-100	-140	180	140	100	60	20
57	-144		-72		0		72		144
90					0				

TABLE 4.1 – Coordonnées en azimut et élévation des 24 haut-parleurs.

Une fois correctement positionné, l'interface était éteinte et le sujet devait rester immobile, dans le noir complet, en l'attente du stimulus. Puis, il devait réaliser la tâche de pointage et valider sa réponse. La principale distinction dans le protocole expérimental de la condition P réside dans le fait que le sujet devait garder la tête immobile (appuyée sur un appui-tête) pendant toute la durée du test. Au contraire, pour les protocoles H et G, les participants étaient invités à utiliser la rotation de la chaise tournante sur laquelle ils étaient assis.

Plus de détails sur le dispositif expérimental peuvent être trouvés dans la section 2.2 de l'article reproduit en annexe B.



(a) (b)

FIGURE 4.1 – Interface graphique de repositionnement schématisant la position (x, y) de la tête du participant à gauche et son orientation (azimut, élévation) à droite. Le centre de chaque interface, matérialisé par le rond blanc, indique la position de référence. Le curseur du milieu indique la position z (hauteur). L'interface (a) invite le sujet à se replacer selon la position et l'orientation de référence. Une fois celles-ci atteintes, l'interface (b) s'éclaire et invite le sujet à maintenir cette position jusqu'au stimulus.

4.2.3 Stimulus

Le stimulus devait être suffisamment court pour éviter les mouvements de la tête pendant la durée d'émission. Selon Blauert [Bla97], les mouvements de la tête prendraient entre 200 et 300 ms à être initialisés. D'après les études [MM90, VVO04, HVO98], la localisation auditive nécessite au moins 80 ms pour offrir une estimation stable de l'élévation de la source sonore. De plus, Katz et Parseihian [KP12] ont montré que la localisation en azimut d'un train de bruits blancs successifs était améliorée par rapport à un bruit continu (composé d'une seule impusion). Pour répondre à ces différents critères, le stimulus utilisé était composé d'un train de bruits gaussiens d'un durée de 230 ms : 4 répétitions de 50 ms incluant 10 ms de rampes cosinusoïdales montantes et descendantes et avec 10 ms de pause entre chaque répétition. Le niveau de pression sonore était de 65 ± 2 dBA.

4.2.4 Méthodes de report testées

Méthodes de pointage avec la tête et au pistolet Dans la condition de pointage avec la tête (H), les participants devaient indiquer la direction perçue avec le nez (c.f. figure 4.2(a)). Une fois en face de la direction perçue, ils devaient valider leur réponse en double-cliquant sur le bouton d'une manette tenue dans la main de leur choix. Dans la condition de pointage avec un pistolet en plastique (G), les participants devaient indiquer la direction perçue en pointant avec le pistolet, bras tendu, et en positionnant le pistolet dans l'alignement du bras afin d'éviter les mouvements du poignet (c.f. figure 4.2(b)). Les participants étaient libres de choisir la main dans laquelle ils tenaient le pistolet. La réponse était validée en appuyant deux fois sur la gachette du pistolet en plastique. Dans ces deux conditions, la direction pointée était calculée comme l'intersection de l'orientation de la tête, ou du pistolet, avec une sphère virtuelle de mêmes centre et rayon que la sphère de haut-parleurs. Le point d'intersection était ensuite converti en coordonnées sphériques. Les participants étaient invités à utiliser la rotation azimutale de la chaise tournante sur laquelle ils étaient assis pour s'orienter vers la direction perçue.

(a) Pointage avec la tête (H)

(b) Pointage au pistolet (G)

FIGURE 4.2 – Photos de sujets réalisant la tâche de pointage avec la tête (a) et au pistolet (b). Dans les deux conditions, la position et l'orientation de la tête du sujet est repérée en temps réel par un système de suivi de position, grâce à l'intérieur d'un casque de chantier placé sur la tête du sujet et équipé de billes réfléchissantes. Dans la condition G (figure (b)), la position et l'orientation du pistolet en plastique tenu dans la main est également repéré en temps réel grâce à des billes réfléchissantes placées sur l'objet.

Méthode de pointage proximale Dans la condition de pointage proximale (P), les participants doivent appuyer leur tête sur un appui-tête pendant toute la durée de l'expérience, de sorte à éviter tout mouvement. Dans chaque main, ils tiennent un stylet composé d'une tige au bout de laquelle se trouve une bille de référence qui doit être placée au bout des doigts. Les participants étaient invités à placer la bille de référence à la direction perçue du son, dans la région proximale de la tête (c.f. figure 4.3) et à valider leur réponse en appuyant sur une pédale avec leur pied. Ils étaient libres de choisir la main permettant d'atteindre la direction perçue le plus confortablement (naturellement, la main droite pour une direction perçue à droite et réciproquement). Aucune contrainte n'était imposée quant à la distance de pointage. Bien que les sujets étaient invités à ne pas bouger la tête, le système de suivi de position a mis en évidence des mouvements inconscients durant la phase de pointage (déviation moyenne par rapport à la position/orientation initiale de 1cm/1°). La direction pointée était calculée comme l'intersection du "centre de la tête" et de la bille de référence. Le "centre de la tête" était défini au moment de la calibration par l'intersection des plans frontal (axe interaural) et médian, à l'aide des faisceaux lasers, et coïncide avec le centre de la sphère de haut-parleurs. Notons que plus la distance de pointage est faible, plus la direction indiquée est sensible à de faibles déplacements de la main ou de la tête.

4.2.5 Déroulement du test

Chronologiquement, le déroulement du report de la localisation pour chaque stimulus consiste en 4 étapes : (1) le sujet vérifie que sa tête est bien en position initiale avec ses bras reposés sur les accoudoirs ; (2) une fois correctement positionné pendant au moins 2 secondes, l'interface s'éteint ; (3) 2 secondes plus tard, le stimulus est émis dans une des 24 directions de l'espace, sélectionnée de manière aléatoire ; (4) le

FIGURE 4.3 – Photos d'un sujet réalisant la tâche de pointage proximale. Le sujet porte un objet de pointage dans chaque main avec le marqueur de référence tenu au bout des doigts. La tête est supportée par un appui-tête (c.f figure de droite) pour éviter les mouvements de la tête, et est repérée en temps réel par un système de suivi de position, de même que les objets de pointage. Le sujet utilise la main de son choix pour pointer en laissant l'autre au repos (bras posé sur l'accoudoir).

participant indique la direction perçue selon la consigne du groupe auquel il appartient (H), (G) ou (P) et valide sa réponse (manette pour H, gachette pour G, pédale pour P).

4.2.6 Session d'entraînement et expérience

Un total de 192 essais (8 répétitions de chacune des 24 directions test) est réalisé au sein d'une expérience, pour une durée d'environ 40 minutes. Chaque sujet était soumis à une session d'entraînement de 10 directions, lui permettant de se familiariser avec les tâches de pointage et de repositionnement. Le stimulus était le même que pour le reste de l'expérience et les directions avaient été sélectionnées de sorte à tester au moins une direction par quadrant. A aucun moment le sujet ne recevait de retour de l'expérimentateur sur ses performances.

4.3 Résultats

4.3.1 Définition des critères de comparaison et analyse statistique

Nous présentons ici les variables étudiées (et méthodes de calcul associées) pour réaliser la comparaison des 3 méthodes de pointage. Puis, nous définissons le principe des analyses statistiques (ANalyses Of VAriance, ANOVA) réalisées pour quantifier les effets de différents facteurs, tels que la méthode de pointage.

Définition des erreurs Pour analyser les performances de localisation, les erreurs horizontales et verticales ont été calculées. Elles ont été définies de la même façon que dans l'article de Makous et Middlebrooks [MM90] et Gilkey et al. [GGE+95]. Une discussion, donnée dans l'article en section 3.1, justifie ce choix par rapport aux autres définitions de la littérature (erreurs latérales/polaires [MA84, MGL10] ou erreurs moyennes d'angle [WK89]). Ce choix est guidé par les mouvements rotatifs impliqués dans les tâches de pointage avec la tête et au pistolet, et repose sur le système de coordonnées sphériques. L'erreur verticale est calculée comme la distance angulaire entre l'élévation de la cible et l'élévation pointée. Comme illustré figure 4.4, l'erreur horizontale est donnée par l'angle entre le vecteur associé à la direction cible et le vecteur formé par le centre de la sphère et un point sur la surface de la sphère de latitude égale à la latitude de la direction cible et de longitude égale à la longitude du jugement (cercle noir sur la figure). Contrairement à une définition plus standard d'erreur d'azimut, cette définition de l'erreur horizontale a l'avantage de prendre en compte la compression de l'azimut au pôle. Ainsi, pour une source située au pôle, l'erreur horizontale est nulle. Par conséquent, le haut-parleur situé au zénith ne sera pas considéré dans l'analyse des erreurs horizontales. Nous avons distingué les erreurs signées et non signées : les valeurs signées traduisent un biais systématique de localisation et les valeurs non signées fournissent une information de précision de localisation. Une valeur d'erreur horizontale positive indique que le sujet a trop tourné sur sa

chaise afin d'aller pointer dans la direction perçue (le zéro étant défini par la position initiale). Une erreur verticale positive indique qu'il a pointé trop haut par rapport à la direction cible.

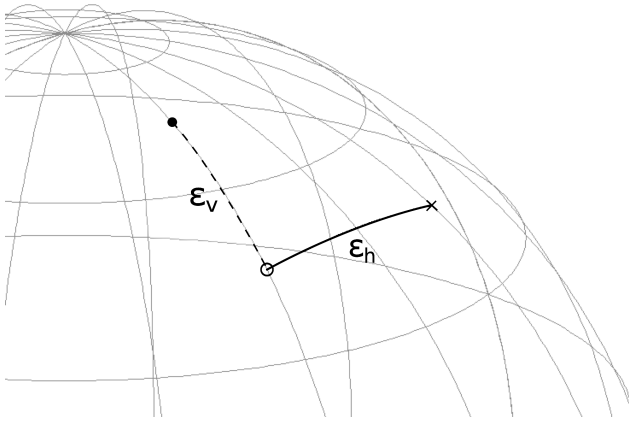


FIGURE 4.4 – Illustration des erreurs horizontale (ϵ_h) et verticale (ϵ_v) telles que définies par Makous et Middlebrooks [MM90] et Gilkey et al. [GGE+95]. Les lignes grises représentent les lignes de longitude et de latitude d'une partie de la sphère, par incrément de 20°. Le point noir indique la direction pointée et la croix, la direction cible. Le point indiqué par le cercle noir est défini par la longitude de la direction pointée et la latitude de la direction cible.

Confusions avant-arrière Etant donné que les mouvements de la tête n'étaient pas autorisés pendant l'émission du stimulus, des confusions avant-arrière sont apparues. Elles sont définies de la même manière que dans Wightman et Kistler [WK89] : si l'angle absolu entre la direction cible et le jugement est réduit en réfléchissant le jugement par symétrie autour du plan frontal, alors ce jugement est considéré comme une confusion avant-arrière. Les sources proches du plan frontal sont exclues de la détection des confusions, afin de ne pas confondre confusion et dispersion de pointage [CLH97], et concernent les directions à (az, el)=(±80°, −5°) ; (±100°, 26°) ; (±72°, 57°) et (0°, 90°).

Dispersion des jugements La dispersion des jugements est un indicateur de la consistance des réponses. D'après [WK89], le paramètre de dispersion κ^{-1} est estimé a partir de la longueur du vecteur résultant R de la somme des vecteurs unité associés aux N jugements :

$$\kappa = \frac{(N-1)^2}{N(N-R)} \tag{4.1}$$

La dispersion κ^{-1} varie entre 0 (pas de dispersion) et 1 (directions uniformément distribuées sur la sphère) pour N tendant vers l'infini.

Temps de réponse Le temps de réponse correspond au temps écoulé entre l'extinction du stimulus et la réponse du sujet. Ce facteur peut être déterminant pour le choix de la méthode de pointage à utiliser lors de la mise en place d'un test de localisation. Il fournit un indicateur du nombre de réponses qui peuvent être collectées en un temps donné.

Main dominante Etant donné que la méthode proximale implique l'usage des deux mains, nous avons examiné si l'effet de la main utilisée pour pointer (dominante, non dominante) avait un impact sur la précision des réponses. Notons que les sujets ont toujours utilisé la main située du côté de la direction cible. Pour les directions du plan médian cependant, la main utilisée pour pointer une même direction peut varier au sein des répétitions (pour un même sujet).

Analyses ANOVA Afin de quantifier l'impact des différents facteurs sur les performances de localisation (par exemple l'effet de l'hémi-champ avant-arrière), des analyses de variances (ANOVA) sont réalisées (logiciel Statistica). Pour tenir compte de la variabilité inter-sujets, et dans la mesure où chacun a testé un même ensemble de conditions, les analyses sont de type ANOVA à mesures répétées. Les différentes variables dépendantes étudiées ont été listées ci-dessus (erreur horizontale, verticale, signée ou non signée, dispersion, taux de confusion, etc.). Le facteur inter-groupes (aussi appelés facteur catégoriel) est défini par la méthode de pointage (H, G, P), aussi référencée par le terme de condition. Suivant la définition du facteur inter-groupe, les individus ont été affectés aléatoirement à chaque groupe, dont la condition a été définie en amont de l'expérience. Les facteurs intra-groupes (ou facteurs de mesures répétées) sont définis par : l'indice de la répétition, entre 1 et 8 ; l'hémi-champ de la source, "avant" ("*Front*") pour les sources

situées aux azimuts [−80°,+80°] et "arrière" ("*Back*") pour les sources situées aux azimuts [−100°, +100°] ; l'élévation cible (−5°, 26°, 57° et parfois 90°) ; la catégorie d'élévation "moyenne" ("*Mid*"), pour les sources situées à élévation −5° ou 26°, et "haute" ("*High*"), pour les sources situées à élévation 57° ou 90° ; l'angle cible absolu défini par l'angle séparant la direction cible de la direction au repos (az.,el.)=(0°,0°) et enfin le facteur de dominance latérale (main dominante ou non-dominante) utilisé dans le cadre unique de la condition proximale.

Les facteurs testés seront d'autant plus significatifs qu'ils sont responsables d'une quantité importante de variance dans les résultats. Un niveau de p à $p = 0.05$ signifie qu'il existe une probabilité de 5% que la relation entre les variables trouvées dans notre échantillon soient dues à la chance. Suivant les recommandations de Johnson [Joh13], un effet sera considéré significatif si la valeur de p est inférieure à 0.005 et marginalement significatif si $0.005 < p < 0.05$. Des tests post-hoc (ou analyses de contraste) seront réalisés pour interpréter les interactions en utilisant la méthode Tukey-Kramer [MGL10].

4.3.2 Résultats

Tout d'abord, les confusions avant-arrière ont été analysées. D'après nos observations, les jugements qui apparaissent dans l'hémi-champ opposé à la source traduisent un véritable phénomène perceptif plutôt qu'une dispersion dans les jugements : soit le sujet a pointé dans le mauvais quadrant pour toutes les répétitions, soit il s'est trompé de quadrant pour un faible nombre de répétitions (voir figure 3 de l'article en annexe B). Par conséquent, les jugements identifiés comme des confusions avant-arrière ont été corrigés pour la suite des analyses (par symétrie autour du plan frontal).

Etant donné les performances aberrantes d'un participant à la condition H (directions pointées à basses élévations seulement), ses données expérimentales ont été supprimées de l'analyse des résultats. Par conséquent, seules les données de 12 participants sont utilisées pour analyser les résultats de la méthode de pointage avec la tête.

Confusions avant-arrière Une analyse ANOVA à mesures répétées avec les facteurs inter-groupes *condition* (H, G, P) et intra-groupes *répétition* (1 à 8), *hémi-champ* (avant, arrière) et *élévation cible* (−5°, 26°, 57°) est réalisée sur le taux de confusion avant-arrière. L'élévation 90° est exclue de l'analyse étant donné qu'elle n'appartient à aucun hémi-champ (elle est située sur le plan frontal séparant les hémi-champs avant et arrière) et qu'elle n'est pas considérée dans la détection des confusions. Premièrement, l'analyse révèle que les facteurs *condition* et *répétition* ne sont pas significatifs sur le taux de confusions ($F(7, 245) = 1.53$; $p = 0.16$ et $F(2, 35) = 0.46$; $p = 0.63$, respectivement). En effet, les confusions sont causées par un phénomène perceptif et ne sont pas liées à la méthode de pointage. Ce résultat permet de s'assurer que le traitement des confusions n'affectera pas la comparaison des méthodes de pointage. Deuxièmement, le taux de confusion apparaît significativement plus élevé à l'arrière qu'à l'avant ($F(1, 35) = 63.54$; $p < 0.001$) ce qui signifie que les confusions sont principalement des confusions arrière-avant. Troisièmement, le facteur *élévation cible* a un impact significatif sur le taux de confusions ($F(2, 70) = 49.75$; $p < 0.001$) et son interaction avec le facteur *hémi-champ* est également significatif ($F(2, 70) = 52.07$; $p < 0.001$). Le test post-hoc associé révèle que la différence significative entre les hémi-champs avant et arrière n'est valable que pour l'élévation 57°, comme on peut l'observer figure 4.5.

Erreurs horizontales Une analyse ANOVA à mesure répétées est réalisée sur les erreurs horizontales non signées avec le facteur inter-groupes *condition* (H, G, P) et les facteurs intra-groupes *répétition* (1 à 8) et *hémi-champ* (avant, arrière). Premièrement, le facteur *répétition* et son interaction avec le facteur *condition* ne sont pas significatifs ($F(7, 245) = 1.40$; $p = 0.21$ et $F(14, 245) = 1.08$; $p = 0.37$, respectivement). Deuxièmement, le facteur *condition* a un impact significatif ($F(2, 35) = 11.63$; $p < 0.001$). La condition de pointage proximale présente en effet des erreurs horizontales non signées significativement plus élevées que les autres conditions (P vs. H : $p = 0.004$; P vs. G : $p < 0.001$; G vs. H : $p = 0.58$). Cependant, si l'on s'intéresse au test post-hoc associé à l'interaction entre les facteurs *condition* et *hémi-champ* ($F(2, 35) = 6.37$; $p = 0.004$), on remarque que ce n'est vrai que pour l'hémi-champ arrière. Le facteur *hémi-champ* a par ailleurs un effet significatif ($F(1, 35) = 167.95$; $p < 0.001$).

Une analyse avec les même facteurs est menée sur les erreurs horizontales signées. Tout d'abord, le biais horizontal présente une réduction marginalement significative ($\approx 2°$) avec l'indice de la répétition ($F(7, 245) = 2.93$; $p = 0.006$). L'interaction non significative entre les facteurs *répétition* et *condition* ($F(14,245) = 0.69$; $p = 0.78$) met en évidence que la tendance est observée pour toutes les méthodes de pointage. L'effet du facteur *condition* est significatif sur le biais horizontal ($F(2, 35) = 9.91$; $p < 0.001$) avec des biais plus importants pour la méthode proximale que pour les autres méthodes (P vs. H : $p = 0.005$, P vs. G : $p < 0.001$). Ici encore, ce n'est vrai que pour l'hémi-champ arrière, comme le révèle le test post-hoc associé à l'interaction entre les facteurs *hémi-champ* et *condition* ($F(2, 35) = 6.90$; $p = 0.003$). La figure 4.6 illustre cette interaction. On observe qu'à l'avant, le biais horizontal est nul pour la condition proximale et positif pour les autres méthodes. A l'arrière, le biais est négatif pour toutes les méthodes, avec une tendance plus marquée pour la méthode proximale.

Erreurs verticales Une analyse ANOVA est réalisée sur les erreurs verticales non signées avec le facteur inter-groupes *condition* (H, G, P) et les facteurs intra-groupes *répétition* (1 à 8) et *catégorie*

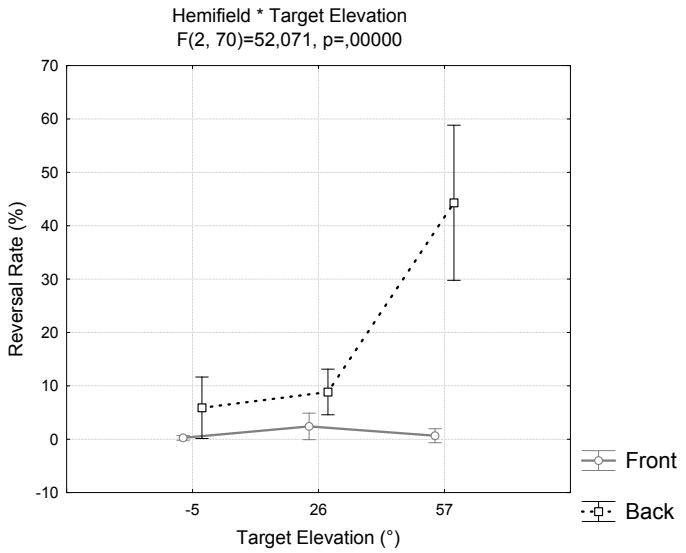


FIGURE 4.5 – Taux de confusions avant-arrière (%) moyen observé sur tous les sujets pour les élévations $-5°$, $26°$ et $57°$, et pour les hémi-champs avant (trait gris) et arrière (trait noir pointillé). La significativité de l'interaction entre les facteurs *hémi-champ* et *élévation* est visible en haut de la figure par les valeurs de F et p. Les barres verticales représentent les erreurs-types.

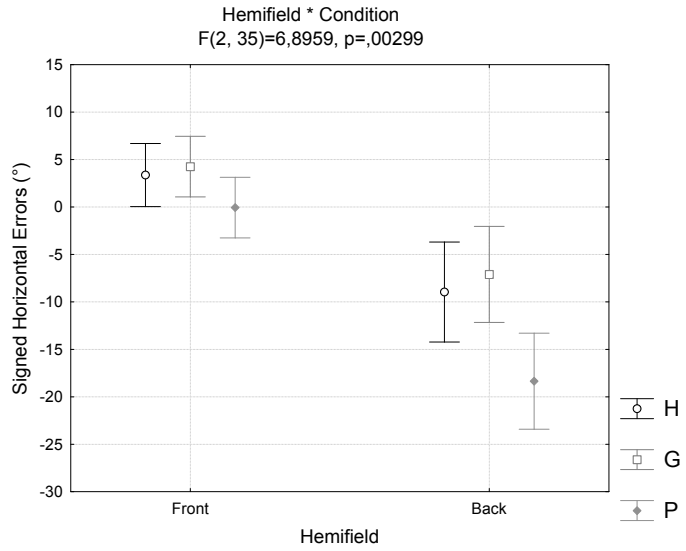
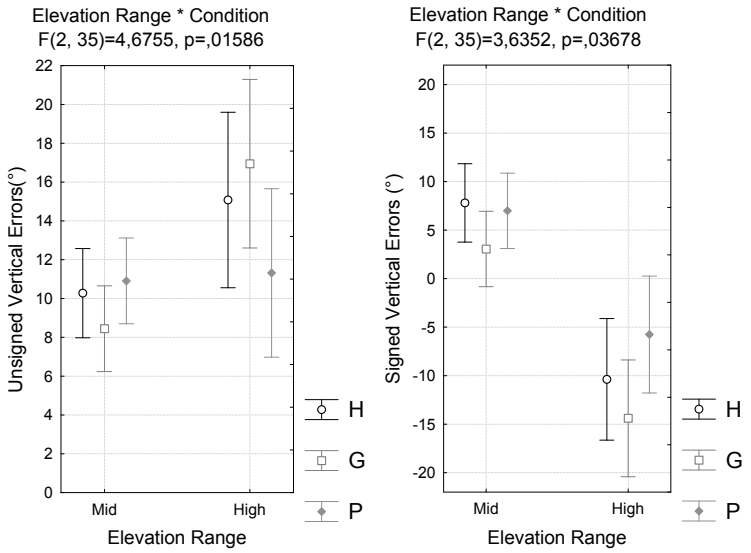


FIGURE 4.6 – Erreurs horizontales signées, moyennées sur les répétitions et les sujets, au sein de chaque condition de pointage (H, G P). Les résultats sont présentés séparément pour les hémi-champs avant et arrière (*Hemifield : Front, Back*).

d'élévation ("moyenne", "haute"). Le facteur *répétition* et son interaction avec le facteur *condition* ne sont pas significatifs ($F(7, 245) = 1.67$; $p = 0.12$ et $F(14, 245) = 1.50$; $p = 0.11$, respectivement). Les erreurs verticales non signées ne se distinguent pas significativement entre les conditions ($F(2, 35) = 1.35$; $p = 0.27$) mais sont significativement plus importantes aux élévations "hautes" ($F(1, 35) = 17.37$; $p < 0.001$). D'après le test post-hoc associé à l'interaction entre les facteurs *condition* et *catégorie d'élévation* ($F(2, 35) = 4.68$; $p = 0.02$), cela n'est vrai que pour la condition de pointage au pistolet ($p < 0.001$). Comme on peut le voir dans la figure 4.7(a), la méthode de pointage proximale ne présente presqu'aucune différence de précision verticale entre les élévations "moyennes" et "hautes". Par ailleurs, cette méthode offre une précision verticale meilleure que la méthode de pointage au pistolet pour les sources élevées ($p = 0.02$).

La même analyse est réalisée sur les erreurs verticales signées. Le facteur *répétition* n'est pas significatif ($F(7, 245) = 1.55$; $p = 0.15$), contrairement à son interaction avec le facteur *condition* ($F(14, 245) = 4.69$; $p < 0.001$). En effet, la méthode de pointage au pistolet présente un biais négatif croissant avec les répétitions (augmentation jusqu'à $6°$), ce qui peut se traduire par un symptôme de fatigue. Cette méthode montre globalement un biais vertical négatif contrairement aux autres méthodes de pointage dont le biais vertical est proche de zéro (effet *condition* : $F(2,35) = 3.95$; $p = 0.03$). Le facteur *catégorie d'élévation* montre un effet significatif sur les biais verticaux ($F(1, 35) = 322.93$; $p < 0.001$) avec des biais positifs de l'ordre

de 6° pour les élévations moyennes contre des biais négatifs de l'ordre −10° pour les élévations "hautes". Son interaction avec le facteur *condition* est marginalement significatif ($F(2, 35) = 3.64$; $p = 0.037$) : la méthode proximale présente un biais vertical moins prononcé que la méthode de pointage au pistolet pour les élévations "hautes" ($p = 0.015$), comme illustré figure 4.7(b).



(a) (b)

FIGURE 4.7 – Erreurs verticales non signées (a) et signées (b) pour chacune des conditions de pointage (H, G, P), en fonction de la catégorie d'élévation "moyenne" ("*Mid*" : el. −5 et 26°) ou "haute" ("*High*" : el. 57° et 90°).

Dispersion des réponses Une analyse ANOVA est réalisée sur les valeurs de dispersion κ^{-1} (calculées selon l'équation 4.1) avec le facteur inter-groupes *condition* (H, G, P) et les facteurs intra-groupes *élévation cible* (−5°, 26°, 57°) et *hémi-champ* (avant, arrière). Le facteur *condition* n'a pas d'effet significatif sur la dispersion des jugements ($F(2, 35) = 0.75$; $p = 0.48$). La dispersion est significativement plus élevée à l'arrière qu'à l'avant ($F(1,35) = 40.05$; $p < 0.001$) et augmente significativement avec l'élévation ($F(2,70) = 16.94$; $p < 0.001$). L'interaction entre les facteurs *condition* et *élévation cible* est marginalement significative ($F(4, 70) = 2.72$; $p = 0.04$). Le test post-hoc associé révèle que cette augmentation significative de dispersion entre l'hémi-champ avant et arrière n'est associé qu'aux méthodes H et G. De plus, l'interaction croisée entre les facteurs *condition*, *hémi-champ* et *élévation cible* ($F(4, 70) = 9.43$; $p < 0.001$) révèle une dispersion significativement plus faible pour la méthode proximale à l'élévation 57° dans l'hémi-champ arrière par rapport à la méthode de pointage au pistolet ($p = 0.002$). Cette observation est illustrée figure 4.8. Notons que l'élévation 90° a été exclue de l'analyse étant donné qu'elle n'appartient à aucun hémi-champ. Cependant, une analyse ANOVA réalisée avec les facteurs *condition* et *élévation* (incluant l'élévation 90°) sur les valeurs de dispersion ne montre pas de différence significative entre les méthodes de pointage à cette élévation.

Temps de réponse Une analyse ANOVA est à présent effectuée sur le temps de réponse avec le facteur inter-groupes *condition* (H, G, P) et les facteurs intra-groupes *répétition* (1 à 8) et *angle cible absolu* (13 valeurs d'angles). Le facteur *condition* a un effet marginalement significatif sur le temps de réponse ($F(2,35) = 5.46$; $p = 0.009$) : on note un temps de réponse moyen de 3.5 sec., 3.1 sec. et 2.5 sec. pour les méthodes H, G et P, respectivement. De plus, le temps de réponse décroît significativement avec les répétitions ($F(7 ; 245) = 15.22$; $p < 0.001$) ce qui traduit une baisse du temps de réaction au fur et à mesure du déroulement du test. Cependant, d'après le test post-hoc associé à l'interaction entre les facteurs *répétition* et *condition* ($F(14, 245) = 2.80$; $p < 0.001$), ce n'est vrai que pour les méthodes H et P, comme on peut le voir sur la figure 4.9. Le facteur *angle cible absolu* a un impact significatif sur le temps de réponse ($F(12,420) = 43.54$; $p < 0.001$), i.e. plus la source est éloignée par rapport à la position de référence plus le sujet met du temps à répondre. Le test post-hoc associé à l'interaction avec le facteur *condition* ($F(24,420) = 6.40$; $p < 0.001$) révèle que cette tendance est particulièrement prononcée pour les méthodes de pointage avec la tête et au pistolet, alors que la méthode proximale ne montre pas de tendance claire.

Effet de la main utilisée pour pointer (méthode P) L'analyse présentée ici ne considère qu'une seule méthode de pointage, la méthode proximale, étant donné qu'elle seule est susceptible d'impliquer l'usage d'une main ou de l'autre. Par conséquent, le facteur inter-groupes *condition* est éliminé. Une ANOVA à mesures répétées est réalisée avec les facteurs *dominance latérale* (main dominante ou non

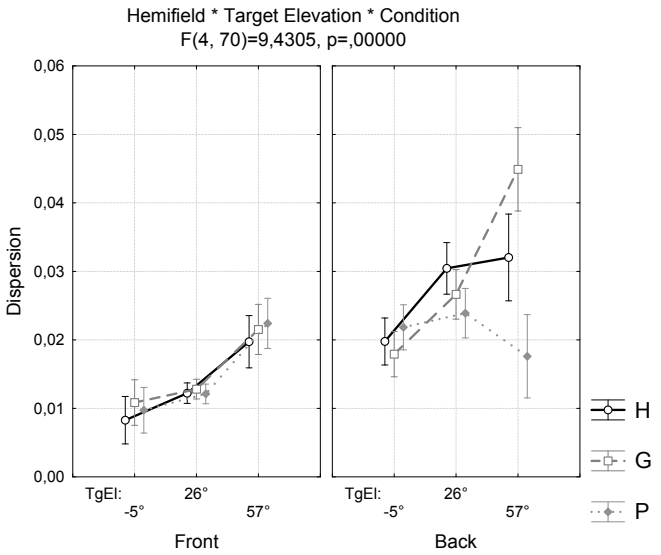


FIGURE 4.8 – Dispersion des jugements en fonction de la condition (H, G, P), de l'hémi-champ avant (à gauche) ou arrière (à droite), et de l'élévation cible (*TgEl :* 5°, 26°, 57°).

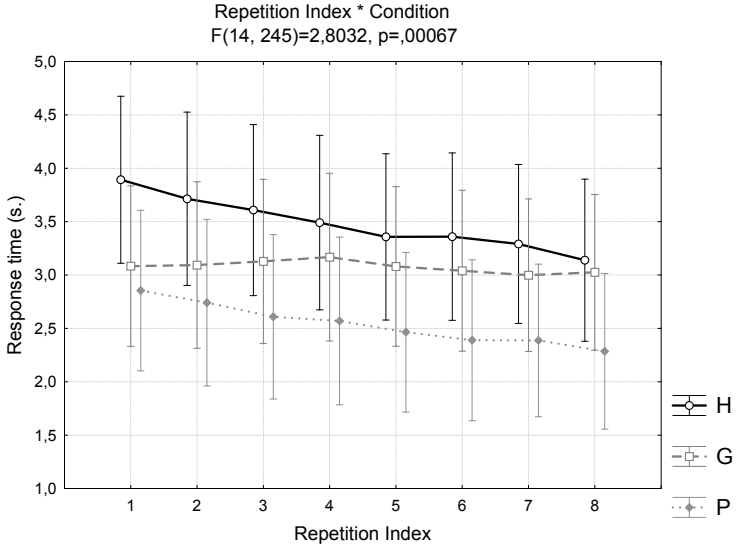


FIGURE 4.9 – Evolution du temps de réponses (en secondes) à travers les 8 répétions d'une même source sonore. Les données sont affichées séparément pour chaque condition de pointage (H, G, P).

dominante) et *hémi-champ* (avant, arrière). Etant donné que les effets du facteur *dominance latérale* ne sont pas significatifs sur les erreurs verticales signées et non signées ($F(1, 12) = 0.24$; $p = 0.64$ et $F(1, 12) = 0.37$; $p = 0.55$, respectivement), l'analyse se focalise sur les erreurs horizontales. L'effet du facteur *dominance latérale* est significatif sur les erreurs horizontales non signées ($F(1,12) = 14.78$; $p = 0.002$) de même que l'effet du facteur *hémi-champ* ($F(1, 12) = 103.4$; $p < 0.001$), comme mentionné plus haut. L'interaction de ces deux facteurs n'est pas significative ($F(1, 12) = 2.06$; $p = 0.18$). Concernant les erreurs signées, le facteur *hémi-champ* est significatif ($F(1, 12) = 188.6$; $p < 0.001$) mais le facteur *dominance latérale* et l'interaction des deux facteurs ne sont pas significatifs ($F(1, 12) = 4.34$; $p = 0.06$ et $F(1, 12) = 0.81$; $p = 0.39$, respectivement). L'utilisation de la main non-dominante affecte donc seulement la précision horizontale, sans présenter de biais particulier, et indépendamment de hémi-champ considéré. Les erreurs horizontales non signées moyennes pour la main dominante et non-dominante diffèrent de 2.5° (avec des valeurs moyennes de 12.3° et 14.8° respectivement).

4.4 Discussion

L'objectif premier de cette étude était de comparer les mérites de la méthode proximale en comparaison avec les méthodes plus conventionnelles de pointage avec la tête et avec un pistolet en plastique. Etant donné le faible nombre de directions testées (au nombre de 24), il est tout d'abord essentiel de vérifier la fiabilité de nos résultats en les comparant directement à ceux de la littérature.

4.4.1 Comparaison avec la littérature

Tout d'abord, une comparaison directe de nos résultats avec ceux de Makous et Middlebrooks [MM90] est rendue possible par : (1) l'utilisation commune de la méthode de pointage avec la tête ; (2) la distribution tridimensionnelle des sources sonores et la proximité des directions testées ; (3) la définition identique des erreurs de localisation. La principale différence réside dans la phase d'entraînement intensif de leurs participants. Les résultats de Gilkey et al. [GGE+95] peuvent également être intégrés à la comparaison, bien que leur étude repose sur la méthode de report GELP (exocentrique).

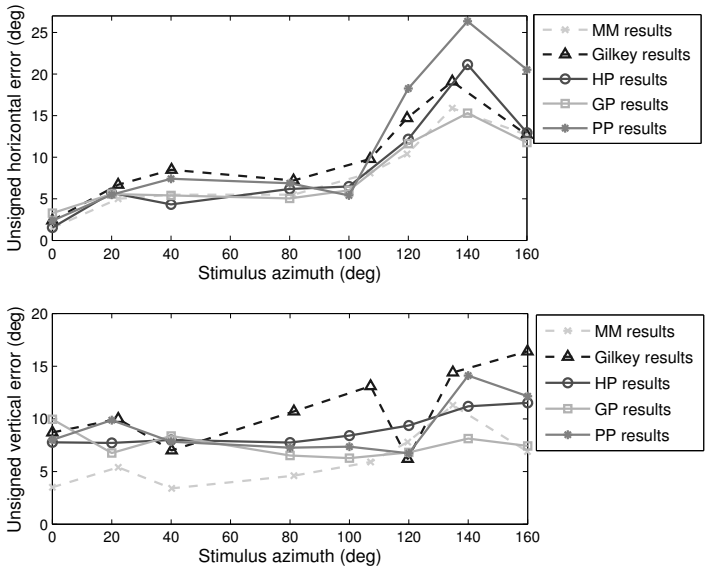


FIGURE 4.10 – Comparaison de nos résultats (pour les 3 conditions de pointage H, G, et P) avec ceux de Makous et Middlebrooks [MM90] et Gilkey et al. [GGE+95] représentés en traits pointillés. Les figures du haut et du bas présentent respectivement les erreurs horizontales et verticales non signées en fonction de l'azimut (en degrés). Les résultats ont été moyennés sur les hémi-espaces gauche et droit et les confusions ont préalablement été supprimées.

La figure 4.10 présente les erreurs horizontales et verticales non signées obtenues dans ces deux études en comparaison de la présente expérience, pour les 3 méthodes de pointage testées. Une tendance très similaire est observée avec des erreurs qui augmentent dans l'hémi-champ arrière, en particulier pour les erreurs horizontales entre les azimuts 110° et 160° (observation en accord avec [OP84]). On remarque qu'en moyenne, nos résultats se situent entre ceux des deux études. Les erreurs relevées par Gilkey et al. [GGE+95] sont globalement supérieures à celles de Makous et Middlebrooks [MM90]. Cette disparité confirme l'avantage des méthodes égocentriques par rapport aux méthodes exocentriques en termes de précision, comme précédemment mentionné dans [GGE+95]. Les erreurs horizontales observées pour la méthode de pointage avec la tête et au pistolet correspondent parfaitement aux résultats de Makous et Middlebrooks [MM90], excepté pour la méthode de pointage avec la tête à azimut 140°. Ces observations permettent de s'assurer de la fiabilité de nos résultats, d'autant que nos sujets ont été soumis à une durée d'entraînement très limitée (10 directions testées en approximativement 2 minutes) en comparaison de l'étude de Makous et Middlebrooks (1000 directions, jusqu'à deux heures d'entraînement). La principale exception à ce comportement réside dans des erreurs horizontales plus élevées pour la méthode proximale dans l'hémi-champ arrière (à partir de l'azimut 120°).

En section 4.3.2, nous avons noté un biais horizontal positif dans l'hémi-champ avant et négatif dans l'hémi-champ arrière. Cette tendance à sur-latéraliser les sources est en accord avec la littérature [OP84, CLH97, MGL10]. De plus, le biais vertical positif observé pour les sources élevées à l'arrière avait déjà été relevé par Oldfield et Parker [OP84].

Le taux de confusions avant-arrière s'est avéré particulièrement important pour les sources élevées à l'arrière. Cette observation avait déjà été faite dans [MM90] et [CLH97]. De manière plus générale, nous

avons remarqué des taux de confusions plus importants à l'arrière qu'à l'avant traduisant une majorité de confusions arrière-avant. Ce résultat est en accord avec l'étude de Bronkhorst [Bro95] qui repose sur la méthode de pointage avec la tête et une localisation à l'aveugle. Cet auteur a mis en évidence que cette prédominance de confusions arrière-avant était spécifique à la localisation de sources réelles, le phénomène étant majoritairement représenté par des confusions avant-arrière et diagonales dans le cas de sources virtuelles (i.e. présentés au casque).

Une autre comparaison peut être menée avec les résultats de Djelani et al. [DPSB00], qui ont comparé les méthodes GELP, de pointage avec la tête et avec le doigt. Ces auteurs ne fournissent cependant que des valeurs globales d'erreur d'angle absolu et de dispersion (κ^{-1}) moyennées sur 12 directions et 8 sujets. L'observation selon laquelle les méthodes égocentriques (pointage avec le doigt ou avec la tête) sont plus précises que la méthode GELP, est en accord avec nos résultats. Cependant, leurs valeurs de dispersion sont bien plus grandes (entre 0.04 et 0.08, dépendamment de la méthode de report) que les nôtres (entre 0.01 et 0.04, dépendamment de la direction et de la méthode). Cette disparité peut s'expliquer par le fait qu'ils ont utilisé des sources virtuelles délivrées au casque alors que nous avons utilisé des sources réelles. De plus, nous avons observé des erreurs horizontales dans l'hémi-champ arrière plus élevées avec la méthode proximale qu'avec la méthode de pointage avec la tête, ce qui n'est apparemment pas le cas dans Djelani et al. [DPSB00]. Nous émettons l'hypothèse que cette différence est liée au fait qu'ils ont utilisé une condition dite "*closed-loop*" (stimulus continu pendant la phase de pointage) pour les méthodes GELP et de pointage avec le doigt, contre une condition dite "*open-loop*" (stimulus éteint avant la phase de pointage) pour la condition de pointage avec la tête. Au contraire, nous avons appliqué la condition "*open-loop*" pour toutes les méthodes testées, de manière à rendre la comparaison plus rigoureuse. La condition "*closed-loop*" avait été exclue de notre expérience étant donné qu'elle sous-entend la mise en œuvre des indices dynamiques uniquement pour les méthodes impliquant les mouvements du corps, rendant ainsi les performances de localisation difficilement comparables avec la méthode proximale, pour laquelle la tête du sujet reste fixe pendant la phase de pointage. De plus, la tâche de localisation se terminerait toujours par localiser la direction frontale [DPSB00]. Les résultats de Djelani et al. suggèrent cependant que les faiblesses de la méthode proximale, concernant le pointage des sources situées à l'arrière, pourraient être réduites en condition "*closed-loop*", comme observé pour d'autres tâches de la localisation auditive [BBB+13]. Elle pourrait également bénéficier d'une session d'entraînement plus longue.

4.4.2 Bilan sur les méthodes

Les résultats ont montré que la méthode proximale souffrait d'un manque de précision dans la dimension horizontale à l'arrière, proche du plan médian. Ce biais horizontal négatif à l'arrière traduit une difficulté motrice à indiquer ces directions (tâche qualifiée d'inconfortable, dans l'article [DPSB00]). Au contraire, l'augmentation des erreurs horizontales à l'arrière pour les méthodes de pointage avec la tête et au pistolet s'expliquent par une imprécision induite par les larges mouvements du corps.

En comparaison avec la méthode proximale, la méthode de pointage au pistolet a montré une imprécision verticale plus importante pour les sources sonores élevées et une dispersion significativement plus grande pour les sources élevées à l'arrière. Cette observation reflète une difficulté dans la tâche motrice induite par la combinaison de la rotation du corps et des mouvements du bras. Plus généralement, les résultats obtenus avec les méthodes de pointage avec la tête et au pistolet ont mis en évidence une tendance à ne pas pointer assez haut, qui pourrait être liée à la contrainte physique éprouvée en indiquant les sources sonores élevées avec ces méthodes [DPSB00]. Un phénomène de fatigue a d'ailleurs été mis en évidence pour la méthode de pointage au pistolet avec un biais vertical négatif qui croît en valeur absolue au cours des répétitions.

Enfin, la méthode de pointage proximale est caractérisée par un temps de réponse plus court que les autres : il est réduit de 30% et 20% par rapport aux méthodes H et G, respectivement. Ces pourcentages sont divisés de moitié si l'on considère le temps de repositionnement. En effet, la méthode proximale comprenait non seulement le repositionnement de la tête, suivant l'interface présentée figure 4.1 et commun à toutes les méthodes, mais également la procédure supplémentaire de repositionnement des bras sur les accoudoirs (c.f. article en annexe B). Etant donné que (1) le repositionnement des bras ne requiert aucune interface particulière, (2) qu'après quelques essais la tâche devient automatique pour le participant, et que (3) la méthode n'implique pas les mouvements de la tête, on pourrait imaginer alléger cette procédure de repositionnement complexe pour la méthode proximale (par exemple, en allumant l'interface uniquement si besoin). Il est tout de même indispensable de s'assurer du positionnement exact de la tête au moment de l'émission du stimulus, étant donné que nous avons observé de légers mouvements de la tête durant la phase de pointage et que les sources sont disposées physiquement autour de l'auditeur. Cependant, dans le cadre d'un test de localisation de sources virtuelles synthétisées au casque de manière statique (i.e. sans rafraichissement de la position des sources en fonction des mouvements de la tête), on pourrait envisager la suppression de cette procédure de repositionnement.

Pour finir, le temps de réponse associé à une méthode de pointage est à prendre en compte lors de la mise en place d'un test de localisation. Un temps de réponse court permet de réduire la durée de l'expérience, et limite ainsi l'effet de fatigue des participants. De plus, il peut permettre de collecter davantage de réponses dans une période de temps équivalente et de renforcer ainsi les analyses statistiques.

Conclusion

La distribution des erreurs de localisation pour chacune des trois méthodes de pointage testées présentent des tendances similaires : une augmentation progressive des erreurs horizontales avec l'azimut, un biais horizontal négatif à l'arrière et un biais vertical négatif pour les sources sonore élevées. Cependant, ces différentes méthodes égocentriques entraînent le mouvement de différentes parties du corps, dont le contrôle moteur affecte différemment les performances de localisation. L'étude a en effet mis en évidence des limitations spécifiques à chaque méthode de pointage. Pour les sources situées à hautes élévations, les méthodes de pointage avec la tête et au pistolet présentent une plus large dispersion des réponses à l'arrière et des biais verticaux plus importants que la méthode proximale. Pour les sources localisées en proximité du plan médian à l'arrière, la méthode proximale présente un biais horizontal négatif marqué. Cependant, cette dernière méthode possède des avantages pratiques et certaines perspectives d'amélioration peuvent être envisagées pour des expériences futures. La méthode proximale présente un temps de réponse plus court que les autres méthodes, ce qui signifie que davantage de réponses peuvent être collectées pour une même durée expérimentale. De plus, cet avantage pourrait être renforcé par une réduction de la complexité de la tâche de repositionnement. Contrairement aux deux autres méthodes, la méthode proximale peut également être utilisée en condition *closed-loop*, i.e. où le stimulus resterait émis pendant la phase de pointage. Combiné à un entraînement des sujets plus intensif, on peut espérer réduire les biais observés pour les sources arrières, en particulier avec la main non-dominante. Pour finir, il serait intéressant de tester la méthode de pointage proximale dans le cadre d'un test de localisation de sources virtuelles impliquant le report de la distance perçue.

En vertu de ces observations, la méthode de pointage proximale sera utilisée pour la mise en place du test de localisation de sources virtuelles synthétisées au casque (chapitre 7). Ce choix est motivé par le fait que, bien que les réponses relevées pour les sources situées dans l'hémi-champ arrière présentent un important biais horizontal, elles sont caractérisées par une faible dispersion. Il semble donc possible d'envisager la correction de ce biais systématique. Nous proposerons une méthode de correction de ce biais sur les réponses dans le chapitre 6, et qui sera appliquée dans le chapitre 7. Les données de localisation collectées dans cette étude sur sources réelles nous permettront à caractériser ce biais. De plus, la suite du travail nous permettra d'exploiter les bénéfices de cette méthode n'impliquant pas les mouvements du corps. Tout d'abord, nous pouvons envisager la suppression de la phase de repositionnement et donc espérer collecter plus de jugements dans un temps donné. Aussi, l'immobilité de la tête pendant la tâche de pointage, en parallèle d'une synthèse statique des sources virtuelles, permettra une mise en pratique en condition *closed-loop*.

Cette étude sur les méthodes de pointage a permis de guider le choix de la méthode qui sera utilisée lors de la mise en place du test de localisation en binaural au chapitre 7, à savoir la méthode proximale. Nous pourrons dès lors exploiter certains de ses avantages mis en évidence dans ce chapitre. Les données de ce test sur haut-parleurs seront également utilisées pour caractériser le biais et la dispersion de pointage afin d'envisager la prédiction des directions pointées relevées dans le test de localisation de sources virtuelles. Le chapitre suivant s'intéresse à l'approche signal de la localisation auditive sur laquelle repose le modèle de prédiction de la localisation auditive, à savoir la métrique spectrale.

Chapitre 5

Métriques et pertinence perceptive

Nous avons vu que le recours à une mesure de dissimilarité ou similarité objective entre deux filtres HRTFs est fréquent et couvre différents objectifs. La composante spectrale des HRTFs possède à la fois une information fortement individuelle et une information spatiale primordiale pour la localisation auditive. Ce chapitre expose les différentes méthodes de caractérisation de la différence spectrale entre deux HRTFs qui ont été utilisées dans la littérature et offre une première analyse comparative de quelques métriques d'intérêt en termes de capacité à identifier les caractéristiques individuelles pertinentes.

La technique binaurale reproduit les indices de localisation sur lesquels se base l'auditeur pour déterminer la direction du son. Les fonctions de transfert relatives à la tête (HRTFs) contiennent ces indices et sont donc souvent utilisées pour générer du son 3D dans les systèmes audio immersifs. La qualité de la reproduction dépend de la précision avec laquelle sont reproduits les indices de localisation, en particulier spectraux. Cela dit, la localisation est robuste à certaines approximations, fréquentielles et spatiales. Cette faculté est liée à la résolution du système auditif. Les études ayant étudié le degré d'approximation (e.g. ordre des filtres, parcimonie des directions de mesure) qu'il est possible d'atteindre sans générer d'artefacts perceptibles s'accordent difficilement. Cela est principalement dû à la méthode d'évaluation perceptive utilisée pour valider ces modélisations. Certaines études se limitent en effet à une évaluation objective, i.e. réalisée au niveau du signal, ce qui sous-entend l'utilisation d'un critère d'erreur qui peut être très variable. En effet, il n'existe pas de critère d'erreur standard. D'autres études vont plus loin dans l'évaluation en réalisant des tests perceptifs. Ils peuvent concerner des tests de localisation ou bien des tests d'évaluation subjective. Cependant, les précautions à prendre lors de la mise en œuvre d'un test perceptif sont nombreuses, telles que la calibration du casque qui est indispensable pour ne pas risquer d'introduire de distorsions dans le spectre. Aussi, les jugements souffrent souvent d'une large variabilité inter-sujets mais aussi intra-sujet, qui dépend de son niveau d'expertise [SK12]. De plus, les directions indiquées dans les tests de localisation sont inévitablement biaisées par la méthode de report. Comme nous l'avons vu au chapitre précédent, les méthodes égocentriques sont plus précises mais requièrent généralement un système de suivi des mouvements, rendant la mise en œuvre complexe. Les tests d'évaluation subjectifs de la qualité du rendu sont eux rendus difficiles par la définition des attributs à évaluer et de la précision des échelles de notation associées. Etant donné les difficultés sous-jacentes aux tests perceptifs, la définition d'une métrique objective, capable de mettre en évidence des distorsions perceptibles, apparaît indispensable pour de nombreuses applications.

Cette nécessité s'inscrit également dans le cadre de la problématique d'individualisation de HRTFs. En effet, comme nous l'avons vu au chapitre 3, la sélection guidée d'un jeu de HRTFs pour un nouvel individu est facilitée par l'identification de HRTFs représentatives d'un ensemble d'individus. La tâche de classification d'une base de données de HRTFs se base sur une mesure de similarité entre les HRTFs. Celle-ci doit être représentative de la variabilité inter-individuelle. De plus, si la base de données est composée d'un ensemble de mesures issues de différents laboratoires et systèmes de mesure qui ont été préalablement standardisées, la métrique de similarité doit s'affranchir le plus possible de la composante intrinsèque au système de mesure car elle n'est pas pertinente d'un point de vue perceptif. Outre la classification de HRTFs, la sélection guidée d'un jeu d'HRTF à partir d'un nombre de réduit de mesures se base également sur une mesure de similarité entre les HRTFs mesurées sur l'individu et les jeux de HRTFs complets de la base de données aux directions correspondantes. La métrique doit alors décrire au mieux la similarité perceptive entre deux HRTFs.

Enfin, les modèles de prédiction de la localisation auditive se basent sur l'hypothèse que la direction perçue d'une source virtuelle présentée au casque est guidée par une similarité importante entre la HRTF utilisée pour la synthèse et celle de l'auditeur dans cette direction. Pour prédire les directions perçues, ces modèles ont donc recours à des métriques qui doivent révéler le mieux possible les similarités entre les indices de localisation de différentes HRTFs, à partir d'une analyse au niveau du signal.

Ce chapitre présente puis analyse les principales métriques qui ont été utilisées dans la littérature. On ne s'intéresse ici qu'aux métriques basées sur des représentations relatives aux spectres d'amplitude des HRTFs car ils contiennent les indices spectraux pertinents pour la localisation auditive en élévation et présentent une importante variabilité inter-sujets.

5.1 Mode de représentation des HRTFs et critères associés

Les différences de temps d'arrivée de l'onde sonore ainsi que l'enveloppe spectrale de chacune des HRTFs gauche et droite, liée au filtrage fréquentiel résultant des réflexions subies par l'onde sonore au contact des éléments du corps, véhiculent une information spatiale. Les indices spectraux sont très importants pour la localisation en élévation. Ils peuvent être représentés de plusieurs façons : par les spectres d'amplitude des HRTFs, en échelle linéaire ou logarithmique (dB), représentés en fonction des fréquences ou des directions de l'espace, par une dérivée des spectres à l'ordre 1 ou 2, ou encore par les coefficients cepstraux, en échelle linéaire ou Mel. On peut mentionner également les représentations par les poids associés à la décomposition en composantes principales ou encore à la décomposition en harmoniques sphériques. Un lissage spectral ou spatial, tenant compte de la résolution du système auditif, est souvent appliqué pour réduire la complexité des HRTFs aux seules caractéristiques pertinentes pour le système auditif (c.f. section 2.3).

Quantifier la différence entre deux HRTFs est un problème qui se pose dans de nombreuses applications : pour l'évaluation de la modélisation, approximation ou interpolation de HRTFs, pour la sélection de HRTFs non-individuelles, la classification de HRTFs ou plus globalement d'individus, etc. Diverses formulations pour caractériser la différence (ou similarité) entre deux HRTFs, représentées selon un des modes de représentation mentionnés ci-dessus, peuvent alors être utilisées. Ces formulations, que l'on désignera par un terme de critère, recourent principalement à des calculs de différences arithmétiques ou quadratiques, moyennées ou sommées sur les fréquences ou directions de mesure, ou encore par une mesure de l'inter-corrélation.

Cette section présente, pour chaque mode de représentation, un ensemble de critères de distance utilisés dans la littérature pour quantifier la différence entre deux HRTFs α et β (dont chaque filtre est associé à une oreille seulement), ainsi que le cadre dans lequel ils ont été appliqués. Nous préciserons si les métriques citées sont dépendantes d'un gain global ($d(\alpha, \alpha + g) > 0$) et si elles sont asymétriques ($d(\alpha, \beta) \neq d(\beta, \alpha)$).

5.1.1 Spectre d'amplitude linéaire

Le spectre d'amplitude est défini par le module de la HRTF représentée en fréquence :

$$\mathrm{mag}\,(f, \theta, \phi) = |H\,(f, \theta, \phi)| \tag{5.1}$$

où $H\,(f, \theta, \phi)$ est la HRTF à la direction (θ, ϕ).

Nicol et al. (2006) ont utilisé plusieurs métriques associées à cette représentation : la différence quadratique moyenne (MSE), le critère de Fahn et d'Avendano.

- **Moyenne de la différence quadratique (MSE)**

  La différence quadratique moyenne (MSE) entre deux spectres d'amplitude linéaire est définie par :

  $$D_{magMSE}(\alpha, \beta) = \frac{1}{K} \sum_{k=1}^{K} \left(|H^{\alpha}(f_k, \theta, \phi)| - \left|H^{\beta}(f_k, \theta, \phi)\right| \right)^2 \tag{5.2}$$

  où $|H^{\alpha}(f_k, \theta, \phi)|$ est le spectre d'amplitude de la HRTF α et $\left|H^{\beta}(f_k, \theta, \phi)\right|$ est le spectre d'amplitude de la HRTF β à la fréquence f_k et à la direction (θ, ϕ), et K le nombre de points fréquentiels utilisés pour la comparaison.
  Notons que la différence MSE prend en compte une différence de gain global entre les HRTFs α et β.

- **Critère de Fahn**

  Dans l'objectif d'évaluer la qualité des HRTFs interpolées, Fahn et al. [CSL03] ont défini le critère de distance suivant :

  $$D_{magFahn}(\alpha, \beta) = \frac{\sum_{k=1}^{K} \left(|H^{\alpha}(f_k, \theta, \phi)| - \left|H^{\beta}(f_k, \theta, \phi)\right| \right)^2}{\sum_{k=1}^{K} |H^{\alpha}(f_k, \theta, \phi)|^2} \tag{5.3}$$

  Ce critère correspond, à un facteur près, à la distance MSE : elles sont toutes deux définies par la différence quadratique pondérée, soit par le nombre de bin fréquentiels dans le cas de la MSE soit par l'énergie d'une HRTF dans le cas du critère de Fahn. Ce dernier est asymétrique, étant donné qu'il dépend du choix de la HRTF utilisée comme référence.

Outre l'étude [NLBB06], Hugeng et Gunawan [HG10] ont utilisé un critère similaire, exprimé en pourcentage, pour évaluer des HRTFs estimées à partir de la régression linéaire entre des mesures anthropométriques et les coefficients issus de l'ACP.

- **Critère d'Avendano**

Le critère d'Avendano est défini de la façon suivante [ADA99] :

$$D_{magAvendano}(\alpha,\beta) = 10 \cdot \log_{10} \left\{ \frac{\sum_{k=1}^{K} \left(|H^{\alpha}(f_k,\theta,\phi)| - |H^{\beta}(f_k,\theta,\phi)| \right)^2}{\sum_{k=1}^{K} |H^{\alpha}(f_k,\theta,\phi)|^2} + 1 \right\} \tag{5.4}$$

Ce critère correspond au critère de Fahn mais exprime la distance en dB et est également asymétrique. Il a été repris dans l'étude de Nicol et al. (2006), où il a montré un caractère effectivement très similaire au critère de Fahn. De plus, l'étude a révélé qu'il présentait un certain avantage par rapport aux métriques MSE et Durant (présenté ci-dessous) en terme d'erreur de reconstruction à partir d'un nombre réduit de HRTFs. Notons que dans l'étude originale d'Avendano, les auteurs ont appliqué un banc de filtres pour lisser les réponses en fréquence en amont du calcul des distances. Selon ce critère, une différence nulle correspond à une distance de 0 dB et non l'infini, grâce au terme (+1).

- **Rapport signal-à-distorsion (SDR)**

Afin d'évaluer l'erreur d'interpolation en azimut des HRTFs, Zhong et Xie [ZX09] utilisent la métrique suivante :

$$D_{SDR}(\alpha,\beta) = 10 \cdot \log_{10} \left\{ \frac{\sum_{k=1}^{K} |H^{\alpha}(f_k,\theta,\phi)|^2}{\sum_{k=1}^{K} |H^{\alpha}(f_k,\theta,\phi) - H^{\beta}(f_k,\theta,\phi)|^2} \right\} \tag{5.5}$$

Cette définition est très proche du critère d'Avendano et s'exprime également en dB. Le critère SDR a souvent été appliqué aux HRIRs dans des problématiques d'interpolation [TUN99, KD08, LpL11]. La différence est alors effectuée directement sur les valeurs complexes et non sur les modules. Par conséquent, comme le fait remarquer [LpL11], il est très sensible à des différences de temps d'arrivée de l'onde.

5.1.2 Spectre d'amplitude en dB

Le niveau sonore perçu est relié de façon logarithmique à l'intensité acoustique mesurée du signal sonore. Pour cette raison, le spectre d'amplitude des HRTFs est souvent exprimé en dB, une unité plus proche de la perception humaine (voir section 1.2.4.1) :

$$\text{log-mag}\,(f,\theta,\phi) = 20 \cdot \log_{10} |H\,(f,\theta,\phi)| \tag{5.6}$$

- **Moyenne de la différence**

Dans le cadre de l'estimation de HRTFs à partir d'un nombre réduit de directions, Lemaire et al. [LCB+05] et Nicol et al. [NLBB06] ont évalué l'erreur de reconstruction suivant une moyenne de la différence absolue en fréquence (*Log Mean Error, LME*) :

$$D_{LME}(\alpha,\beta) = \frac{1}{K} \sum_{k=1}^{K} 20 \cdot \log_{10} \left(\frac{|H^{\beta}(f_k,\theta,\phi)|}{|H^{\alpha}(f_k,\theta,\phi)|} \right) \tag{5.7}$$

- **Moyenne de la différence quadratique (MSE)**

Dans cette même étude, le critère utilisé pour réaliser la classification des HRTFs repose sur la MSE des spectres d'amplitude en dB (ou *Log-Mean Squared Error*, *LMSE*) :

$$D_{LMSE}(\alpha,\beta) = \frac{1}{K} \sum_{k=1}^{K} \left[20 \cdot \log_{10} \left(\frac{|H^{\beta}(f_k,\theta,\phi)|}{|H^{\alpha}(f_k,\theta,\phi)|} \right) \right]^2 \tag{5.8}$$

- **Racine carrée de la moyenne de la différence quadratique (RMSE)**

  Hu et al. [HZZ+06] ont évalué l'erreur engendrée par l'estimation de HRTFs issues de la régression à partir de paramètres anthropométriques sur la base de la racine carrée de la MSE :

$$D_{RMSE}(\alpha,\beta) = \sqrt{D_{LMSE}} \tag{5.9}$$

  Ce critère a également été employé par [SC05] et [Bre13] pour quantifier l'erreur de reconstruction par l'utilisation d'un nombre réduit de composantes principales issues de l'ACP.

- **Variance de la différence ou critère de Durant**

  Le critère d'erreur utilisé par Durant et Wakefield [DW02] pour évaluer/optimiser une approximation pôles-zéros des HRTFs consiste à mesurer la variance de la différence entre les spectres d'amplitudes en dB :

$$D_{magdBVar}(\alpha,\beta) = \frac{1}{K}\sum_{k=1}^{K}\left\{20\log 10\left(\frac{\left|H^{\beta}(f_k,\theta,\phi)\right|}{\left|H^{\alpha}(f_k,\theta,\phi)\right|}\right) - \overline{d}\right\}^2 \tag{5.10}$$

  où $\overline{d} = \frac{1}{K}\sum_{k=1}^{K} 20\log 10\left(\frac{\left|H^{\beta}(f_k,\theta,\phi)\right|}{\left|H^{\alpha}(f_k,\theta,\phi)\right|}\right)$.
  Avec l'utilisation de la variance, une différence de gain global est négligée. Ce critère d'erreur a notamment été réutilisé par Nicol et al. [NLBB06] et Lee et Lee [LpL11].

5.1.3 Spectre lissé en fréquences ou profil spectral en dB

L'échelle perceptive d'amplitude est logarithmique, d'où l'utilisation de spectres d'amplitude exprimés en dB. Les spectres sont issus de la transformée de Fourier et sont échantillonnés linéairement sur l'axe des fréquences. Cela dit, la résolution fréquentielle du système auditif n'est pas constante et l'échantillonnage en fréquences des HRTFs peut être modifié pour mieux s'approcher de l'échelle perceptive. Comme nous l'avons vu section 1.2.4, le filtrage spectral de la cochlée se modélise par un banc de filtres et résulte en un lissage fréquentiel. Les bancs de filtres utilisés à cet effet sont très variables d'une étude à l'autre, notamment en termes de largeurs de bande et fréquences centrales, du nombre, de la forme et du degré de chevauchement des filtres. Certaines études réalisent même ce lissage fréquentiel sans passer par un banc de filtres, en ré-échantillonnant simplement les spectres d'amplitude en fréquence suivant une échelle logarithmique (voir par exemple [MPC05]). Cela dit, les bancs de filtres gammatone suivant les largeurs de bandes ERB ont montré s'approcher au plus près de la réponse cochléaire [MG96].

Il existe plusieurs façons de procéder pour obtenir une valeur par bande. Certaines études procèdent au filtrage des HRTFs en fréquence puis représentent les niveaux de chaque bande par la moyenne ou la moyenne quadratique des spectres d'amplitude à l'intérieur de chaque bande fréquentielle (voir équation 2.4, [LB02]). D'autres procèdent dans le domaine temporel, et les niveaux par bande sont obtenus par la moyenne quadratique des réponses impulsionelles filtrées, tel que :

$$\tilde{H}(b,\theta,\phi) = 20\cdot\log_{10}\sqrt{\frac{\sum_{i=1}^{L}\left[h(n,\theta,\phi) * h_G(b)\right]^2}{L}} \tag{5.11}$$

où b est la bande fréquentielle considérée, $b = 1\ldots B$, et L le nombre de points de la réponse impulsionnelle $h(n,\theta,\phi)$, $n = 1\ldots L$ convoluée par la réponse impulsionnelle du filtre gammatone $h_G(b)$. L'ensemble des valeurs RMS obtenues par bande sur l'intervalle fréquentiel considéré résulte en un profil spectral en dB qui s'apparente à un spectre d'amplitude lissé. Un exemple peut être visualisé dans la figure 5.1 (2^{eme} ligne). Les deux méthodes exposées (filtrage en fréquence ou en temps) offrent les mêmes profils spectraux à un gain global près.

Notons que les études citées ci-dessous ont utilisé différentes méthodes pour lisser les spectres d'amplitude selon la résolution fréquentielle du système auditif. Cependant, étant donné que l'objectif des traitements est le même, nous n'en présenterons pas les détails. Plus d'éléments peuvent être trouvés dans les articles correspondants.

- **Moyenne de la différence absolue**

  Minaar et al. [MPC05] ont cherché à quantifier les artefacts audibles introduits par l'interpolation des HRTFs à partir d'un nombre réduit de mesures. Pour cela, ils utilisent la moyenne fréquentielle de la différence absolue des spectres ré-échantillonnés logarithmiquement en fréquence :

$$D_{GTLME}(\alpha,\beta) = \frac{1}{B}\sum_{b=1}^{B}\left|\tilde{H}^{\beta}(b,\theta,\phi) - \tilde{H}^{\alpha}(b,\theta,\phi)\right| \tag{5.12}$$

Les auteurs mettent en évidence un lien entre cette mesure objective de l'erreur d'interpolation, et une mesure subjective de l'audibilité des artefacts liés à l'interpolation. Selon cette étude, ce critère aurait une pertinence perceptive.

- **Variance de la différence**

Le terme "ISSD" (*Inter-Subject Spectral Difference*), introduit par Middelbrooks [Mid99a], définit un critère de distance entre HRTFs basé sur la variance de la différence entre les spectres d'amplitude lissés par bandes fréquentielles. Le choix de la variance est motivé par la volonté d'extraire des disparités spectrales indépendantes d'un gain global. Notons que ce critère se distingue du critère de Durant uniquement par l'application du banc de filtres en amont du calcul.

$$D_{GTVar}(\alpha,\beta) = \frac{1}{B}\sum_{b=1}^{B}\left\{d_b - \overline{d}\right\}^2, \qquad d_b = \left[\tilde{H}^{\beta}(b,\theta,\phi) - \tilde{H}^{\alpha}(b,\theta,\phi)\right] \tag{5.13}$$

Dans l'objectif de trouver le facteur d'échelle fréquentiel optimal entre les HRTFs de deux sujets, Middlebrooks [Mid99a] utilise ce critère pour identifier le facteur qui minimise la différence inter-spectres. L'auteur valide son choix en mettant en évidence une corrélation entre l'amplitude de la différence et l'erreur de localisation. En effet, il observe que plus la distance D_{ISSD} est faible, plus le taux de confusions avant-arrière est faible et plus la localisation en élévation est précise.

De façon similaire à Middlebrooks [Mid99a], Guillon [Gui09] cherche à identifier les paramètres optimaux d'ajustement (rotation et *scaling* fréquentiel) entre les jeux de HRTFs de deux individus en termes de minimisation de la différence inter-spectrale (selon le critère d'ISSD). Ce critère de distance a également été repris par Rugeles et al. [FMB14] afin de comparer différentes méthodes de lissage. Ces auteurs mettent en évidence une corrélation entre la distance et la dégradation perçue. Enfin, Andéol et al. [AMS13] ont utilisé ce critère pour quantifier à la fois la force des détails spectraux contenus dans les HRTFs (ISSD entre le spectre étudié et un spectre plat) et la différence spectrale entre une HRTF non-individuelle et une HRTF individuelle. Cependant, et contrairement à Middlebrooks, leurs résultats ne montrent pas de lien entre les erreurs de localisation à l'écoute de HRTFs non-individuelles et l'amplitude de la différence entre la HRTF individuelle de l'auditeur et la HRTF présentée.

- **Déviation standard de la différence**

Plusieurs modèles de localisation [LB02, BML13, BML14, MBL14, MNFC14] ont utilisé la déviation standard de la différence spectrale calculée entre deux HRTFs lissées, en référence à Middlebrooks soit $D_{GTSTD} = \sqrt{D_{GTVar}}$. Langendijk et Bronkhorst [LB02] mettent en évidence l'avantage de cette métrique par rapport à l'inter-corrélation pour la prédiction de la localisation perçue. La déviation standard, comme la variance, s'affranchissent d'une différence globale de gain entre les profils spectraux.

- **Coefficient de l'inter-corrélation normalisée**

Le coefficient d'inter-corrélation normalisée définit la similarité entre 2 HRTFs α et β suivant la formule du taux de corrélation linéaire de Pearson :

$$S_{GTCorr}(\alpha,\beta) = \frac{\sum_{k=1}^{K}\left[\tilde{H}^{\alpha}(f_k,\theta,\phi) - \overline{H}^{\alpha}(\theta,\phi)\right]\cdot\left[\tilde{H}^{\beta}(f_k,\theta,\phi) - \overline{H}^{\beta}(\theta,\phi)\right]}{\sqrt{\left\{\sum_{k=1}^{K}\left[\tilde{H}^{\alpha}(f_k,\theta,\phi) - \overline{H}^{\alpha}(\theta,\phi)\right]^2\right\}\cdot\left\{\sum_{k=1}^{K}\left[\tilde{H}^{\beta}(f_k,\theta,\phi) - \overline{H}^{\beta}(\theta,\phi)\right]^2\right\}}} \tag{5.14}$$

avec

$$\overline{H}^{\alpha}(\theta,\phi) = \frac{1}{K}\sum_{k=1}^{K}\tilde{H}^{\alpha}(f_k,\theta,\phi) \tag{5.15}$$

Par définition, $0 \leq |S_{GTCorr}| \leq 1$. Ce critère est insensible à une différence globale de gain.

Cette mesure de similarité a notamment été utilisé dans les modèles de localisation de Middlebrooks [Mid92], Hofman et Van Opstal [HVO98] et Langendijk et al. [LKW01]. Aussi, dans le but de réaliser la classification de sujets, Xie et al. [XZZ13, XZH15] utilisent ce critère pour évaluer la similarité entre paires de sujets. Pour ce faire, le coefficient est moyenné sur toutes les directions et les 2 oreilles et transformé en critère de distance suivant :

$$D_{GTCorr}(\alpha,\beta) = \sqrt{2\times(1 - S_{GTCorr}(\alpha,\beta))} \tag{5.16}$$

5.1.4 Gradient des profils spectraux

Plusieurs auteurs ont mis en évidence l'importance de certaines caractéristiques spectrales contenues dans les HRTFs, telles que les pics ou les creux, pour la localisation sonore (voir section 1.1.2). Certains émettent l'hypothèse que ce sont les fréquences centrales de ces détails spectraux qui jouent un rôle dans la localisation auditive, étant donné qu'elles se déplacent en fonction de la position de la source sonore. D'autres assurent que ce sont les pentes montantes et descendantes associées à ces caractéristiques qui sont prépondérantes. En se basant sur cette dernière hypothèse, Zakarauskas et Cynader [ZC93] proposent d'extraire les dérivées première et seconde des spectres d'amplitude représentés par bandes de fréquence. De plus, dans une étude neurophysiologique du chat, Reiss et Young [RY05] se sont intéressés au rôle du DCN (Dorsal Cochlear Nucleus) dans l'extraction des indices acoustiques et ont mis en évidence une sensibilité particulière pour les gradients spectraux positifs.

Le gradient est défini par :

$$\tilde{H}'(f,\theta,\phi) = \tilde{H}(f_{k+1}) - \tilde{H}(f_k) \tag{5.17}$$

La restriction au gradient positif s'écrit :

$$\tilde{H}'_{pos}(f,\theta,\phi) = max\left[\tilde{H}(f_{k+1}) - \tilde{H}(f_k), 0\right] \tag{5.18}$$

La figure 5.1 permet de visualiser l'ensemble des étapes qui mènent au gradient et au gradient positif à partir du spectre d'amplitude d'une HRTF donnée. On note que les 3 creux principaux du spectre original aux fréquences 1.1, 2.5 et 9.6 kHz se retrouvent dans le profil spectral aux bandes 4, 11, et 22. Le gradient a pour effet de traduire ces minima locaux par des maxima.

Il s'agit du seul mode de représentation fréquentiel indépendant du gain global.

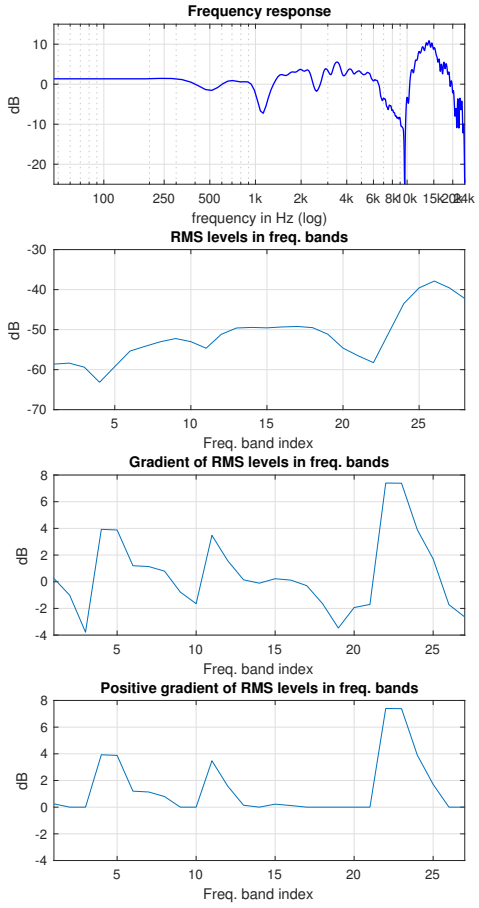


FIGURE 5.1 – (1) Spectre d'amplitude en dB d'une HRTF mesurée à la direction frontale (2) Profil spectral obtenu par l'application d'un banc de 28 filtres gammatone $\in [700, 18000]$ Hz sur le spectre d'amplitude (3) Gradient du profil spectral (4) Gradient positif du profil spectral.

- **Norme L1**

  Dans un modèle de prédiction de la localisation, Zakarauskas et Cynader utilisent la norme L1 des dérivées premières et secondes des profils spectraux :

$$D_{gradL1}(\alpha,\beta) = \sum_{b=2}^{B} \tilde{H}'^{\alpha}(b,\theta,\phi) - \tilde{H}'^{\beta}(b,\theta,\phi) \tag{5.19}$$

  avec $\tilde{H}''$ au lieu de $\tilde{H}'$ pour la dérivée seconde. Selon les résultats des simulations obtenues dans ce papier, la dérivée seconde apparaît plus précise et robuste que la dérivée première.

- **Moyenne de la différence**

  Baumgartner et al. [BML14] proposent un modèle basé sur la moyenne de la différence des dérivées premières positives :

$$D_{posGradMean}(\alpha,\beta) = \frac{1}{B}\sum_{b=2}^{B} max\left[\ \tilde{H}'^{\alpha}(b,\theta,\phi) - \tilde{H}'^{\beta}(b,\theta,\phi),\ 0\ \right] \tag{5.20}$$

5.1.5 Coefficients cepstraux

La représentation cepstrale est principalement utilisée dans le domaine du traitement de la parole et de la reconnaissance vocale automatique (ASR, *Automatic Speech Recognition*) pour dissocier la source du filtre et ainsi extraire les informations relatives à la fréquence fondamentale et aux formants. Plus généralement, cette représentation permet de séparer l'enveloppe spectrale des détails fins du spectre. La décomposition en coefficients cepstraux est également utilisée pour réduire la dimensionnalité des données. En effet, la réduction du nombre de coefficients cepstraux permet d'obtenir une représentation lissée du signal, comme on peut le voir figure 5.2.

Les coefficients cepstraux sont définis par la transformée de Fourier inverse du logarithme de la puissance du spectre d'amplitude (en échelle linéaire) [CSL03] :

$$CC = \mathcal{F}^{-1}\left\{\log\left(|H|^2\right)\right\} \tag{5.21}$$

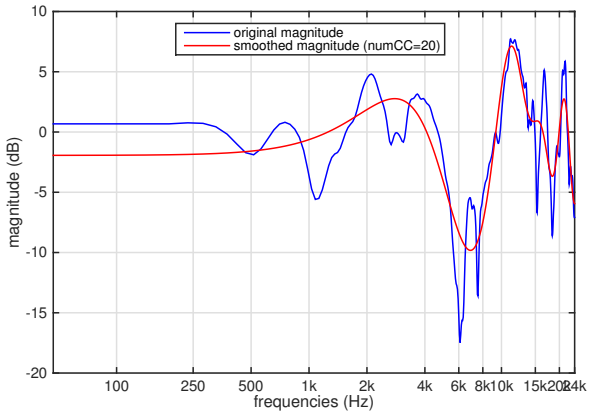


FIGURE 5.2 – Spectre d'amplitude lissé après décomposition en coefficients cepstraux, et recomposition à partir des 20 premiers coefficients.

Cette représentation a été plusieurs fois utilisée pour représenter les HRTFs dans le cadre de travaux de classification [SHH94, CSL03, ARB13].

- **Distance euclidienne (norme L2)**

  Shimada et al. [SHH94] ont réalisé la classification d'un ensemble de HRTFs. Les auteurs réduisent tout d'abord la taille de leur données en décomposant les HRTFs en coefficients cepstraux et en ne conservant que les 16 premiers coefficients. La classification et l'identification de HRTFs prototypiques sont ensuite effectuées sur la base d'une distance euclidienne des cepstres. De plus, ils mettent en évidence un lien entre la mesure objective de distance et les jugements subjectifs en termes de fidélité du rendu spatial.

- **Différence absolue**

  La représentation cepstrale des HRTFs a également été utilisée Fahn et al. [CSL03] pour la mise en évidence de directions représentatives au sein d'un jeu d'HRTF. Lors de la classification, une HRTF représentative est sélectionnée au sein de chaque classe en termes de la minimisation de la distance par rapport aux autres éléments de sa classe. Pour ce faire les auteurs semblent utiliser un simple critère de différence absolue.

- **Inter-corrélation**

  Andreopoulou et Roginska [AR14] ont tenté de mettre en évidence le lien entre une mesure objective de similarité entre HRTFs et des jugements subjectifs sur la qualité du rendu spatial. Pour cela, plusieurs jeux de HRTFs sont classés par ordre de similarité croissante par rapport aux HRTFs d'un sujet donné, sur la base de la corrélation entre leurs 20 premiers coefficients cepstraux [ARB13]. Les auteurs montrent que ce critère est globalement lié à la qualité sonore perçue des HRTFs non-individuelles, en particulier quand il s'agit d'un jugement d'externalisation.

5.1.6 Mel-Frequency Cepstral Coefficient (MFCC)

Les coefficients MFCCs sont équivalents aux coefficients cepstraux à la différence près que le calcul des MFCCs se base sur une échelle fréquentielle perceptive, appelée échelle MEL, présentée section 1.2.4. Pour le calcul des MFCCs, un banc de filtres triangulaires espacés sur une échelle MEL est utilisé. Le spectre d'amplitude est ensuite intégré sur les bandes MEL pour obtenir un spectre $\tilde{H}(b,\theta,\phi)$ par bande b ($b = 1 \dots B$). Les MFCCs sont obtenues par la transformée en cosinus inverse du logarithme de $\tilde{H}(b,\theta,\phi)$ [LpL11] :

$$MFCC(p) = DCT\left[log\left(\tilde{H}\right)\right] = \sum_{b=1}^{B} \log\left(\tilde{H}(b,\theta,\phi)\right) \cos\left[p\left(b-\frac{1}{2}\right)\frac{\pi}{B}\right] \tag{5.22}$$

où B est le nombre de bandes, p est l'ordre du coefficient MFCC, $p = 0 \dots P-1$ avec généralement, $P \leq B$. Comme pour les coefficients cepstraux, l'utilisation d'un nombre réduit de coefficients P permet de s'affranchir de la structure fine de la réponse en fréquence des HRTFs. Le premier coefficient MFCC (ordre 0) quantifie principalement l'énergie contenue dans la HRTF, soit le gain global.

- **MSE**

  Lee et Lee (2011) proposent de quantifier la différence entre deux HRTFs α et β par le calcul de la MSE des $P = 20$ premiers coefficients MFCC :

$$D_{MFCCMSE} = \frac{1}{P}\sum_{p=0}^{P-1}\left(MFCC^{\alpha}(p) - MFCC^{\beta}(p)\right)^2 \tag{5.23}$$

  Lee et Lee montrent que cette métrique est pertinente d'un point de vue perceptif (prédit la fidélité du rendu spatial) et qu'elle équivaut au critère de Durant, basé sur une représentation complète du spectre d'amplitude des HRTFs. Ainsi, le profil global (premiers coefficients seulement) de la représentation du spectre d'amplitude par bandes Mel suffirait à extraire les informations pertinentes liées à la localisation auditive.

5.1.7 SFRS

Les SFRS (*Spherical Frequency Response Surfaces*) sont définies par la dépendance spatiale des HRTFs sur la sphère pour une fréquence donnée f.

- **Inter-corrélation normalisée**

  Guillon [Gui09] utilise l'inter-corrélation normalisée des SFRS sur la sphère :

$$S_{SFRSCorr}(SFRS^{\alpha}, SFRS^{\beta}) = \frac{\sum_{i=1}^{M}\left[SFRS^{\alpha}(f,\theta_i,\phi_i) - \overline{SFRS}^{\alpha}(f)\right]\cdot\left(\left[SFRS^{\beta}(f,\theta_i,\phi_i) - \overline{SFRS}^{\beta}(f)\right]\right)}{\sqrt{\left\{\sum_{i=1}^{M}\left|SFRS^{\alpha}(f,\theta_i,\phi_i) - \overline{SFRS}^{\alpha}(f)\right|^2\right\}\cdot\left\{\sum_{i=1}^{M}\left|SFRS^{\beta}(f,\theta_i,\phi_i) - \overline{SFRS}^{\beta}(f)\right|^2\right\}}} \tag{5.24}$$

avec $\overline{SFRS}$ la moyenne spatiale,

$$\overline{SFRS}(f) = \frac{1}{4\pi} \sum_{i=1}^{M} SFRS(f, \theta_i, \phi_i) \tag{5.25}$$

5.1.8 Discussion

5.1.8.1 Analyse des métriques

Nous avons vu que de nombreuses propositions de caractérisations de la différence entre deux HRTFs ont été testées dans la littérature. Le choix de la métrique dépend de ce que l'on veut mettre en évidence. Les deux principaux types d'études ayant recours aux métriques concernent (1) la mise en évidence de distorsions dans le signal reconstruit, interpolé ou modélisé et (2) l'évaluation du degré de similarité entre HRTFs ou plus globalement entre individus pour l'interpolation ou la classification. Nous tentons ici de dévoiler certaines caractéristiques des métriques spectrales à partir des formulations présentées ci-dessus.

Intégration, facteur, exposant et information de gain global La plupart des métriques somment les différences point à point (sur les bins fréquentiels, temporels, les bandes fréquentielles, coefficients ou encore les directions, selon le domaine de représentation) afin de déterminer une valeur unique de similarité ou dissimilarité entre deux HRTFs. Cette somme peut être appliquée à la fois aux dimensions spatiale et fréquentielle, et combinée à droite et à gauche, ce qui permet d'obtenir une valeur globale de distance entre deux jeux de HRTFs complets, traduisant alors la dissimilarité entre deux individus. Elle peut être pondérée soit par le nombre d'échantillons considérés (moyenne) soit par l'énergie d'une des deux HRTFs. Dans ce dernier cas, le critère possède un caractère asymétrique, i.e. qu'il dépend de la HRTF prise pour référence. Nous avons pu noter que la différence point à point est le plus souvent élevée au carré. Les différences sont parfois centrées de manière à retirer l'information relative au gain global, information effectivement peu pertinente en termes de localisation directionnelle ou de distorsion spectrale (intervient cependant dans la perception de la distance). Les critères d'inter-corrélation, de variance ou de déviation standard ainsi que la représentation par le gradient spectral s'affranchissent naturellement d'une différence de gain.

Pondération suivant des considérations perceptives Les moyennes ou les sommes peuvent être pondérées afin d'introduire une information liée à la pertinence perceptive relative des fréquences ou des directions. En fréquence, des études ont proposé d'introduire des poids inversement proportionnels à la résolution fréquentielle du système auditif (i.e. plus élevés en basse qu'en haute fréquence), en particulier lorsque l'échelle fréquentielle est échantillonnée linéairement [HZK99, NLBB06]. Les pondérations associées aux échelles Barks et ERB sont visibles en figure 5.3. Spatialement, la pondération peut concerner (1) la surface d'angle solide représentée par chaque point, (2) des poids plus importants aux directions où la reconstruction du signal est plus critique, comme la direction frontale [RE02], ou (3) relever de la résolution spatiale humaine, avec des poids plus importants pour les directions frontales que derrière ou sur les côtés. Cela dit, le caractère non-uniforme de la résolution fréquentielle du système auditif est le plus souvent intégré à la métrique par un lissage en amont des réponses en fréquences. Aussi, la prise en compte de l'échelle perceptive d'amplitude (logarithmique) directement dans le mode de représentation a l'avantage de quantifier de manière perceptive les différences qui existent. En effet, nous cherchons une métrique dont les valeurs sont proportionnelles aux différences perceptibles. Enfin, l'intervalle fréquentiel considéré doit être défini avec soin. Il doit prendre en compte à la fois la bande passante limitée des éléments composants le système de mesure des HRTFs (généralement dans l'intervalle $[0.2 - 20]$kHz) mais aussi la bande fréquentielle audible (notre perception sonore commence à se dégrader vers 16kHz, parfois à des fréquences inférieures). De plus, la comparaison soit se limiter aux hautes fréquences si l'on ne s'intéresse qu'aux indices spectraux.

Similarité à une rotation près Il est important de noter que la différence point à point n'est pas toujours bien représentative des similarités qui existent entre deux jeux de HRTFs. En effet, la plupart des métriques ne tiennent pas compte des similarités qui existeraient à un décalage fréquentiel près ou à une rotation spatiale près, notamment à cause de la variabilité dans le placement des individus au moment des mesures de HRTFs (en particulier liée à l'inclinaison de la tête).

5.1.8.2 Etudes comparatives de métriques

Parmi les études qui ont été amenées à quantifier la différence entre deux HRTFs, certaines ont essayé de mettre en évidence un lien entre la mesure de dissimilarité objective et une caractéristique perceptive : liée à des artefacts perceptibles [MPC05], liée à la localisation (en termes de qualité, [SHH94, HZK99], d'erreur [Mid99a], de capacité de prédiction [LB02]), ou liée à la qualité sonore perçue (globale ou spécifique à certaines caractéristiques comme l'externalisation [AR14], ou le timbre [HZK99]). Peu d'études ont réalisé une comparaison directe de plusieurs métriques en termes de pertinence perceptive. On peut cependant citer deux études dont les méthodes et résultats sont exposés ici.

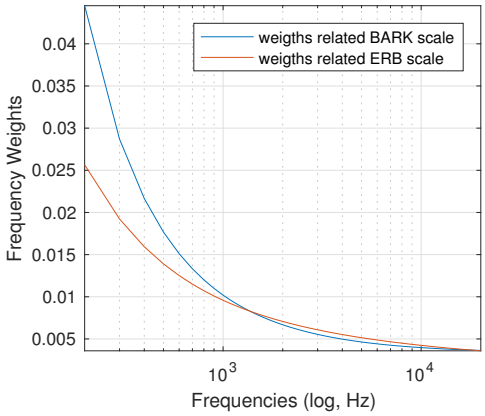


FIGURE 5.3 – Pondération en fonction de la fréquence suivant une échelle perceptive Bark ou ERB.

Dans le cadre d'une étude sur la classification de HRTFs, Nicol et al. [NLBB06] ont comparé 5 critères de distances basés sur le spectre d'amplitude des HRTFs : les critères MSE, Avendano, Fahn, et Durant. La première évaluation menée dans leur étude concerne la capacité des métriques à évoluer de façon proportionnelle à une différence d'azimut dans le plan horizontal. Ils observent que les distances augmentent globalement linéairement avec l'azimut et que les critères de Fahn et d'Avendano atteignent un plateau à partir d'une différence angulaire de 100°. Ces deux critères montrent aussi un plus faible ambitus pour une même différence d'azimut par rapport aux autres critères. Aussi, leurs valeurs croissent plus rapidement que les autres. Compte tenu de ces observations, leur étude montre *a priori* un avantage pour les critères Avendano et Fahn. La deuxième évaluation menée dans leur étude repose sur la classification d'un ensemble de HRTFs (de 23 individus dans 1250 directions) et l'identification de HRTFs représentatives. Les 5 critères sont jugés en termes de minimisation de l'erreur de quantification en fonction du nombre de classes. L'erreur de quantification est identique pour toutes les métriques et définie selon la différence absolue entre les spectres d'amplitude en dB de la HRTF reconstruite et la HRTF originale. Le critère de Fahn est alors abandonné étant donné qu'il donne exactement les même résultats de classification que le critère d'Avendano. Enfin, les critères MSE et Avendano montrent offrir le meilleur compromis entre une faible erreur de reconstruction et un faible nombre de classes. Toutefois, nous savons que la localisation en azimut est principalement guidée par les indices interauraux. En conséquence, l'hypothèse selon laquelle la différence spectrale devrait refléter au mieux une différence azimutale est discutable. Enfin, aucun test d'écoute n'est réalisé pour évaluer ces métriques.

Lee et Lee [LpL11] ont comparé plusieurs critères de mesure des distorsions liées à l'interpolation de HRTFs dans le plan horizontal. Les métriques testées sont les suivantes : critère SDR sur les HRIRs, critère de Durant sur les HRTFs, MSE sur les 20 premières MFCCs, MSE sur les spectres interauraux de phase, différence quadratique d'ITD, MSE sur les ILDs. Elles sont évaluées selon leur capacité à prédire les distorsions perceptibles mesurées au sein d'un test d'écoute. Ce test consiste en une comparaison par paire, où les participants doivent indiquer si les HRTFs interpolées sont perçues aux même positions que les HRTFs de référence, i.e issues de la mesure. Les auteurs notent tout d'abord que le mérite des différentes métriques dépend du type d'HRTF étudié. Dans le cas de HRTFs humaines, les métriques basées sur les MFCCs, l'ILD et les HRTFs montrent les meilleurs résultats. Dans le cas de HRTFs issues d'une tête artificielle KEMAR, c'est la métrique basée sur l'ITD qui se démarque des autres. Les auteurs concluent en recommandant d'utiliser une combinaison des distances basées sur l'ITD et les MFCCs.

5.2 Caractérisation des métriques

Nous menons ici une étude comparative des métriques en analysant leur capacité à extraire les indices spectraux d'intérêt. Les métriques testées ont été sélectionnées à partir des études de la littérature à savoir : les métriques de l'étude de Nicol et al. [NLBB06] ; les métriques ayant suscité un intérêt marqué, en particulier dans les modèles de prédiction de la localisation auditive où elles ont montré de bons résultats [Mid99a, LB02, BML14] ; les métriques basées sur une représentation cepstrale, étant donné les résultats prometteurs des études [LpL11] et [AR14]. L'ensemble des métriques étudiées ici sont répertoriées dans le tableau suivant. Les abréviations employées utilisent la règle [*mode de représentation des HRTFs - critère de distance*].

Au moment d'appliquer ces métriques, nous considèrerons uniquement la bande de fréquence audible et contenant les indices spectraux variables selon les individus (i.e. relatifs à la morphologie) soit l'intervalle fréquentiel $[0.7 - 18]$kHz. Algazi et al. [AAD01] ont en effet montré que les réflexions du torse interviennent à partir de 700Hz (fréquence la plus basse affectée par la morphologie) et 18kHz correspond à peu près à la

Mode de représentation	Critère de distance	Abréviation	Références bibliographiques
Spectres d'amplitude linéaire	MSE	magMSE	[NLBB06]
Spectres d'amplitude linéaire	critère de Fahn	magFahn	[CSL03, HG10]
Spectres d'amplitude linéaire	critère d'Avendano	magAvendano	[ADA99]
Spectres d'amplitude en dB	variance (Durant)	magdBvar	[DW02, LpL11, NLBB06]
Profil spectraux	variance	GTVar	[Mid99a, Mid99b]
Profil spectraux	standard déviation	GTSTD	[LB02, BML13, BML14, MNFC14]
Profil spectraux	inter-corr. normalisée	GTCorr	[Mid92, HVO98, XZZ13, XZH15]
Dérivée 1^{ere} des profils spectraux	moyenne de la diff. abs.	gradMean	[ZC93]
Dérivée 1^{ere} pos. des profils spectraux	moyenne de la diff. abs.	posGradMean	[RY05, BML14]
[1 : 20] premiers coeff. cepstraux	inter-corr. normalisée	cepsCorr	[AR14]
[1 : 20] premiers MFCCs	MSE	mfccMSE	[SHH94, LpL11]

fréquence maximale audible. De cette façon, les métriques basées sur les spectres d'amplitude n'effectuent la comparaison que sur cette bande fréquentielle et les bancs de filtres utilisés pour déterminer les profils spectraux ont des fréquences centrales minimales et maximales de 700Hz et 18000Hz respectivement. Notons que l'application d'un banc de filtres sera réalisée suivant l'équation 5.11.

Nous tentons ici de caractériser ces métriques spectrales. Tout d'abord nous nous intéressons à leur robustesse vis-à-vis du bruit de mesure (i.e. la composante indépendante de l'individu). Pour cela, nous tirons avantage du fait qu'il existe au sein de la base de données BiLi des sujets qui ont été mesurés à la fois à l'IRCAM et à Orange Labs (référencés par l'étiquette IRCAM et ORANGE respectivement). Nous évaluons dans quelle mesure les métriques sont capable d'identifier les jeux de HRTFs correspondant au même sujet, et provenant de systèmes de mesure différents. En effet, une étude a montré que la composante liée au système de mesure est perceptivement non pertinente pour un auditeur devant juger ses propres HRTFs issues de deux bases de données différentes en comparaison à des jeux de HRTFs non-individuels [AK15]. Puis, nous étudions leur capacité à extraire des HRTFs les informations spécifiques à l'individu, ou autrement dit, leur caractère discriminant envers les individus. En effet, les travaux de classification ou d'individualisation nécessitent une métrique qui dégage principalement la composante inter-individuelle. Enfin, nous tentons d'identifier si une métrique est représentative des autres en étudiant les similitudes dans le caractère intrinsèque des métriques.

5.2.1 Standardisation

Dans le cadre de l'étude menée ici, nous utilisons toutes les HRTFs "humaines" de la base de données BiLi mesurées à l'IRCAM (au nombre de 53) ainsi que les HRTFs mesurées à Orange Labs des sujets ayant déjà fait l'objet de mesures à l'RCAM (on compte 10 sujets communs entre les deux systèmes de mesure). Les traitements appliqués aux HRTFs mesurées dans les deux laboratoires ont été appliqués indépendemment par chacun d'entre eux. La principale différence réside dans le fait que les HRTFs IRCAM ont été aplaties en dessous de 200Hz, sans traitement en hautes fréquences, alors que les HRTFs ORANGE n'ont pas de subi de traitement en basses fréquences et ont été filtrées passe-bas à 20kHz. Afin de standardiser les HRTFs, nous appliquons un traitement identique consistant à aplatir les spectres d'amplitude en-dessous de 200Hz et au-delà de 20kHz. Il est vrai que ces post-traitements concernent des bandes fréquentielles qui ne sont pas utilisées par les métriques 1 à 9, basées sur les spectres d'amplitude et les profils spectraux. Cependant, ils peuvent avoir un impact sur les métriques cepstrales, bien que la zone fréquentielle concernée ne contienne pas d'indices spectraux d'intérêt. C'est ce que nous étudions ici.

La figure 5.4 permet de visualiser l'effet d'un tel traitement (filtrage passe-bas à 20kHz) sur les coefficients cepstraux et les MFCCs obtenus. On note que le filtrage passe-bas à un impact très important sur le calcul des coefficients cepstraux alors que les MFCCs sont très peu affectées par ce traitement. Cette différence est due à l'échelle fréquentielle considérée dans chacune de ces représentations. Les coefficients cepstraux considèrent une échelle linéaire alors que les MFCCs considèrent une échelle MEL, plus proche de la perception. Cette échelle permet de délaisser les informations spectrales très hautes fréquences, qui n'ont aucune pertinence perceptive. Pour conclure, la représentation MFCC est plus robuste aux post-traitements appliqués sur les HRTFs en très hautes fréquences.

5.2.2 Robustesse vis-à-vis du bruit de mesure

Les degrés de dissimilarité entre les HRTFs d'un sujet, mesuré deux fois, et tous les autres jeux de HRTFs de la base sont déterminés. Une métrique robuste aux variations imputables aux systèmes de mesure devrait trouver une distance minimale pour la paire de HRTFs IRCAM et ORANGE du même individu mesurés dans les deux bases. Plus le degré de dissimilarité sera faible comparativement à celui considérant deux individus distincts, plus cela signifie que la métrique est robuste et qu'elle extrait les informations pertinentes, capables de discriminer les individus.

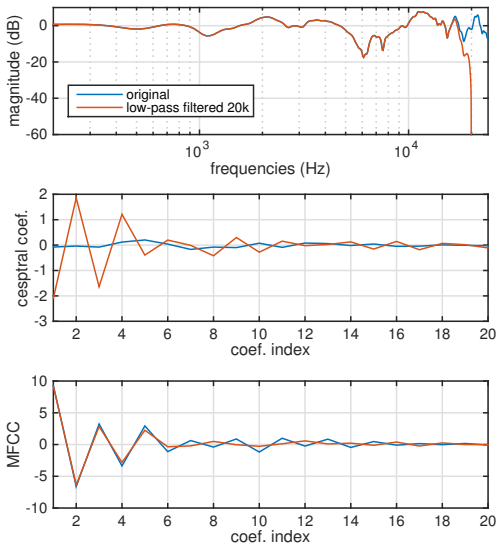


FIGURE 5.4 – Effet du filtrage passe-bas à 20kHz appliqué sur le spectre d'amplitude, sur le calcul des 20 premiers coefficients cesptraux et des MFCCs.

La procédure consiste à comparer le jeu des HRTFs IRCAM de chacun des 10 sujets communs, d'une part avec les autres HRTFs de la base IRCAM (53 sujets) et, d'autre part, avec les 10 jeux de HRTFs de ces mêmes sujets issus de la base ORANGE (soit avec un total de 63 HRTFs). La comparaison est effectuée en chacune des 1500 directions communes entre les deux systèmes de mesure puis moyennée sur toutes les directions. Cela se traduit en une valeur unique pour chacune des comparaison inter-individuelle, ou pour chaque couple de sujets. Enfin, la liste des 63 valeurs de distances est ordonnée par ordre croissant et la procédure est répétée pour chacune des 11 métriques testées. On obtient donc $11 \times 10 = 110$ listes ordonnées. Les valeurs de distance sont brutes, sauf pour le coefficient d'inter-corrélation dont nous prenons le complément à 1 de sorte à retrouver une échelle de dissimilarité croissante, de manière similaire à Xie et al. [XZZ13] : $d_{corr} = (1 - S_{corr})$.

Un exemple de listes ordonnées pour un des 10 sujets et chacune de 11 métriques est présenté figure 5.5. Par définition, la distance entre le jeu de HRTFs de référence, pour lequel on calcule les distances par rapport à tous les autres jeux, et lui-même est égale à zéro. Il s'agit du premier point sur l'axe x. La plupart du temps, les métriques trouvent leur minimum pour le jeu de HRTFs du même sujet mesuré dans la base ORANGE. Quelques exceptions ont par ailleurs été observées. Pour 2 des 10 sujets, aucune métrique n'arrive à identifier le jeu de HRTFs équivalent. Cela met en évidence des disparités importantes entre les HRTFs issues des deux systèmes de mesure pour ces 2 sujets qui peuvent être dues à une différence dans le placement du sujet pendant la mesure. En effet, les deux bases de données utilisent le seul repère de l'axe interaural visualisé à l'aide de deux lasers pointant sur chacun des canaux auditifs des sujets. Aucune référence commune concernant l'inclinaison de la tête (hauteur du nez) n'a été définie. Cela peut générer un décalage en terme d'angle polaire entre les mesures issues des deux systèmes. Une manière d'identifier ce décalage serait de trouver la rotation permettant d'optimiser la corrélation entre les deux jeux de HRTFs, représentés sous forme de SFRS à une fréquence donnée [Gui09]. Une différence de placement des micros dans les canaux auditif pourrait également expliquer ces disparités. Pour deux autres sujets, seules quelques métriques parviennent à identifier son *alter-ego*.

La figure 5.5 présente, sous forme de symboles superposés, les différents sujets communs aux deux bases. Pour ces sujets la mesure IRCAM est symbolisée par une croix et un indice "i", et la mesure ORANGE par un rond et un indice "o". On peut ainsi juger si la métrique les classe immédiatement voisins ou non. On note tout d'abord que les métriques sont très sélectives. En effet, l'évolution des valeurs de distance est peu progressive. On observe que les premiers 50% de l'échelle discriminent les 2 à 4 premiers sujets (dont l'*alter-ego*), tandis que la deuxième moitié de l'échelle sert à classer les autres, avec par conséquent un gradient très faible qui ne permet pas de maintenir leur identité sous forme de voisinage immédiat. Cependant, on peut noter qu'en considérant le classement moyen de chaque sujet l'ordre global est respecté. Par exemple, le sujet 52i-52o (symboles magenta) est systématiquement classé plus proche par l'ensemble des métriques que le sujet 16i-16o (symboles orange). On peut également noter que le sujet 37 est systématiquement classé comme le plus proche.

Cette observation peut être mise en parallèle des résultats subjectifs obtenus par Andreopoulou et Katz [AK15]. Premièrement, ces auteurs ont observé que les jugements d'un sujet à l'écoute de ses HRTFs issus de deux bases différentes (Listen et BiLi) sont fortement corrélés et correspondent à des scores élevés de satisfaction. La présente étude sur le degré de similarité objective présente également cette caracté-

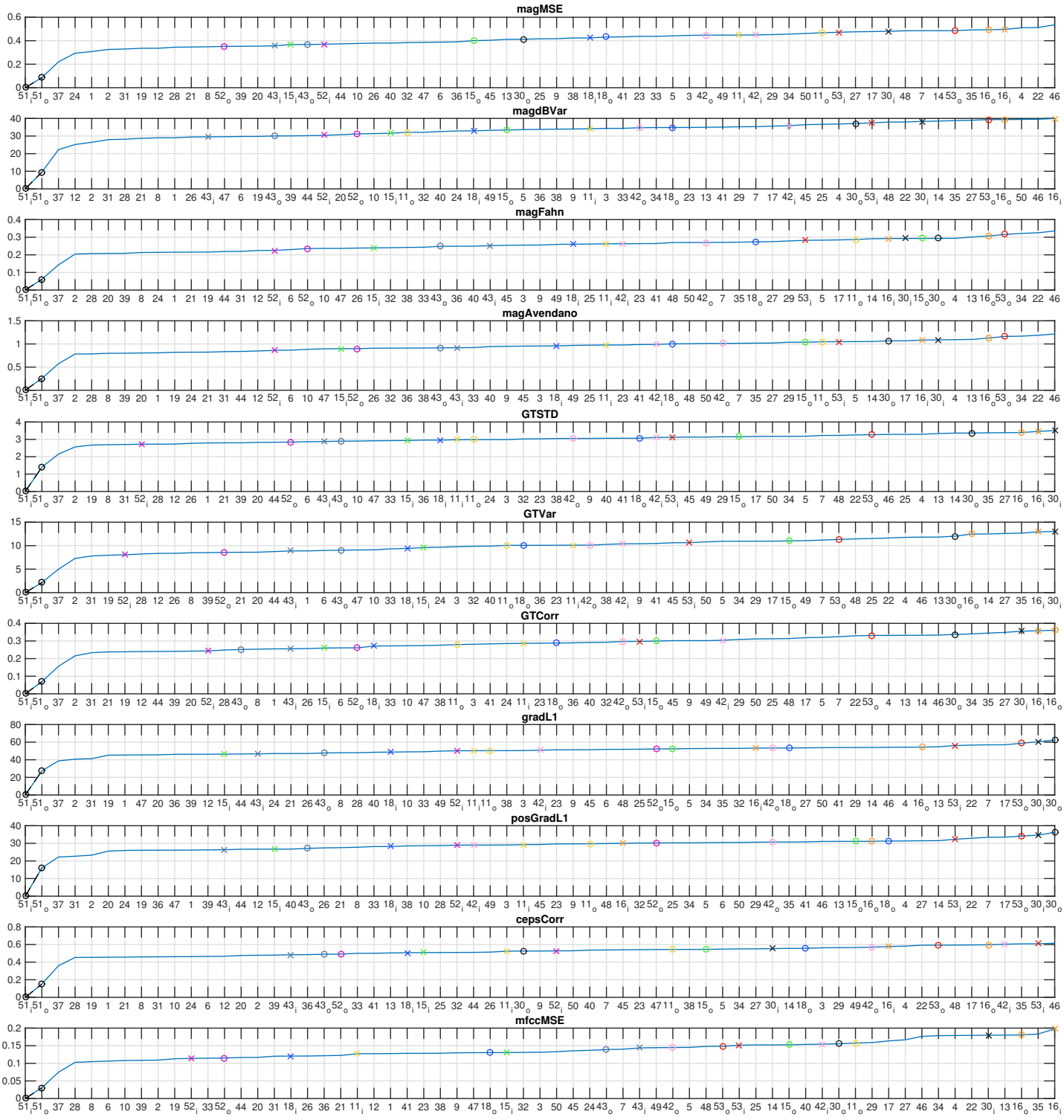


FIGURE 5.5 – Ordonnancement des jeux de HRTFs par ordre croissant de dissimilarité par rapport au jeu de HRTFs IRCAM du sujet 51i. Toutes les métriques testées trouvent une similarité maximale pour le jeu de HRTFs ORANGE du sujet 51o. Les deux jeux de HRTFs de sujets communs sont mis en évidence par les croix et ronds de même couleur, ainsi qu'un indice i et o associé au même numéro (respectivement HRTFs IRCAM et ORANGE).

ristique dans le sens où la similarité maximale apparaît entre les HRTFs d'un même individu mesurées dans deux laboratoires différents. Deuxièmement, dans le cas où un sujet est invité à juger deux jeux de HRTFs d'un même individu mais mesurés dans deux laboratoires, Andreopoulou et Katz observent que la corrélation entre les jugements associés à chacune est bien présente mais moins marquée. Leur interprétation consiste à dire que le degré de qualité spatiale offert par les HRTFs d'un autre sujet prévaut sur les caractéristiques dépendantes de la base de données mais qu'elles ont un impact plus important sur les

jugements de HRTFs non-individuelles. L'étude objective menée ici montre une tendance similaire : les HRTFs non-individuelles issues d'un même individu mais de deux systèmes différents présentent globalement le même degré de similarité avec la HRTF de référence bien que la distance sur l'échelle ne soit pas minimale. Les auteurs suggèrent également qu'une diminution du corpus de HRTFs utilisé pour réaliser l'étude comparative subjective réduirait les différences de jugements observées pour deux jeux de HRTFs non-individuelles issus d'un même individu.

Nous nous intéressons à présent à la corrélation entre les listes ordonnées obtenues pour le jeu de HRTFs IRCAM et le jeu de HRTFs ORANGE d'un même sujet, mesurés dans les deux bases IRCAM et ORANGE. Les métriques qui offrent des listes similaires pour ces deux jeux de HRTFs sont supposées être robustes aux variations inter-mesures. Pour chacun des 10 sujets, le coefficient de Pearson est calculé entre les deux listes obtenues pour chaque jeu de mesure, et ce pour chaque métrique. Les résultats pour chaque sujet sont donnés dans la figure 5.6. On voit que la tendance varie beaucoup entre les sujets étudiés. On remarque que pour le sujet 15 (bleu) les valeurs de corrélations sont très faibles. Il s'agit d'un des deux sujets pour lesquels aucune des métriques ne classe son *alter-ego* comme le plus similaire. Globalement, on note un avantage de la métrique *mag-MSE*.

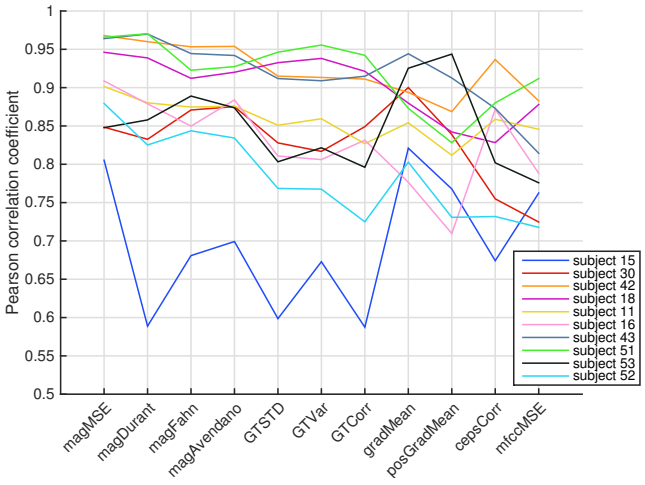


FIGURE 5.6 – Coefficient de corrélation entre les listes ordonnées obtenues à partir des HRTFs IRCAM et ORANGE d'un même sujet, pour chaque métrique. Les résultats sont donnés pour chacun des 10 sujets étudiés (une couleur par sujet, identique à la figure 5.5).

5.2.3 Caractère discriminant entre individus

Nous déterminons ici, pour chaque métrique, le degré de dissimilarité affecté aux différences entre systèmes de mesure par rapport à une dissimilarité inter-individus. Pour ce faire, nous comparons la valeur de distance spectrale minimale, associéa au couple de jeux de HRTFs IRCAM-ORANGE du même individu, par rapport à la valeur maximale de distance spectrale observée avec le reste de la base de données. Cela nous donne une idée de la précision avec laquelle chaque métrique quantifie les différences inter-individuelles.

Nous focalisons notre analyse uniquement sur les sujets communs pour lesquels toutes les métriques ont réussi à identifier un maximum de similarité entre leur deux jeux de mesure (soit 6/10 sujets). La procédure consiste à mesurer le pourcentage du rapport de la distance minimale sur la distance maximale (multipliée par 100). Plus le pourcentage est faible, plus la métrique assigne aux différences inter-individuelles une large portion de son échelle de valeurs, par rapport à la différence inter-mesures (intra-sujet). Le tableau 5.7 permet de visualiser la tendance. On voit que le pourcentage atteint son maximum pour la métrique *GT-STD* et son minimum pour la métrique *mag-MSE*. Ici encore, la métrique *mag-MSE* présente un avantage.

5.2.4 Corrélation entre les métriques

Nous évaluons ici les similitudes dans les comportements des différentes métriques. Pour cela, nous calculons le coefficient de corrélation de Pearson entre les listes ordonnées de HRTFs données par chaque métrique. Les résultats, moyennés sur les 10 sujets d'étude, peuvent être visualisés en figure 5.8. Une corrélation de 0.997 apparaît entre les métriques Avendano et Fahn, ce qui n'est pas étonnant puisqu'elles reposent sur le même calcul (la seule différence est que le critère d'Avendano converti la mesure de distance en dB). Cette observation est en accord avec l'étude de Nicol et al. [NLBB06], où les résultats de classification (représentants de chaque classe) étaient identiques pour ces deux critères. On observe également

Method	Percent ambitus(%)	std(%)
magMSE	17.3286	5.0776
magdBVar	27.9901	6.6865
magFahn	19.0337	4.6618
magAvendano	22.3463	5.1078
GTSTD	43.8805	5.0454
GTVar	20.2186	4.5766
GTCorr	21.8188	3.9494
gradMean	51.4312	5.6972
posGradMean	52.7257	6.2542
cepsCorr	25.6413	7.5991
mfccMSE	21.2530	5.0848

FIGURE 5.7 – Rapport entre la valeur minimale de distance, correspondant à la distance inter-mesures, et la valeur maximale de distance inter-sujets, en pourcentage. Le rapport est calculé pour chacun des 6 sujet pour lesquels les 11 métriques trouvent leur minimum de dissimilarité pour les *alter-ego*.

que la métrique *mag-MSE* est très proche de ces deux métriques. En effet, si on s'attarde sur les équations qui leur sont associées, on voit qu'elles reposent toutes 3 sur une différence quadratique entre les deux spectres HRTF de référence et qu'elles se distinguent à un facteur près. La deuxième association que l'on peut faire au vu de la figure 5.8 concerne les métriques *GT-STD*, *GT-Var* et *GT-Corr* : les métriques *GT-STD* et *GT-Var* sont corrélées à hauteur de 0.993 et la métrique *GT-Corr* est corrélée à 0.973 et 0.968 avec ces deux métriques, respectivement. En effet, ces métriques se basent sur le même mode de représentation. Troisièmement, on note un coefficient de corrélation de 0.98 entre les métriques *grad-Mean* et *posGrad-Mean* qui se basent également sur la même représentation du contenu spectral. Enfin, on voit que les 4 dernières métriques se démarquent globalement des 7 permières (effet encore plus marqué pour les métriques *posGrad-Mean* et *grad-Mean*). Cette observation est due aux différences dans les modes de représentation. En effet, le gradient spectral et la représentation cepstrale se distinguent des représentation plus classiques du spectre d'amplitude, qu'il soit lissé ou non. Pour finir, aucune métrique ne semble être vraiment représentative des autres, bien que les coefficients de corrélation soient relativement élevés (> 0.8).

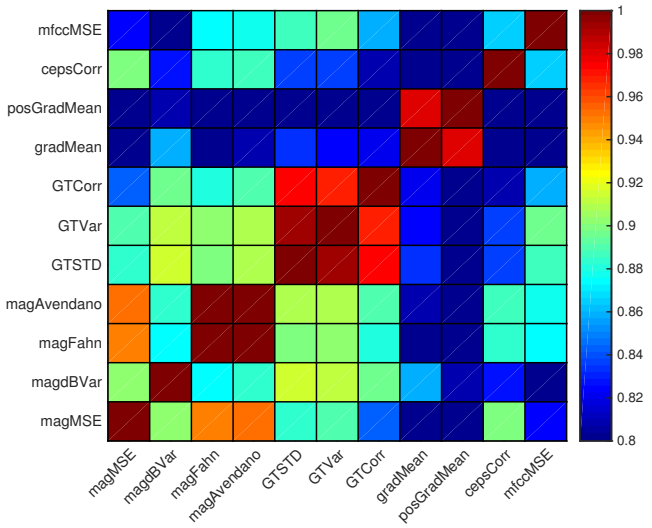


FIGURE 5.8 – Coefficient de corrélation entre les listes ordonnées obtenues pour chaque métrique. Les résultats ont été moyennés sur les corrélation obtenues pour les listes ordonnées de chacun des 10 sujets étudiés.

Conclusion

Ce chapitre a tout d'abord permis d'obtenir une vue d'ensemble des métriques spectrales utilisées dans la littérature et de prendre conscience de la variabilité importante dans la définition des critères. Nous avons ainsi entrevu l'éventail des possibilités et les paramètres qui permettent de considérer différents aspects perceptifs. Les éléments qui définissent les caractéristiques de sensibilité au gain et d'asymétrie ont été mis en évidence. Les études ayant eu recours à l'utilisation des métriques spectrales concernent différents domaines : les problématiques d'interpolation, de modélisation ou d'approximation des HRTFs, les problématiques de classification et d'individualisation et enfin les modèles de prédiction de la localisation auditive. Dans chacun de ces domaines d'application, la pertinence perceptive des métriques objectives utilisées est primordiale en particulier lorsqu'aucun test perceptif n'est mis en œuvre pour valider les méthodes.

L'étude comparative a permis de caractériser la tendance relative de plusieurs métriques et leur robustesse vis-à-vis de la composante liée au dispositif de mesure. Nous avons noté que des disparités importantes peuvent exister entre deux jeux de HRTFs issus de systèmes de mesure différents pour un même individu. Les métriques ont globalement montré une tendance similaire dans la tâche de classification des individus d'une base de données de HRTFs par rapport à un individu de référence. Les corrélations les plus fortes apparaissent pour des métriques basées sur le même mode de représentation. Cela signifie que les caractéristiques d'une métrique spectrale sont principalement définies par la façon dont elles représentent les HRTFs. Enfin, ces premiers résultats suggèrent que la métrique *mag-MSE* permet de quantifier les différences inter-individuelles avec la plus grande précision.

Cette étude n'a cependant pas permis de juger de la pertinence de ces métriques d'un point de vue perceptif, notamment ce qu'elles nous enseignent sur les éventuels artefacts de localisation. Pour aller plus loin dans cette analyse, nous mettons en place un modèle de prédiction de la localisation auditive dans lequel plusieurs métriques seront comparées. Le but ultime de ce travail est d'identifier une métrique proche de la façon dont le système auditif extrait les indices spectraux de localisation.

L'intérêt porté aux métriques spectrales dans nos travaux concerne plus particulièrement la prédiction de la localisation auditive.

L'intérêt porté aux métriques spectrales dans nos travaux concerne plus particulièrement la prédiction de la localisation auditive. Ce chapitre a permis de nous familiariser avec les différentes métriques qui existent et de guider le choix de celles qui seront testées au sein du modèle de prédiction de la localisation auditive.

Chapitre 6

Modèle de prédiction de la localisation auditive

Ce chapitre aborde la problématique de prédiction de la localisation auditive en profondeur, par une introduction aux modèles existants. Il permet de situer notre étude parmi le large domaine de la modélisation de la localisation auditive. Les principes, hypothèses et objectifs de notre modèle ainsi que l'ensemble de ses éléments, de la définition d'une métrique à son exploitation dans le modèle de prédiction, y sont exposés. Les caractéristiques de certains paramètres du modèle, jugés déterminants pour la prédiction, sont expliquées en détails.

La modélisation de la localisation des sources acoustiques à partir de la connaissance des signaux parvenant aux oreilles de l'auditeur peut adopter différents points de vue. Dans une approche traitement du signal audio, le principe est de rechercher des paramètres pertinents accessibles sans ou avec peu de connaissances a priori sur les signaux et l'environnement. L'approche physiologique entend tirer parti des mécanismes de transductions et d'encodage intervenant à différents étages du traitement auditif depuis le système périphérique jusqu'au cortex auditif (cochlée, noyaux cochléaires et complexe olivaire du tronc cérébral, colliculus inférieur, thalamus). L'approche psycho-physique fournit des mesures de performances de l'audition humaine et tente de les relier aux caractéristiques observées sur les signaux [CK05].

Les modèles présentés ci-dessous, ainsi que notre démarche, relèvent simultanément des approches traitement du signal audio et psycho-physiques. Elles recherchent dans les signaux présentés aux oreilles de l'auditeur des informations dont on peut supposer grâce aux résultats psycho-physiques qu'elles sont pertinentes sur le plan perceptif. Ces modèles considèrent d'une part les indices interauraux de temps et d'intensité et d'autres part les indices spectraux monauraux, tous deux encodés dans les HRTFs du sujet. Une distinction majeure existe entre ces catégories d'indices. Les indices interauraux responsables notamment de la latéralisation de la source, sont accessibles sans nécessiter la préconnaissance du signal émis par la source. En revanche, le traitement des indices spectraux monauraux, déterminants pour la localisation en élévation se heurte à l'ambiguïté entre l'association de ces indices à une caractéristique propre du signal émis par la source ou à la transformation subie par l'onde sonore par effet de diffraction sur la tête.

Un série de modèles de prédiction de la localisation dits "*template-based*" (ou "*template matching*") supposent que la localisation est guidée par une ressemblance entre les indices de localisation délivrés par le stimulus et les indices de localisation stockés par l'auditeur (i.e. appris par l'expérience) à la direction perçue. Associées à ce postulat, les deux hypothèses principales sous-jacentes aux modèles sont les suivantes : (1) à travers l'expérience, le système auditif a appris une carte audio-spatiale reliant les indices acoustiques (contenus dans les HRTFs) à chaque direction de l'espace (2) pour localiser, le système auditif analyse les indices acoustiques qui lui sont présentés et utilise cette carte audio-spatiale pour déterminer la direction de provenance du son. La probabilité que l'auditeur localise dans une certaine direction est alors proportionnelle à la similarité entre le stimulus et les HRTFs de l'auditeur à cette direction, relativement aux autres directions de l'espace. L'hypothèse pour que cela fonctionne sans connaissance *a priori* de la source est en effet que le spectre émis par celle-ci soit suffisamment monotone et avec un support fréquentiel suffisament large bande pour que l'auditeur soit encore capable d'y repérer le relief particulier de telle ou telle HRTF.

Les deux principales étapes des modèles de prédiction fonctionnels sont la simulation des traitements physiques et physiologiques qui mènent à la représentation interne du stimulus (sous forme de ses caractéristiques pertinentes pour la localisation auditive dans la dimension étudiée) et le processus de comparaison *target-templates*. Il existe des modèles de prédiction de la localisation à une dimension (prédiction de l'azimut ou de l'élévation perçu) ou à deux dimensions.

Concernant les modèles dont la prédiction de localisation est réalisée en élévation, la première étape consiste à extraire l'information spectrale pertinente pour le système auditif. Cette étape réside généra-

lement en une représentation du spectre d'amplitude par bande de fréquence de manière à modéliser la sélectivité fréquentielle de la membrane basilaire, mais peut être plus complexe. Ces traitements déduits des observations psycho-physiques sont appliqués à la fois sur le son à localiser, appelé "cible" (ou *target*, soit la HRTF de la direction test convoluée avec le stimulus), et les HRTFs de l'auditeur (ou *templates*, large bande) afin d'approximer leur représentation interne (neuronale). Pour des modèles dont la prédiction de localisation est réalisée en azimut, cette première étape consiste à extraire les indices interauraux. Les caractéristiques extraites des HRTFs pour la prédiction sont donc choisies selon les dimensions mises en jeu. Si la prédiction considère l'espace 3D, alors les indices interauraux et spectraux interviennent de manière conjointe. La deuxième étape des modèles de prédiction consiste à comparer les caractéristiques de la cible (spectrale et/ou interaurale) avec celles de chacune des HRTFs de l'auditeur, selon une certaine métrique, afin de déterminer la probabilité que l'auditeur localise en chacune des directions de l'espace de prédiction.

La technique binaurale est susceptible de restituer la localisation des sources avec une précision identique à celle des conditions d'écoute réelles. Cependant lorsque les indices de localisation délivrés par le stimulus ne correspondent pas à ceux appris par l'expérience (indices non-individuels), les performances de localisation se dégradent. Le système auditif localise alors la source virtuelle de manière floue ou ambiguë. Cette observation est variable selon les individus pour un même jeu de HRTFs non-individuel.

Suivant les hypothèses des modèles de prédiction de la localisation, la direction de localisation peut s'expliquer par une ressemblance objective entre les indices acoustiques délivrés par la HRTF non-individuelle et les indices acoustiques propres à l'auditeur. Le modèle que nous nous proposons d'étudier s'appuie sur cette hypothèse. Une des motivations concerne en particulier la comparaison de différentes métriques spectrales de la littérature en termes de capacité à prédire les directions perçues, i.e. à expliquer les processus d'extraction et de reconnaissance des indices spectraux réalisés par le système auditif pour localiser. Pour mener cette évaluation, cet objectif suppose d'avoir préalablement mesuré les HRTFs des auditeurs dont les réponses à un test de localisation seront analysées. Dans un second temps, une fois avoir identifié la ou les métriques les plus à mêmes de prédire la localisation perçue, on peut envisager traiter le problème inverse. A savoir, trouver au sein d'une base de données de HRTFs celles qui expliquent le mieux les directions désignées par un auditeur dont on ne connait pas les HRTFs et soumis à un test de localisation utilisant des HRTFs quelconques.

Nous présentons dans un premier temps les modèles de prédiction de la localisation développés dans la littérature puis étudions les détails des méthodes adoptées par les modèles ayant fait leurs preuves vis-à-vis à de données réelles. Puis, nous proposons un nouveau modèle de localisation auditive pour la prédiction des directions perçues à l'écoute de stimuli synthétisés avec des HRTFs individuelles et non-individuelles.

6.1 Etat de l'art des modèles de prédiction

Les modèles auxquels nous nous intéressons ici reposent sur les hypothèses présentées en introduction. Il s'agit de modèles basés sur une approche fonctionnelle plutôt que statistique, i.e. faisant appel aux algorithmes d'apprentissage tels que les réseaux de neurones artificiels (voir par exemple [NYS92, JSC00]). En effet, dans de tels modèles, l'interprétation des paramètres sous-jacents n'est pas triviale et ils sont donc moins adaptés à la compréhension des mécanismes de localisation auditive. De plus, nous ne présentons que les modèles mettant en œuvre la prédiction de la localisation en élévation, i.e. utilisant une métrique spectrale. Un état de l'art plus général des modèles de prédiction de la localisation peut être trouvé dans [CK05].

L'étude de Middlebrooks [Mid92] est l'étude pionnière dans la réalisation de modèles de prédiction fonctionnels de la localisation auditive. Cette étude s'intéresse à la localisation de sources sonores réelles à bande fréquentielle étroite et centrée sur différentes fréquences supérieures ou égales à 6 kHz. L'auteur observe que, pour ce type de stimulus, la localisation en azimut peut être assez précisément prédite par l'indice d'ILD. Puis, il remarque que l'élévation perçue dépend non pas de la position de la source mais de la fréquence centrale du stimulus et qu'elle varie d'un auditeur à l'autre. La direction perçue semble en effet être guidée par une similarité spectrale entre le spectre cible et les HRTFs de l'auditeur, quantifiée en termes d'inter-corrélation spectrale. De ces observations est dérivé un modèle prenant en compte d'une part l'ILD et les indices spectraux. Ce modèle se montre capable de prédire les secteurs de l'espace 3D où les jugements de localisation sont les plus susceptibles d'apparaître. Ce modèle, à l'origine de nombreux modèles plus récents, sera détaillé dans la section 6.2.1.

Dans un autre article, Middlebrooks [Mid99b] adapte son modèle à la prédiction de localisation de sources virtuelles large bande synthétisées par des HRTFs non-individuelles brutes et adaptées à l'auditeur par un facteur d'échelle fréquentiel. La localisation en azimut est analysée séparément de la localisation en élévation. Dans un premier temps, l'étude montre que la source est localisée dans le plan sagittal de l'auditeur dont l'ITD correspond à l'ITD encodé dans le signal cible. Dans un deuxième temps, la prédiction

se focalise sur l'élévation perçue. Les résultats montrent que la direction perçue apparaît à la direction (dans ce plan sagittal) où la HRTF individuelle est la plus proche du spectre cible, i.e. où la variance de la différence inter-spectrale est minimisée. Dans plusieurs cas, deux minima locaux apparaissent sur le plan sagittal, permettant d'expliquer les jugements associés à des confusions avant-arrière ou haut-bas.

Le modèle de localisation d'Hofman et Van Opstal [HVO98] se base sur celui de Middlebrooks [Mid92] en apportant une caractéristique temporelle dans l'estimation de l'élévation perçue. L'étude s'intéresse aux effets des caractéristiques spectro-temporelles sur la localisation auditive de sources sonores réelles. La première observation est que l'élévation perçue varie avec le contenu spectro-temporel de la source sonore alors que la localisation en azimut est précise et invariante. L'azimut est alors supposé être extrait des indices interauraux et le modèle se focalise sur la prédiction en élévation. La seconde observation suggère que l'estimation de l'élévation par le système auditif évolue à mesure qu'il reçoit des informations acoustiques. L'étape de comparaison consiste alors à intégrer le spectre cible sur une fenêtre temporelle de quelques millisecondes et à le comparer avec les HRTFs de l'auditeur sur la base du coefficient de l'inter-corrélation normalisée. Puis, l'élévation perçue est estimée à partir de la moyenne des coefficients de corrélation "à court terme". Celle-ci est supposée apparaître dans le voisinage du maximum d'inter-corrélation. La décision doit cependant tenir compte de la consistance des estimations en fonction du temps et de la première impression (*initial percept*). Les résultats du modèle ne sont pas directement confrontés aux données expérimentales.

L'étude de Zakarauskas et Cynader [ZC93] tente d'identifier quelles quantités doivent être extraites des indices spectraux pour que la prédiction de la localisation soit précise. Les auteurs simulent l'effet du spectre de la source sur la capacité de deux métriques à prédire un maximum à la direction test. Les deux stimuli testés sont des sons naturels : le bruit d'un papier déchiré (spectre large bande) et un son de voix prononçant le mot "shock" (enveloppe spectrale présentant des formants caractéristiques). Les métriques utilisent la somme des différences entre les dérivées première ou seconde des spectres par bandes fréquentielles. Un pic dans la réponse en fréquence sera alors traduit par un passage par zéro dans la dérivée première et une valeur négative localisée dans la dérivée seconde. Les résultats montrent un avantage marqué de la dérivée seconde quel que soit le contenu spectral de la source sonore avec une réussite (prédiction du maximum à la direction test) de presque 100% et 50% pour les stimuli de papier et de voix, respectivement. Les performances dégradées pour la voix illustrent le fait qu'il y ait concurrence d'interprétation entre relief spectral lié à la source elle-même ou au filtrage directionnel (HRTF).La réduction de la largeur de bande des filtres semble améliorer les résultats, en renforçant la prise en compte des caractéristiques localisées comme les creux et les pics. Cependant, ces résultats reposent sur l'hypothèse que la localisation des stimuli testés apparaît à la position de la source sonore, hypothèse non validée par la mise en place d'un test de localisation.

Langendijk et Bronkhorst [LB02] ont cherché à identifier les indices spectraux responsables de la localisation auditive en élévation. Un test de localisation est réalisé, avec des sources virtuelles synthétisées avec des DTFs individuelles dont certaines bandes fréquentielles, de largeur et de fréquence centrale variées, ont été aplaties. Un modèle de prédiction probabiliste est élaboré sur la base des indices spectraux délivrés par le stimulus par rapport à ceux stockés dans les HRTFs individuelles, brutes et large bande. Plusieurs paramètres sont testés : la largeur de bande des filtres utilisés dans l'étape de représentation interne, la représentation des spectres par la dérivée 0, 1 ou 2 (en référence à Zakarauskas et Cynader), la métrique utilisée dans l'étape de comparaison (déviation standard de la différence spectrale [Mid99b] ou inter-corrélation spectrale [Mid92]), ainsi qu'un paramètre lié à la transformation des distances en indices de similarité. Les détails de ce modèle seront présentés en section 6.2.2.

Bremen et al. [BvWvO10] ont examiné la capacité du système auditif à discriminer deux sources réelles présentées simultanément sur le plan médian en fonction de la distance séparant les haut-parleurs, de leurs gains respectifs et de leur position en élévation. Ils réalisent en parallèle un modèle de prédiction de la localisation pour tenter d'expliquer les motifs de localisation observés expérimentalement. Partant du principe que les ondes sonores s'additionnent linéairement au tympan, la représentation interne de deux sons émis simultanément est obtenue par la somme linéaire des signaux temporels pondérée par les niveaux relatifs d'émission. Le modèle prédit alors une similarité inversement proportionnelle à la déviation standard de la différence spectrale entre le spectre cible et les DTFs de l'auditeur. De façon intéressante, il apparait que le spectre cible présente une ressemblance avec le spectre d'amplitude de la DTF située spatialement entre les deux sources (observation en contradiction avec [LKW01]). Ce modèle est le premier à prendre en compte le biais de pointage, estimé à partir des erreurs associées aux directions pointées dans la condition de référence (une seule source) et appliqué de manière inverse pour corriger les réponses.

Les modèles de prédiction précédents utilisent une représentation simplifiée du traitement des sons par le système auditif périphérique en considérant uniquement une représentation du spectre par bandes fréquentielles, approximant la réponse de la membrane basilaire au signal sonore. Mattes et al. [MNFC14] proposent un modèle de prédiction de la localisation sur tout l'espace, basé sur un encodage des indices par le système auditif périphérique plus complexe. S'ajoutent à la modélisation de la sélectivité fréquentielle de la membrane basilaire (banc de filtres gammatone), la prise en compte de la réponse de l'oreille moyenne (filtre passe-bande $[1, 4]$ kHz), de la transduction des impulsions nerveuses dans l'organe de Corti

(rectification de moitié d'onde et filtrage passe-bas à 1 kHz [DPK96]) et des non-linéarités de la membrane basilaire (compression de l'enveloppe spectrale d'un facteur 0.5). Afin de réaliser la prédiction sur tout l'espace, les auteurs utilisent des descripteurs à la fois binauraux et monauraux. Ce modèle s'appuie sur une combinaison des modèles de Park et al. [PNK08] et Baumgartner et al. [BML13] (présenté ci-dessous), centrés respectivement sur une prédiction en azimut et en élévation. Le modèle de Park et al., non détaillé ici, se base sur la théorie d'égalisation-annulation pour la représentation interne des indices binauraux sous forme de motifs d'*excitation-inhibition* (en référence à Durlach, 1972). L'analyse des indices spectraux repose sur la déviation standard de la différence spectrale des spectres lissés préalablement par un banc de filtres auditifs [BML13]. Pour évaluer leur modèle, Mattes et al. utilisent les données expérimentales du test de localisation de Makous et Middlebrooks [MM90]. Cependant, les résultats restent qualitatifs et rien n'indique que les HRTFs utilisées pour la prédiction appartiennent aux 6 participants du test de localisation de Makous et Middlebrooks.

La plupart des modèles présentés précédemment ont été développés dans l'objectif de mieux comprendre les mécanismes sous-jacents à la localisation auditive. Les études de Baumgartner et al. (2013 [BML13] et 2014 [BML14]) ont mis l'accent sur la mise au point d'un modèle flexible et performant dans l'optique de pouvoir s'affranchir de tests de localisation longs et complexes. De plus, ils s'intéressent à la prédiction des performances (erreurs) de localisation, contrairement aux autres modèles qui se sont concentrés sur la capacité du modèle à prédire la position spatiale des réponses. Les modèles de Baumgartner et al. adoptent une approche probabiliste et se focalisent sur la prédiction des erreurs de localisation en élévation, sur des plans sagittaux plus ou moins latéralisés. La prédiction repose donc sur l'analyse des indices spectraux à travers une mesure de la déviation standard de la différence spectrale [BML13] ou de la somme des différences entre les dérivées premières positives des profils spectraux [BML14]. La nouveauté de ces modèles réside principalement dans l'introduction d'un paramètre non-acoustique relatif à la précision avec laquelle chaque auditeur localise les sources sonores dans l'espace, et dont l'importance a été mise en évidence dans une étude parallèle [MBL14]. Les auteurs analysent la capacité du modèle à s'adapter à des cas de localisation divers : après une calibration du modèle spécifique à un ensemble de sujets dont ils possèdent les HRTFs individuelles et les données expérimentales de localisation, le modèle est utilisé pour la prédiction des données expérimentales issues de différents tests de localisation de la littérature, incluant parfois d'autres sujets que ceux utilisés pour la calibration. Les résultats montrent que les performances de localisation prédites sont proches de celles mesurées.

Dans le tableau suivant sont résumés les détails relatifs aux différents modèles qui viennent d'être mentionnés. Dans le cas où plusieurs paramètres ont été testés, seuls les paramètres optimaux sont donnés.

Etude	Espace prédiction	Descripteurs	Métrique	Type de stimulus expérimental	Evaluation modèle
Middlebrooks (1992)	Toute la sphère (1652 secteurs spatiaux)	[Monaural] DTFs par bandes fréq. (dB) (40 bandes $\in [3, 16]$kHz) [Binaural] ILD	[Mon.] inter-corr. [Bin.] diff. d'ILD	sources réelles à bande étroite	histogramme des SI aux réponses (paramètre d' et pourcentage correct)
Zakarauskas et al. (1993)	Plan médian (35 pos.) ou sphère (614 pos.)	Dérivée 1 et 2 des DTFs par bandes (banc de filtres $Q_{10} = 10$ constante sur échelle log., freq. min.=1kHz)	norme L1	stimuli : large bande ou de voix (simulations uniquement, pas d'expérience)	erreur moy. de prédiction (position de la dist. min. vs position cible)
Hofman et al. (1998)	Plan médian	DTF par bandes freq. (dB) ($\frac{1}{20}$ octave $\in [2, 16]$kHz)	inter-corr.	sources réelles aux caractéristiques spectro-temporelles variables	observations qualitatives
Middlebrooks (1999b)	Plans sagittaux	DTFs par bandes freq. (dB) (64 bandes, $\in [3.7, 12.9]$kHz)	variance	HRTFs non-indiv. brutes ou adaptées à l'auditeur	observations qualitatives
Langendijk et al. (2002)	Plan médian (53 pos.)	DTFs par bandes freq. (dB) ($\frac{1}{6}$ bandes d'octave $\in [200, 16$kHz$]$)	STD	HRTFs indiv. avec suppression de certains indices spectraux	log-vraisemblance
Bremen et al. (2010)	Plan médian	DTFs (dB) $\in [3, 12$kHz$]$	STD	2 HPs dans le plan médian (var. niveaux relatifs et écart spatial)	observations qualitatives + histogrammes des SIs aux réponses
Baumgartner et al. (2013)	Plans sagittaux $\in \pm 30°$ latéral (≈ 40 pos.)	DTFs par bandes freq. (dB) (banc de filtres de largeur ERB $\in [700, 18$kHz$]$)	STD	données expérimentales de [GML10, Wal10] (manipulation des indices spectraux)	corrélation des performances (PE,QE) prédites et mesurées
Mattes et al. (2014)	Sphère	[Monaural] DTFs par bandes freq. (banc de filtres $\in [0.1,20]$kHz) [Binaural] motifs EI [PNK08]	[Mon.] STD [Bin.] inter-corr.	données expérimentales de [MM90] (sources réelles $[1.8, 16]$kHz)	observations qualitatives
Baumgartner et al. (2014)	Plans sagittaux $\in \pm 60°$ latéral (≈ 40 pos.)	Gradient pos. des DTFs par bandes (dB) (banc de filtres de largeur ERB $\in [700, 18$kHz$]$)	moyenne	données expérimentales de [GML10, MGL10, MWL13] (manipulation des indices spectraux)	corrélation des performances (PE,QE) prédites et mesurées + log-vraisemblance

6.2 Métriques, similarité et critère de log-vraisemblance

Les études auxquelles nous nous intéressons ici ont proposé des modèles de prédiction de la localisation auditive de type *template-based* dont les méthodes de prédiction sont complètes et ont été validées par rapport à des données réelles. Ils constituent par conséquent une référence pour le développement du modèle que nous souhaitons développer. Nous exposons ici les méthodes employées dans chacun des modèles, de la définition d'une métrique à son exploitation dans le modèle de prédiction, à savoir les méthodes : de calcul les distances *target-templates* (i.e. la métrique de comparaison), d'estimation des indices de similarité à partir des valeurs de distances, de combinaison des indices de similarité spectrale gauche et droit et éventuellement des indices de similarité interauraux et spectraux, d'évaluation des performances de prédiction vis-à-vis des observations expérimentales et, pour le dernier modèle présenté, de prise en compte de l'incertitude de report. Si les paramètres du modèle ont fait l'objet d'une optimisation, nous donnons les résultats de l'optimisation des paramètres.

6.2.1 Modèle de Middlebrooks (1992)

Middlebrooks [Mid92] a développé un modèle de prédiction de localisation sur tout l'espace de sources réelles à bande fréquentielle étroite. Les stimuli sont constitués de bruit large bande ([2,15] kHz) filtré dans une bande d'octave (fréquences centrales : 6, 8, 10, ou 12 kHz). La prédiction est basée sur une combinaison d'indices interauraux (ILD) et spectraux. Les performances du modèles sont évaluées à partir des données expérimentales d'un test de localisation sur haut-parleurs ainsi que des DTFs individuelles large bande des sujets.

Modèle – Le modèle de de prédiction compare le signal cible convolué à la DTF de l'auditeur dans la direction test aux DTFs de ce même auditeur sur l'ensemble de la sphère. La prédiction se base sur le calcul d'une distribution d'indices de similarité sur la sphère combinant indices spectraux et interauraux :

1. (a) terme de similarité d'ILD obtenu après multiplication par (-1) des différences d'ILD entre l'ILD cible et les ILDs associées aux DTFs de l'auditeur
   (b) normalisation Z-score du terme d'ILD sur l'espace
2. (a) similarités spectrales gauche et droite obtenues par l'inter-corrélation du spectre du stimulus et des DTFs de l'auditeur, à gauche et à droite (préalablement filtrées par un banc de filtres espacés logarithmiquement en fréquence de sorte à approximer la fonction de transfert de la cochlée)
   (b) combinaison pondérée des similarités gauche et droite selon leur puissance relative
   (c) normalisation Z-score du terme de similarité spectrale sur l'espace
3. addition des termes d'ILD et spectraux puis normalisation Z-score sur l'espace

Evaluation du modèle – L'évaluation du modèle se base sur la théorie de détection du signal et consiste à comparer la distribution des indices de similarité (SI) aux réponses des sujets par rapport à la distribution normalisée des indices de similarité sur tout l'espace. Pour chaque stimulus unique (une fréquence centrale) et chaque sujet, il existe une distribution unique d'indices de similarité. Plus la distribution des SI aux réponses des sujets s'écarte vers les valeurs positives de la distribution normalisée sur tout l'espace (centrée en zéro grâce à la normalisation Z-score), meilleure est la prédiction. L'aire sous la courbe ROC (*Receiving Operating Characteristic*) permet alors de quantifier la séparation entre les deux distributions. Deux critères sont dérivés de l'aire sous la courbe : le "pourcentage correct" (*percent correct*), égal au pourcentage d'aire sous la courbe (ou *Area Under Curve*, AUC) multipliée par 100, et l'indice de sensibilité d' (avec $d' = \sqrt{2} \cdot \text{norminv}(AUC)$). Dans le cas où l'histogramme des indices de similarité aux réponses se superpose avec celui des indices de similarité sur la sphère (cas des réponses données au hasard), le pourcentage correct prend des valeurs autour de 50% et d', autour de zéro. Les résultats de ce modèle donnent des valeurs de pourcentage correct comprises entre 75% et 95% avec une médiane de 90% et des valeurs de d' comprises entre 0.9 et 2.3 avec une médiane de 1.8.

6.2.2 Modèle de Langendijk et Bronkhorst (2002)

L'objectif de l'étude menée par Langendijk et Bronkhorst [LB02] est d'identifier les bandes fréquentielles contenant les indices spectraux responsables de localisation avant-arrière et haut-bas. Ils réalisent un modèle de prédiction de la localisation de sources sonores synthétisées avec des DTFs individuelles dont certaines plages fréquentielles (plus ou moins larges) ont été aplaties, en parallèle d'un test de localisation de ce type de sources. La prédiction se restreint au plan médian (dimension polaire) et se base donc uniquement sur l'analyse des indices spectraux.

Modèle – Le modèle tente de prédire la localisation des stimuli présentés dans 11 positions du plan médian pour chacune des 9 conditions suivantes : DTFs originales (condition de référence) ou modifiées, i.e. dont les indices spectraux ont été supprimés dans une demie bande d'octave (4 conditions selon la bande fréquentielle considérée), une bande d'octave (3 conditions) ou deux bandes d'octaves (1 condition).

Plusieurs paramètres sont testés avant d'identifier ceux qui optimisent la prédiction de l'angle polaire perçu. Tout d'abord, la représentation interne des DTFs consiste à moyenner (en termes de valeur efficace, voir eq. du A1 du papier) les DTFs par bandes d'octave (de largeur de bande constante par rapport à une échelle fréquentielle logarithmique). Le modèle est analysé pour les largeurs de bandes (BW) : $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{12}$ et $\frac{1}{24}$ bandes d'octave. Les profils spectraux obtenus sont ensuite dérivés en fréquence à l'ordre $DO = 0, 1$ ou 2 (en référence à Zakarauskas et Cynader [ZC93]). L'étape de comparaison (CP) entre la DTF cible et les DTFs de l'auditeur consiste à mesurer la similarité en terme d'inter-corrélation ou de déviation standard sur les profils spectraux. Les distances mesurées sont ensuite transformées en indices de similarité entre 0 et 1. Dans le cas de l'inter-corrélation, la transformation est réalisée en ajoutant 1 et en divisant par 2 les coefficients de corrélation. Dans le cas de la déviation standard, les auteurs appliquent une fonction gaussienne centrée en zéro et de déviation standard S, avec S testé dans l'intervalle $[0.5, 4]$. Les indices de similarité sont ensuite moyennés gauche-droite (poids relatifs égaux car il s'agit d'une prédiction sur le plan médian) puis normalisés de sorte à obtenir avec une distribution de probabilité pour chaque stimulus unique (condition×direction).

Evaluation du modèle – Le critère d'évaluation du modèle consiste à utiliser la log-vraisemblance associée à la probabilité prédite par le modèle dans la direction des réponses (*actual likelihood*) :

$$L_a = \sum_{r=1}^{N_R} -2 \cdot ln(p_r) \tag{6.1}$$

où r est l'indice de la réponse, p_r la probabilité prédite par le modèle dans la direction de la réponse r et N_R le nombre total de réponses (ici, $N_R = 11$ positions $\times 5$ répétitions). Selon cette définition, plus la probabilité aux réponses est forte, plus la log-vraisemblance est faible. Notons que, par la suite, nous utiliserons le terme de log-vraisemblance "actuelle" pour désigner L_a afin de conserver la même initiale qu'en anglais. Le terme "actuel" est pris au sens "effectif" en actes (basés sur les observations réelles).

Les auteurs identifient les paramètres du modèle qui minimisent la log-vraisemblance avant de discuter des résultats, soit : $BW = \frac{1}{6}$, $DO = 0$, $CP =$ déviation standard et $S = 2$. Tout d'abord, la largeur de bande identifiée comme optimale s'approche d'une largeur de bande ERB (validé plusieurs fois dans la littérature). De plus, ils réfutent l'hypothèse de Zakarauskas et Cynader selon laquelle le traitement des indices spectraux par le système auditif serait basé sur la dérivée première ou seconde du spectre.

Selon les observations des auteurs, le modèle semble prédire assez précisément la localisation dans toutes les conditions. Dans les conditions de référence (large bande), à une bande d'octave et à une demie bande d'octave, le modèle prévoit une forte probabilité à la direction test (maximum dans le cas large bande), plus étalée sur les positions voisines à 90° (observation en accord avec [HVO98]) et une probabilité non-nulle d'apparition de confusions avant-arrière. Pour la condition à 2 bandes d'octave le modèle prévoit une localisation à seulement une ou deux directions, indépendantes de la direction test et qui coïncident bien avec les réponses.

Pour approfondir l'évaluation du modèle, la log-vraisemblance aux réponses est comparée à d'autres calculs de log-vraisemblance. Tout d'abord, la log-vraisemblance "escomptée" L_e (*expected likelihood*) est calculée à partir d'un tirage aléatoire de 5 réponses en suivant la fonction de distribution de probabilité du modèle associée au stimulus. La procédure est répétée 100 fois puis la moyenne et l'intervalle de confiance à 99% sont déterminés. La log-vraisemblance actuelle L_a devrait alors se trouver dans l'intervalle de confiance de L_e si les réponses suivent la tendance prédite par le modèle, autrement dit si le modèle est performant. La log-vraisemblance associée au seuil de chance est également calculée à partir d'une distribution de probabilité constante sur l'espace mais aucune comparaison des résultats n'est effectuée avec cette donnée. Enfin, les auteurs s'intéressent aux log-vraisemblances associées à un modèle que prédirait une densité de probabilité (1) centrée uniquement à la direction test (distribution unimodale), (2) centrée à la direction test et à la direction symétrique par rapport au plan frontal (distribution bimodale), (3) centrée à la direction test, à la direction symétrique par rapport au plan frontal, et à la direction au pôle (distribution trimodale). Notons que la direction symétrique par rapport au plan frontal correspond à la direction associée à une confusion avant-arrière. Pour modéliser ces différents cas, les densités de probabilité respectives sont dessinées à l'aide de fonctions gaussiennes d'écart-type 17° (imprécision globale des sujets en élévation dans la condition individuelle).

Pour conclure, l'étude a permis de mettre en évidence les paramètres du modèle de prédiction qui permettent d'optimiser les résultats, tels qu'une métrique de comparaison spectrale et une représentation interne basée des largeurs de bande proche des largeurs ERB. Ce modèle permet d'introduidre le critère de log-vraisemblance.

6.2.3 Modèle de Baumgartner, Majdak et Laback (2013)

L'objectif du modèle proposé par Baumgartner et al. en 2013 [BML13] est d'être capable de prédire les performances de localisation en élévation et les taux de confusions quel que soit le contenu spectral du stimulus, de sorte à pouvoir s'affranchir de tests de localisation coûteux. Il se distingue des précédentes études motivées par le désir de comprendre les mécanismes sous-jacents à la localisation auditive. Le modèle se focalise sur la prédiction des performances de localisation sur des plans sagittaux plus ou moins latéralisés. La prédiction est basée sur la comparaison des indices spectraux délivrés par le stimulus et ceux contenus dans les DTFs de l'auditeur de manière à obtenir une distribution de probabilité de réponse à partir de laquelle sont calculées les erreurs de localisation polaires et les taux de confusions. L'évaluation du modèle repose sur une comparaison des erreurs prédites et relevées dans des tests de localisation, et non de la probabilité de réponse prédite aux directions pointées. Un paramètre non-acoustique tenant compte de la capacité de l'auditeur à évaluer les indices spectraux est introduit. La calibration du modèle est donc réalisée de manière spécifique à chaque sujet, à partir de ses performances en condition d'écoute individuelle (stimuli synthétisés avec fonctions de transfert individuelles).

Modèle – **Représentation interne et métrique associée** – Pour modéliser la représentation interne des sons, le stimulus (soit la DTF de la direction de synthèse, modifiée selon la condition expérimentale) et les DTFs de l'auditeur (large bande) sont convolués par un banc de filtres gammatone de largeur de bande égale à un ERB et de fréquences minimale et maximale 700 Hz et 18 kHz. La valeur efficace du signal filtré est ensuite calculée dans chaque bande fréquentielle et le profil spectral obtenu est exprimé en dB (c.f. équation 5.11). Les auteurs considèrent également un modèle de cellules ciliées. Il consiste en une rectification de la forme d'onde suivie d'un filtre passe-bas (Butterworth à l'ordre 2). L'étape de comparaison consiste à calculer des distances entre la cible et les DTFs de l'auditeur sur le plan sagittal de la direction test en termes de la déviation standard de la différence spectrale, à gauche et à droite.

Paramètre non-acoustique – De la même façon que Langendijk et Bronkhorst [LB02], les distances obtenues sont transformées en indices de similarité (SI) par l'intermédiaire d'une fonction gaussienne centrée en zéro, et d'écart-type U. Les auteurs interprètent la largeur de la gaussienne comme la capacité de l'auditeur à localiser ou plus précisément à identifier et discriminer les indices de localisation spectraux. Meilleur est le localisateur, plus la gaussienne sera d'écart-type U faible. L'utilisation d'une telle fonction repose sur l'hypothèse que l'auditeur détecte les fortes similarités spectrales avec une certaine imprécision et que la probabilité de réponse décroît vers zéro avec une baisse de la similarité. Etant donné que la probabilité n'est jamais nulle même lorsque les indices de similarité sont très faibles, cela pourrait mener à une surestimation des erreurs de localisation. Le paramètre non-acoustique U permet donc de calibrer le modèle pour chaque participant. Cela suppose d'avoir récolté les performances de localisation de l'individu (pour lequel on cherche à caractériser U) en condition d'écoute individuelle i.e. à l'écoute de stimuli large bande synthétisés avec ses propres DTFs (autrement appelée condition de référence). La valeur de U optimale est déterminée au travers de la minimisation des différences entre les erreurs polaires PE (*polar error*, équation 6.4) et de quadrant QE (*quadrant error*, équation 6.3) prédites et mesurées.

L'avantage d'individualiser le paramètre U par rapport à une valeur générique telle qu'utilisée par Langendijk et Bronkhorst [LB02] est illustrée par l'amélioration de la prédiction des erreurs PE et QE. L'étude de Majdak et al. [MBL14] confirme ce résultat en montrant que ce paramètre non-acoustique a un impact important sur les performances de localisation prédites par le modèle dans la condition d'écoute individuelle. Cette étude simule également l'effet de la qualité des DTFs individuelles (i.e. la qualité des indices acoustiques de l'auditeur) sur les performances de localisation. Pour cela, le paramètre U individuellement calibré est conservé et les DTFs d'un autre sujet sont utilisées pour effectuer la prédiction. Notons que le changement du jeu de DTFs est alors effectué simultanément pour la cible et les DTFs *templates* (simule une recalibration complète de la carte auditive spatiale). Les résultats montrent un effet relativement faible de la modification du jeu de DTFs sur les performances prédites, comparé à l'effet du paramètre non-acoustique.

Pondération binaurale – Afin de combiner les indices de similarité gauche et droit, Baumgartner et al. introduisent une méthode de pondération binaurale adaptée à la prédiction sur des plans sagittaux plus latéralisés que le plan médian. Elle est basée sur les résultats expérimentaux de Morimoto [Mor01] qui suggèrent que la contribution de l'oreille controlatérale devient négligeable pour des sources situées au-delà d'un angle latéral de 60° et sur ceux de Macpherson et Sabin [MS07] qui ont quantifié la contribution relative de chaque oreille aux positions latérales ±45°. La pondération binaurale est modélisée par une fonction sigmoïde qui dépend de l'angle latéral de la direction cible :

$$w_L = (1 + e^{-\Theta_k/\Omega})^{-1}, \qquad w_R = 1 - w_L \tag{6.2}$$

avec w_L et w_R les poids gauche et droit, Θ_k l'angle latéral de la cible et $\Omega = 13$ le coefficient de pondération binaural, ajusté aux résultats expérimentaux de Macpherson et Sabin (2007).

Enfin, les indices de similarité sont interpolés selon un échantillonnage régulier avant d'être normalisés par la somme des SI sur le plan sagittal. La distribution obtenue peut alors être interprétée comme une

fonction de densité de probabilité de réponse du sujet. Les principales étapes de ce modèle sont résumées dans le schéma figure 6.1.

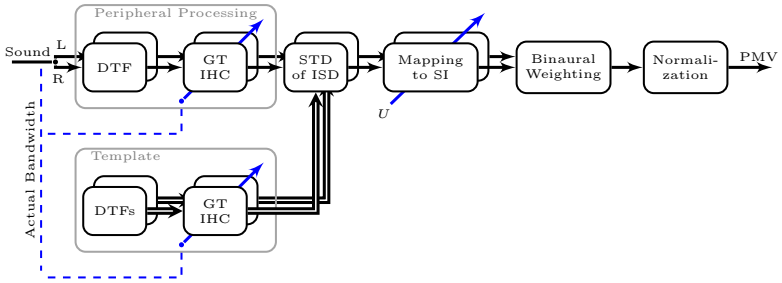


FIGURE 6.1 – Structure du modèle de Baumgartner et al. (2013) (figure issue de l'article [BML13]).

Evaluation du modèle – **Critères** – La méthode d'évaluation du modèle se base sur la comparaison entre les performances de localisation prédites et les performances mesurées. Le pourcentage d'erreurs de quadrant (erreur polaires $\geq 90°$), incluant à la fois les confusions avant-arrière et haut-bas, est déterminé en calculant la somme des probabilités aux points éloignés à ou à plus de 90° de la direction test :

$$QE_{j,k} = \sum_{i \in A_k} p_{j,k}[\Theta_{i,k}] \tag{6.3}$$

où $A_k = \{i = 1 \cdots N_\Theta[k], |\Theta_{i,k} - \Theta_{j,k}| \geq 90°\}$ avec $N_\Theta[k]$ le nombre d'angles polaires du plan sagittal k et $p_{j,k}[\Theta_{i,k}]$ la probabilité de répondre à l'angle polaire $\Theta_{i,k}$ en réponse à une source positionnée à $\Theta_{j,k}$.

La variance et le biais des erreurs locales (erreurs polaires $< 90°$) sont combinées dans une mesure unique d'erreur polaire locale calculée ainsi :

$$PE_{j,k} = \sqrt{\frac{\sum_{i \in B_k} (\Theta_{i,k} - \Theta_{j,k})^2 \cdot p_{j,k}[\Theta_{i,k}]}{\sum_{i \in B_k} p_{j,k}[\Theta_{i,k}]}} \tag{6.4}$$

où $B_k = \{i = 1 \cdots N_\Theta[k], |\Theta_{i,k} - \Theta_{j,k}| < 90°\}$. Le modèle prend donc en entrée le signal à localiser puis renvoie les mesures PE et QE, qui peuvent être directement comparées aux performances de localisation mesurées dans un test de localisation.

6.2.4 Apport du modèle de Baumgartner et al. (version 2014)

Le modèle de Baumgartner et al. a été enrichi dans une étude plus récente [BML14]. Il considère une étape supplémentaire d'extraction des gradients positifs spectraux à la représentation interne des DTFs. De plus, il complexifie la fonction qui transforme les distances en indices de similarité et modélise le comportement de la tâche motrice liée au report de localisation. Les mérites de ces paramètres sont discutés. La structure principale du modèle peut être visualisée en figure 6.2.

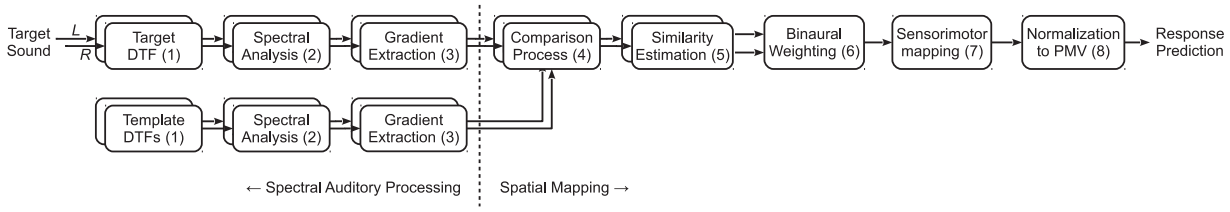


FIGURE 6.2 – Structure du modèle de Baumgartner et al. (2014) (figure issue de l'article [BML14]).

Modèle – **Représentation interne et métrique associée** – Comme dans le précédent modèle, la représentation interne du son cible et des DTFs *templates* de l'auditeur consiste à moyenner le spectre par bande fréquentielle suivant l'application d'un banc de filtres gammatone de largeur de bande ERB. Dans cette version du modèle, les gradients positifs des profils spectraux sont extraits, de façon à modéliser le rôle du DCN (*Dorsal Cochlear Nucleus*) dans la localisation auditive. Cette fonctionnalité a été mise en évidence par Reiss et Young [RY05] dans le cadre d'une étude sur les chats [RY05] et avait déjà été suggérée par Zakarauskas et Cynader [ZC93] (sans restriction aux gradient positifs). L'étape de comparaison consiste à effectuer la moyenne des distances absolues entre les gradients positifs des profils spectraux (c.f. équation 5.20) de la cible et des DTFs de l'auditeur sur le plan sagittal de la direction test. La déviation standard de la différence n'est pas utilisée ici étant donné que l'information de gain global est déjà négligée par le calcul du gradient [1].

Les auteurs comparent les résultats de prédiction en considérant d'un côté la déviation standard de la différence entre les profils spectraux (version précédente du modèle) et de l'autre la moyenne de la différence absolue des gradients positifs (voir section IV. A. de l'article). Lorsque l'extraction des gradients positifs est considérée, un amélioration des résultats de prédiction est observée dans le cadre de la prédiction de localisation de stimuli de voix, de stimuli à spectre ondulé, ou dans le cas de stimuli filtrés passe-bas. Cependant, la métrique basée sur la déviation standard des profils spectraux optimise la prédiction dans le cas de stimuli large bande. L'extraction des gradients positifs semble donc être nécessaire lorsque les stimuli contiennent des variations spectrales importantes. D'après les résultats, cette métrique permet également la prise en compte automatique de la bande fréquentielle du stimulus.

Fonction sigmoïde – Afin de transformer les valeurs de distances en indices de similarité, une fonction non-linéaire sigmoïde est utilisée. La fonction sigmoïde ressemble à une fonction gaussienne à un seul côté, tel qu'utilisé par [BML13] et [LB02], mais présente l'avantage d'être paramétrée par deux variables. L'utilisation de la fonction sigmoïde dans l'objectif de transformer des distances en indices de similarité (inversement proportionnels) consiste à multiplier son expression par -1 [BML14], soit :

$$SI(d) = (-1) \cdot \frac{1}{1 + e^{-\Gamma(d - S_l)}} \tag{6.5}$$

où d est la distance, Γ le degré de sélectivité (pente) et S_l le paramètre de sensibilité (point d'inflexion, seuil de détection). Plus S_l est faible et plus Γ est grand, plus l'auditeur est sensible aux variations spectrales et plus sa localisation est précise. Les paramètres Γ et S_l seront déterminés de sorte à optimiser la prédiction, avec S_l individualisé pour chaque sujet.

Dispersion de pointage – Une partie de l'erreur de localisation observée dans les tests de localisation est liée à l'imprécision introduite par la méthode de report [BCNW16]. Une modélisation de la dispersion de pointage est proposée. Pour ce faire, le vecteur d'indices de similarité est convolué à une fonction gaussienne définie sur l'axe polaire, de concentration κ dépendante du paramètre de dispersion ϵ :

$$\kappa = \frac{cos^2 \Phi_k}{\epsilon^2} \tag{6.6}$$

où Φ_k est l'angle latéral du plan sagittal de prédiction. Notons qu'avec ϵ constant dans l'espace, l'effet de flou résultant est d'autant plus important sur les plans sagittaux latéralisés, ce qui semble refléter la réalité [BML14]. De plus, l'opération de convolution suppose un échantillonnage régulier du plan sagittal (obtenu par interpolation si nécessaire). Le paramètre $\epsilon = 17°$ permet d'optimiser les performances de prédiction, ce qui en accord avec la déviation standard en élévation obtenue dans [LB02]. La prise en compte de la dispersion montre une amélioration de la prédiction des erreurs de localisation, en particulier sur les plans sagittaux latéralisés.

Evaluation du modèle – L'optimisation des paramètres ϵ, Γ et S_l est réalisée en termes de minimisation des différences entre QE et PE (équations 6.3 et 6.4) prédits et mesurés en condition d'écoute individuelle (stimuli synthétisés avec DTFs individuelles) pour 23 sujets dont on possède à la fois les données expérimentales de localisation et les DTFs individuelles. Les données expérimentales sont issues de tests de localisation de la littérature [GML10, MGL10, MIC+13, MWL13].

L'implémentation des modèles de Baumgartner et al. (2013 et 2014) fait l'objet d'une *toolbox* en libre accès sur internet appelée *Auditory Modelling Toolbox*.

1. Une disparité apparaît entre le code open source de l'AMToolbox associée au modèle et l'équation (4) de l'article [BML14]. Dans le code, la moyenne et non la somme (comme indiqué dans l'article) des différences absolues des gradients spectraux positifs par bande fréquentielle est calculée. Après discussion avec les auteurs, c'est bien la moyenne des distances absolues qui a été utilisée et il existe donc une erreur dans l'article.

6.2.5 Conclusion sur les modèles existants

Les modèles de la littérature présentés ici se focalisent principalement sur des cas simples de localisation d'une source unique en environnement anéchoïque. Dans le cas de la présence d'un éventuel effet de salle (e.g. pour les tests sur sources réelles), celui-ci n'est pas pris en compte dans la prédiction.

La structure commune des modèles implique principalement une étape de modélisation de la représentation interne des HRTFs et une étape de comparaison *target-templates* pour la détermination des indices de similarité sur l'espace de prédiction. La prédiction de la localisation auditive repose donc directement sur la définition d'une métrique qui mesure la similarité les HRTFs. Celle-ci est uniquement basée sur les indices spectraux lorsque la prédiction est réduite à la dimension polaire, et inclut également les indices interauraux lorsqu'elle s'étend à tout l'espace. La représentation interne utilisée dans ces modèles s'appuie principalement sur une modélisation de la transduction de la membrane basilaire sous la forme d'un banc de filtres appliqué sur les spectres d'amplitude des HRTFs. Selon les modèles, cette modélisation est variable et plus ou moins proche des aspects physiologiques du système auditif. Le premier modèle de Baumgartner et al. (2013) comprend un modèle supplémentaire de cellules ciliées [BML13]. Cependant, les bénéfices de ce modèle sur la prédiction de la localisation n'a pas été mis en évidence. Par ailleurs, le modèle postérieur (2014) fait alternativement appel à une étape d'extraction du gradient positif spectral [BML14]. L'étude a révélé que ce traitement présentait un avantage significatif dans le cas de stimuli à forte variations spectrales ou à bandes fréquentielles étroites contrairement au cas de stimuli large bande et à spectre plat [BML14]. Les auteurs mettent ainsi en évidence la nécessité d'adapter la représentation des HRTFs au type de stimulus mis en jeu.

6.3 Introduction au modèle proposé

Le modèle que nous proposons est basé sur une approche similaire aux modèles précédents à savoir qu'il suppose que les auditeurs se sont créés une carte audio-spatiale interne basée sur leurs HRTFs à travers un processus d'apprentissage. Il admet que la localisation auditive est guidée par une ressemblance entre les indices acoustiques du signal entrant et les indices acoustiques de localisation individuels contenus dans les HRTFs individuelles. Le principe est étendu au cas particulier de la localisation de sources sonores virtuelles large bande synthétisées avec des HRTFs non-individuelles. Dans une telle condition d'écoute, les indices de localisation délivrés à l'auditeur lui sont par définition étrangers, ce qui peut entraîner divers artefacts de localisation (localisation erronée, floue ou ambiguë).

Dans un premier temps, le modèle suppose que la direction de localisation peut s'expliquer par une similarité entre les indices acoustiques de la HRTF non-individuelle cible et ceux des HRTFs individuelles de l'auditeur à la direction perçue. L'objectif est de mettre au point un modèle capable de prédire les directions indiquées par l'auditeur à l'écoute de sources virtuelles synthétisées avec les HRTFs non-individuelles, issues de différents sujets et différentes directions de mesure. Premièrement, cela sous-entend procéder à un test expérimental pour la collection de données de localisation sur des individus dont nous possédons les HRTFs individuelles. Ce test comprendra la localisation de sources virtuelles large bande synthétisées avec des HRTFs individuelles et non-individuelles. Les données de localisation en condition d'écoute individuelle nous serviront de référence. Deuxièmement, la prédiction des directions sera basée sur le calcul d'une distribution de probabilité de réponse, résultant de la comparaison objective entre la HRTF cible et les HRTFs de l'auditeur. Cette étape de comparaison nécessitera donc la recherche d'une métrique de similarité entre HRTFs qui reflète les mécanismes de localisation auditive, notamment par l'extraction des caractéristiques spectrales pertinentes pour le système auditif. Pour une prédiction en deux dimensions, la métrique devra combiner une mesure de similarité liée à la fois aux indices interauraux et spectraux, tous deux responsables de la localisation en azimut et élévation. Cette étape comprendra également la combinaison des similarités spectrales obtenues à gauche et à droite. De plus, les incertitudes liées à la tâche de report des directions perçues par l'auditeur seront prises en compte dans la prédiction des directions pointées.

Le modèle repose sur une approche probabiliste et une évaluation locale. Une distribution de probabilité de réponse est déterminée sur l'espace de prédiction, sur la base d'une comparaison objective de la HRTF cible et des HRTFs de l'auditeur, et l'évaluation de la prédiction s'appuie sur la probabilité aux directions pointées. Chaque direction pointée est étudiée indépendamment de la direction test (direction théorique de synthèse) et est traitée de manière unique, en lien avec la probabilité de réponse prédite par le modèle à cet endroit.

Si le modèle s'apparente à celui de Baumgartner et al. [BML13, BML14], notre objectif s'en distingue néanmoins. L'objectif de ces auteurs est de rendre possible l'évaluation de la qualité du rendu spatial de systèmes de spatialisation sonore, sans avoir recours à la mise en place d'expériences de localisation auditive couteuses. Les auteurs cherchent ainsi à prédire les directions perçues, ou plus précisément les erreurs de localisation, associées à des stimuli qui n'auraient pas été testés expérimentalement. Leur modèle est tout d'abord calibré individuellement pour chacun des membres d'un ensemble de sujets à partir de leurs données de localisation en condition d'écoute individuelle. Le modèle ainsi optimisé est ensuite utilisé pour prédire les performances de localisation qui auraient été observées dans d'autres conditions d'écoute (par exemple, pour des stimuli spatialisés avec des HRTFs non-individuelles). Au contraire, notre objectif est

d'appliquer le modèle pour des individus dont on ne possède pas les HRTFs individuelles et pour lesquels le modèle ne peut donc être calibré. Notre objectif est donc centré principalement sur la condition d'écoute non-individuelle. Il sera intéressant d'étudier si un modèle calibré dans la condition individuelle et appliqué à la condition non-individuelle est aussi performant qu'un modèle calibré spécifiquement à cette condition d'écoute non-individuelle. Par ailleurs, la prédiction des erreurs de localisation du modèle de Baumgartner et al. repose sur une approche globale. Les erreurs sont calculées par l'intégration des erreurs de localisation (distance par rapport à la direction test) associées à chaque point de l'espace de prédiction, pondérées par leur probabilité d'apparition (voir équations 6.4 et 6.3). Une manière alternative consiste à générer des réponses "virtuelles" à partir de la distribution de probabilité du modèle et de les traiter de la même manière que des réponses issues d'un test de localisation réel. Contrairement à cette approche, nous ne cherchons pas à estimer directement les directions pointées ou les erreurs de localisation mais à identifier les HRTFs *templates* qui les expliquent le mieux, i.e. celles qui maximisent la probabilité aux directions pointées. L'évaluation de la prédiction est donc basée sur une approche locale plutôt que globale.

Contrairement au modèle de Baumgartner et al. ainsi qu'à d'autres modèles de la littérature, l'espace de prédiction considéré ici n'est pas restreint à un plan sagittal en particulier. L'objectif est de ne pas omettre la composante latérale de la localisation. Cependant, l'espace de prédiction ne couvre pas non plus l'ensemble de la sphère comme proposé par Middlebrooks [Mid92]. En effet, nous savons *a priori* que la probabilité de réponse dans l'hémi-espace opposé à la cible (i.e. symétrique par rapport au plan médian) est proche de zéro. L'espace de prédiction est donc centré sur le plan sagittal le plus vraisemblable, i.e. où les indices interauraux de la HRTF cible et des HRTFs de l'auditeur correspondent, avec une tolérance suffisamment large pour prendre en considération les variations individuelles d'ITD, les variations des lignes d'iso-ITD et la dispersion causée par la méthode de report. Une étape importante du développement de ce modèle consiste par conséquent à déterminer la largeur latérale de l'espace de prédiction et à modéliser la dispersion de pointage liée à la méthode de report utilisée.

Etant donné que l'espace de prédiction comprend la dimension latérale, le modèle se doit de considérer à la fois les indices interauraux, responsables de la localisation latérale, et spectraux, responsables de la localisation en élévation et de la discrimination avant-arrière. Ces deux types d'indices varient entre les individus et sont à l'origine des erreurs de localisation observées dans les deux dimensions lors de l'utilisation de HRTFs non-individuelles. Dans le cadre de cette étude, l'accent est mis sur la comparaison de plusieurs métriques spectrales, définies suivant différents modes de représentation des HRTFs et différents critères de distance. Les métriques testées sont tirées de la littérature. Le but est d'identifier la métrique la plus proche des processus d'extraction et de reconnaissance des indices spectraux. En effet, peu d'études ont évalué ce paramètre. Baumgartner et al. [BML14] ont mis en évidence l'avantage d'une représentation des HRTFs par le gradient spectral, par rapport au profil spectral, dans le cadre de stimuli dont le spectre est à bande fréquentielle étroite ou composé d'ondulations plus ou moins marquées. Cependant, nous nous intéressons ici à des stimuli à bande fréquentielle large et à spectre plat. Langendijk et Bronkhorst [LB02] ont comparé deux critères de distance (déviation standard et inter-corrélation) basés sur une représentation des HRTFs par le profil spectral. Cependant, leurs résultats sont peu détaillés et l'équivalence de la mise en pratique des deux métriques est discutable. Dans notre étude, la comparaison de différentes métriques sera précédée d'une étape de standardisation rendant la comparaison possible.

6.4 Structure et paramètres du modèle de prédiction

Nous présentons ici les principaux éléments qui définissent le modèle de prédiction proposé. Tout d'abord, la structure chronologique du modèle est détaillée. Elle permet de comprendre comment sont obtenues les distributions de probabilités de réponse du modèle et d'appréhender le rôle de chacun des paramètres. Puis, l'espace de prédiction ainsi que les contraintes qui y sont associées sont exposées. Une prise en main des critères d'évaluation est ensuite proposée en parallèle de la simulation de cas simplifiés de prédiction à une dimension. Enfin, nous présentons chacun des paramètres du modèle.

6.4.1 Structure

Le modèle se décompose en 7 étapes, optionnellement 8 si on applique un modèle de dispersion de pointage. Tout abord, les indices de similarité sont obtenus à partir de la transformation d'une combinaison de distances spectrales et interaurales issues de la comparaison *target-templates*. Comme effectué par les modèles précédents (c.f. section 6.2), les distances spectrales reposent la comparaison spectrales entre les DTFs associées au stimulus (*target*) et les DTFs de l'auditeur (*templates*). L'utilisation des DTFs, plutôt que des HRTFs, est justifié par le fait que nous nous intéressons à la composante directionnelle des fonctions de transfert. Les DTFs sont obtenues par une égalisation champ diffus des HRTFs, i.e. dont la composante commune à toutes les directions de mesure a été compensée (c.f. section 2.2.3.1). Elles confèrent l'information spectrale dépendante de la direction. Par abus de language, nous continuerons à utiliser le terme HRTFs, bien qu'il s'agisse en réalité des DTFs.

Le calcul de la distribution de probabilité de réponse relève d'une étape de normalisation de ces indices de similarité par leur somme sur l'espace de prédiction, i.e. l'échantillonnage sur lequel est réalisée la prédiction. La normalisation des indices de similarité suppose un échantillonnage spatial régulier. C'est pourquoi l'espace de prédiction est défini par un échantillonnage régulier et qu'une étape d'interpolation des distances spectrales et interaurales sur cet espace est réalisée en amont du calcul des indices de similarité et de leur normalisation sur l'espace. Les différentes étapes de calcul sont illustrées en figure 6.3.

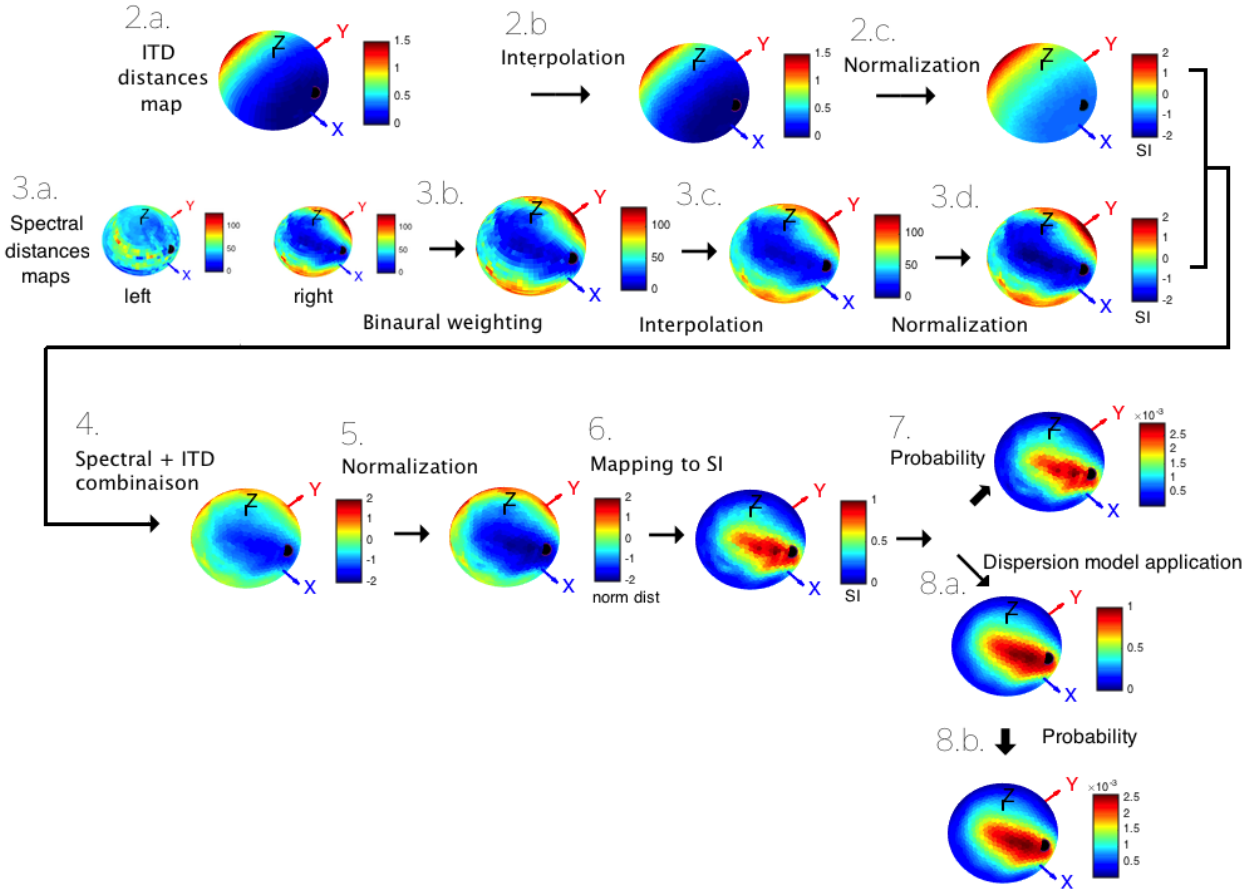


FIGURE 6.3 – Diagramme illustrant les principales étapes du modèle, du calcul des distances spectrales et interaurales *target-templates* à l'obtention d'une distribution de probabilité de réponse sur l'espace (ici, toute la sphère).

1. Définition de la cible, i.e. une paire de HRTFs $H^{\beta}_{targ}(f, \theta_t, \phi_t)$ d'un individu β à une direction test (θ_t, ϕ_t)
2. Calcul des distances interaurales :
   a. Distances d'ITD D_{ITD_i} entre l'ITD de la HRTF cible $H^{\beta}_{targ}(f, \theta_t, \phi_t)$ et les ITDs correspondants aux HRTFs de l'auditeur $H^{\alpha}_{temp}(f, \theta_i, \phi_i)$ (HRTFs *templates*) selon la métrique de l'équation 6.16, avec i l'indice de la direction, $i = 1 \dots N_S$ et N_S le nombre de points de mesure des HRTFs α
   b. Interpolation des distances d'ITD sur l'espace de prédiction défini par un échantillonnage régulier (θ_j,ϕ_j), $j = 1, \dots N_d$ et N_d le nombre de points de l'espace de prédiction, par la méthode des 4 plus proches voisins
   c. Normalisation Z-score des valeurs sur l'espace de prédiction, selon l'équation :
   $$D_{j_{norm}} = \frac{D_j - \mu(D)}{\sigma(D)} \tag{6.7}$$
   où $\mu(D)$ est la moyenne des N_d distances d'ITD sur l'espace de prédiction et $\sigma(D)$ la déviation standard.
3. Calcul des distances spectrales :
   a. Distances spectrales à gauche $D_{S_{L_i}}$ et à droite $D_{S_{R_i}}$ (définie selon la métrique spectrale choisie, voir section 6.4.5) entre la HRTF cible β gauche et droite à (θ_t, ϕ_t), et les HRTFs de l'auditeur α gauches et droites aux directions i, $i = 1 \dots N_S$, respectivement.
   b. Combinaison pondérée des distances binaurales :
   $$D_{S_i} = w_{L_i} \cdot D_{S_{L_i}} + w_{R_i} \cdot D_{S_{R_i}} \tag{6.8}$$
   avec w_{L_i} et w_{R_i} les poids des distances spectrales gauche et droite à la direction i, respectivement.
   c. Interpolation des distances spectrales sur l'espace de prédiction défini par un échantillonnage régulier (aux directions (θ_j,ϕ_j), $j = 1, \dots N_d$)
   d. Normalisation Z-score des distances spectrales D_{S_j} sur l'espace de prédiction (équation 6.7).
4. Combinaison pondérée des distances spectrales et interaurales :
   $$D_j = w_{ITD} \cdot D_{ITD_j} + w_S \cdot D_{S_j} \tag{6.9}$$
   avec w_{ITD} et w_S les poids relatifs donnés à la distance d'ITD et à la distance spectrale.
5. Normalisation Z-score des distances combinées D_j sur l'espace de prédiction (équation 6.7).
6. Transformation des valeurs de distances en indices de similarité (SI) compris entre 0 et 1 par l'application d'une fonction sigmoïde définie par ses paramètres Γ et S :
   $$\mathrm{SI}_j = (-1) \cdot \frac{1}{1 + e^{-\Gamma(D_j - S)}} \tag{6.10}$$
7. Calcul des probabilités de réponse p_j en chaque point (θ_j, ϕ_j) de l'espace de prédiction par une normalisation des SI par la somme :
   $$p_j = \frac{\mathrm{SI}_j}{\sum_{j=1}^{N_d} \mathrm{SI}_j} \tag{6.11}$$
8. (optionnel)
   a. Application du modèle de dispersion sur la carte des indices de similarité obtenus par l'équation 6.10 (voir section 6.4.7.2).
   b. Normalisation des SI par la somme (eq. 6.11) pour obtenir les probabilités *a posteriori* de l'application du modèle de dispersion (sous-entendu, l'étape 7 est ignorée dans le cas où l'on applique le modèle de dispersion).

On appellera "carte d'indices de similarité" ou "carte de similarité" l'ensemble des indices SI sur l'espace de prédiction. Notons que pour un sujet donné, il existe autant de cartes de similarité que de stimuli proposés. Chaque stimulus présenté à l'auditeur fait l'objet d'un calcul d'une carte de similarité de ce stimulus par rapport aux HRTFs de l'individu. Par la suite, l'indice q désignera l'indice de la carte de similarité avec $q = 1 \cdots Q$. P_q représentera alors la distribution de probabilité associée à la q-ième carte unique et $N_d|_q$ le nombre de points composant la q-ième carte.

Suivant la structure qui vient d'être présentée, les paramètres variables du modèle sont les suivants :

— les paramètres (Γ, S) de la fonction sigmoïde qui transforment les distances en indices de similarité
— la métrique utilisée dans l'étape de comparaison spectrale (D_S)
— les poids de la combinaison des indices de similarité interauraux et spectraux (w_{ITD} et w_S)
— la méthode de pondération binaurale pour le calcul des poids w_L et w_R de la combinaison des indices de similarité spectraux gauche et droit
— la valeur de concentration κ associée au modèle de dispersion de pointage

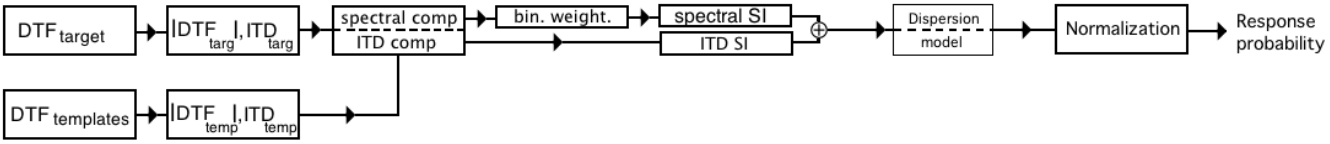


FIGURE 6.4 – Diagramme synthétisant les principales étapes du modèle de prédiction.

6.4.2 Espace de prédiction

L'espace de prédiction correspond à l'espace sur lequel est effectuée la prédiction de localisation. Rappelons tout d'abord ce qui a été réalisé dans les modèles de la littérature. La plupart de ces modèles ont effectué la prédiction de localisation sur des plans sagittaux. Tous négligent alors les erreurs latérales. Hofman et Opstal [HVO98] s'intéressent à la localisation de stimuli au contenu temporel et spectral varié et se limitent à la prédiction sur le plan médian. Ils supposent que l'azimut de la source a déjà été extrait des indices binauraux et se limitent à l'estimation de l'élévation au regard des indices spectraux. Langendijk et Bronkhorst [LB02] ont étudié la localisation de sources virtuelles dont les indices spectraux de différentes bandes fréquentielles ont été supprimés. Il se limitent au plan médian "par simplicité". Middlebrooks [Mid99b] a tenté de prédire la localisation de HRTFs non-individuelles avant et après adaptation des HRTFs à l'auditeur (i.e. dans les conditions dites "*other-ear*" et "*scaled-ear*"). Il traite le problème de localisation en deux temps : premièrement, il cherche à corréler les réponses latérales avec les positions latérales où l'ITD cible correspond à celui de l'auditeur ; deuxièmement, il néglige les erreurs latérales et ne s'intéresse qu'au plan sagittal de localisation théorique (*"correct lateral plane"*). Baumgartner et al. [BML13, BML14] se sont intéressés à la localisation sur des plans sagittaux plus ou moins latéralisés, de largeur $\pm 5°$. Les réponses qui apparaissent en dehors de cet intervalle sont négligées. Enfin, Bremen et al. [BvWvO10] s'intéressent uniquement à la localisation d'un ou deux sons émis par des haut-parleurs situés sur le plan médian. Middlebrooks [Mid92] a étudié la prédiction sur tout l'espace en prenant en compte à la fois les indices interauraux et spectraux. Dans cette étude sur la localisation de stimuli à bande fréquentielle étroite, il compare les indices de similarité à l'endroit des réponses par rapport aux indices de similarité sur toute la sphère. Cependant, cette méthode tend à sur-estimer la capacité de prédiction du modèle étant donné que la similarité du côté controlatéral est très faible (indices interauraux en opposition).

Pour cette raison, nous limiterons l'espace de prédiction à un intervalle latéral restreint et centré sur l'angle latéral de localisation théorique (i.e. centré sur l'ITD cible), que l'on appellera "intervalle latéral théorique". Cette méthode pose alors la question du traitement des réponses en dehors de cet intervalle latéral. De façon similaire à Baumgartner et al., ces réponses seront considérées comme aberrantes et seront ignorées. Cependant, la largeur latérale de l'espace de prédiction devra être ajustée de manière à ne pas supprimer trop de réponses. Au moment d'appliquer la prédiction à des données réelles, nous étudierons le taux de réponses concernées en fonction de la largeur latérale de l'espace de prédiction (voir section 7.2.3 du chapitre suivant).

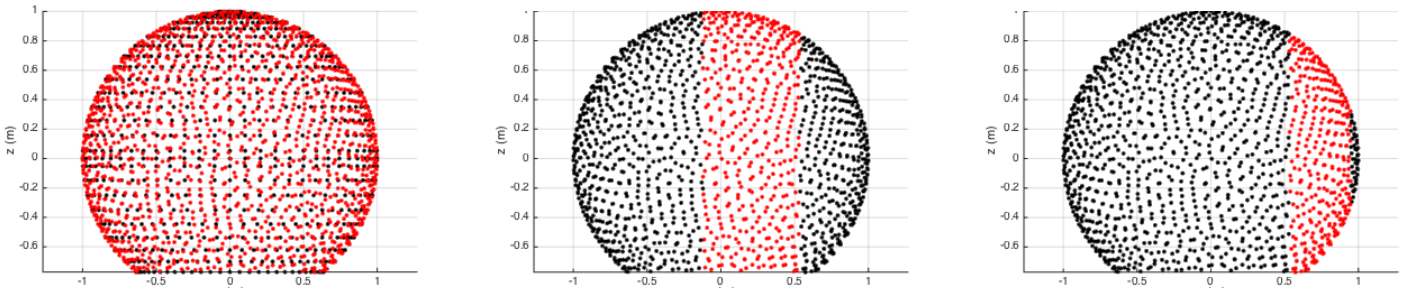
Il faut avoir conscience que la largeur de l'intervalle latéral sur lequel est effectué la prédiction détermine le poids de l'information interaurale à considérer en parallèle de l'information spectrale dans le calcul des indices de similarité. Pour une prédiction sur un plan sagittal, seule la dimension polaire est considérée et l'analyse des indices spectraux suffit pour réaliser la prédiction. Cependant, si la largeur latérale de l'espace de prédiction dépasse le seuil de discrimination latéral (capacité du système auditif à détecter une différence d'angle latéral), alors il sera nécessaire de combiner les indices de similarité spectraux à des indices de similarité interauraux, tels que l'ITD pour le cas particulier des sources large bande (soit $w_{ITD} > 0$ selon l'équation 6.9).

Afin de satisfaire le critère du maillage régulier, notamment imposé par le calcul des probabilités et l'application du modèle de dispersion (plus de détails section 6.4.7.2), les distances D_i calculées sur les $N_S = 1500$ points de la grille originale de mesure sont automatiquement interpolées sur un maillage plus régulier avant la normalisation Z-score. Nous utiliserons une grille de type HyperInterpolation, quasi-régulière, comprenant le vide polaire respectif à la grille de mesure. Ce vide polaire correspond à la zone spatiale de la calotte inférieure de la sphère dépourvue de points, à cause des contraintes pratiques imposées par les systèmes de mesures de HRTFs. Comme présenté section 2.2.2, le système de mesure de l'IRCAM ne permet pas de mesurer d'élévations inférieures à $50.5°$. Ce vide polaire correspond également la zone où il n'est pas possible de pointer. Il est donc vraisemblable de ne pas interpoler les distances D_i sur cette zone de l'espace.

Les grilles de type HyperInterpolation peuvent être définie pour différents nombres de points. Pour éviter de procéder à un sur-échantillonnage spatial, nous utiliserons un échantillonnage dont le nombre de points équivaut à celui de la grille de départ. La grille d'HyperInterpolation à l'ordre 40 (suivant la théorie liée à la décomposition en harmoniques sphériques), théoriquement composée de 1681 points, se présente comme un bon compromis. Une fois la calotte inférieure supprimée (soit l'ensemble des points à élévation

$< 50.5°$), la grille comprend un total de 1488 points contre 1500 pour la grille originale. La superposition de ces deux grilles est donnée figure 6.5(a).

Etant donné que la prédiction se concentre sur un intervalle latéral, on extraira la tranche latérale de cette grille qui correspond à l'intervalle latéral d'étude (intervalle centré sur l'angle latéral théorique de localisation, comme expliqué ci-dessus). Selon le centre latéral de cet espace et pour un intervalle latéral théorique de largeur 40°, le nombre de points de l'espace de prédiction varie entre 150 et 450 points. En effet, un plus grand nombre de points est considéré pour des tranches latérales autour du plan médian. Les figures 6.5(b) et 6.5(c) présentent respectivement deux tranches latérales de la grille d'HyperInterpolation centrées respectivement sur les angles latéraux 8° et 42°.



(a) (b) (c)

FIGURE 6.5 – (a) Grille originale de mesure (points noirs) et grille d'interpolation complète c'est-à-dire la grille d'HyperInterpolation à l'ordre 40 théorique (points rouges). (b) Grilles d'interpolation complète (points noirs, équivalents aux points rouge de la figure (a)) et intervalle latéral de prédiction centré sur angle latéral 8° (points rouges). (c) Grille d'interpolation complète et intervalle latéral de prédiction centré sur angle latéral 42° (points rouges).

6.4.3 Evaluation du modèle selon le critère de log-vraisemblance

Les critères d'évaluation du modèle sont tirés des précédents modèles présentés section 6.2. Nous utilisons la log-vraisemblance aux réponses (*actual likelihood*) en parallèle de la log-vraisemblance associée à la loi de probabilité du modèle (*expected likelihood*) comme introduit par Langendijk et Bronkhorst [LB02].

Définition du critère de la log-vraisemblance La log-vraisemblance mesure la correspondance entre une distribution observée d'un échantillon et la loi de probabilité supposée décrire cette distribution. Plus la log-vraisemblance est faible meilleur est le modèle. La log-vraisemblance est notamment utilisée pour déterminer les paramètres optimaux qui offrent un modèle dont les sorties s'approchent au mieux des observations. De la même façon que Langendijk et Baumgartner ([LB02] p.1590, [BML14] p.799), nous calculons la log-vraisemblance actuelle (au sens de effective) à partir des probabilités prédites par le modèle à l'endroit des réponses $p(r)$ [LB02] :

$$L_a = \sum_{r=1}^{N_R} -2 \cdot \ln\left[p(r)\right] \tag{6.12}$$

avec N_R le nombre total de stimuli testés. Pour rappel (équation 6.11), la probabilité correspond aux indices de similarité normalisés par la somme de tous les indices de similarité sur l'espace de prédiction. Ainsi, plus le nombre de points sur l'espace de prédiction est grand, plus les valeurs de probabilité sont faibles, bien que les écarts relatifs soient conservés. Nous étudierons plus en détails ce phénomène un peu plus bas.

La log-vraisemblance actuelle L_a est comparée à la log-vraisemblance escomptée L_e (ou *expected likelihood*), calculée comme la moyenne de T tirages aléatoires (typiquement $T = 100$ [LB02]) de réponses $\tilde{r}$ qui suivent la loi de probabilité du modèle définie sur l'espace de prédiction. Elle peut se résumer selon l'équation suivante :

$$L_e = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{q=1}^{Q} -2 \cdot \ln\left[p(\tilde{r}_q^t)\right] \right) \tag{6.13}$$

où $\tilde{r}_q^t$ correspond aux indices des directions du t-ième tirage suivant la loi de probabilité P_q. L_a et L_e sont ensuite normalisés par la log-vraisemblance liée au hasard L_{chance} (ou *chance likelihood*) calculée à partir d'une loi de probabilité constante sur tout l'espace ($\frac{1}{N_{d|q}}$) :

$$L_{chance} = -2 \cdot N_R \cdot \ln(\frac{1}{N_{d|q}}) \tag{6.14}$$

Les log-vraisemblances L_a et L_e utilisées par la suite correspondent aux log-vraisemblances normalisées par L_{chance}.

Si la log-vraisemblance actuelle L_a est comprise dans l'intervalle de confiance à 99% de la log-vraisemblance escomptée L_e (noté aussi $CI_{99\%}(L_e)$), cela signifie que les réponses r suivent la loi de probabilité prédite par le modèle. L'intervalle de confiance est défini par :

$$CI_{99\%}(L_e) = \left[L_e - 2.58 \cdot \frac{\sigma(L_e)}{\sqrt{T}}; L_e + 2.58 \cdot \frac{\sigma(L_e)}{\sqrt{T}} \right] \tag{6.15}$$

Notons que L_a est basée sur des probabilités qui ont été interpolées aux directions pointées à partir des probabilités calculées en tous points de l'espace de prédiction alors que L_e est directement basée sur les probabilités de l'espace de prédiction.

Lors de l'analyse, nous aurons l'occasion d'observer plusieurs cas de figure :

— $L_a > CI_{99\%}(L_e)$: le sujet a pointé trop peu souvent dans les zones de forte similarité prédites par le modèle ou le modèle prévoit des zones de forte similarité trop concentrées dans l'espace

— $L_a \in CI_{99\%}(L_e)$: les réponses du sujets suivent la loi de probabilité du modèle, le modèle est capable de prédire les réponses des sujets

— $L_a < CI_{99\%}(L_e)$: le sujet a pointé trop souvent dans la zone de forte similarité prédite par le modèle (i.e. le sujet est plus précis que ce que prévoit le modèle), ou le modèle prévoit des zones de similarité trop dispersées dans l'espace

— $L_a \geq 1$: le sujet a pointé au hasard dans l'espace, ou le modèle prévoit une distribution de probabilité constante dans l'espace

Dépendance vis-à-vis du nombre de points de l'espace La difficulté de ce critère d'évaluation est que l'ordre de grandeur des valeurs dépend intimement du nombre de points considérés. En effet, la log-vraisemblance est basée sur des valeurs de probabilité issues de la division des indices de similarité par leur somme sur l'espace de prédiction. La comparaison directe des valeurs de L_a obtenues dans chaque étude est rendue difficile par ce fait. En effet, les études [LB02] et [BML14] considèrent des espaces à une dimension (prédiction dans la dimension polaire uniquement) dont le nombre de points est de l'ordre de 50 alors que nous effectuons une prédiction sur des espaces à $N_d|_q$ points où $N_d|_q$ varie globalement entre 150 et 450 selon la carte de similarité q considérée (plus de précisions seront données en section 6.4.2).

La figure 6.6 illustre cette caractéristique à travers une simulation de la prédiction de localisation sur la dimension polaire. La figure du haut présente la distribution de similarité sur le plan sagittal, avec un maximum localisé à l'angle polaire 30°. Les croix rouges indiquent la position de 5 réponses. La figure du milieu affiche la distribution de probabilité correspondante pour un espace plus ou moins discrétisé (entre 20 et 800 points). Enfin, la dernière figure offre un aperçu de l'influence du degré de discrétisation spatial sur les valeurs de log-vraisemblance associée aux réponses. On voit que plus le nombre de points est grand, plus les valeurs de probabilité sont faibles et donc plus la log-vraisemblance augmente, étant donné qu'elle est inversement proportionnelle à la probabilité aux réponses.

Dépendance vis-à-vis de la sélectivité spatiale de la distribution de probabilité La log-vraisemblance est également sensible au contraste, ou sélectivité spatiale, de la distribution de probabilité. Dans le cas idéal où les réponses apparaissent aux régions de forte probabilité, la log-vraisemblance actuelle atteindra des valeurs proches du seuil de chance ($L_a = 1$) pour une distribution lissée, i.e. avec des valeurs de probabilité peu contrastées dans l'espace. Au contraire, elle sera d'autant plus faible que les régions de forte probabilité sont en faible nombre et concentrées dans l'espace.

Pour illustrer ce phénomène, la figure 6.7 présente différentes distributions de probabilité avec un maximum plus ou moins marqué à $\Phi = 30°$, i.e. avec un écart-type plus ou moins important. La simulation concerne ici aussi une prédiction réalisée sur la dimension polaire uniquement. La log-vraisemblance escomptée permet de caractériser la sélectivité de chaque distribution. Elle est représentée par le point central des barres verticales oranges sur les figures du bas. On note qu'elle augmente avec l'écart-type σ c'est-à-dire que plus la distribution est sélective plus L_e est faible. En effet, L_e est obtenue par des tirages "aléatoires" suivant la loi de probabilité du modèle. Plus la région de forte probabilité est concentrée dans l'espace, plus la probabilité à cet endroit est élevée (relativement au reste de l'espace) donc plus on a de chance d'aller tirer des points à cet endroit et d'obtenir des probabilités élevées. La log-vraisemblance étant inversement proportionnelle aux probabilités issues des tirages, cela explique qu'elle diminue lorsque la concentration spatiale du maximum augmente. L'intervalle de confiance associé (taille des barres verticales) ne dépend pas du paramètre σ.

Les figures (a), (b) et (c) simulent respectivement des cas où le sujet pointerait autour du maximum de probabilité avec une certaine dispersion, exactement à la position du maximum (cas idéal) et de manière aléatoire. Les positions angulaires des réponses sont indiquées sur les distributions (figures du haut) par les barres verticales rouges pointillés. Les figures du bas permettent d'observer les résultats de log-vraisemblance associées à ces réponses (courbe bleue), selon la sélectivité de la distribution (définie par le paramètre σ). Premièrement, dans la figure (a), on voit que L_a se trouve au-dessus et hors de l'intervalle de confiance de L_e pour une distribution très sélective ($\sigma = 10$). Cela signifie que les réponses ont été

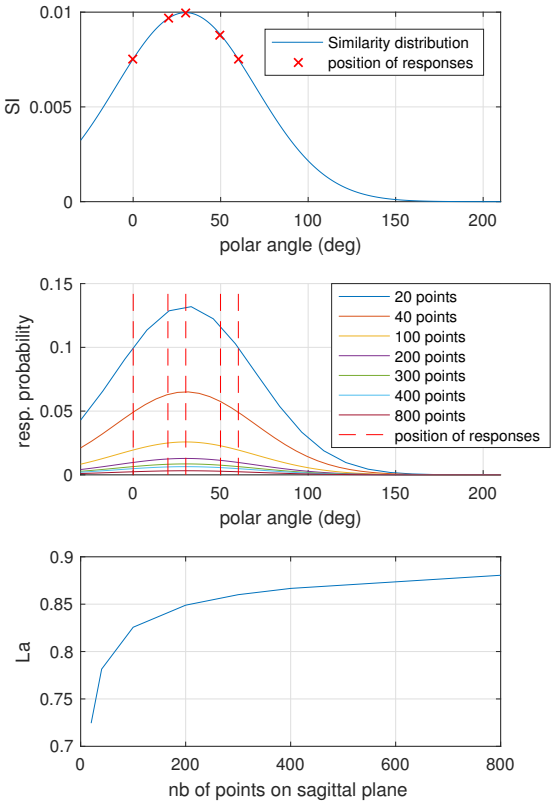
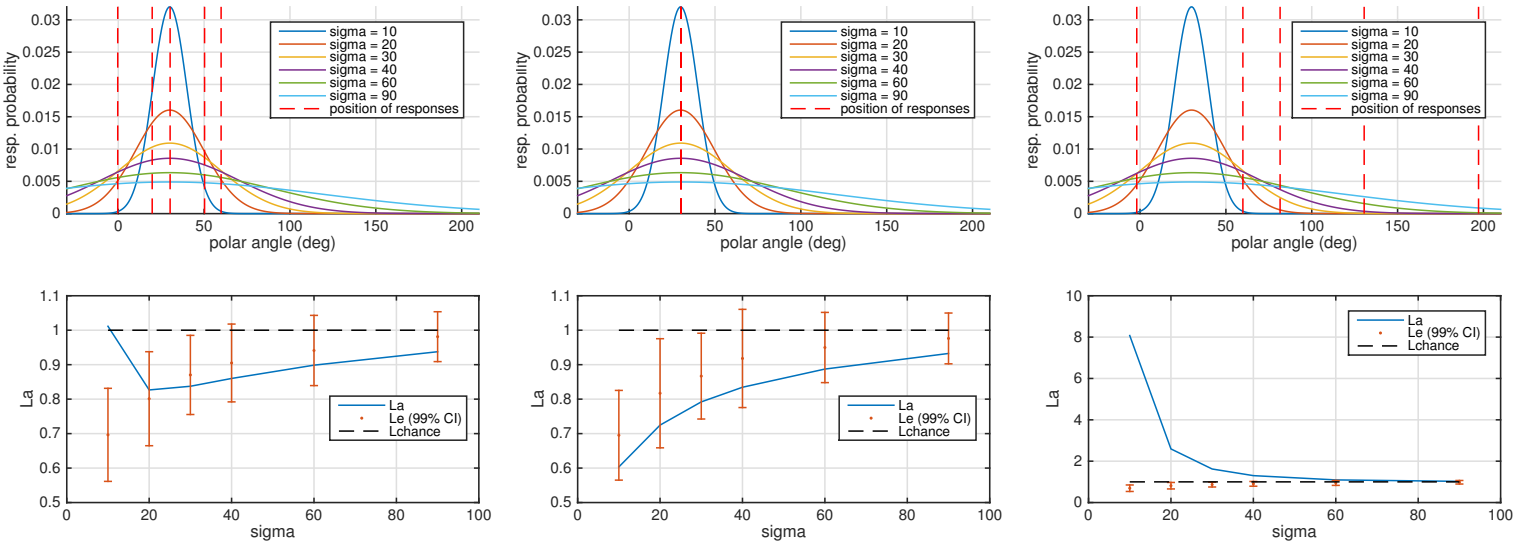


FIGURE 6.6 – Evolution des valeurs de log-vraisemblance actuelle L_a en fonction du nombre de points considéré sur l'espace de prédiction. Dans cette simulation, l'espace de prédiction est réduit à un plan sagittal dont la dimension polaire est matérialisée par l'axe x. De haut en bas sont présentées la distribution d'indices de similarité, les distributions de probabilité de réponse pour différents nombres de points et l'évolution de la log-vraisemblance associées au réponses.

données trop peu souvent dans la zone de forte probabilité. De plus, L_a se situe au niveau du seuil de chance ($L_a = 1$) alors que les réponses apparaissent autour du maximum. Ces deux observations indiquent que la distribution à $\sigma = 10$ est trop sélective vis-à-vis des réponses du sujet "normal". Pour $\sigma \geq 20$, la log-vraisemblance est comprise dans l'intervalle de confiance ce qui signifie que le modèle prédit correctement les réponses. Le minimum de log-vraisemblance associée aux réponses est atteint pour la distribution de probabilité à $\sigma = 20$. Cette distribution optimise donc les résultats de prédiction. A mesure que la distribution devient de plus en plus lissée (pour $\sigma > 20$), la log-vraisemblance actuelle se rapproche du seuil de chance. La figure (b) illustre le fait que lorsque les réponses apparaissent exactement au maximum de similarité, la log-vraisemblance actuelle est plus faible que dans le cas où les réponses apparaissent autour de ce maximum (figure (a)). Pour que cet avantage soit visible, il faut que la distribution soit suffisamment sélective. En effet, les résultats pour $\sigma = 90$ sont identiques dans les figures (a) et (b). La figure (c) illustre le fait que si les réponses sont données de manière aléatoire sur l'espace de prédiction alors on obtient $L_a \geq 1$, avec des résultats globalement hors et au-dessus de l'intervalle de confiance de la log-vraisemblance escomptée.

Sélectivité spatiale des distributions et mise en évidence des paramètres optimaux
La simulation précédente a permis de prendre conscience que si la sélectivité spatiale était mal ajustée, i.e. trop ou trop peu sélective, les résultats pouvaient être affectés, comme par exemple dépasser le seuil de chance alors que les réponses apparaissent autour du maximum de probabilité. Il est donc nécessaire d'ajuster la sélectivité spatiale des distributions de probabilité à la dispersion des réponses des sujets, sans quoi nous ne pourrons comparer correctement les valeurs de log-vraisemblance actuelle entre les paramètres du modèle.

Un des paramètres d'intérêt dans cette étude concerne la métrique spectrale. Nous souhaitons mettre en évidence la métrique qui prédit aux mieux les directions pointées, i.e. celle qui prédit la plus importante probabilité aux directions pointées. Cela revient à identifier la métrique spectrale qui minimise la log-vraisemblance actuelle. Comme nous le verrons dans la section 6.4.5.4, les métriques possèdent un caractère plus ou moins sélectif spatialement. L'objectif de l'étude n'est pas de comparer les métriques par rapport à ce point mais bien par rapport à la correspondance entre les zones de forte probabilité et les réponses des sujets. Si une métrique est par nature trop ou trop peu sélective, il se peut que sa fiabilité soit masquée



(a) Moyen (b) Très bon (c) Très mauvais

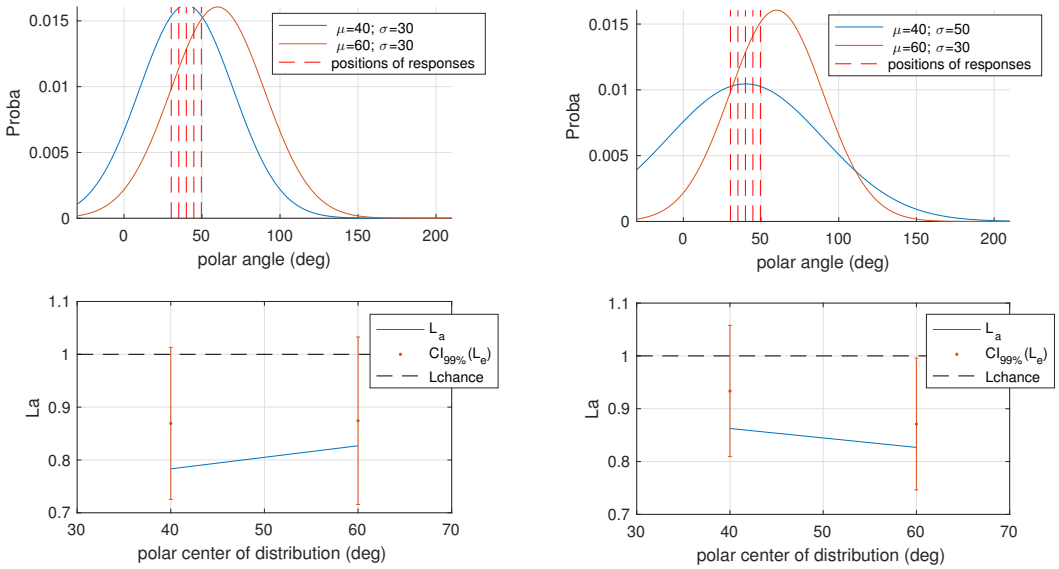
FIGURE 6.7 – Evolution des valeurs de log-vraisemblance en fonction de la sélectivité de la distribution de probabilité, pour trois cas de localisation simulés dans la dimension polaire (sujet moyen, idéal ou très mauvais). En haut, figurent les distributions spatiales de probabilité avec une sélectivité variable (plusieurs couleurs) et inversement proportionnelle à σ. Les barres verticales rouges en pointillés indiquent les directions pointées. En bas, la courbe bleue représente l'évolution de L_a et les barres verticales oranges, celle de L_e (ainsi que son intervalle de confiance à 99%). Le seuil de chance est présenté en pointillés noirs.

par cette propriété, comme on a pu l'observer figure 6.7(a).

La figure 6.8 illustre ce phénomène. Imaginons que les distributions bleue et orange représentées dans les figures du haut soient associées aux métriques spectrales A et B, respectivement. En haut de la figure, on observe que le maximum de probabilité de la distribution associée à la métrique A est localisé à l'endroit même des réponses. La métrique B, elle, prédit un maximum de probabilité légèrement décalé par rapport aux réponses. La métrique A permet donc de prédire les directions pointées de manière plus précise que la métrique B. Dans la figure 6.8(a), la sélectivité spatiale (caractérisée par σ) des distributions associées aux deux métriques sont identiques. On observe alors que la log-vraisemblance actuelle est minimisée avec la métrique A. Dans la figure 6.8(b), la sélectivité spatiale des distributions associées aux métriques A et B diffèrent : la distribution bleue est moins sélective que la distribution orange. On observe alors que la log-vraisemblance actuelle est minimisée avec la métrique B. La log-vraisemblance actuelle obtenue dans le cas (a) traduit correctement l'avantage de la métrique A par rapport à la métrique B. Cependant, le cas (b) met en évidence l'importance d'adapter la sélectivité spatiale à la métrique afin d'obtenir des résultats fiables.

Conclusion Les simulations menées dans cette section ont permis d'illustrer le comportement du critère de log-vraisemblance en fonction de différentes propriétés des métriques et performances de localisation des sujets. La première simulation a révélé une dépendance de l'ordre de grandeur des valeurs de log-vraisemblance actuelle selon le nombre de points considérés sur l'espace. Cela met en évidence que si les modèles n'utilisent pas le même nombre de points, leurs résultats respectifs peuvent ne pas être comparables. La seconde simulation a montré que la valeur de L_a dépend de la correspondance spatiale entre les zones de forte probabilité et les directions pointées. C'est exactement ce que nous cherchons à quantifier. L'idée est d'identifier un modèle qui prédise des distributions spatiales de probabilité en adéquation avec les réponses des sujets, i.e. qui minimise la log-vraisemblance actuelle. Cette simulation a également mis en évidence la nécessité d'adapter la sélectivité spatiale des distributions à la dispersion des réponses des sujets, et que la sélectivité spatiale pouvait être caractérisée par la valeur moyenne de la log-vraisemblance escomptée.

La log-vraisemblance actuelle dépend donc de trois facteurs : (1) la capacité du sujet à identifier les régions de forte similarité et à les indiquer avec précision ; (2) la capacité du modèle (selon les paramètres qui le constituent e.g. les métriques spectrales) à définir des zones de fortes probabilité localisées aux directions pointées ; (3) la sélectivité des distributions de probabilité. Enfin, la troisième simulation a mis en évidence que, pour que la log-vraisemblance actuelle soit un critère fiable, la sélectivité spatiale de chaque distribution doit être ajustée de sorte à minimiser la log-vraisemblance actuelle aux réponses.



(a) Sélectivité égale

(b) Sélectivité variable

FIGURE 6.8 – Evolution des valeurs de log-vraisemblance actuelle en fonction de la correspondance spatiale entre le maximum de probabilité et les réponses. En haut, figurent les distributions spatiales de probabilité dans la dimension polaire centrées respectivement sur l'angle polaire $\Phi = 40°$ et $\Phi = 60°$ (courbes bleue et orange, respectivement) avec une sélectivité identique ($\sigma = 30$) dans la figure (a) ou variable ($\sigma = 50$ et $\sigma = 30$) dans la figure (b). Les barres verticales rouges en pointillés indiquent les directions pointées. En bas, la courbe bleue représente la valeur de L_a associée à chaque distribution et les barres verticales oranges, les valeurs de L_e associées (avec son intervalle de confiance à 99%). Le seuil de chance est présenté en pointillés noirs.

De plus, la comparaison entre la log-vraisemblance actuelle et l'intervalle de confiance de la log-vraisemblance escomptée nous a permis de juger de la capacité du modèle à reproduire les directions pointées. Ce critère est indispensable si l'objectif est de calibrer un modèle pour prédire ensuite la localisation d'autres stimuli qui n'ont pas fait l'objet d'un test expérimental. Cependant, la présente étude s'intéresse plutôt à identifier les paramètres qui expliquent au mieux les observations expérimentales. Le problème d'optimisation est donc simplifié par l'estimation des paramètres du modèle qui minimisent la log-vraisemblance actuelle, i.e. qui maximisent la probabilité aux directions pointées. Le nombre de fois où la log-vraisemblance actuelle est comprise dans l'intervalle de confiance de la log-vraisemblance escomptée pour l'ensemble des sujets sera étudiée plus particulièrement dans le cadre de la comparaison de notre modèle avec les modèles précédents [LB02, BML14]. Nous parlerons alors de pourcentage de réussite pour désigner le pourcentage de fois où $L_e - CI_{99\%}(L_e) < L_a < L_e + CI_{99\%}(L_e)$.

Pour finir, comme on l'a vu en section 6.1, les modèles de prédiction n'ont pas toujours utilisé le critère de la log-vraisemblance actuelle pour quantifier le score de prédiction. Baumgartner et al. [BML13, BML14] se sont principalement focalisés sur la corrélation entre les résultats PE (*polar error*) et QE (*quadrant error*) prédits par le modèle et mesurés dans des tests de localisation. Les valeurs prédites sont obtenues par une intégration des erreurs quadratiques multipliées par leur probabilité d'apparition (voir équations 6.3 et 6.4). Ces critères sont peu adaptés à notre modèle étant donné que la prédiction n'est pas à une seule dimension. De plus, nous considérons une approche locale, où l'évaluation de la prédiction se concentre sur la probabilité à chacune des directions pointées, plutôt que globale. Par ailleurs, Middlebrooks [Mid99b] a utilisé les critères du pourcentage correct et du d'. Ces paramètres n'offrent pas la notion de probabilité de réponse et permettent plus difficilement d'identifier lorsque les réponses s'approchent du seuil de chance. Ils ne seront donc par utilisé pour la paramétrisation du modèle. Cependant, nous les calculerons dans le cadre d'une comparaison des résultats avec ceux de Middlebrooks.

6.4.4 Des distances aux indices de similarité et fonctions sigmoïdes

Les valeurs de distances calculées selon les différentes métriques spectrales s'expriment dans différentes unités et sont distribuées sur des intervalles de largeur divers. Afin de pouvoir les comparer, les valeurs des distances sont normalisées selon la méthode de standardisation Z-score (équation 6.7) qui permet de les re-dimensionner de telle sorte à ce que leur distribution soit centrée en 0 avec une déviation standard de 1. Notons que l'intervalle des valeurs de distances normalisées reste variable, que ces valeurs sont sans dimension et que la normalisation doit se faire sur l'espace de prédiction. La combinaison des distances interaurales et spectrales est alors rendue possible par la normalisation et c'est pour cette raison qu'elle est effectuée en amont de la combinaison. Cette méthode de normalisation a notamment été utilisée par Middlebrooks [Mid92].

Les distances normalisées sont ensuite transformées en indices de similarité compris entre 0 et 1. Bremen et al. [BvWvO10] ont utilisé une fonction de transformation linéaire, qui consiste à assigner une similarité de 0 à la distance maximale et de 1 à la distance minimale. Ce type de transformation est très dépendante des valeurs de distances prises sur l'espace de prédiction et donc très sensible aux valeurs aberrantes. Par exemple, si les distances normalisées se regroupent autour de 0 mais qu'il existe quelques valeurs plus élevées qui se détachent de la tendance globale, la présence de ces valeurs affectera la transformation et concentrer la majorité des indices de similarité vers 1. Dans le modèle de Middlebrooks, les valeurs de distances (en terme d'ILD) sont multipliées par (-1) ce qui correspond à une transformation linéaire mais non réduite à l'intervalle [0,1]. Les modèles de Baumgartner et al. [BML14] et Langendijk et Bronkhorst [LB02] utilisent par ailleurs des fonctions de transformation non-linéaires, gaussiennes ou sigmoïdes. La fonction sigmoïde est souvent utilisée comme fonction psychométrique afin de modéliser la probabilité de détection d'un stimulus. Elle ressemble à une fonction gaussienne à un seul côté mais est définie par deux paramètres Γ et S (voir équation 6.5) dont les effets respectifs peuvent être visualisés en figure 6.9. Le paramètre Γ détermine la pente et le paramètre S, le point d'inflexion. Ce type de transformation modélise le processus de localisation auditive en faisant l'hypothèse qu'il existe un seuil de dissimilarité à partir duquel l'auditeur perçoit une différence et que, pour des dissimilarités croissantes, la probabilité de réponse décroît lentement vers zéro. Il a l'avantage d'être indépendant de la distribution des valeurs de distances contrairement à une transformation linéaire.

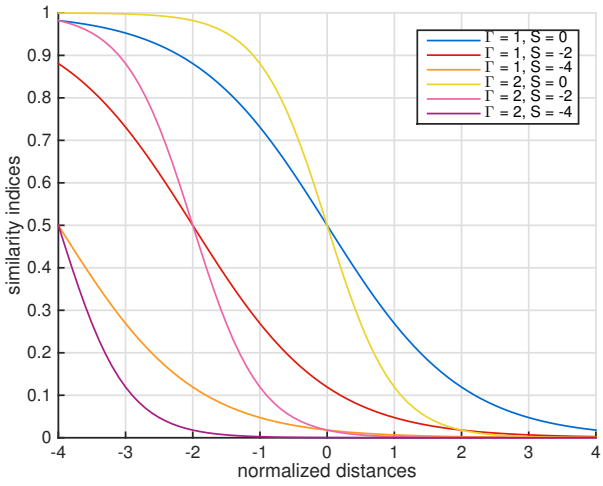


FIGURE 6.9 – Fonctions sigmoïdes à paramètres variables Γ et S, utilisées pour transformer les distances normalisées (axe x) en indices de similarité (axe y).

Dans la figure 6.11, on peut observer l'influence des fonctions sigmoïdes de la figure 6.9 sur la distribution des valeurs de distances et des indices de similarité sur l'espace de prédiction. Par colonne, sont représentées : (1) la distribution des valeurs de distances spectrales pour chacune des 10 métriques testées, (2) l'histogramme des valeurs de distances normalisées (entre -5 et 5), puis (3 et $+$) l'histogramme des indices de similarité compris entre 0 et 1 selon la fonction sigmoïde utilisée.

La fonction sigmoïde permet d'ajuster la sélectivité spatiale des distributions de probabilité. La figure 6.10 illustre le caractère de sélectivité du modèle associé à chaque fonction sigmoïde de la figure 6.9. Comme présenté dans la section précédente, celui-ci peut être quantifié au moyen de la log-vraisemblance associée à des tirages suivants la loi de probabilité du modèle, i.e. la log-vraisemblane escomptée L_e, dont l'intervalle de confiance à 99% et la moyenne sur les T tirages sont respectivement représentés par les barres d'erreur et le centre de celles-ci. Pour rappel, plus L_e est faible plus la sélectivité spatiale est grande. On observe que, selon les paramètres Γ et S utilisés, les distributions de probabilité deviennent plus ou moins sélectives. Pour un Γ fixé, la sélectivité augmente avec S décroissant. Lorsque $\Gamma = 2$, cette tendance est

beaucoup plus marquée que lorsque $\Gamma = 1$. L'effet de S est donc davantage marqué que Γ est grand. Cette observation traduit une dépendance de l'effet associé à chacun des paramètres Γ et S. Pour cette raison, ils ne seront pas étudiés de manière indépendante.

Comme nous l'avons vu dans la section précédente, pour que le critère de log-vraisemblance actuelle soit un critère fiable d'évaluation des paramètres du modèle, il est nécessaire d'ajuster la sélectivité spatiale de chaque distribution, de manière à minimiser la log-vraisemblance aux réponses. La fonction sigmoïde fera donc l'objet d'une paramétrisation spécifique à chaque métrique spectrale. Nous étudierons également l'intérêt d'adapter la fonction sigmoïde à chaque individu dont nous avons collecté les données expérimentales de localisation, tel que suggéré par Baumgartner et al. Cependant, comme mentionné dans l'introduction section 6.3, il n'est pas souhaitable de devoir systématiquement individualiser la fonction sigmoïde à chaque nouvel individu. En effet, bien que le modèle sera établi à partir de données de localisation de sujets dont on possède les HRTFs, le but applicatif correspond au cas on ne connait pas les HRTFs du sujet pour lequel des données de localisation ont été relevées.

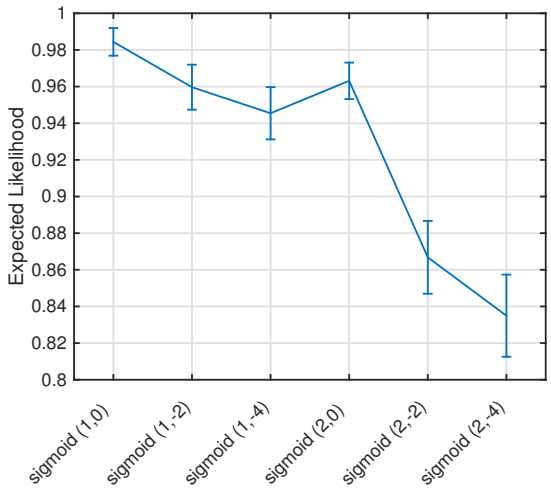


FIGURE 6.10 – Log-vraisemblance escomptée associée au modèle pour chaque fonction de sigmoïde utilisée. La log-vraisemblance escomptée L_e a été moyennée sur les $T = 100$ tirages (c.f. équation 6.13), les 10 métriques spectrales et les sujets ayant participé à l'expérience décrite au chapitre 7. Les barres verticales représentent l'intervalle de confiance moyen associé aux tirages.

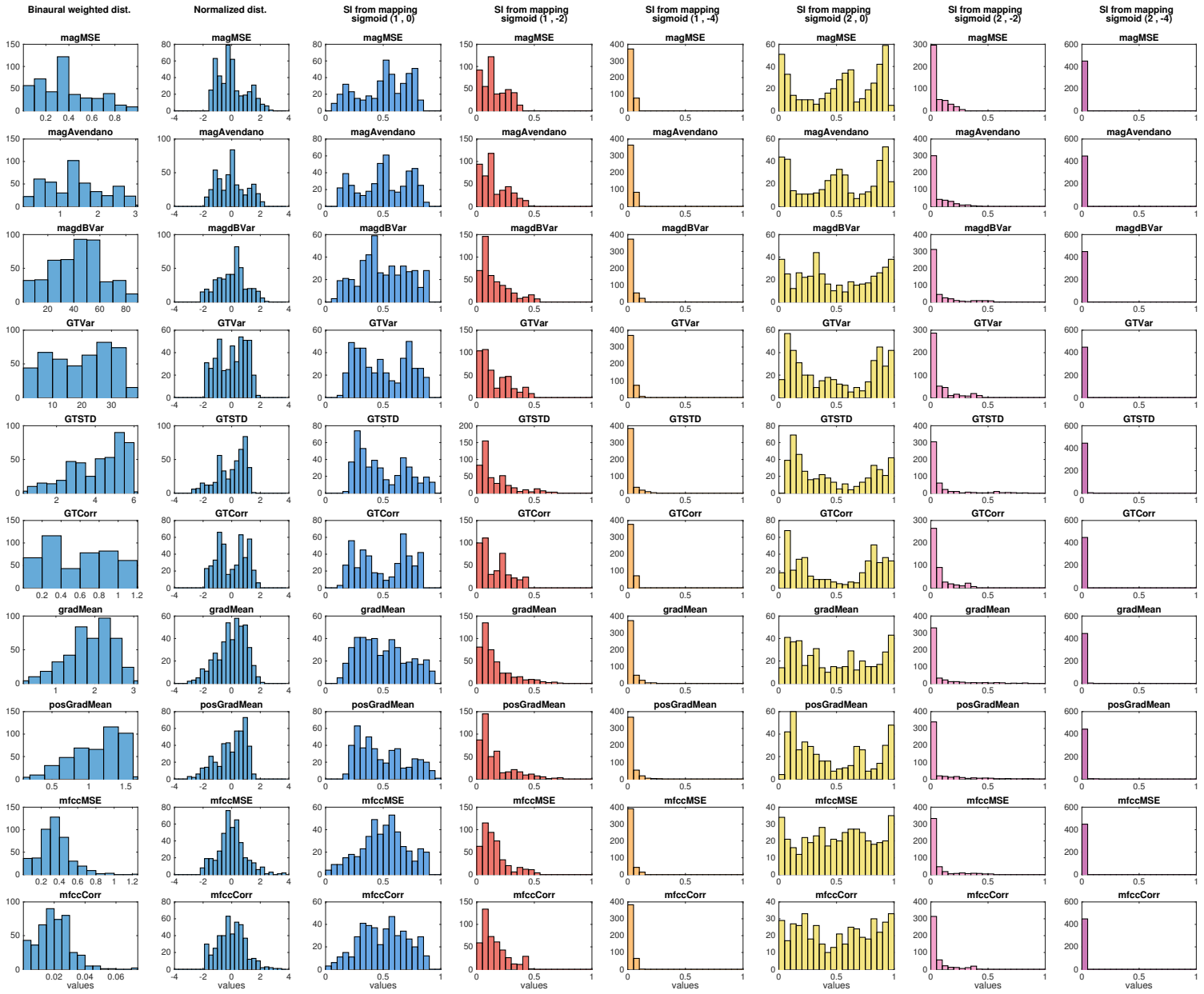


FIGURE 6.11 – Illustration de l'évolution des distributions de valeurs de distances et d'indices de similarité pour les différentes métriques et les différentes fonctions sigmoïdes. Dans la première colonne figure l'histogramme de distances spectrales (brutes, combinées gauche et droite) entre la HRTF mesurée à la direction (30°,67.5°) et les autres HRTFs de l'espace de prédiction (cas d'une écoute en condition individuelle). La deuxième colonne présente l'histogramme des valeurs de distances normalisées. Les colonnes 3 à 8 correspondent aux histogrammes des indices de similarité obtenus après application des différentes fonctions sigmoïdes (indiquées en haut de chaque colonne). Par ligne, sont présentés les résultats pour chacune des 10 métriques spectrales testées.

6.4.5 Métriques de similarité

Afin de déterminer la probabilité de réponse associée à chaque point de l'espace de prédiction, les HRTFs utilisées pour la synthèse du stimulus sont comparées aux HRTFs de l'auditeur sur la base d'un calcul de similarité entre les HRTFs. Nous introduisons ici les métriques spectrales qui seront testées au sein du modèle ainsi que certaines caractéristiques qui les distinguent et qu'il est nécessaire de prendre en main afin d'appréhender la comparaison. Les HRTFs sont également comparées selon leurs indices interauraux. Le traitement des indices spectraux et interauraux est réalisée séparément puis une combinaison pondérée des indices de similarité issus de chaque type de comparaison (interaurale et spectrale) est effectuée. Le choix des poids associés à chacun des indices est l'un des paramètres du modèle de prédiction.

6.4.5.1 Métriques spectrales

La définition d'une métrique de comparaison spectrale entre HRTFs consiste à combiner un mode de représentation des HRTFs et un critère de distance. Les principales métriques ayant déjà été appliquées dans la littérature ont été présentées en section 5.1. On se rapportera à cette section pour plus d'informations sur les métriques testées ici, notamment au niveau des équations associées.

Le choix des métriques étudiées suit les premières observations effectuées en section 5.2. Dans cette analyse, nous avions observé une très forte ressemblance entre les critères de Fahn et d'Avendano. De la même façon que Nicol et al. [NLBB06], il semblerait donc raisonnable de continuer l'analyse des métriques en supprimant le critère de Fahn qui offre des résultats identiques au critère d'Avendano. Bien qu'elles présentent des caractères proches, les métriques *GT-STD*, *GT-Var* et *GT-Corr* ainsi que *grad-Mean* et *posGrad-Mean* seront étudiées car elles ont fait l'objet d'un intérêt particulier dans les précédents modèles de prédiction. Concernant la représentation par le cepstre, celle-ci avait été jugée trop peu robuste à des traitements appliqués hors de la bande fréquentielle audible. Par conséquent, nous ne conserverons que la représentation MFCC, basée sur une échelle perceptive. Il sera alors intéressant d'étudier l'effet de deux critères de distance tels que l'inter-corrélation et la distance MSE.

Les métriques spectrales utilisées par la suite sont répertoriées dans le tableau 6.1. Les abréviations employées utilisent la règle [*mode de représentation des HRTFs - critère de distance*].

Mode de représentation	Critère de distance	Abréviation
Spectres d'amplitude linéaire	MSE	*mag-MSE*
Spectres d'amplitude linéaire	critère d'Avendano	*mag-Avendano*
Spectres d'amplitude en dB	variance (Durant)	*magdB-Var*
Profil spectraux	variance	*GT-Var*
Profil spectraux	standard déviation	*GT-STD*
Profil spectraux	inter-corrélation normalisée	*GT-Corr*
Dérivée première des profils spectraux	moyenne	*grad-Mean*
Dérivée première positive des profils spectraux	moyenne	*posGrad-Mean*
[1 : 20] premiers MFCCs	MSE	*mfcc-MSE*
[1 : 20] premiers MFCCs	inter-corrélation normalisée	*mfcc-Corr*

Table 6.1 – Métriques spectrales comparées au sein du modèle de prédiction et définition des abréviations utilisées.

Voici quelques précisions quant à l'utilisation de ces diverses métriques.

- Les profils spectraux correspondent aux HRTFs moyennées (moyenne RMS) par bande fréquentielle (c.f. équation 5.11) et sont exprimés en dB. Les bandes fréquentielles sont définies par un banc de 28 filtres gammatone de largeur de bande suivant l'échelle ERB et espacés en fréquence par 1 ERB. Les fréquences minimale et maximale de définition du banc de filtres sont $f_{min} = 700\,\text{Hz}$ et $f_{max} = 18\,\text{kHz}$ en référence à [BML13] et [MBL14], et correspondent approximativement à la fréquence minimale à laquelle intervient les réflexions du torse [AAD01] et à la fréquence maximale audible.
- Pour les 3 métriques basées sur les spectres d'amplitude (en échelle linéaire ou dB), la comparaison sera limitée au même intervalle fréquentiel que pour la représentation par bande fréquentielle soit [700 Hz, 18 kHz].
- La dérivée première positive des profils spectraux correspond aux uniques valeurs positives de la dérivée première des profils spectraux comme suggéré dans [BML14] et présenté figure 5.1 (valeurs négatives = 0).
- Lors de la comparaison des coefficients MFCCs, seuls les $P = 20$ premiers coefficients sont conservés (en référence à Lee et Lee 2011).
- Les valeurs d'inter-corrélation normalisée ($NCC \in [-1, 1]$) sont transformées en valeurs de distances comprises sur l'intervalle borné $D \in [0, 2]$ en appliquant $D = (NCC + 1) \times (-1)$.

6.4.5.2 Métrique interaurale

La prédiction de l'angle latéral perçu est basée sur le calcul de la différence entre les indices interauraux délivrés par la paire de HRTFs cible et les indices interauraux relatifs à l'auditeur. Etant donné que les sources sonores utilisées dans le test de localisation sont à bande fréquentielle large, l'ITD est l'indice dominant la localisation dans la dimension horizontale [WK92, MM02a]. La prédiction de localisation latérale est donc basée sur une analyse de l'indice d'ITD. Comme précédemment observé par Middlebrooks [Mid99b], l'auditeur localise à la position latérale où son ITD correspond à celui de la cible. Comme proposé par Lee et Lee [LpL11], la métrique utilisée pour calculer les différences d'ITDs est définie par la différence quadratique :

$$D_{ITD_i} = (\tau_\alpha - \tau_\beta)^2 \tag{6.16}$$

où τ_α et τ_β sont les ITDs des paires de HRTFs α et β à comparer, en échantillons.

Suivant les modèles neurologiques de Jeffress [Jef48] et plus récemment de Fischer et al. [FCP08], l'information d'ITD serait extraite des signaux reçus aux oreilles gauche et droite selon un mécanisme d'inter-corrélation. Par conséquent, l'ITD τ est ici estimé avec la méthode du maximum de l'inter-corrélation des HRIRs gauche et droite (*MaxIACCr*). Cette méthode, également utilisée par Lee et Lee [LpL11], sera comparée à d'autres méthodes d'estimation au moment de l'analyse des résultats.

6.4.5.3 Métriques spectrales et information interaurale

A l'exception du gradient spectral, tous les modes de représentation des HRTFs sont dépendants d'un gain global. Comme on l'a vu dans la section 5.1, les premiers coefficients MFCCs encodent cette caractéristique. Cela signifie que la différence (ou la différence quadratique) entre une HRTF ispsilatérale et une HRTF controlatérale sera importante, même si le spectre est très semblable. Ainsi, les métriques qui sont sensibles à la différence de gain encodent intrinsèquement une information relative à une différence qui s'apparente à l'ILD. Le critère de distance joue aussi un rôle important dans la prise en compte du gain. En effet, un critère tel que l'inter-corrélation ne s'intéresse qu'à la ressemblance dans le profil des items qui sont comparés. C'est le cas également de la variance ou déviation standard de la différence qui quantifie l'écart à la moyenne, peu importe la valeur de celle-ci.

Cette caractéristique est importante à prendre en compte dans le cadre du modèle de prédiction. L'ILD est, comme l'ITD, porteuse d'information latérale. Cela signifie que la combinaison des indices de similarité issus des métriques spectrales et interaurales n'aura pas le même impact selon la sensibilité de la métrique spectrale à une différence de gain global. Pour une métrique spectrale qui encode l'information de l'ILD, i.e. qui est sensible aux différences de gain global, l'information interaurale apportée lors de la combinaison sera redondante.

Pour caractériser cette redondance d'information, nous évaluons ici la corrélation entre les cartes de distances spectrales et interaurales (basée sur l'ITD), pour chacune des métriques spectrales testées. Cette méthode a déjà été utilisée dans l'analyse du modèle de Middlebrooks [Mid92] pour caractériser la dépendance de chacun des indices de localisation. Les coefficients de corrélation sont moyennés un ensemble de cartes de similarité (plus précisément, l'ensemble des Q cartes uniques associées au test de localisation du chapitre 7) afin d'obtenir une valeur typique par métrique. Les résultats sont visibles en figure 6.12. Plus la corrélation est importante, plus la métrique spectrale suit la tendance liée à l'indice interaural. Tout d'abord, la caractéristique extraite des HRTFs (spectre d'amplitude, MFCC, etc.) joue un rôle important sur le coefficient de corrélation. La métrique MSE appliquée sur les MFCCs présente un coefficient de corrélation maximal de 0.85. En effet, les premiers coefficients MFCCs encodent par défaut la caractéristique du gain global. Lorsque le premier coefficient MFCC est retiré, on note une baisse du coefficient de corrélation (résultats non présentés). On note aussi une baisse de 0.65 à 0.38 lorsque la variance est appliquée respectivement sur les niveaux par bande de fréquence (*GT-Var*, en dB) ou sur le spectre d'amplitude en dB (*magdB-Var*). Les niveaux par bande de fréquence encodent donc davantage le gain global par rapport au spectre d'amplitude en dB. Le critère de distance a également un impact important sur le coefficient de corrélation. On voit que lorsqu'on utilise l'inter-corrélation sur les MFCCs, le coefficient de corrélation est réduit à 0.65 par rapport à 0.85 avec le critère MSE. Enfin, les métriques *grad-Mean* et *posGrad-Mean* sont les métriques les moins corrélées à l'indice d'ITD. En effet, la représentation par le gradient élimine la caractéristique de gain global.

6.4.5.4 Sélectivité spatiale des métriques spectrales

Chaque métrique spectrale se distingue également par son caractère de sélectivité spatiale. Certaines métriques, très sensibles à de faibles variations spectrales, sont très discriminantes. Par conséquent, les zones de forte similarité sont très concentrées dans l'espace, la distance spectrale croît rapidement avec la distance angulaire et les cartes de similarité sont très contrastées. Les caractéristiques de sélectivité spatiale de 3 métriques spectrales peuvent être visualisées en terme d'étalement des zones de similarité (zones rouges) sur la figure 6.13. Dans cet exemple, les indices de similarité sont calculés exceptionnellement selon $(-1) \times D_{norm}$, i.e. sans le recours à une fonction sigmoïde quelconque.

Une autre manière de visualiser le caractère sélectif de chaque métrique spectrale est d'étudier les distributions de probabilité de réponse sur le plan médian, comme présenté figure 6.15, et de manière analogique aux articles de la littérature (voir par exemple la figure 7 de [LB02] ou la figure 3 de [BML14]).

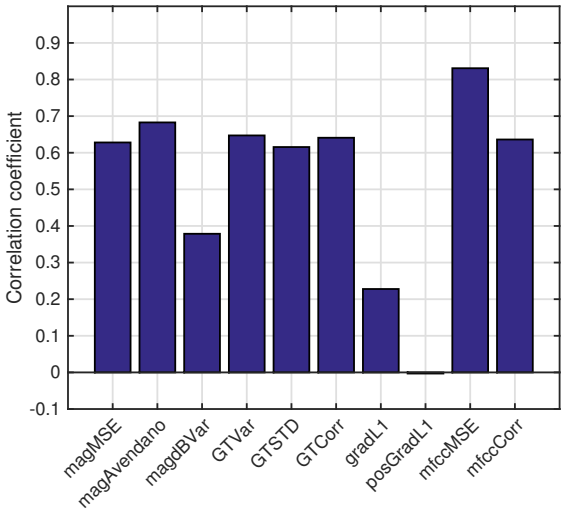
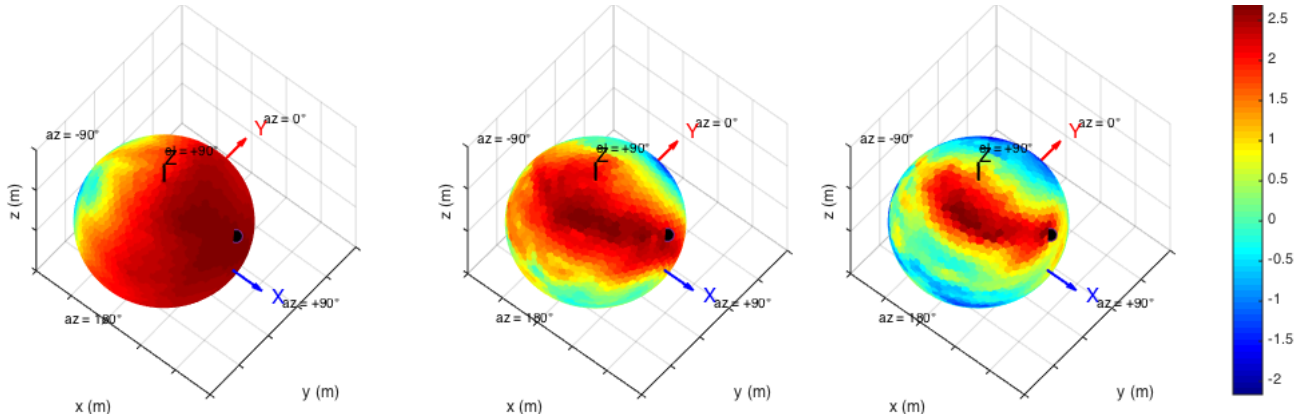


FIGURE 6.12 – Coefficients de corrélation entre les cartes spatiales de distances spectrales, obtenues pour chacune des 10 métriques spectrales, et la carte spatiale de distance d'ITD. Ils se présentent comme des indicateurs de la dépendance de chaque métrique spectrale envers les différences de gain global entre HRTFs et de la quantité d'information interaurale portée par chaque métrique spectrale.



(a) *mfcc-MSE* (b) *magdB-Var* (c) *grad-Mean*

FIGURE 6.13 – Exemple de cartes d'indices de similarité normalisés sur tout l'espace pour 3 métriques à caractère de sélectivité spatiale croissant de (a) vers (c). La direction test, située à $(78°, 26.5°)$, est représentée par le point noir.

Pour obtenir de telles figures, les HRTFs de l'auditeur α et celles d'un sujet cible β sont comparées en chaque point $(0, \Phi_t)$ du plan médian, défini par l'angle polaire $\Phi_t \in [-50°, +250°]$. Les distances obtenues sont transformées en indices de similarité à travers l'utilisation d'une fonction sigmoïde (dans ces exemples, (Γ, S)=(1,-4)). La matrice carrée est ensuite utilisée pour calculer les distributions de probabilité associées à chaque HRTF cible du plan médian appartenant au sujet β. Ainsi, pour chaque point en abscisse, la distribution verticale indique la probabilité que le sujet a de répondre en chaque point du plan médian.

L'exemple présenté figure 6.15(a) concerne les distributions de probabilité en condition individuelle, i.e. pour un auditeur à l'écoute de ses propres HRTFs ($\alpha = \beta$). Le caractère sélectif des métriques est évident : plus la similarité est concentrée sur la diagonale, plus la métrique est sélective. Notons que par définition, la comparaison spectrale entre une HRTF individuelle à une direction (Θ_t, Φ_t) et les HRTFs de l'auditeur est maximale à la direction (Θ_t, Φ_t). Les métriques *grad-Mean* et *posGrad-Mean* sont les plus sélectives, la probabilité que le sujet réponde en une autre direction que la direction test est quasi nulle. Au contraire, les métriques spectrales *mag-MSE* ou *mfcc-MSE*, présentent, pour les directions autour du pôle $\Phi = 90°$, une probabilité très diffuse spatialement. Cette propriété avait déjà été observée par Langendjk et Bronkhorst [LB02] pour la métrique *GT-STD*.

Les figures 6.15(b), 6.15(c) et 6.15(d) présentent les distributions de probabilité pour des HRTFs cibles issues des sujets β_1, β_2 et β_3. On observe tout d'abord que les motifs de probabilité sont plus étalés

et plus complexes. Les diagonales inverses indiquent une forte probabilité de confusions avant-arrière. Aussi, on note quelques différences entre les métriques qui ne sont pas basées sur le même mode de représentation. Cette particularité est intéressante étant donné que nous souhaitons élire la métrique qui prédise un maximum de probabilité localisé aux directions pointées. Le fait que les maxima ne soient pas systématiquement localisés aux mêmes endroits selon les métriques justifie l'intérêt cette comparaison. De plus, on observe que les prédictions varient d'un sujet à l'autre (différences entre figures (b), (c) et (d)). Cette observation est intéressante car le but de cette étude est d'étudier la prédiction des directions perçues à l'écoute de HRTFs non-individuelles issues de différents individus. Le choix des HRTFs non-individuelles testées au sein du test de localisation de sources virtuelles, présenté au chapitre 7, visera à maximiser les cas où le maximum de probabilité est marqué et éloigné de la diagonale, ainsi qu'à solliciter la plus grande variété de cartes de similarité pour une même direction test (maxima distribués à différentes directions du plan sagittal). Nous en reparlerons section 7.1.5.

Les distributions de probabilité de réponses associées aux métriques basées sur le même mode de représentation, telles que *mag-MSE* et *mag-Avendano* ou *mfcc-MSE* et *mfcc-Corr*, sont très semblables. On remarque notamment que les métriques *GT-Var* et *GT-STD* prédisent exactement les mêmes zones de forte et faible probabilité de réponse. En effet, la métrique *GT-STD* se distingue de *GT-Var* uniquement par l'application d'une racine carrée, ce qui explique notamment que cette métrique possède un caractère plus contrasté que la métrique *GT-Var*. Après adaptation de la sélectivité spatiale des distributions de probabilité pour chacune de ces métriques, celles-ci devraient offrir les mêmes résultats de prédiction.

Pour finir, la sélectivité spatiale des métriques spectrales peut être quantifiée en termes de la log-vraisemblance escomptée associée au modèle pour chacune des métriques spectrales. La figure 6.14 présente les valeurs de log-vraisemblance escomptée L_e. Plus L_e est faible, plus la sélectivité spatiale de la métrique spectrale est prononcée. On voit que les métriques basées sur le gradient spectral sont les plus sélectives et que celles basées sur les MFCCs sont les moins sélectives, ce qui est en accord avec les observations précédentes.

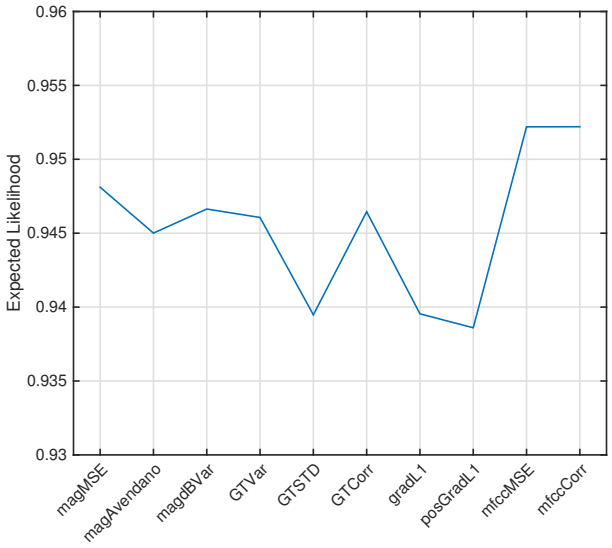
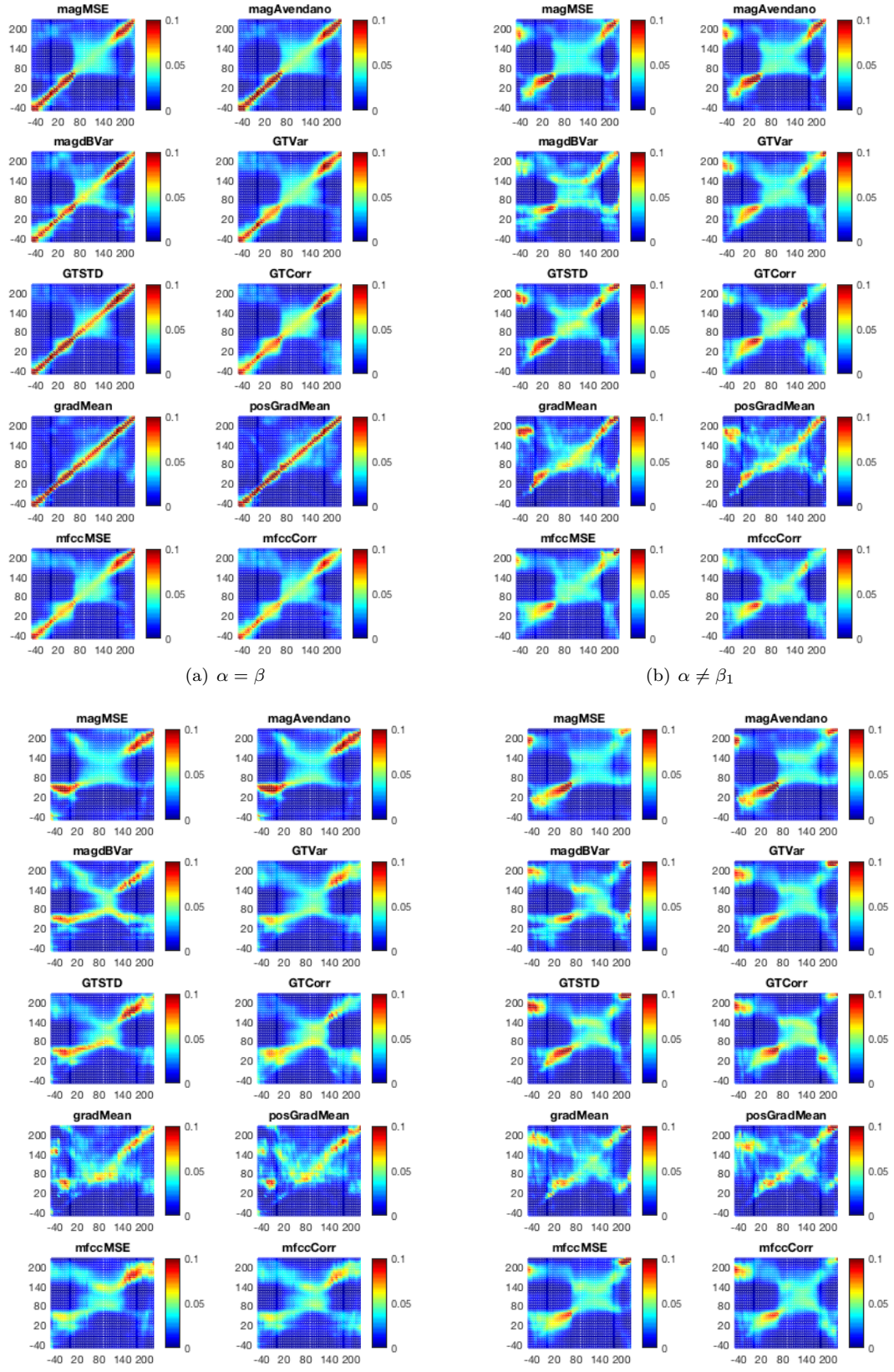


FIGURE 6.14 – Log-vraisemblance escomptée moyenne (sur les $T = 100$ tirages et les 12 sujets) associée au modèle pour chaque métrique spectrale.



(a) $\alpha = \beta$ (b) $\alpha \neq \beta_1$

(c) $\alpha \neq \beta_2$ (d) $\alpha \neq \beta_3$

FIGURE 6.15 – Exemples de cartes de prédiction sur le plan médian offertes par chacune des 10 métriques spectrales pour des HRTFs cibles présentées en chaque direction du plan médian $(\Theta_t, \Phi_t) = (0, \Phi_t)$ où $\Phi_t \in [-50, +250]$: (a) en condition individuelle ; (b, c, d) en condition non-individuelle pour un même auditeur α et des HRTFs cibles appartenant à 3 sujets différents $(\beta_1, \beta_2, \beta_3)$.

6.4.6 Pondération binaurale

Après avoir effectué la comparaison spectrale entre la paire de HRTFs cible avec les HRTFs de l'auditeur séparément pour chaque oreille, une combinaison pondérée des distances gauche et droite (équation 6.8) est réalisée. D'après la littérature, l'oreille ipsilatérale joue un rôle plus important dans le processus de localisation en élévation que l'oreille controlatérale [HVO03], et il faut donc pondérer l'information relative aux deux oreilles. Nous présentons et comparons ici les méthodes de calcul des poids utilisées dans les modèles de Middlebrooks [Mid92] et Baumgartner et al. [BML13, BML14].

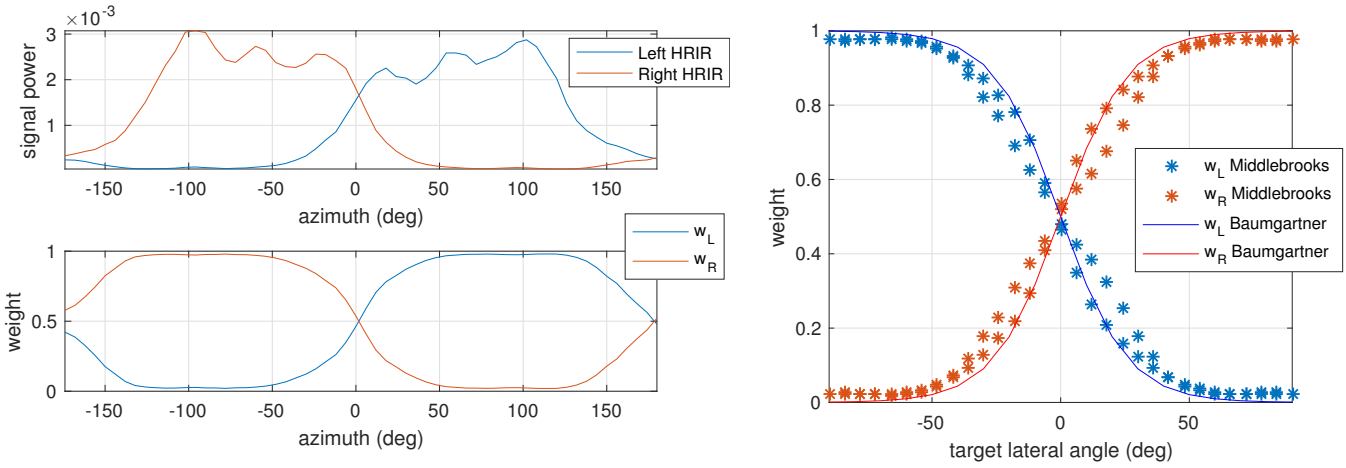
Dans le modèle de prédiction de la localisation en 3 dimensions, Middlebrooks pondère les distances spectrales gauche et droite par l'énergie des signaux cibles gauche et droit :

$$w_\lambda = \frac{P_\lambda}{\sum_{\lambda=L,R} P_\lambda} \tag{6.17}$$

avec

$$P_\lambda = \frac{\sum_{i=1}^{N} |h_\lambda(i,\theta_t,\phi_t)|^2}{N}, \tag{6.18}$$

Cette méthode de pondération binaurale dépend donc de la mesure des HRIRs cibles gauche et droite $h_L(i,\theta_t,\phi_t)$ et $h_R(i,\theta_t,\phi_t)$. Le calcul des poids selon cette méthode est illustrée dans la figure 6.16(a).



(a) (b)

FIGURE 6.16 – (a) Pondération binaurale en fonction de l'azimut selon la méthode de Middlebrooks (1992). Les poids (en bas) sont obtenus depuis le calcul de l'énergie relative des HRIRs gauche et droite d'un sujet dans le plan horizontal (en haut). (b) Superposition des courbes de pondération binaurale selon les méthodes de Middlebrooks (étoiles, [Mid92]) et Baumgartner et al. (lignes, [BML14]). Notons qu'il existe 2 points mesurés par angle latéral dans le cas de la méthode de Middlebrooks car les mesures des poids ont été effectuées pour un intervalle d'azimut compris entre $-180°$ et $180°$.

La méthode de pondération binaurale utilisée par Baumgartner et al. est basée sur les résultats expérimentaux de Morimoto [Mor01], selon lesquels la contribution de l'oreille controlatérale deviendrait négligeable pour des sources situées au-delà d'un angle latéral de $60°$, ainsi que de ceux de Macpherson et Sabin [MS07] qui ont quantifié la contribution relative de chaque oreille aux positions latérales $\pm 45°$. La pondération binaurale est modélisée par une fonction sigmoïde qui dépend de l'angle latéral de la direction cible :

$$w_L = (1 + e^{-\Theta_k/\Omega})^{-1}, \qquad w_R = 1 - w_L \tag{6.19}$$

avec w_L et w_R les poids gauche et droit, Θ_k l'angle latéral de la cible et $\Omega = 13$ le coefficient de pondération binaural, ajusté aux résultats expérimentaux de Macpherson et Sabin [MS07].

Sur la figure 6.16, on voit que les deux méthodes de pondération sont très semblables. En modélisant le comportement, Baumgartner [BML14] assurent une pondération plus continue, qui ne risque pas d'être altérée par une mesure aberrante. Cependant, la méthode de Middlebrooks est plus réaliste car elle est basée sur ce qui est entendu par le sujet amené à localiser la source sonore virtuelle.

6.4.7 Modélisation du biais et de la dispersion de pointage

Le modèle de prédiction est basé sur une comparaison des indices acoustiques de localisation entre la HRTF cible et les HRTFs de l'auditeur. Il cherche à prédire la direction perçue des sources sonores virtuelles. Son évaluation consiste à évaluer la probabilité de réponse aux directions indiquées par le sujet dans l'expérience. Etant donnés le biais et la dispersion introduits dans les réponses par la méthode de pointage [BCNW16], il apparaît nécessaire de les intégrer au modèle si l'on souhaite prédire les directions pointées.

L'idée est ici de corriger les réponses des sujets, en amont de l'application du modèle, par le biais de pointage inverse spécifique à la méthode de pointage afin d'en supprimer l'effet sur les réponses. Cette méthode a déjà été appliquée pour la prédiction de localisation [BvWvO10]. Dans cette étude, le biais de pointage (méthode de pointage avec la tête) a été caractérisé de manière individuelle à partir des jugements de localisation sur haut-parleurs (90 réponses par haut-parleur). Concernant la prise en compte de la dispersion de pointage, celle-ci sera appliquée directement dans le modèle afin de modéliser l'incertitude sur la prédiction.

Dans notre étude, nous collectons des données en condition individuelle mais il s'agit de sources virtuelles et le nombre de réponses collectées par direction test n'est que de 5. Pour caractériser le biais et la dispersion de pointage proximal, nous avons donc utilisé les données du précédent test de localisation sur haut-parleurs [BCNW16]. Lors de ce test, nous avons collecté $13 \times 8 = 104$ réponses (13 participants, 8 répétitions) pour chacune des 24 directions test réparties dans l'espace. Depuis ces données, nous avons utilisé des outils de statistiques sphériques afin de caractériser le biais et la dispersion des réponses associés à chaque direction test. Notons que ces résultats ne sont donc pas individuels.

6.4.7.1 Biais de pointage proximal

Au chapitre 4, une étude comparative de 3 méthodes de pointage, dont la méthode proximale, a été réalisée à partir d'un test de localisation sur sources réelles. Elle a permis de mettre en évidence que la méthode de pointage proximale présente un biais horizontal négatif à l'arrière significativement supérieur à ceux des méthodes de pointage au pistolet ou avec la tête. Cette observation indique que la tendance est spécifique à la méthode de pointage proximale. De plus, étant donné que le biais horizontal est significativement supérieur à l'arrière qu'à l'avant (en valeur absolue), la prise en compte de la dépendance spatiale de ce biais est nécessaire. Dans la dimension verticale, les erreurs et biais associés à la méthode de pointage proximale ne diffèrent pas significativement des deux autres méthodes. Au contraire, la méthode proximale a montré offrir une précision en élévation meilleure que la méthode de pointage au pistolet pour les sources élevées. Au vu de ces résultats, seule la correction du biais horizontal, et non vertical, semble nécessaire.

Pour caractériser le biais de pointage proximal associé à chacune des 24 directions du test de localisation sur haut-parleurs, nous utilisons simultanément les $13 \times 8 = 104$ réponses disponibles. Le centroïde des jugements associé à chacune des directions est calculé comme la direction résultante de la somme des vecteurs associés aux jugements uniques. Les confusions avant-arrière sont ignorées du caclul. La figure 6.17 permet de visualiser la position des centroïdes associé à chaque direction.

Le biais est ensuite calculé comme la différence d'angle signée entre le centroïde et la direction test, et une moyenne des biais gauches et droits est effectuée pour supprimer l'influence de la main dominante (pour rappel, les sources latéralisées à gauche et à droite sont symétriques par rapport au plan médian). Suivant la définition du biais de pointage relatif au chapitre 4, le biais en azimut est négatif dans le cas où le sujet n'est pas allé pointer suffisamment vers l'arrière et positif si il est allé trop loin. Cette méthode de calcul implique un biais en azimut systématiquement positif (ou nul) pour les cibles situées sur le plan médian à l'avant et systématiquement négatif (ou nul) à l'arrière. Cependant, contrairement au biais associé aux sources arrières du plan médian, le biais associé aux sources frontales du plan médian n'est pas significatif (de l'ordre de 2°). Celui-ci étant peu représentatif d'un biais de pointage systématique puisque très faible et pour éviter de modéliser une tendance qui serait liée à la méthode de calcul, ce biais a été fixé à zéro. La figure 6.18(a) permet de visualiser le biais de pointage en azimut ainsi estimé. Les croix indiquent les biais mesurés et les lignes en pointillés présentent le résultats de l'interpolation du biais par la méthode des 4 plus proches voisins aux azimuts intermédiaires (non mesurées) et aux élévations $-5°$, $26°$ et $57°$. On note que le principal défaut de la méthode de pointage proximal réside dans une tendance à ne pas pointer suffisamment vers l'arrière pour les cibles de l'hémi-champ arrière ($|\theta|>90°$). On observe également que, grâce à la continuité du biais mesuré en fonction de l'azimut, l'interpolation aux azimuts intermédiaires fonctionne bien. De plus, étant donné que la tendance associée à chaque élévation est très semblable, cela permet d'être confiant quant à l'interpolation aux élévations intermédiaires. La figure 6.18(b) présente le résultat de l'interpolation sur toute la sphère. Le biais en azimut est globalement négatif (tendance à ne pas pointer assez vers l'arrière) et augmente en valeur absolue plus on s'approche des directions arrières du plan médian.

Pour en revenir au modèle de prédiction de la localisation, l'idée serait de corriger les réponses des sujets du test en binaural par l'inverse de ce biais global en azimut mesuré dans la précédente étude sur haut-parleurs. Il faut alors avoir conscience que cette correction concerne non seulement une tendance liée à la méthode de pointage mais également un phénomène qui peut être perceptif, les deux étant indissociables. Pour procéder à la correction, l'interpolation de l'inverse du biais en azimut est réalisée aux directions

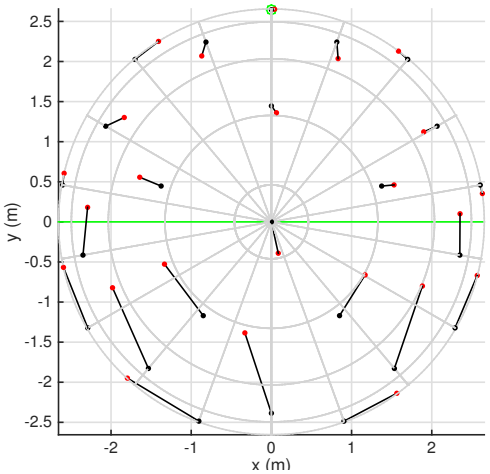
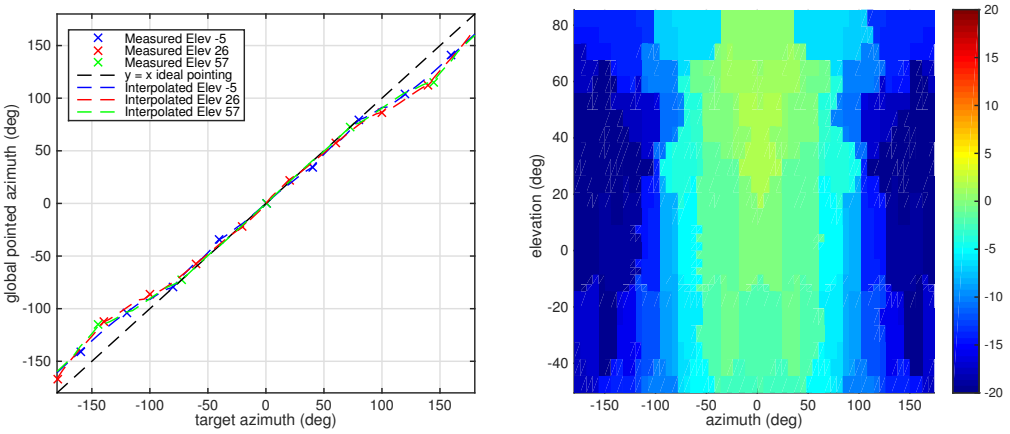


FIGURE 6.17 – Centroïdes des réponses (points rouges) associés aux 24 directions test (points noirs) du test sur haut-parleurs [BCNW16]. La sphère est représentée en vue de dessus (plan XY). Le trait plein (noir) lie le centroïde à la direction test associée. La ligne verte représente l'axe interaural, et le point vert, la direction frontale (0°, 0°). Les azimuts et élévations sont indiqués tous les 20° par les lignes et cercles gris, respectivement.



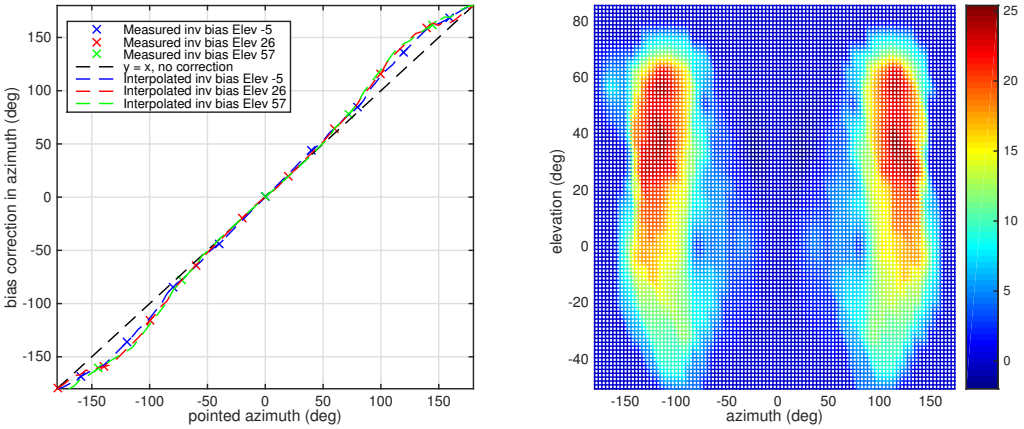
(a) (b)

FIGURE 6.18 – (a) Azimut pointé en fonction de l'azimut cible pour chaque élévation mesurée ($-5°$, $26°$, $57°$). Les croix représentent la mesure et les traits en pointillés, l'interpolation entre les points mesurés, pour chaque élévation. La déviation par rapport à la diagonale en noir pointillés (d'équation $y = x$) indique le biais global en azimut (symétrisé gauche/droite), spécifique à la méthode proximale et obtenu à partir des centroïdes des jugements présentés figure 6.17. (b) Cartographie du biais en azimut (degrés) sur toute la sphère après interpolation par les 4 plus poches voisins.

pointées par la méthode des 4 plus proches voisins. Basé sur l'hypothèse que le sujet n'utilise jamais la main opposée à l'hémi-espace de la direction perçue pour pointer, le biais inverse en azimut est forcé à zéro sur le plan médian, comme on peut le voir dans la figure 6.19(a). La figure 6.19(b) permet de visualiser quelle correction sera appliquée à chaque direction de l'espace.

6.4.7.2 Dispersion de pointage

La méthode de pointage introduit également une dispersion dans les réponses qu'il est nécessaire de prendre en compte. Par exemple, dans le cas où les HRTFs de deux directions spatiales proches, présentent une similarité respectivement forte et faible avec la HRTF cible, il se peut que le sujet ne soit pas en mesure de pointer correctement à la direction de forte similarité bien qu'il distingue ces deux directions. La dispersion de pointage est liée à une imprécision motrice à indiquer les directions de l'espace. Celle-ci peut être modélisée par un processus aléatoire de type gaussien. Afin de l'introduire dans les prédictions



(a) (b)

FIGURE 6.19 – (a) Correction du biais proximal en azimut à appliquer en fonction de l'azimut pointé pour les élévations $-5°$, $26°$, $56°$. (b) Corrections du biais en azimut interpolé sur la grille de mesure de 1680 points.

du modèle, la carte d'indices de similarité obtenue pour chaque cible unique est convoluée par un noyau de convolution gaussien de concentration $\kappa(\theta, \phi)$ variable dans l'espace. Sous l'hypothèse que la dispersion de pointage est isotrope, nous utilisons la distribution gaussienne de Von Mises-Fisher (VMF) définie sur une sphère de dimension $(p-1)$ par l'équation suivante [DS03] :

$$M_p(x, \mu, \kappa) = c_p(\kappa) e^{\kappa \mu^T x} \tag{6.20}$$

avec μ est le centroïde de la distribution ($\|\mu\| = 1$), κ sa concentration ($\kappa \geq 0$) et $x \in S^{p-1}$.
$c_p(\kappa)$ correspond au coefficient de normalisation défini par :

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \tag{6.21}$$

où $I_{p/2-1}(\kappa)$ est la fonction de Bessel modifiée de première espèce. Dans le cas d'une sphère de dimension 2 ($p = 3$), on peut utiliser l'identité suivante :

$$I_{\frac{1}{2}}(\kappa) = \sqrt{\frac{2}{\pi\kappa}} \sinh(\kappa). \tag{6.22}$$

L'expression de $c_p(\kappa)$ devient alors

$$c_3(\kappa) = \frac{\kappa}{4\pi \sinh(\kappa)} \tag{6.23}$$

L'opération $\mu^T x$ de l'équation 6.20 correspond au produit scalaire entre les vecteurs unitaires μ, le centroïde de la distribution, et x, les points d'échantillonnage de la sphère. Sa valeur est donc comprise entre -1 et $+1$. La multiplication par κ fait varier le résultat du produit scalaire entre $-\kappa$ et κ.

Afin d'estimer la concentration κ à partir d'une distribution de points, l'approximation suivante est couramment utilisée :

$$\hat{\kappa} = \frac{\bar{R}p - \bar{R}^3}{1 - \bar{R}^2} \tag{6.24}$$

avec $\bar{R}$ la résultante moyenne des N_R directions pointées $\overrightarrow{d_r}$, soit :

$$\bar{R} = \frac{\left\| \sum_{r=1}^{N_R} \overrightarrow{d_r} \right\|}{N_R} \tag{6.25}$$

Notons qu'en réalité, la dispersion de pointage se modélise plus exactement par une distribution de Kent, i.e. par une gaussienne anisotrope [Kas15]. Cependant, étant donné le faible nombre de directions pour lesquelles nous sommes en mesure d'estimer la dispersion de pointage (au nombre de 24), il est apparu difficile de réaliser une interpolation sur l'orientation des ellipses.

Nous estimons la concentration κ à partir des jugements de localisation du test sur haut-parleurs avec la méthode de pointage proximale. En procédant par une moyenne sur les 13 sujets ayant participé à cette expérience, nous supposons que la concentration estimée est représentative et caractéristique de la méthode de pointage proximale. L'estimation de κ est tout d'abord réalisée sur les jugements de localisation de chacun des sujets en réponse aux 24 directions test. Etant donnée la variabilité dans la position

des centroïdes des jugements en fonction de sujets, une estimation basée sur la combinaison de tous les jugements conduirait à une sur-estimation de la dispersion. C'est pourquoi elle est réalisée sujet par sujet puis une moyenne sur l'ensemble des sujets est réalisée pour chacune des 24 directions test. Les confusions avant-arrière sont supprimées en amont de l'estimation pour éviter que l'ambiguïté perceptive ne mène à une sur-estimation de la dispersion de pointage. Si le taux de confusions est supérieur ou égal à 50% des réponses (soit $\geq \frac{4}{8}$) alors les jugements associés au sujet et à la direction test concernée ne sont pas pris en compte dans la moyenne. Les valeurs de concentration obtenues pour chacune des directions test sont ensuite moyennées sur les hémi-espaces gauche et droit afin d'obtenir une distribution spatiale de la concentration symétrique par rapport au plan médian. Enfin, elles sont interpolées par la méthode de l'interpolation spline (assure une meilleure continuité par rapport à la méthode des k plus proches voisins) afin d'en déduire une distribution de κ sur toute la sphère. Par la suite, nous utiliserons le terme $\kappa_{[R]}$ pour désigner la distribution spatiale de concentration κ estimée à partir du test de localisation sur haut-parleurs. Le résultat est illustré en figure 6.20 et montre une concentration plus importante devant qu'à l'arrière (ce qui est en accord avec les conclusions du papier [BCNW16]). Notons qu'en moyenne, $\bar{\kappa}_{[R]} = 144$.

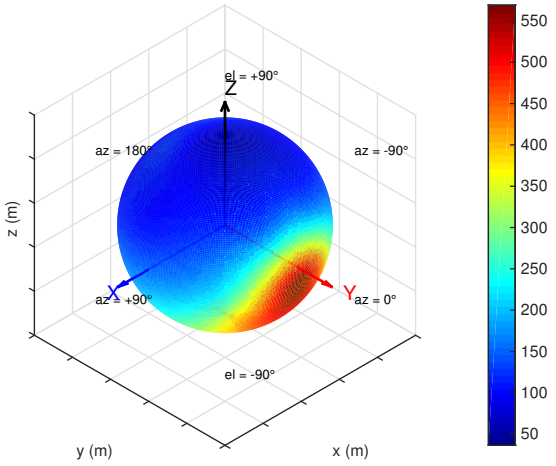
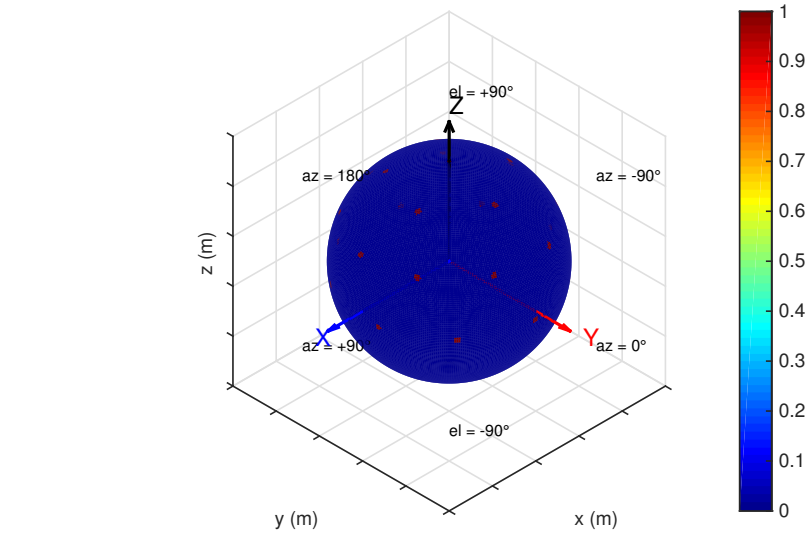


FIGURE 6.20 – Concentration $\kappa_{[R]}$ obtenue à partir d'une moyenne des jugements de chacun des 13 sujets du test sur haut-parleurs puis interpolée sur toute la sphère.

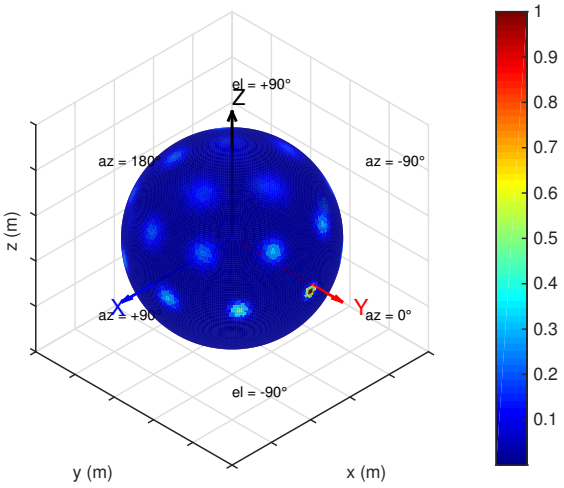
La convolution des données par le noyau gaussien doit être effectuée sur un maillage régulier et fin. La grille de mesure des HRTFs suit un maillage de type Gaussien qui présente une compression des points aux pôles. Les indices de similarité (SI) ainsi que la concentration $\kappa_{[R]}$ sont donc interpolés sur une grille d'échantillonnage sphérique plus régulière de type Hyperinterpolation à l'ordre 40, comme présenté en section 6.4.2.

Pour tester ce modèle de dispersion de pointage, prenons un signal test de valeur zéro sur l'ensemble des points de la sphère à l'exception des 24 directions du test sur haut-parleurs qui ont pour valeur 1. Sur la figure 6.21 peut être observé le signal test avant et après convolution avec les distributions de Von Mises-Fisher dont la concentration varie selon la concentration $\kappa_{[R]}$ estimée. On observe que le signal devient plus flou à l'arrière qu'à l'avant ce qui est cohérent avec la distribution visualisée en figure 6.20.

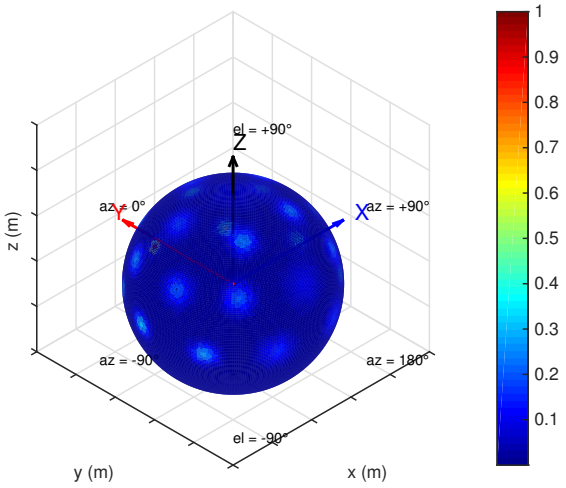
Pour finir, lorsqu'il s'agira d'appliquer ce modèle de dispersion sur les cartes d'indices de similarité, nous testerons différents modèles de dispersion : le cas $\kappa_{[R]}$ que l'on vient de voir ; le cas où la concentration $\kappa_{[R]}$ est constante sur l'espace soit $\bar{\kappa}_{[R]} = 144$; ainsi que plusieurs cas de figure où la concentration diminue jusqu'à une valeur proche de zéro, soit $\kappa = 60, 20, 10, 5$ et 0.001 de manière à évaluer l'effet d'une dispersion croissante. Le cas $\kappa = 0.001$ correspond à une probabilité uniforme sur la sphère. Notons qu'une concentration $\kappa = 0$ conduirait à une division par zéro ($\sinh(0) = 0$, équation 6.23).



(a)



(b)



(c)

FIGURE 6.21 – Signal test (a) convolué par le noyau gaussien de concentration $\kappa_{[R]}$: vue frontale (b) et arrière (c).

Conclusion

Ce chapitre a permis de situer le modèle de prédiction de la localisation auditive proposé par rapport à la littérature. Nous avons vu que l'approche de modélisation adoptée dans ce travail a déjà fait l'objet de plusieurs études. Elle repose sur l'hypothèse que les auditeurs tendent à localiser la source sonore vers la direction où les indices acoustiques qu'il reçoit aux tympans ressemblent le plus à ses HRTFs. La notion même de similarité acoustique entre ces informations reçues et celles apprises par l'expérience introduit le recours à une métrique. Les études précédentes, détaillées à la section 6.2, ont proposé des méthodes permettant d'accéder à une distribution d'indices de similarité, autrement dit de probabilité de réponse, à partir de cette mesure de similarité objective. Nous avons tiré avantage de ces méthodes pour mettre au point notre modèle.

Après avoir exposé la structure complète du modèle et introduit la notion d'espace de prédiction, nous avons présenté le critère de log-vraisemblance qui sera utilisé pour évaluer les performances de prédiction. Afin de prendre en main ce critère et de bien comprendre le sens des résultats qu'il nous confère, nous avons simulé différents cas de figure. Les observations effectuées ont permis de prendre conscience de l'importance de la sélectivité des distributions spatiales de probabilité. Nous avons mis en évidence la nécessité d'ajuster la sélectivité de ces distributions aux différents cas de figure étudiés afin d'accéder à des résultats pertinents. Puis, nous avons compris que la paramétrisation de la fonction sigmoïde utilisée pour transformer les distances (issues de l'étape de comparaison *target-templates*) en indices de similarité nous permettrait de réaliser cet ajustement de sélectivité spatiale.

Chacun des paramètres qui composent notre modèle de prédiction de la localisation auditive a été exposé. Tout d'abord, les métriques spectrales, permettant de calculer les indices de similarité spectraux, ont été présentées conjointement à la métrique interaurale associée aux indices de similarité interauraux. Ces deux types d'indices font l'objet d'une combinaison pondérée à partir de laquelle sont obtenues les distributions de probabilité de réponse. Certaines caractéristiques des métriques spectrales ont attiré notre attention. Dans un premier temps, les informations véhiculées par les métriques spectrales contiennent une part d'information interaurale qui s'est avérée variable selon les métriques. Cette observation nous permet d'entrevoir la nécessité d'ajuster les poids relatifs attribués respectivement aux indices de similarité spectraux et interauraux, ou du moins de s'attendre à des divergences dans la tendance des résultats obtenus pour chacune des métriques en fonction de ces poids relatifs. Dans un deuxième temps, nous avons pu observer que les distributions de probabilité obtenues à partir de chacune des métriques spectrales étaient caractérisées par un degré de sélectivité spatiale variable. En parallèle des remarques précédentes sur la log-vraisemblance, cela suggère de devoir ajuster la sélectivité spatiale, par l'intermédiaire des paramètres de la fonction sigmoïde, spécifiquement pour chacune des métriques afin de rendre la comparaison des métriques possible.

Parmi les autres paramètres du modèle, les méthodes de pondération binaurale permettant de combiner les indices de similarité spectrale gauche et droit ont été tirées de la littérature. Enfin, nous avons utilisé les données expérimentales du précédent test sur haut-parleurs (chapitre 4) afin de caractériser le biais et la dispersion de pointage associés à la méthode de report proximale. Cette étude a permis de proposer une méthode de correction du biais de pointage sur les réponses qui seront collectées dans le test de localisation de sources sonores virtuelles ainsi qu'un modèle de dispersion de pointage dont l'utilité au sein du modèle pourra être évaluée en termes des performances de prédiction.

Dans le chapitre suivant, nous confronterons le modèle à des données réelles et analyserons la probabilité aux directions pointées afin de dégager les paramètres optimaux. Une comparaison avec les résultats des études de référence, présentées au début de ce chapitre, permettra de valider la méthode mise en œuvre. Enfin, nous pourrons évaluer la pertinence de la sélection d'un jeu de HRTFs dans la base de données pour un nouvel individu au travers de la modélisation de ses réponses avec différentes HRTFs *templates*.

Pour mettre en place notre modèle de prédiction de la localisation auditive, nous nous sommes inspirés de modèles existants pour définir les éléments clés menant à l'obtention d'une probabilité de réponse à partir d'une évaluation de la similarité entre HRTFs. L'ensemble des étapes du modèle proposé ainsi que les critères d'évaluation de la prédiction ont été détaillés et permettent d'appréhender son évaluation vis-à-vis de données réelles, réalisée au chapitre suivant.

Chapitre 7

Prédiction de la localisation en synthèse binaurale non-individuelle

Ce chapitre évalue les performances de prédiction de la localisation auditive du modèle proposé au chapitre précédent, pour des sources virtuelles synthétisées avec des HRTFs non-individuelles. Il présente tout d'abord la mise en place d'un test de localisation auditive en synthèse binaurale non-individuelle pour la collection de données réelles, à partir desquelles est menée une optimisation de l'ensemble des paramètres du modèle. La méthode de sélection guidée d'un jeu de HRTFs par la modélisation des réponses de localisation d'un auditeur à différentes sources virtuelles, synthétisées avec des HRTFs non-individuelles, est examinée en fin de chapitre.

Le modèle de localisation auditive proposé suppose que la localisation d'une source virtuelle synthétisée avec une paire de HRTFs non-individuelle peut être prédite par un calcul de similarité objective entre les indices acoustiques délivrés par la HRTF cible et les indices acoustiques propres à l'auditeur. La prédiction est réalisée sur un espace à deux dimensions. Suivant notre connaissance des mécanismes de localisation auditive humaine, présentés notamment au chapitre 1, la prédiction de la localisation repose à la fois sur l'analyse des indices interauraux, responsables de la localisation latérale, et des indices spectraux, déterminants pour la localisation en élévation et la discrimination avant-arrière. A travers ce modèle, différentes méthodes de calcul de la similarité spectrale entre les HRTFs seront comparées. Comme nous avons pu l'identifier tout au long des chapitres précédents, la définition d'une métrique spectrale a une importance dans de nombreuses applications en synthèse binaurale, comme par exemple pour la sélection guidée d'un jeu de HRTFs dans une base de données, l'évaluation de méthodes de modélisation ou d'interpolation des fonctions de transfert. L'identification d'une métrique spectrale proche de l'extraction et la reconnaissance des indices spectraux pertinents pour le système auditif est l'un des objectifs de la présente étude.

Ce chapitre décrit la mise en oeuvre et la validation du modèle de prédiction de la localisation auditive sur la base de la comparaison entre le stimulus présenté à l'auditeur et ses HRTFs propres. Le déroulement de la mise en œuvre et de l'étape de validation sont décrits respectivement par les figures 7.11 et 7.2.

Premièrement, la mise en œuvre du modèle repose initialement sur la réalisation d'un test de localisation auditive de sorte à collecter les données qui serviront à l'optimisation des différents paramètres du modèle (fonction sigmoïde, poids relatifs des indices interauraux et spectraux, méthode de pondération binaurale, modèle de dispersion de pointage, métrique spectrale). Ces données recueillies sur un ensemble de participants comprend essentiellement les réponses de localisation perçues avec des HRTFs non-individuelles. Cependant, pour s'assurer du comportement du modèle, le test comprend quelques réponses de localisation perçues avec les HRTFs propres de chaque participant. Conformément aux résultats du chapitre 6, la méthode de report adoptée pour le test est la méthode de pointage proximale. Cette méthode a l'avantage de présenter une meilleure précision pour les sources élevées et de ne pas susciter les mouvements du corps. Le chapitre précédent a montré comment tirer avantage du test de localisation sur sources réelles pour caractériser le biais et la dispersion associés à la méthode de report. La méthode de correction du biais sur les réponses mise au point sera évaluée dans ce chapitre, et la modélisation de la dispersion de pointage sera injectée en tant que paramètre d'optimisation du modèle.

Deuxièmement, l'optimisation du modèle de prédiction vise à identifier les paramètres qui prédisent au mieux les observations expérimentales du test de localisation. Ce travail d'optimisation porte en particulier sur le choix des paramètres de la fonction sigmoïde, des poids relatifs des indices inteurauraux et spectraux, de la méthode de pondération binaurale, de l'utilisation du modèle de dispersion de pointage et de la métrique spectrale. Des analyses de variance permettront de caractériser la sensibilité des résultats de prédiction à la variation de chaque paramètre. L'optimisation est conduite principalement par l'observation de la log-vraisemblance qui mesure la probabilité du modèle prédite aux directions indiquées par les sujets dans le test. Cette procédure d'optimisation est menée séparément pour les conditions d'écoute individuelle

et non-individuelle. L'optimisation en condition d'écoute individuelle est réalisée uniquement dans un souci d'évaluation de la consistance du modèle. Nous nous intéressons cependant de manière privilégiée à l'optimisation en condition d'écoute non-individuelle, puisque cela correspond au cas pratique visée par l'étude.

Pour l'étape de validation, nous nous plaçons dans le cadre d'une procédure d'individualisation reposant sur une sélection dans une base de données de HRTFs guidée par l'observation des directions perçues par l'auditeur soumis à une série de HRTFs non-individuelles et présentées selon un ensemble de directions. La démarche de validation consiste à repartir des réponses collectées lors du test de localisation, et d'appliquer le modèle précédemment optimisé en lui soumettant tour à tour chacun des jeux d'une base de données de HRTFs. Selon les hypothèses du modèle, celui-ci devrait donc être capable de sélectionner au sein de cette base le jeu de HRTFs prédisant au mieux les réponses du sujet. Si le modèle est consistant, le jeu sélectionné devrait être celui appartenant à l'auditeur, ou du moins le jeu sélectionné devrait se trouver parmi les tous premiers élus.

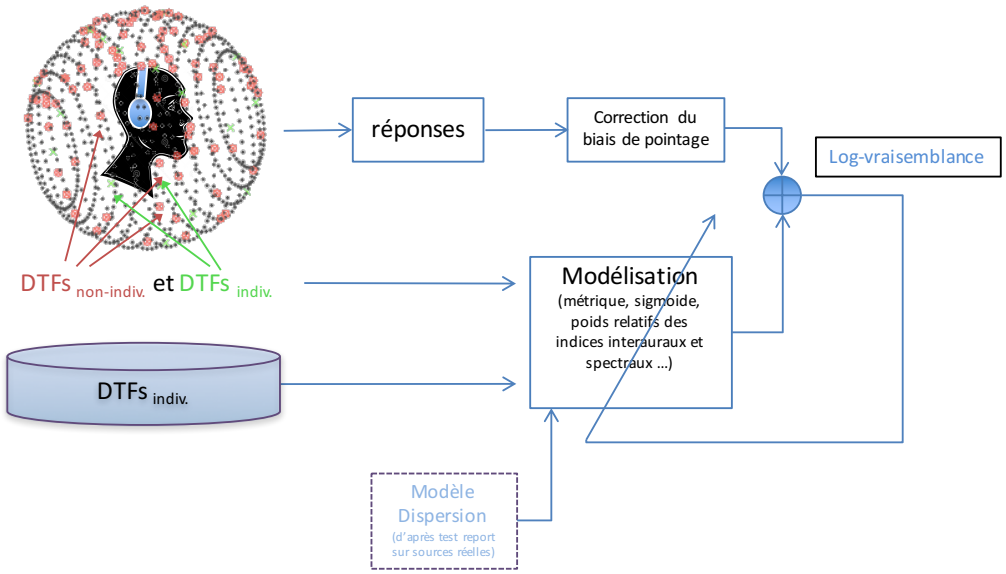


FIGURE 7.1 – Principe de l'optimisation des paramètres du modèle de prédiction de la localisation auditive.

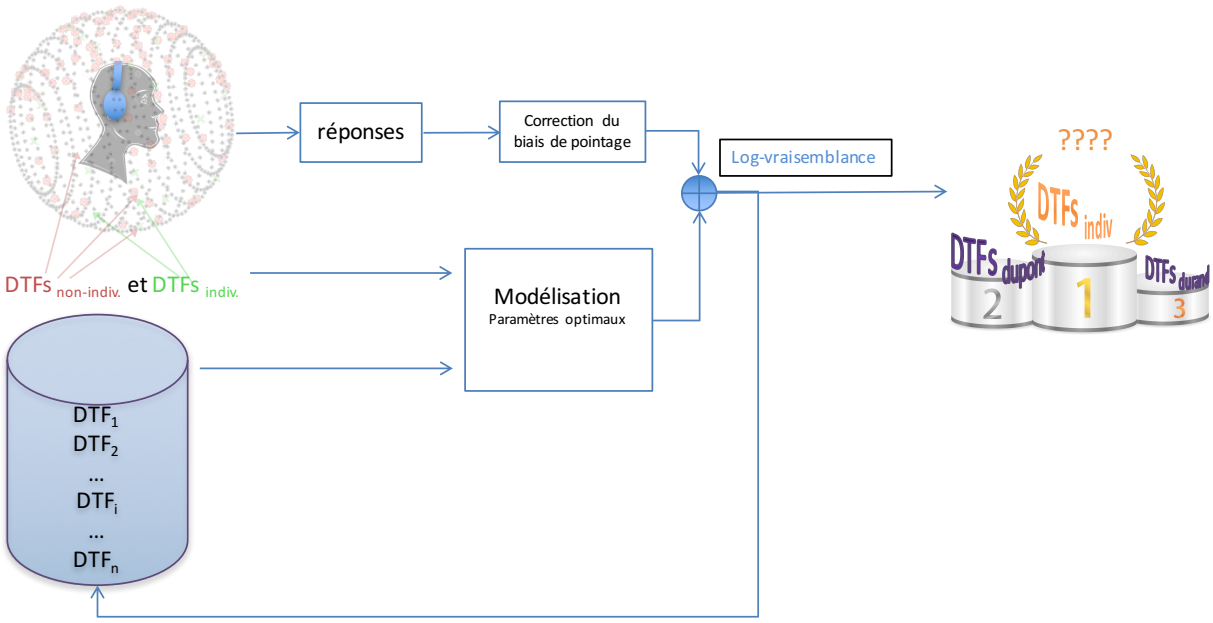


FIGURE 7.2 – Validation de la méthode de sélection guidée dans une base de données par l'observation des réponses d'un individu à un test de localisation de sources virtuelles synthétisées par des HRTFs quelconques.

7.1 Test de localisation de HRTFs individuelles et non-individuelles

La collection de données expérimentales est la première étape indispensable pour l'optimisation de la paramétrisation du modèle. Un test de localisation de sources virtuelles synthétisées avec des HTFs individuelles et non-individuelles a donc été réalisé. Notons qu'en réalité, ce sont les HRTFs égalisées champ diffus (i.e. les DTFs) qui ont été utilisées pour effectuer la synthèse des sources au casque. Etant donné que la réponse de celui-ci a été égalisée (c.f. section 7.1.3), le contenu spectral reçu à l'entrée des canaux auditifs de l'auditeur correspondent exactement aux DTFs. Par abus de language, nous utiliserons le terme HRTFs pour désigner les DTFs.

Nous présentons ici les différentes procédures expérimentales et éléments techniques qui le composent ainsi que le raisonnement ayant mené au choix des HRTFs non-individuelles et des directions testées.

7.1.1 Procédure et dispositif expérimental

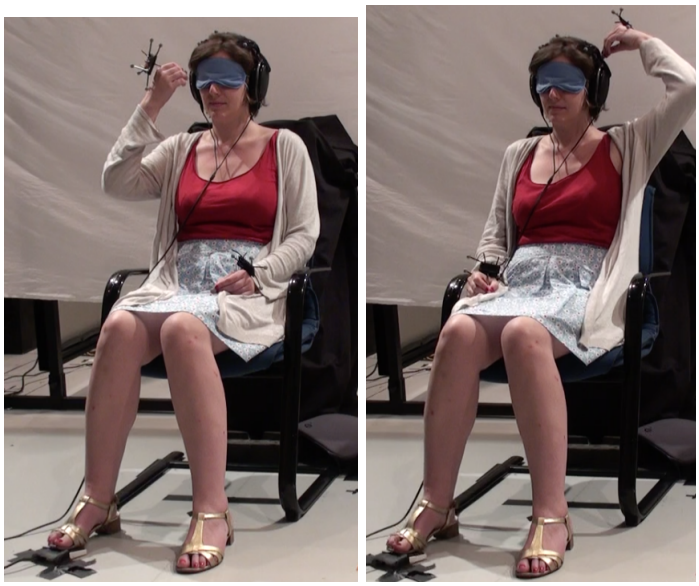
Sujets 12 sujets ont participé au test de localisation en binaural (2 femmes, 10 hommes). Les participants possèdent une audition normale et sont droitiers. Excepté l'un d'entre eux, tous les participants travaillent dans le domaine de l'audio. Tous ont déjà participé à des expériences psycho-acoustiques, y compris pour certains à des test de localisation auditive et ont connaissance du principe de la synthèse binaurale. Il s'agit donc dans l'immense majorité d'auditeurs experts. Cependant, aucun n'avait de pré-connaissance des motivations du test ni de sa composition. Conformément aux règles d'usage, les participants ont signé un formulaire de consentement après avoir pris connaissance du déroulement du test et ont reçu une gratification en fin d'expérience.

Expérience Le test de localisation s'est déroulé dans un studio insonorisé de l'IRCAM. Plusieurs sources virtuelles étaient présentées au casque (casque audio de type Sennheiser HD 650 circumaural) et synthétisées avec des HRTFs individuelles, i.e. mesurées sur l'auditeur, ou non-individuelles, i.e. mesurées sur d'autres individus. Les HRTFs non-individuelles avaient été sélectionnées en amont au sein de la base de données BiLi (la méthode de sélection est présentée en section 7.1.5) et pouvaient avoir été mesurées à l'IRCAM ou à Orange Labs. L'expérience était constituée de 108 stimuli sonores répétés 5 fois et présentés de manière aléatoire. Chaque stimulus correspond à un train de bruits blancs (c.f. section 7.1.2) filtré par une paire de HRTFs gauche-droite ainsi que le filtre d'égalisation du casque audio mesuré de manière individuelle (c.f. section 7.1.3). Le test était divisé en 3 sessions de 17 minutes environ (180 stimuli par session) avec une pause entre chaque, soit une durée expérimentale totale d'environ 50 minutes (sans compter les pauses).

Méthode de pointage Les participants étaient assis sur une chaise placée au milieu de la pièce et un appui-tête permettait d'éviter les mouvements de le tête. Ils portaient un bandeau sur les yeux pendant la phase expérimentale afin d'éviter que les ancrages visuels ou l'incohérence entre les indices visuels et auditifs n'affectent la localisation des sources sonores virtuelles. Les participants reportaient la direction perçue de la source virtuelle avec la méthode de pointage proximale, présentée au chapitre 4. Comme indiqué à la fin du chapitre 4, le choix de cette méthode de pointage est ici motivé par plusieurs points, notamment par le fait qu'elle n'implique pas les mouvements du corps et qu'elle offre ainsi un temps de réponse relativement faible et la possibilité d'une mise œuvre en condition *closed-loop*. De plus, elle se présente comme une méthode de report intuitive dans le cadre de sources virtuelles perçues relativement proches de la tête (en particulier en l'absence d'effet de salle, c.f. section 2.1.3). Pour rappel, cette méthode de report comprend l'utilisation de deux objets de pointage, un tenu dans chaque main, avec la bille de référence placée au bout des doigts. Avant chaque émission du stimulus, les sujets doivent poser les mains sur les cuisses. Durant l'émission du stimulus, ils sont invités à indiquer la direction perçue en plaçant un des deux objets dans la région proximale de la tête, en utilisant la main de leur choix et en gardant l'autre au repos. Enfin, ils doivent valider leur réponse en appuyant sur la pédale avec le pied. Le stimulus restait émis durant la phase de pointage, conformément aux conclusions de article [BBB+13] et aux perspectives de [BCNW16]. Deux photos illustrent la tâche de pointage en figure 7.3. Au début de l'expérience, les sujets étaient soumis à une session d'entrainement de 18 stimuli synthétisés avec leurs HRTFs individuelles aux directions test, afin de se familiariser avec la tâche de pointage.

Repérage spatial et calibration Un système de tracking de type OptiTrack constitué de 6 caméras infrarouge permettait de récupérer la position et l'orientation des objets de pointage ainsi que celles de la tête. Les deux objets de pointage ainsi que le casque audio étaient équipés de billes réfléchissantes dans l'infrarouge, et étaient définis au début de l'expérience dans le logiciel de repérage spatial. Les directions pointées par le sujet étaient calculées comme l'intersection de la droite sous-tendue par la bille de référence de l'objet de pointage, tenue au bout des doigts, et le centre de la tête. Une procédure de calibration était menée au début de l'expérience afin de définir le "centre de la tête" comme le point à équidistance des entrées des conduits auditifs, situés sur l'axe interaural. Pour ce faire, le participant devait rester immobile sur la chaise, les yeux bandés, et un laser croix était alors allumé face à lui. Le plan horizontal était matérialisé par le faisceau laser horizontal et le plan médian, par le faisceau vertical. La position du bout du nez était définie par l'intersection des deux faisceaux et l'entrée des canaux auditifs était alignée avec le plan horizontal. La position et l'orientation de la tête du participant étaient ajustées par l'expérimentateur

suivant ces repères visuels. A l'aide d'un objet repéré par le système de tracking, la position spatiale de l'entrée de chacun des canaux auditifs était définie par l'expérimentateur. De manière automatique, la position du centre de la tête était alors déduite de ces positions comme le point équidistant, et une procédure de translation de l'objet de repérage de la tête (défini par le casque audio) permettait ensuite de récupérer directement la position spatiale du "centre de la tête". Suivant cette définition, la direction pointée par un objet placé en face du bout du nez correspondait à la direction frontale $(0°, 0°)$. Cette procédure rapide (quelques minutes) était effectuée à chaque début ou redémarrage de l'expérience (i.e. après chaque pause). Le participant était ensuite invité à garder la tête fixe pendant la phase expérimentale. Contrairement au test sur haut-parleurs, aucune interface de repositionnement n'était utilisée. En effet, les sources virtuelles sont synthétisées de manière statique, i.e. qu'elles restent à des positions fixes par rapport la tête. Une interface expérimentateur permettait cependant de vérifier que la tête du sujet ne se décale de trop de la position de référence définie au moment de la calibration. L'émission du stimulus n'était activée que si plusieurs conditions étaient respectées : les deux mains devaient être posées sur les cuisses et la position et l'orientation de la tête devaient respecter un certain intervalle de tolérance par rapport à la position de référence.



(a) (b)

FIGURE 7.3 – Photo d'une participante réalisant la tâche de pointage proximal pendant le test de localisation. La participante porte un bandeau sur les yeux et un casque audio. Elle tient un objet de pointage dans chaque main, avec la bille de référence placée au bout des doigts. Figure (a), la participante indique une source virtuelle qu'elle perçoit légèrement à droite et au-dessus du plan horizontal, avec sa main droite. Figure (b), la participante indique une source virtuelle perçue vers l'arrière, du côté gauche et élevée, avec sa main gauche. Une fois sa main positionnée, la participante valide sa réponse en appuyant sur la pédale avec son pied. La main qui n'est pas utilisée pour le pointage reste au repos.

7.1.2 Stimulus

Le stimulus consistait en un train continu de bruits gaussiens de 100 ms, espacés de 10 ms de silence, et de transitoires de début et fin avec une rampe en cos^2 de 10 ms. La structure temporelle de ce stimulus peut être observé en figure 7.4. Le stimulus restait émis pendant toute la phase de pointage, jusqu'à validation de la réponse du sujet. Bien que le sujet avait pour consigne de garder la tête fixe durant toute la durée du test, de légers mouvements de la tête pouvaient avoir lieu. Cependant, la synthèse du stimulus étant statique et non dynamique (i.e. où la spatialisation s'adapterait en temps réel aux mouvements de la tête), cela permettait d'assurer l'absence des indices dynamiques de localisation.

Les stimuli étaient filtrés par le filtre de compensation individuel du casque audio et la paire de HRTFs cible gauche-droite. Le niveau sonore était d'environ 61 dBA à la sortie du casque audio. Ce niveau moyen a été estimé à partir des mesures (sur plaque) gauches et droites du niveau sonore de 4 jeux de HRTFs individuels égalisés par la mesure individuelle du casque sur les 18 directions test.

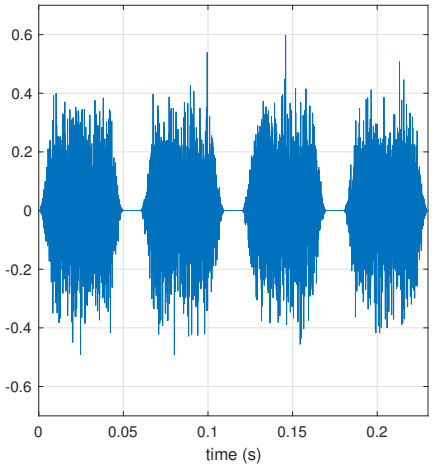


FIGURE 7.4 – Train de 4 bruits blancs gaussiens représentant la structure temporelle du stimulus expérimental.

7.1.3 Egalisation du casque audio

Comme présenté en section 2.2.3, la réponse en fréquence du casque varie en fonction de l'auditeur. Cela est dû à la présence du pavillon d'oreille qui colore de manière individuelle le spectre reçu aux oreilles du participant. Ces colorations spectrales apparaissent au delà de 4 kHz et sont d'amplitude comparables aux indices spectraux. Afin d'éviter que ces caractéristiques spectrales ne viennent entacher la spatialisation des sources virtuelles et de manière à contrôler précisément le spectre émis, une égalisation de la réponse en fréquence du casque audio a été réalisée de manière individuelle. Le casque utilisé pour la reproduction est un casque circumaural égalisé champ diffus par le fabriquant industriel. Ce type de casque permet de minimiser les problèmes de répétabilité de la mesure à chaque repositionnement du casque mais possède une composante individuelle forte étant donné qu'il capte l'effet du pavillon [PC96].

Le filtre d'égalisation a été obtenu à partir d'une moyenne de 5 mesures de réponse en fréquence réalisées après replacement du casque sur les oreilles de l'auditeur, de manière à tenir compte de la faible répétabilité de cette mesure [KC00]. Des capsules microphoniques Knowles (de type FG26107 C34) ont été placées à l'entrée des conduits auditifs de l'auditeur suivant la méthode du conduit auditif bloqué, en ré-utilisant les moules individuels fabriqués à l'occasion de la constitution de la base de données. Ces microphones possèdent une réponse en fréquence jugée suffisamment plate pour ne pas être compensée (voir figure 2.8). Un exemple de mesures de réponse en fréquence de casque peut être visualisé figure 7.5. Une fois ces 5 mesures réalisées, une étape de vérification permettait de retirer les mesures dont la réponse en fréquence est non représentative des autres. Une moyenne sur les mesures valides était ensuite effectuée ainsi qu'un lissage et une symétrisation gauche-droite, afin d'éviter d'incorporer une différence interaurale de niveau (notons que les réponses fréquentielles du casque à gauche et droite sont en général très proches). Le filtre d'égalisation individuel résultant correspond à l'inverse de la réponse en fréquence ainsi générée et était appliqué à tous les stimuli expérimentaux présentés au sujet. Cette procédure était réalisée avant le début de l'expérience.

7.1.4 Directions test

La méthode de sélection des directions test a été choisie de sorte à couvrir le mieux possible l'ensemble des directions de la grille de mesure de HRTFs, en évitant les directions trop latéralisées où les plans sagittaux sont compressés, et avec une sélection de points plus dense autour du plan médian. Les 18 directions test ainsi sélectionnées résultent d'un compromis entre le nombre de directions, le nombre de HRTFs non-individuelles et la durée du test. Le système de coordonnées utilisé correspond aux coordonnées latérales-polaires (Θ,Φ).

Les directions test ont été sélectionnées au sein de chacun des 13 plans sagittaux définis dans l'intervalle latéral $[-60°, +60°]$, séparés d'un pas latéral de 10° et d'une marge latérale de $\pm 5°$. Chaque direction test correspond à un point de mesure de la grille de mesures de HRTFs de la base BiLi-IRCAM. Pour les plans sagittaux proches du plan médian (les 5 plans sagittaux de l'intervalle latéral $[-20°, 20°]$), 2 directions ont été sélectionnées par plan sagittal. Pour ce faire, l'intervalle polaire disponible a été divisé en $2 \times 5 = 10$ parties polaires. Puis, chacune d'entre elle a été assignée à 2 de ces 5 plans sagittaux, de façon aléatoire. Le point situé au milieu de la partie polaire sur le plan sagittal sélectionné est ainsi déterminé. Pour les plans sagittaux plus latéralisés (les 8 plans sagittaux centrés sur $[\pm 60°, \pm 50°, \pm 40°, \pm 30°]$), 1 direction a été sélectionnée par plan sagittal. L'intervalle polaire a alors été divisé en 8 et chacune des parties polaires ainsi définies a été assignée aléatoirement à chacun des 8 plans sagittaux latéralisés. Dans la figure 7.6 est

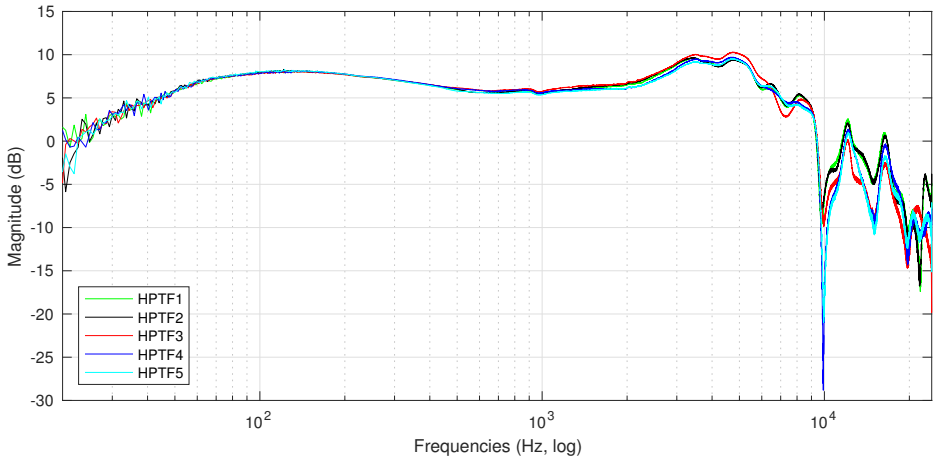
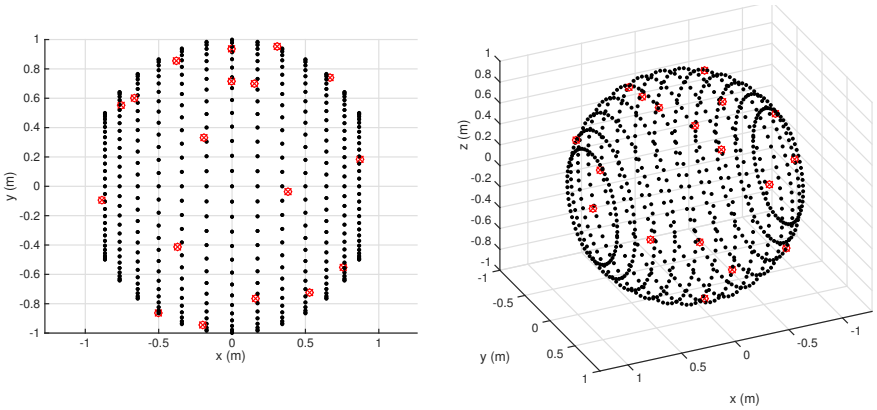


FIGURE 7.5 – Variabilité intra-sujet dans les mesures de réponses en fréquence du casque audio.

représentée la distribution spatiale des ces 18 directions test.



(a) Vue XY (b) Vue de côté

FIGURE 7.6 – Les 18 directions test (en rouge) disposées sur une grille matérialisée par les 13 plans sagittaux (points noirs) sur lesquels les directions ont été sélectionnées.

7.1.5 HRTFs cibles

A chacune des directions test sont présentées les HRTFs individuelles de l'auditeur ainsi qu'une sélection de 5 HRTFs non-individuelles. Toutes ces HRTFs appartiennent à la base de données BiLi et sont extraites à la direction mesurée la plus proche de la direction test. Celle-ci comprend (à la date de la phase expérimentale) un total de 82 HRTFs mesurées avec les systèmes de mesure de l'IRCAM ou d'Orange Labs. Les 1500 directions communes mesurées avec chacun de ces systèmes de mesure correspondent exactement en azimut et s'écartent de moins de 0.4° en élévation. Le post-traitement des HRTFs appliqué est également similaire et consiste en une égalisation champ diffus des HRTFs mesurées (pour plus de détails sur les mesures, voir l'article [CBNW14] et la section 2.2.2).

La démarche de sélection des HRTFs non-individuelles vise à fournir pour chaque participant un ensemble de HRTFs balisant son secteur sagittal, i.e. pointant *a priori* dans des directions variées plus ou moins éloignées de la direction test (direction de mesure de la HRTF non-individuelle). Les HRTFs non-individuelles sélectionnées sont donc différentes pour chacun des sujets.

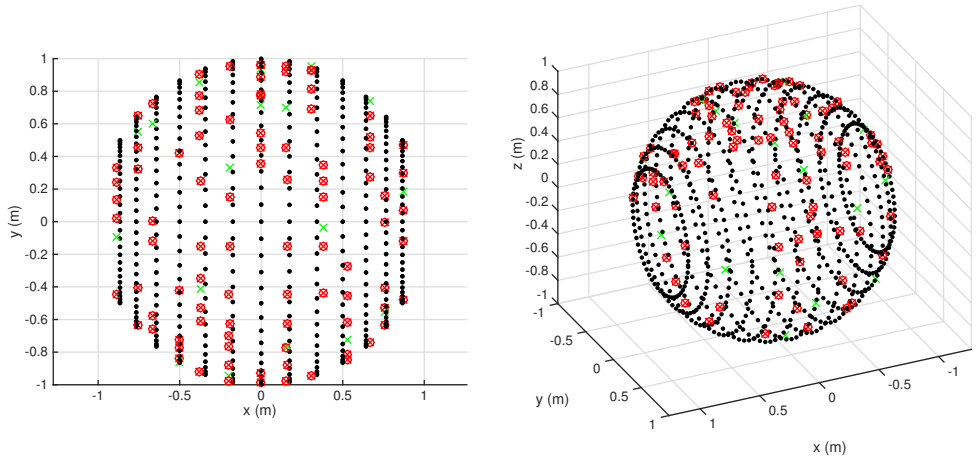
Le choix des 5 HRTFs non-individuelles pour chaque direction test repose sur une sélection par l'ITD et les résultats *a priori* du modèle. La première étape consiste à pré-sélectionner les HRTFs non-individuelles qui, à la direction test, ont un ITD proche de celui de l'auditeur (<3 échantillons soit $< 62\,\mu$s). Pour ce faire, une comparaison de l'ITD des 81 HRTFs non-individuelles disponibles à la direction test et de l'ITD des HRTFs de l'auditeur à cette direction test est réalisée au sens de la méthode d'estimation de l'ITD "MaxIACCr" (soit le maximum de l'inter-corrélation interaurale des HRIRs gauche et droite). Cette pré-sélection par l'ITD se justifie par le fait que nous souhaitons étudier les directions perçues dans la

dimension polaire (sur le plan sagittal de la direction test) de HRTFs au spectre d'amplitude variable. La méthode des HRTFs hybrides, composées de l'ITD de l'auditeur et du spectre d'amplitude associé à une autre HRTF, a été délaissée afin d'éviter la contradiction des indices d'ITD et d'ILD.

La deuxième étape consiste à appliquer le modèle de prédiction avec la métrique spectrale *mag-MSE* ayant présenté certains avantages, notamment en termes de l'évaluation des différences inter-individuelles, dans la section 5.1.8.2. Les distances spectrales entre chaque paire de HRTFs gauche-droite non-individuelle pré-sélectionnée à la direction test et l'ensemble des HRTFs de l'auditeur sont calculées au sens de la différence MSE des spectres d'amplitude linéaire des HRTFs. Une pondération binaurale des distances à gauche et à droite est réalisée en utilisant la méthode de Middlebrooks (c.f. section 6.4.6). La carte spatiale des distances spectrales ainsi obtenues est alors interpolée sur une grille spatiale régulière définissant le plan sagittal centré latéralement sur la direction test et de largeur latérale 10° (±5°). Les valeurs de distances sont alors normalisées sur ce plan sagittal suivant la normalisation Z-score, et transformées en indices de similarité par une simple multiplication par -1 (transformation linéaire, pas de fonction gaussienne ou sigmoïde appliquée ici). La position polaire du maximum de similarité est repéré pour chacune des HRTF non-individuelles potentielles, i.e. issues de la pré-sélection par l'ITD. Une méthode de *clustering* spatial de tous les maxima permet d'identifier les HRTFs non-individuelles qui prévoient des maxima de similarité répartis le plus uniformément possible sur le plan sagittal théorique.

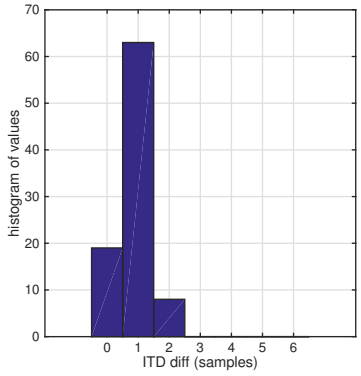
La méthode de sélection des HRTFs prévoit ainsi l'observation d'une variété de motifs de localisation dans la dimension polaire pour une même direction test. Notons qu'étant donné le caractère unique du motif induit par chaque HRTF cible, on parlera de cible unique pour mentionner une HRTF à une direction test. Notons que $Q = 108$ cibles uniques sont présentées à chaque sujet (6 HRTFs pour chacune des 18 directions test). Chaque cible unique est répétée 5 fois. La figure 7.7 illustre les directions perçues les plus probables tel que prédit par le modèle pour l'ensemble des HRTFs non-individuelles sélectionnées pour un sujet donné. La méthode de sélection maximise l'homogénéité de la distribution des directions perçues les plus probables sur le plan sagittal. Notons que dans la condition d'écoute individuelle (stimulus synthétisé avec les HRTFs de l'auditeur), le maximum apparaît par définition à la direction test.

Enfin, la figure 7.8(a) présente un exemple d'histogramme des différences d'ITD entre les HRTFs de l'auditeur et les HRTFs non-individuelles sélectionnées aux directions test. La figure 7.8(b) présente les cartes de prédiction sur le plan sagittal de la direction test associées aux 5 HRTFs non-individuelles sélectionnées pour un auditeur donné à la direction test $(\Theta,\Phi)=(-11°, 70.3°)$.

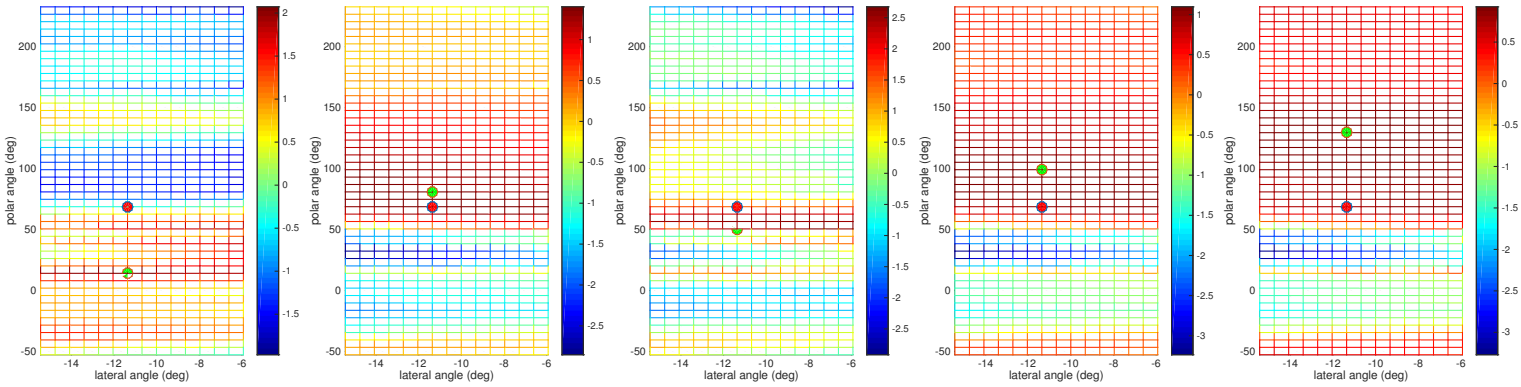


(a) Vue XY (b) Vue de côté

FIGURE 7.7 – Directions perçues les plus probables (en rouge) à l'écoute des HRTFs non-individuelles sélectionnées aux directions test (en vert), selon le modèle de prédiction basé sur la métrique *mag-MSE*. Les points noirs matérialisent les plans sagittaux espacés de 10° sur l'intervalle latéral $[-60°, +60°]$.



(a)



(b)

FIGURE 7.8 – (a) Histogramme des différences d'ITD entre les HRTFs individuelles et les HRTFs non-individuelles sélectionnées pour un auditeur donné, sur l'ensemble des directions test. (b) Cartes de prédiction sur le plan sagittal de la direction test associées aux 5 HRTFs non-individuelles sélectionnées pour ce même auditeur à la direction test (Θ,Φ)=$(-11°, 70.3°)$, indiquée par le point rouge. Le point vert indique le maximum de similarité prédit par le modèle pour chaque HRTF non-individuelle.

7.2 Observations sur les données de localisation

7.2.1 Premières observations

Nous nous intéressons ici à plusieurs données relatives à la localisation des sources Virtuelles présentées au casque (référencé par [V]) en comparaison avec la localisation des sources Réelles (référencé par [R]). La figure 7.9 présente ces données et distingue les résultats relatifs aux stimuli filtrés avec les HRTFs individuelles de l'auditeur (condition d'écoute individuelle) de ceux relatifs aux stimuli filtrés avec des HRTFs non-individuelles (condition non-individuelle), au sein du test en binaural. Les barres en bleu foncé et rouge correspondent aux valeurs moyennes pour chaque sujet du test en binaural [V] pour les conditions d'écoute individuelle et non-individuelle, respectivement. Les barres d'erreur associées représentent l'écart-type au sein de chacun des sujets. Les barres en bleu clair et rose indiquent la moyenne sur ces sujets pour les conditions individuelle et non-individuelle, respectivement. Les barres d'erreur associées présentent l'écart-type sur les sujets. La barre verte, présente sur certaines figures, correspond à la valeur moyenne pour les sujets du test de localisation sur haut-parleurs [R] et la barre d'erreur associée correspond à l'écart-type sur les sujets.

- Confusions avant-arrière en condition individuelle

  On s'intéresse ici aux taux de confusions avant-arrière observé dans la condition d'écoute individuelle du test [V]. Le taux de confusions en condition d'écoute non-individuelle n'est pas étudié ici car cela supposerait qu'il y ait une "bonne réponse" (typiquement la direction de synthèse). Or, l'étude de la condition d'écoute non-individuelle s'intéresse justement aux directions perçues, indépendamment de la direction associée à la HRTF non-individuelle cible.

  La figure 7.9(a) présente le taux de confusions avant-arrière en condition d'écoute individuelle pour les tests [V] et [R] (barres bleues et verte respectivement). Le taux de confusion est calculé de la même manière que dans le test sur haut-parleurs (c.f. section 4.3.1). On voit que le taux de confusion pour chacun des sujets est très variable et de 13% en moyenne, contre 7% en moyenne pour le test sur haut-parleurs. Ceci est en accord avec de précédentes études ayant mis en évidence un doublement du taux de confusions avant-arrière dans le cas d'une écoute binaurale par rapport à la localisation de sources réelles [WK89, Bro95]. On note aussi que les sujets 1, 3 et 9 produisent significativement plus de confusions avant-arrière que les autres, ce qui pourrait s'expliquer par une difficulté à localiser leurs propres HRTFs, ou bien d'un défaut de mesure lié aux HRTFs ou à l'égalisation du casque.

- Dispersion des réponses

  La dispersion des réponses au sein des 5 répétitions est définie de la même façon qu'au chapitre sur les méthodes de pointage (c.f. équation 4.1). Le paramètre κ^{-1} varie entre 0 (pas de dispersion) et $\kappa^{-1} = 1$ (réponses distribuées uniformément sur la sphère pour N tendant vers l'infini).

  La figure 7.9(b) permet d'observer la dispersion des réponses dans le cas de la condition d'écoute individuelle (bleu) ou non-individuelle (rouge). On remarque tout d'abord une augmentation de la dispersion des réponses dans la condition non-individuelle par rapport à la condition individuellle pour tous les sujets : la dispersion est de 0.1 en moyenne en condition non-individuelle contre 0.06 en condition individuelle. Par comparaison avec le test sur haut-parleurs, la dispersion en condition individuelle est 3 fois plus importante (0.06 pour [V] contre 0.02 pour [R]). Cette observation est principalement liée au taux de confusions donné en figure 7.9(a)). En effet, la dispersion au sein des répétitions (au nombre de 5 dans le cas [V], 8 dans le cas [R]) est calculée sans retrait ou résolution des confusions.

  Par ailleurs, on note que le sujet 4 présente la plus grande dispersion, que ce soit en condition individuelle ou non-individuelle. Cette observation témoigne d'une difficulté globale à localiser les sources virtuelles. Le sujet 3 présente également une dispersion particulièrement marquée dans le cas non-individuel.

- Erreurs latérales par rapport à la direction cible

  Dans la figure 7.9(c), la différence latérale moyenne (non signée) entre l'angle latéral de la direction cible et l'angle latéral des directions pointées est donnée pour chaque sujet, de manière distincte pour les conditions d'écoute individuelle et non-individuelle. On observe que les erreurs latérales sont légèrement plus importantes pour la localisation en condition non-individuelle par rapport à la localisation en condition individuelle. Une différence d'ITD à la direction de synthèse entre la HRTF non-individuelle et la HRTF de l'auditeur pourrait expliquer ces erreurs en condition non-individuelle. En condition individuelle, les erreurs latérales s'approchent de celles du test sur haut-parleurs.

- Déviation standard latérale au sein des répétitions

  La figure 7.9(d) présente la dispersion latérale des réponses au sein des répétitions (ou déviation standard latérale). Elle renseigne sur la consistance des sujets à pointer dans la dimension latérale.

Bien que les sujets 1, 3, 4 et 12 montrent une dispersion latérale remarquablement élevée, la déviation standard latérale moyenne relative au test [V] est du même ordre que celle obtenue dans le test [R]. De la même façon que pour les erreurs latérales, la dispersion latérale ne semble pas être liée au type de sources sonores (réelles ou virtuelles). Cependant, on note une légère augmentation de la dispersion latérale lorsque des HRTFs non-individuelles sont utilisées.

- ITD à la direction pointée

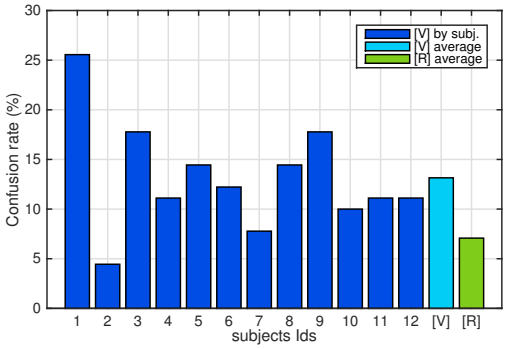
  La figure 7.9(e) présente la différence moyenne entre l'ITD aux directions pointées (ITDs associés aux HRTFs du sujet et interpolé aux directions pointées) et l'ITD cible (associé à la HRTF cible). De la même façon, la figure 7.9(f) présente le coefficient de corrélation entre les ITDs aux directions pointées et les ITDs cibles. Ces mesures sont des indicateurs à la fois de la capacité du sujet à identifier les indices de localisation interauraux (principalement l'ITD dans le cas de sources larges bandes) et de sa capacité à pointer correctement latéralement. Ces deux quantités sont liées : on remarque par exemple que le sujet 9 qui offre la plus faible différence d'ITD dans la figure 7.9(e) est celui dont la corrélation en figure 7.9(f) est la plus forte.

  L'ITD aux directions pointées a été déterminé à partir des ITDs de l'auditeur estimés sur les 1500 points de mesures avec la méthode "MaxIACCr" puis interpolée à l'endroit des réponses. La fréquence d'échantillonnage étant de 48 kHz, une différence d'un échantillon représente une différence d'ITD de 20 μs. La différence d'ITD est en moyenne de 5 échantillons ce qui correspond donc à une différence d'ITD de 10 ms. Cette valeur est bien au-dessus du seuil de discrimination (JND) de l'ITD, de l'ordre de 10 μs [Bla97]. Les erreurs latérales semblent donc plutôt refléter une imprécision de pointage. La figure 7.9(f) montre par ailleurs une forte corrélation (supérieure à 0.9 pour tous les sujets) entre l'ITD cible et l'ITD aux directions pointées. De plus, on observe sur cette figure une légère baisse du taux de corrélation en condition non-individuelle qui est particulièrement marquée pour les sujets 1, 4, et 10. Cela traduit une localisation dans la dimension latérale plus confuse lorsque des HRTFs non-individuelles sont utilisées.

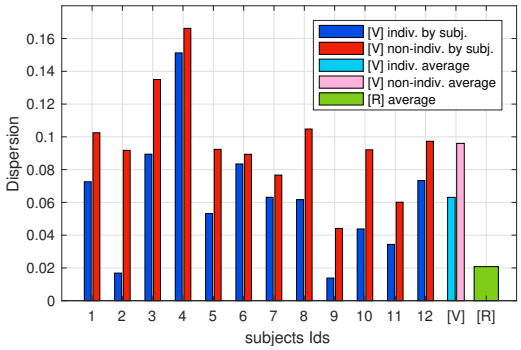
- Distance moyenne de pointage et temp de réponse

  Pour finir, la distance de pointage moyenne pour chaque sujet est donnée en figure 7.9(g). Elle de 23 cm en moyenne (toutes conditions d'écoute confondues) contre 26 cm dans le cas du test sur haut-parleurs. Cette faible différence peut s'interpréter par une externalisation des sources sonores plus faible dans le cas de la localisation de sources virtuelles et donc une tendance à pointer plus proche de la tête. Cela dit, rien n'indique ici que cette différence est significative.

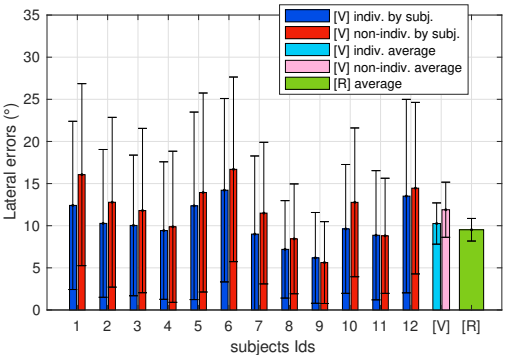
  Enfin, la figure 7.9(h) présente le temps de réponse moyen (entre le début et la fin du stimulus) pour chacun des sujets. Il est de 3.57 sec. dans le cas du test en binaural (toutes conditions d'écoute confondues) contre 2.47 sec. dans le cas du test sur haut-parleurs. Cette différence peut être liée au fait que le son restait émis en continu jusqu'à réponse du sujet dans le cas du test en binaural. Elle peut également s'expliquer par le fait que les sources virtuelles rendent la localisation plus floue, en particulier en condition non-individuelle, et donc que les sujets prennent plus de temps avant de valider leur réponse.



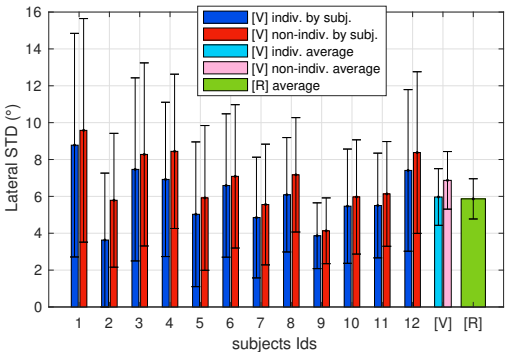
(a) Taux de confusions avant-arrière en cond. indiv.



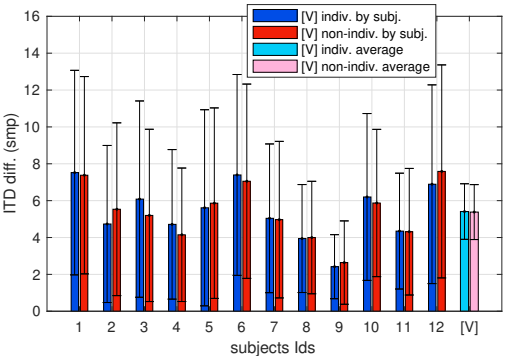
(b) Dispersion κ^{-1} dans les répétitions (indiv./non-indiv.)



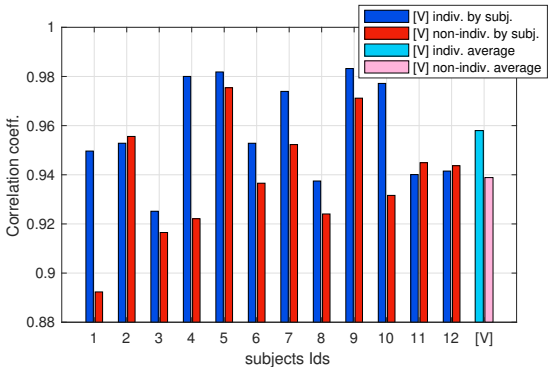
(c) Erreur latérale par rapport à la direction cible



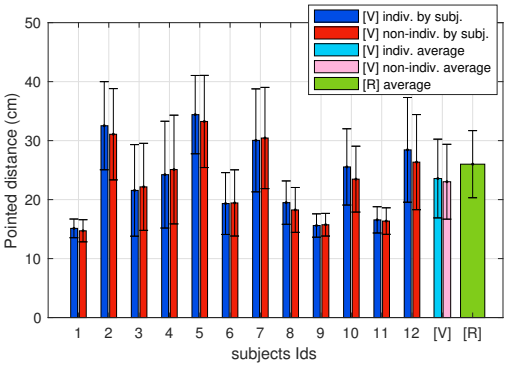
(d) Dispersion latérale



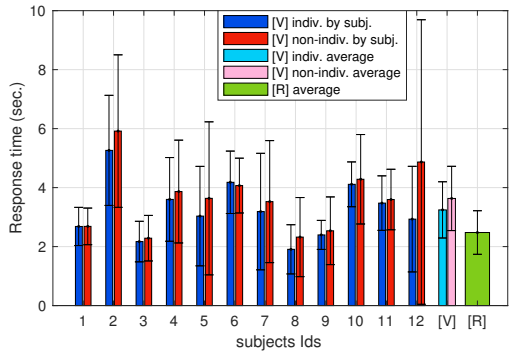
(e) Différence d'ITD aux réponses par rapport à l'ITD cible



(f) Corrélation entre l'ITD cible et l'ITD aux directions pointées



(g) Distance moyenne de pointage par rapport à la position du "centre de la tête"



(h) Temps de réponse moyen

FIGURE 7.9 – Moyenne et écart-type de plusieurs quantités estimées sur les données test de localisation en binaural ([V]) pour chaque sujet dans les conditions d'écoute individuelle et non-individuelle (barres en bleu foncé et rouge respectivement), ainsi que moyenné sur les 12 sujets (barres bleu clair et rose). Comparaison avec la moyenne estimée sur les 13 sujets du test de localisation sur haut-parleurs ([R]) avec la méthode de pointage proximale (barre verte).

7.2.2 Correction du biais de pointage

Pour rappel, le biais de pointage en azimut a été caractérisé à partir des données de chacun des 13 sujets du test sur haut-parleurs puis moyenné sur les sujets de sorte à obtenir un biais global représentatif de la méthode de pointage (voir section 6.4.7.1). Nous vérifions ici que la correction des réponses par l'inverse du biais de pointage en azimut permet de réduire les erreurs latérales du test en binaural en condition individuelle. La figure 7.10(a) présente l'erreur latérale moyenne de chacun des sujets en condition individuelle avant et après correction du biais de pointage en azimut sur les réponses. On observe une réduction globale des erreurs latérales grâce à la correction du biais sauf pour les sujets indiqués par les étoiles vertes ($\frac{3}{12}$), qui ne semblent pas nécessiter cette correction.

7.2.3 Largeur latérale de l'espace de prédiction

Comme présenté dans la section 6.4.2, réduire l'espace de prédiction à un intervalle latéral soulève les questions de la largeur latérale de l'intervalle et du traitement des réponses qui apparaissent en dehors. Les réponses éloignées de l'intervalle de localisation théorique (centré sur l'ITD cible) ne peuvent s'expliquer par un processus de localisation basé sur les indices interauraux. Ces réponses semblent donc représenter des réponses aberrantes. Comme effectué par Baumgartner et al., ces réponses seront donc supprimées. Cette section présente le nombre de réponses concernées en fonction de la largeur de l'intervalle latéral considéré. En effet, la largeur de l'intervalle latéral détermine le nombre de réponses aberrantes.

L'intervalle latéral de prédiction est centré sur la position latérale théorique de localisation. Celle-ci est définie par l'angle latéral moyen où la différence d'ITD entre la cible et les ITDs de l'auditeur est minimale (inférieure ou égale à 1 échantillon, soit 20μs). La figure 7.10(b) présente le pourcentage de fois (sur les 540 réponses) où les réponses apparaissent en dehors de l'intervalle latéral théorique défini par une largeur de $20°, 30°, 40°$ ou $50°$. Les boîtes à moustache indiquent la médiane, les quartiles inférieur et supérieur, le minimum et le maximum associés aux pourcentages obtenus pour les 12 sujets. Les boîtes bleues et rouges présentent ce pourcentage avec ou sans correction du biais de pointage en azimut, respectivement (voir section 6.4.7.1). On voit que le nombre de réponses aberrantes est réduit lorsque les réponses sont corrigées par le biais de pointage inverse. A partir des données de la figure 7.10(b), il apparaît acceptable de définir un intervalle latéral théorique de largeur 40°. En effet, cela revient à supprimer autour de 10% des réponses, dans le cas où les réponses sont corrigées par un biais en azimut (cette correction paraît indispensable au vu des résultats). De plus, il semblerait que la tendance progressive décroissante du nombre de valeurs aberrantes atteigne un plateau à partir d'une largeur latérale de 40° (les résultats sont peu améliorés pour une largeur de 50°).

Pour rappel et par comparaison, dans le modèle de prédiction sur plans sagittaux, Baumgartner et al. avaient utilisé une largeur latérale de 10°. Utiliser une telle définition reviendrait à supprimer plus de la moitié des réponses des sujets, ce qui n'est pas souhaitable.

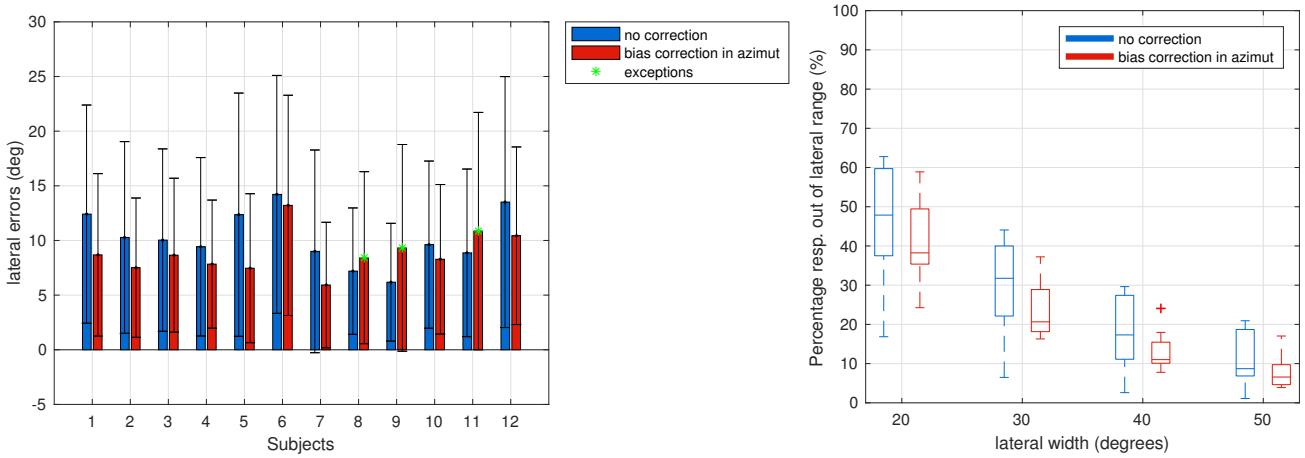


FIGURE 7.10 – (a) Erreur latérale moyenne et écart-type associé, pour chaque sujet du test en binaural, avant (bleu) et après (rouge) correction des réponses par le biais inverse de pointage en azimut. Les étoiles vertes indiquent les sujets pour lesquels la correction augmente l'erreur latérale moyenne. (b) Pourcentage de réponses en dehors de l'intervalle latéral centré sur l'ITD cible et de largeur latérale variable : $20°, 30°, 40°, 50°$. Les boîtes à moustache indiquent les quartiles inférieur et supérieur des pourcentages obtenus pour les 12 sujets. Le trait à l'intérieur des boîtes représente la médiane et les limites inférieure et supérieure des barres d'erreur, les valeurs min. et max.

7.3 Paramétrisation du modèle

Nous nous intéressons à présent à la prédiction de localisation des données expérimentales du test en binaural. Les paramètres du modèle ont été présentés en section 6.4 du chapitre précédent. Le schéma figure 7.11 permet de visualiser le principe de la paramétrisation.

Etant donné la multitude de paramètres du modèle et leurs interactions, l'analyse des résultats de prédiction est divisée en 4 analyses simplifiées où l'effet des paramètres est étudié pas par pas en parallèle d'une optimisation des paramètres réalisée au fur et à mesure. Cette procédure sous-entend fixer certains paramètres *a priori*. Dans la première analyse par exemple, la contribution relative des indices interauraux et spectraux pour la prédiction est fixée à égalité ($50\% - 50\%$). En effet, ces deux types d'indices sont complémentaires pour la localisation. De plus, bien que nous travaillions avec des espaces restreints à une tranche latérale de la sphère (c.f. section 6.4.2), la largeur latérale de ces espaces est supérieure au seuil de discrimination de l'ITD. Il apparaît donc nécessaire de considérer l'indice interaural à hauteur de l'indice spectral. Dans les deux premières analyses, le modèle de dispersion n'est pas appliqué et son apport est étudié à la troisième étude. Les paramètres fixés sont spécifiés dans un tableau au début de chaque section.

Les paramètres du modèle sont étudiés conjointement avec des analyses statistiques de variance (ANOVAs) à mesures répétées. Ce type d'analyse permet d'évaluer la significativité statistique des effets observés. Plus d'informations sur le principe des analyses ANOVAs peut être trouvé en section 4.3.1 du chapitre sur les méthodes de pointage. Pour rappel, un effet sera considéré comme significatif si la valeur de $p < 0.005$ [Joh13] et marginalement significatif si $0.05 > p > 0.005$. Aucun facteur inter-groupe n'est ici considéré et les facteurs intra-groupes seront définis à chaque début d'analyse. L'optimisation des paramètres du modèle est basée sur la minimisation de la log-vraisemblance actuelle L_a, comme présenté section 6.4.3. Ce critère permet de quantifier la probabilité de réponse prédite par le modèle en chacune des directions pointées. Pour mener correctement les analyses ANOVA, les conditions individuelle et non-individuelle sont étudiées séparément car elles ne sont pas basées sur le même nombre d'observations ($\frac{1}{6}$ versus $\frac{5}{6}$ respectivement).

Les différentes étapes de ce travail d'optimisation sont illustrées dans le schéma suivant.

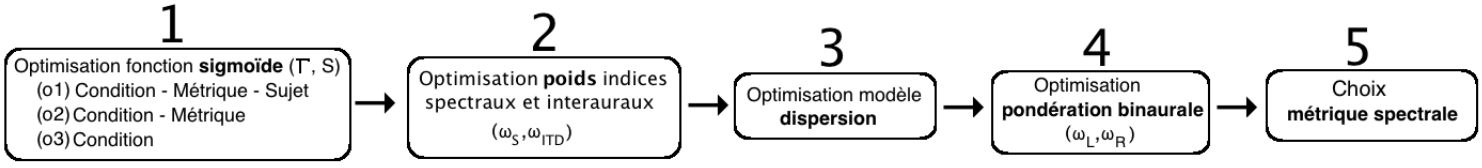


FIGURE 7.11 – Étapes principales de paramétrisation.

7.3.1 Etude des fonctions sigmoïdes

Cette première analyse se penche sur l'étude des fonctions sigmoïdes qui transforment les distances normalisées en indices de similarité et permettent d'obtenir une distribution de probabilité de réponse sur l'espace de prédiction. Pour rappel, la fonction sigmoïde est définie par deux paramètres, Γ et S qui influent respectivement sur la pente et le point d'inflexion. L'effet de chacun des deux paramètres n'est pas indépendant : l'effet de S est d'autant plus marqué que Γ est grand. Pour cette raison, nous n'étudierons pas leurs effets séparément. Chaque fonction sigmoïde testée sera caractérisée par son degré de sélectivité spatial tel que présenté en section 6.4.4 : plus la pente Γ est forte et le point d'inflexion S est faible, plus les zones de similarité seront concentrées dans l'espace de prédiction. Le degré de sélectivité spatial du modèle associé à chaque fonction sigmoïde est quantifié par la valeur moyenne de la log-vraisemblance escomptée L_e tel que présenté section 6.4.3.

Nous cherchons ici à identifier les fonctions sigmoïdes qui permettent d'optimiser les résultats de prédiction, i.e. de minimiser la log-vraisemblance associée aux directions pointées L_a. La nécessité d'adapter la fonction sigmoïde à chaque sujet du test expérimental tel que suggéré dans de précédentes études [BML13, BML14] est étudiée en parallèle de la log-vraisemblance actuelle et du pourcentage de réussite du modèle. Pour rappel, le pourcentage de réussite correspond au pourcentage de fois (sur les 12 sujets) où $L_a \in CI_{99\%}(L_e)$ et permet d'évaluer la capacité du modèle à prédire les observations réalisées dans les tests de localisation. Ce critère correspond à une variable dichotomique ($L_a \in CI_{99\%}(L_e)$ ou $L_a \notin CI_{99\%}(L_e)$) et non à une variable continue. Par conséquent, des analyses de variance ne peuvent être réalisées sur ce critère. Nous aborderons alors une autre type d'analyse, le test d'indépendance du χ^2. Les paramètres du modèle de cette première analyse sont répertoriés dans le tableau 7.1.

Correction du biais	correction du biais en azimut
Condition d'écoute	individuelle et non-individuelle, séparément
Espace de prédiction	intervalle latéral de largeur 40°
Fonction sigmoïde (γ, S)	toutes combinaisons $\Gamma = 1, 1.25, 1.5, 1.75, 2$ et $S = 0, -1, -2, -3, -4, -5$ (soit 30 fonctions sigmoïdes)
Dispersion	sans dispersion
w_{ITD}	$w_{ITD} = 0.5$
Métrique spectrale	les 10 du tableau 6.1
Pondération binaurale	méthode de Baumgartner
Evaluation du modèle	log-vraisemblance actuelle L_a et pourcentage de réussite $L_a \in CI_{99\%}(L_e)$

TABLE 7.1 – Paramètres de l'analyse sur les fonctions sigmoïdes

7.3.1.1 Effet de la fonction sigmoïde

Nous réalisons dans un premier temps une analyse ANOVA à mesures répétées avec les facteurs fonction sigmoïde, métrique et condition d'écoute. Notons que ce dernier facteur est utilisé exceptionnellement dans cette analyse pour mettre en évidence les interactions qui existent entre les différents facteurs. Etant donné que les conditions d'écoute individuelle et non-individuelle ne sont pas basées sur le même nombre d'observations, nous séparerons par la suite la condition non-individuelle de la condition individuelle.

Selon l'analyse ANOVA à mesures répétées, la fonction sigmoïde a un effet significatif sur la log-vraisemblance L_a ($F(29, 319) = 16.70$; $p < 0.001$). Les figures 7.12 et 7.13 permettent de visualiser l'évolution des valeurs de L_a en fonction de la fonction sigmoïde utilisée. Les fonctions sigmoïdes sont ordonnées par sélectivité spatiale croissante (suivant la décroissance de L_e, voir section 6.4.4). Les figures indiquent pour quelles fonctions sigmoïdes L_a est minimisée, en fonction de la métrique et du sujet. Elles offrent une comparaison du minimum de L_a avec l'intervalle de confiance de la log-vraisemblance L_e. Les résultats sont présentés séparément pour les conditions d'écoute individuelle et non-individuelle (figures 7.12 et 7.13, respectivement). La fonction sigmoïde optimale, soit celle qui offre un minimum de log-vraisemblance L_a, est indiquée par un rond rouge.

Sur ces figures, d'après la comparaison entre les valeurs de L_a minimales et l'intervalle de confiance de la log-vraisemblance L_e (représenté par les barres verticales noires), on remarque que le minimum de L_a se situe la plupart du temps dans l'intervalle de confiance de L_e. Cette caractéristique est valable pour cette valeur minimale de L_a, mais ne se généralise pas pour le reste de la courbe (résultats non présentés). Cela signifie que le modèle prédit correctement les réponses des sujets, lorsqu'il est paramétré par la fonction sigmoïde qui optimise les résultats. On observe également que l'intervalle de confiance associé à la log-vraisemblance L_e est plus large en condition individuelle que non-individuelle. En effet, un sixième des cas seulement correspondent à la condition individuelle ce qui explique que la déviation standard liée aux multiples tirages de L_e soit plus importante.

Fonction sigmoïde et sujet Sur ces figures 7.12 et 7.13, on observe également que la fonction sigmoïde optimale varie selon le sujet. Ces observations confirment les résultats de Baumgartner et al. [BML14]. Ces auteurs ont montré que la fonction du modèle de prédiction qui transforme les distances en indices de similarité doit être individualisée pour chaque sujet. Celle-ci reflète en effet la capacité de chacun à discriminer les indices de localisation. Comme présenté section 6.4.3, les scores de prédiction d'un sujet qui identifierait systématiquement les régions de forte probabilité seraient d'autant améliorés que la fonction sigmoïde est sélective. Au contraire, la log-vraisemblance associée aux réponses d'un sujet "moins bon localisateur" atteindra le seuil de chance si la fonction sigmoïde est trop sélective. Pour rappel, Baumgartner et al. [BML14] individualisent uniquement le paramètre S de la fonction sigmoïde (point d'inflexion) avec Γ (pente) fixé pour tous les sujets. Le paramètre S est déterminé de sorte à optimiser les critères PE et QE, non utilisés ici (voir la section 6.2.3 qui présente ce modèle), et pour la condition individuelle uniquement. Nos résultats suggèrent qu'il est nécessaire d'individualiser à la fois les paramètres S et Γ. De plus, il semblerait qu'une optimisation de la fonction sigmoïde à partir des données du sujet en condition individuelle ne permette pas d'optimiser les résultats en condition non-individuelle car les fonctions optimales associées à un même sujet diffèrent entre les conditions d'écoute. Dans la section suivante, nous évaluerons l'impact d'une optimisation associée à la condition d'écoute individuelle sur les résultats de prédiction en condition non-individuelle et inversement.

Fonction sigmoïde et condition d'écoute Les fonctions sigmoïdes présentent une interaction significative avec la condition d'écoute ($F(29, 319) = 25.58$; $p < 0.001$). A partir des figures 7.12 et 7.13, on observe globalement que les résultats en condition individuelle sont optimisés avec des fonctions sigmoïdes plus sélectives qu'en condition non-individuelle. En effet, si l'on s'intéresse aux fonctions sigmoïdes qui optimisent le plus grand nombre de cas au sein de chaque condition d'écoute on trouve que les paramètres (1,-5) et (2,-2) optimisent la majorité des résultats en condition non-individuelle et individuelle, respectivement. Cela peut s'expliquer par le fait que certains sujets présentent une difficulté à localiser les sources virtuelles

lorsque des HRTFs non-individuelles sont utilisées. En effet, les observations effectuées en section 7.2.1, notamment figure 7.9(b), montrent une dispersion des réponses en condition non-individuelle qui est plus importante qu'en condition individuelle, ce qui traduit une localisation plus confuse dans le cas de HRTFs non-individuelles. Le modèle doit donc s'adapter à cette tendance en offrant des zones de forte probabilité plus étalées dans l'espace (i.e. en utilisant des fonctions sigmoïdes moins sélectives).

Fonction sigmoïde et métrique L'interaction entre les fonctions sigmoïdes et les métriques est significative ($F(261, 2871) = 32.72$; $p < 0.001$). Les métriques spectrales se distinguent non seulement par les régions spatiales de forte similarité spectrale mais également par leur caractère de sélectivité spatial. En effet, elles présentent des zones de forte similarité qui sont plus ou moins concentrées dans l'espace et la fonction sigmoïde va jouer un rôle sur l'étalement de ces zones. Si une métrique présente un caractère très sélectif, alors une fonction sigmoïde moins sélective sera mieux adaptée. La comparaison des métriques est ici menée uniquement sur la correspondance entre les régions spatiales de forte similarité spectrale et les réponses des sujets. Par conséquent, il est important d'adapter la fonction sigmoïde à chaque métrique, comme nous l'avions mentionné section 6.4.3. La figure 7.13 confirme ce phénomène. On observe que les métriques basées sur le gradient spectral (*grad-Mean* et *posGrad-Mean*), qui possèdent un caractère plus sélectif que les autres, requièrent des fonctions sgimoïdes moins sélectives pour optimiser les résultats de prédiction en comparaison aux métriques *mfcc-MSE* ou *mfcc-Corr*.

Enfin, l'interaction entre les 3 facteurs fonction sigmoïde, métrique et condition d'écoute est également significative ($F(261, 2871) = 5.62$; $p < 0.001$). Cela suggère que la fonction sigmoïde optimale doit s'adapter à chaque métrique au sein de chaque condition d'écoute, séparément.

Pour résumer, nous avons observé que :

- pour L_a minimisée, $L_a \in CI_{99\%}(L_e)$
- la fonction sigmoïde qui optimise la prédiction des données en condition individuelle n'optimise pas la prédiction en condition non-individuelle
- les fonctions sigmoïdes sont plus sélectives en condition individuelle que non-individuelle, ce qui peut être lié aux différences de dispersion dans les jugements entre les deux conditions d'écoute
- les fonctions sigmoïdes optimales diffèrent d'une métrique à l'autre, résultat pouvant être interprété par le caractère de sélectivité spatial variable des métriques spectrales

Dans le cadre de la sélection d'un jeu de HRTFs pour un nouveau sujet, l'adaptation de la fonction sigmoïde au sujet n'est pas envisageable. On peut alors se demander quel est l'impact d'une optimisation globale de la fonction sigmoïde (à chaque métrique spectrale au sein de chaque condition d'écoute) par rapport à une optimisation individualisée (i.e. adaptée à chaque particulier : sujet $\times$ métrique spectrale $\times$ condition d'écoute) sur les résultats de prédiction. Les fonctions sigmoïdes qui optimisent globalement les résultats par métrique spectrale et condition d'écoute, sans prise en compte du sujet, peuvent être visualisées figures 7.14. Les ronds rouges indiquent les fonction sigmoïdes optimales et les barres horizontales de couleur indiquent les fonctions sigmoïdes qui offrent des résultats de log-vraisemblance non significatifs par rapport aux résultats des fonctions sigmoïdes optimales.

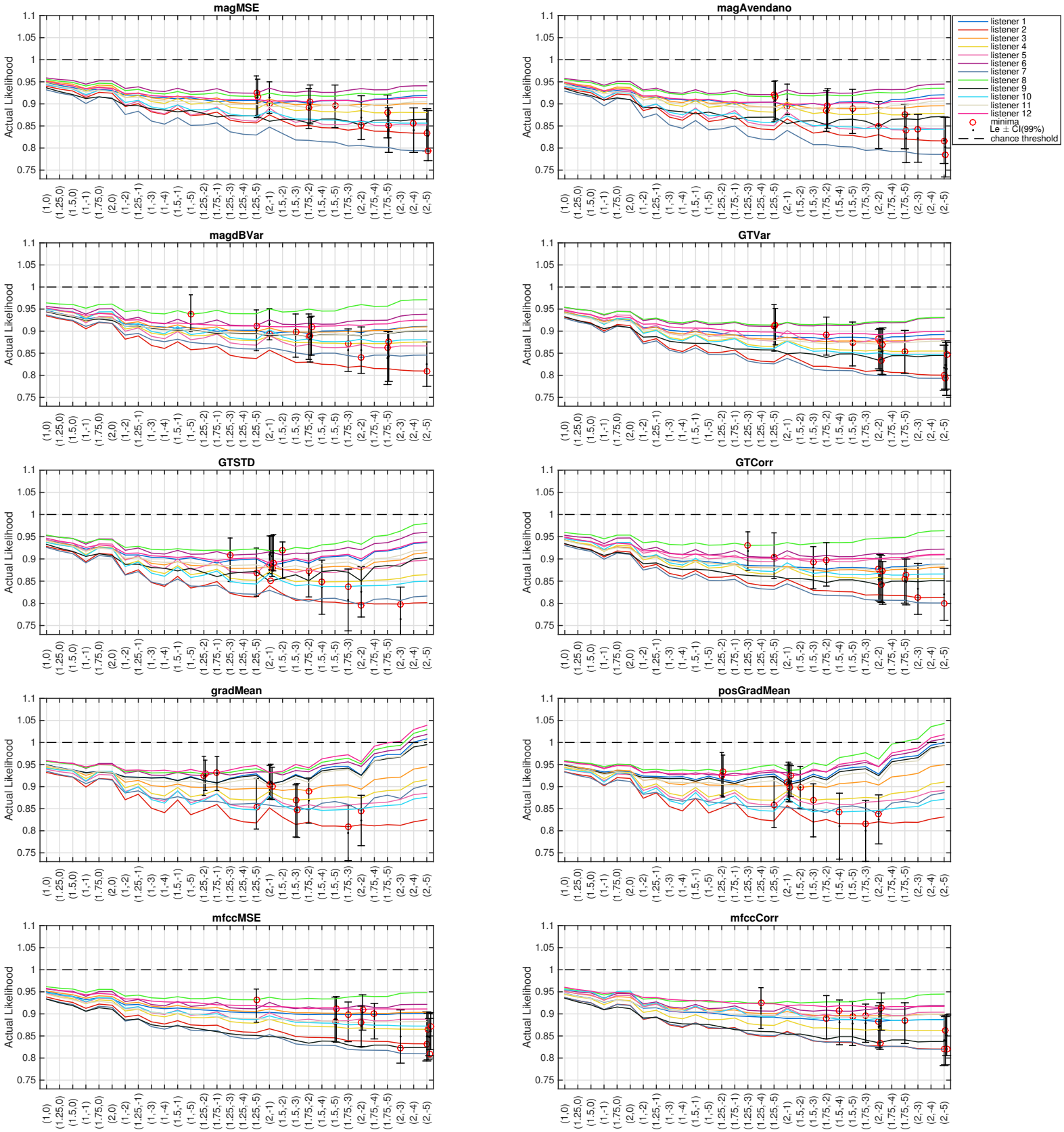


FIGURE 7.12 – Evolution de la log-vraisemblance L_a en fonction de la sélectivité de la fonction sigmoïde. Le minimum de log-vraisemblance est indiqué par un rond rouge pour chaque sujet (traits de couleur) en condition individuelle et séparément pour chaque métrique. Les barres verticales noires indiquent l'intervalle de confiance à 99% de la log-vraisemblance escomptée L_e associée au modèle paramétré avec la fonction sigmoïde qui optimise les résultats.

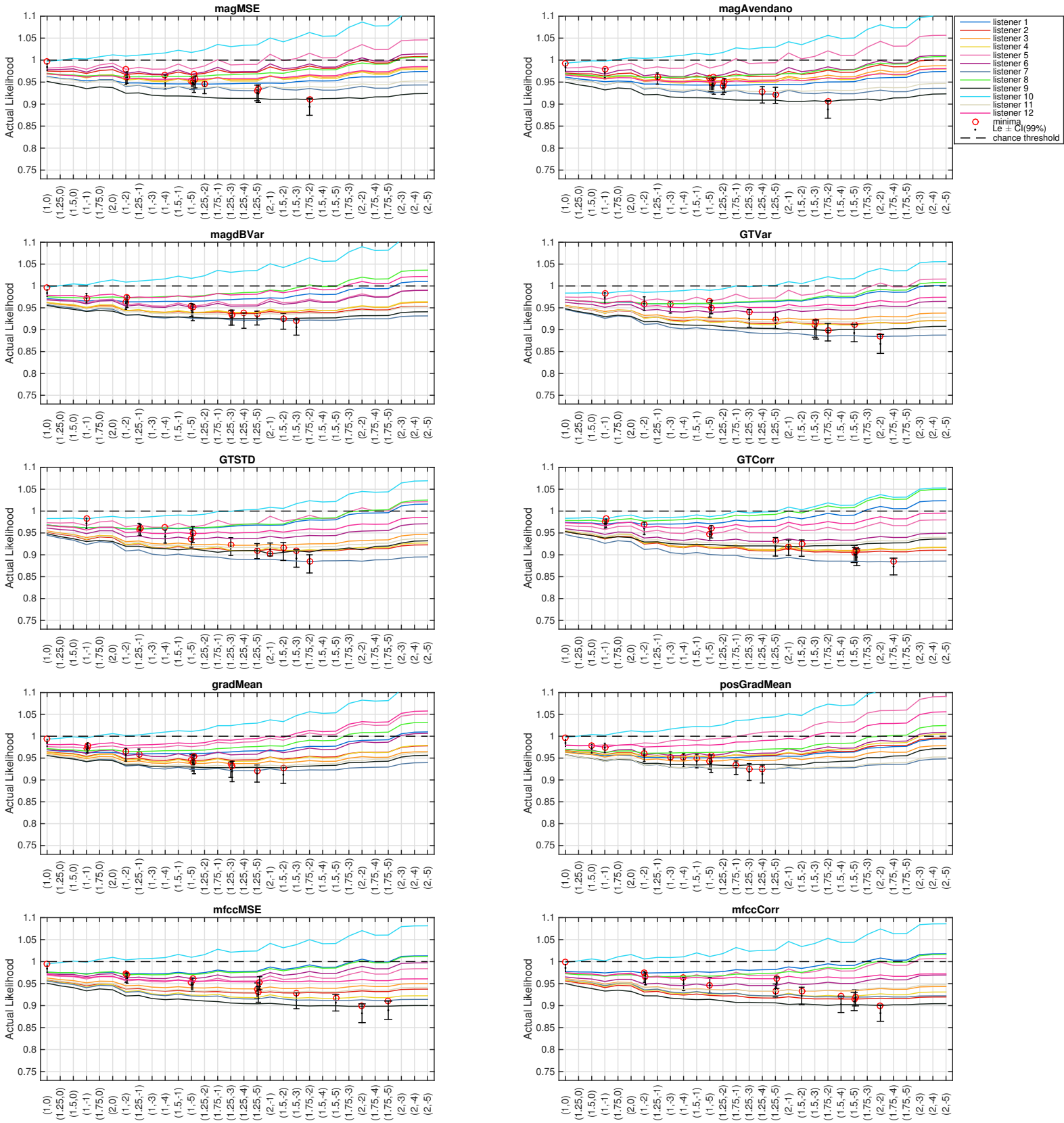


FIGURE 7.13 – Evolution de la log-vraisemblance L_a en fonction de la sélectivité de la fonction sigmoïde. Le minimum de log-vraisemblance est indiqué par un rond rouge pour chaque sujet (traits de couleur) en condition non-individuelle et séparément pour chaque métrique. Les barres verticales noires indiquent l'intervalle de confiance à 99% de la log-vraisemblance escomptée L_e associée au modèle une fois paramétré par la fonction sigmoïde qui optimise les résultats.

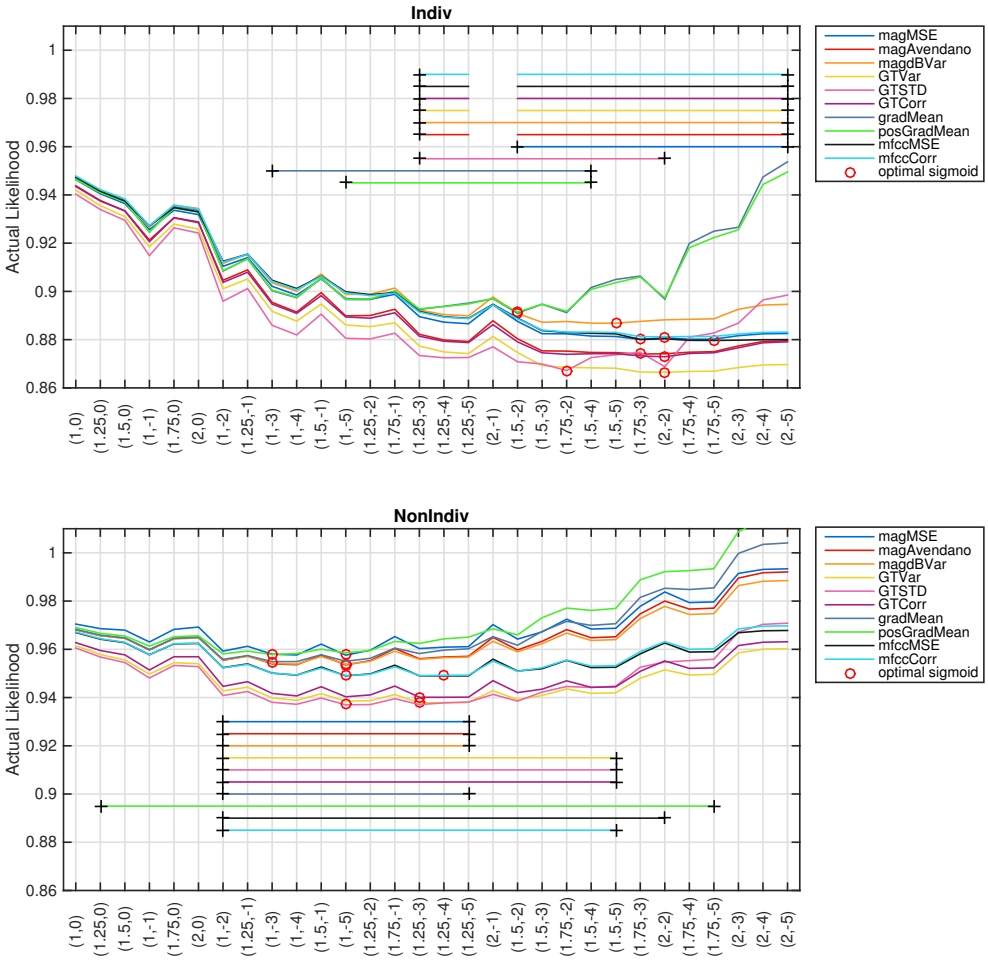


FIGURE 7.14 – Log-vraisemblance actuelle moyennée sur les sujets selon la fonction sigmoïde, pour chaque métrique spectrale (courbes de couleurs) et les conditions individuelle (en haut) et non-individuelle (en bas). Les ronds rouges indiquent pour quelle fonction sigmoïde L_a est en moyenne minimisée pour chacune des métriques spectrales. Les barres horizontales indiquent les fonctions sigmoïdes offrant des résultats de log-vraisemblance non significatifs des résultats associés à la fonction sigmoïde optimale (résultats issus du test post-hoc de l'ANOVA).

7.3.1.2 Facteurs pris en compte dans l'optimisation de la fonction sigmoïde

Pour évaluer la nécessité d'adapter la fonction sigmoïde à la condition d'écoute, à la métrique et au sujet, nous étudions ici l'impact de 3 types d'optimisation de la fonction sigmoïde :

(o1) une adaptation de la fonction sigmoïde spécifique à chaque triplet [condition d'écoute ; métrique ; sujet]. Ce type d'optimisation doit théoriquement conduire aux meilleurs résultats puisqu'elle correspond à une optimisation pour chaque cas de figure.

(o2) une adaptation spécifique au doublet [condition d'écoute ; métrique] (i.e. recherche de la fonction qui minimise L_a en moyenne sur les sujets). Ce type d'optimisation illustre le cas réel d'une prédiction menée sur les données d'un nouvel auditeur, dont les HRTFs ne sont pas disponibles.

(o3) l'utilisation de la fonction sigmoïde associée à l'optimisation de type (o1) pour la condition d'écoute opposée. Ce dernier type d'optimisation modélise ce qui est proposé par Baumgartner et al. [BML14], soit une fonction sigmoïde optimisée à partir des résultats en condition individuelle uniquement, pour chaque sujet (et une métrique spécifique), et appliquée dans d'autres cas d'écoute. Il s'agit de vérifier la généricité de l'optimisation. Dans notre cas nous voulons vérifier si une optimisation menée sur des HRTFs individuelles peut s'appliquer à des HRTFs non-individuelles.

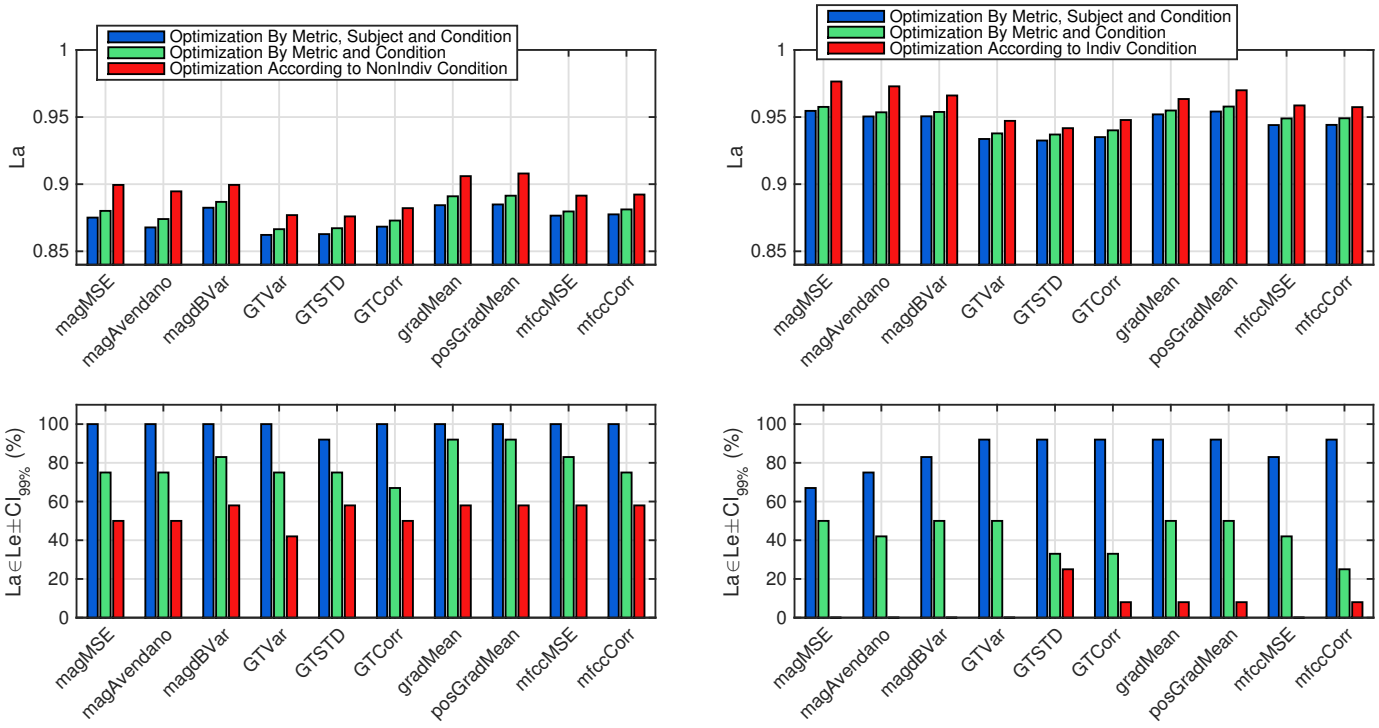
La figure 7.15 permet de visualiser l'effet de chacune de ces optimisations sur les résultats de prédiction pour chaque métrique spectrale, tous sujets confondus. La valeur moyenne de log-vraisemblance L_a ainsi que le pourcentage de fois où elle apparaît dans l'intervalle de confiance de la log-vraisemblance escomptée par le modèle L_e sont obtenus par une moyenne et un pourcentage sur les 12 sujets, respectivement. Une analyse ANOVA à mesures répétées est réalisée en parallèle sur la log-vraisemblance L_a pour évaluer l'effet des facteurs intra-groupe type d'optimisation et métrique spectrale, au sein des conditions d'écoute individuelle et non-individuelle séparément. Par ailleurs, le pourcentage de fois où $L_a \in CI_{99\%}(L_e)$ correspond à un pourcentage sur les sujets (i.e. 100% correspond au cas où nous observons 12 fois la tendance $L_e - CI_{99\%} \leq L_a \leq L_e + CI_{99\%}$). Dans ce cas, la variable n'est pas continue et une analyse de variance ne peut être appliquée. Par conséquent, nous réalisons un test d'indépendance du χ^2, adapté aux variables dichotomiques. Cette analyse permet de tester l'indépendance entre la variable pourcentage de réussite du modèle (2 niveaux : $L_a \in CI_{99\%}(L_e)$ et $L_a \notin CI_{99\%}(L_e)$) et le type d'optimisation (3 niveaux) ou encore la métrique spectrale (10 niveaux). Ce type d'analyse ne permet cependant pas d'analyser l'interaction entre plusieurs facteurs, à savoir l'interaction entre les facteurs métrique spectrale et type d'optimisation. Les conditions individuelle et non-individuelle sont analysées séparément.

Analyse de la log-vraisemblance actuelle Tout d'abord, dans la condition individuelle, le facteur type d'optimisation et son interaction avec le facteur métrique spectrale sont significatifs ($F(2, 22) = 7.44$; $p = 0.003$ et $F(18, 198) = 2.42$; $p = 0.001$, respectivement). En condition non-individuelle, le type d'optimisation a un effet seulement marginalement significatif sur les résultats ($F(2, 22) = 6.37$; $p = 0.007$) et son interaction avec le facteur métrique spectrale est significatif ($F(18, 198) = 2.85$; $p < 0.001$). Les tests post-hoc associés à ces interactions révèlent tout d'abord que l'optimisation (o1) offre des résultats significativement meilleurs par rapport à l'optimisation (o3) pour toutes les métriques et conditions d'écoute. Ce résultat remet en cause le recours à la méthode d'optimisation proposée par Baumgartner et al. [BML14] dans le cas d'une écoute ne se restreignant pas aux HRTFs individuelles. En effet, optimiser la fonction sigmoïde au sujet et à la métrique dans la condition individuelle pour une prédiction dans la condition non-individuelle offre des résultats significativement moins bons que si l'optimisation avait été réalisée spécifiquement à la condition d'écoute. Deuxièmement, les résultats de l'optimisation (o1) ne diffèrent pas significativement de l'optimisation (o2), quelle que soit la condition d'écoute et la métrique spectrale. Cela signifie que l'individualisation de la fonction sigmoïde à chaque sujet n'améliore pas significativement les résultats de prédiction. Enfin, même si cela nous intéresse moins ici, les différences entre les optimisations (o2) et (o3) varient selon les métriques et les conditions d'écoute. En condition individuelle, les différences entre les optimisations (o2) et (o3) sont significatives sauf pour la métrique *GT-STD* et les métriques *GT-Var* et *GT-Corr*, pour lesquelles les différences sont marginalement significatives. En condition non-individuelle, les différences entre les optimisations (o2) et (o3) sont significatives sauf pour les métriques *GT-STD* et *GT-Corr* et les métriques *GT-Var*, *grad-Mean* et *mfcc-Corr* pour lesquelles les différences sont marginalement significatives.

Analyse du pourcentage de réussite En bas des figures 7.15(a) et 7.15(b), on note que le pourcentage de réussite du modèle dépend beaucoup du type d'optimisation. On observe globalement une dégradation des pourcentages de réussite de l'optimisation (o1) à l'optimisation (o3). On voit également qu'en condition non-individuelle (à droite), les pourcentages de réussite sont proches de 0 pour l'optimisation de type (o3).

Les tests χ^2 révèlent en effet que le type d'optimisation a un effet significatif sur le pourcentage de réussite du modèle, quelle que soit la condition d'écoute ($p < 0.001$ pour les deux conditions). Le chi-2 de Pearson est égal à 69.96 et 155.7 pour les conditions individuelle et non-individuelle respectivement. Le facteur métrique spectrale n'est cependant pas significatif ($p = 0.95$, $\chi^2 = 3.33$ en condition individuelle et $p = 0.98$, $\chi^2 = 2.57$ en condition non-individuelle). Notons que toutes les métriques spectrales et tous

les sujets sont considérés de manière distincte pour tester l'indépendance du type d'optimisation et du pourcentage de réussite du modèle (tableau à $12 \times 10 \times 3$ lignes et 3 colonnes). Même si l'analyse statistique diffère, il semble que le type d'optimisation ait un effet davantage marqué sur le pourcentage de réussite du modèle que sur la log-vraisemblance actuelle ($p = 0.003$ et $p = 0.007$ pour les conditions individuelle et non-individuelle, respectivement).



(a) Condition Individuelle

(b) Condition Non-individuelle

FIGURE 7.15 – Log-vraisemblance actuelle et pourcentage de réussite du modèle en fonction du type d'optimisation de la fonction sigmoïde, pour chacune des métriques spectrales. Les conditions individuelle et non-individuelle sont analysées séparément (figures (a) et (b) respectivement).

Conclusion sur la comparaison (o3) vs (o1)-(o2) Pour conclure, il apparaît primordial, à la fois pour minimiser L_a et maximiser la situation où $L_a \in CI_{99\%}(L_e)$, de tenir compte de la condition d'écoute pour déterminer la fonction sigmoïde optimale. Une prédiction optimisée pour le sujet, la métrique et la condition d'écoute individuelle ne permet pas de prédire les résultats en condition non-individuelle (optimisation o3). Les résultats sont significativement moins bons que ceux obtenus avec une adaptation réalisée indépendamment pour chaque condition d'écoute. Cela remet en cause la méthode proposée dans les modèles précédents [BML13, BML14]. Ces modèles suggèrent en effet que l'optimisation de la fonction sigmoïde doit se baser sur les données expérimentales de la condition d'écoute individuelle afin de prédire la localisation pour d'autres conditions d'écoute. Cependant, il faut avoir conscience que les critères d'optimisation entre leur étude et la nôtre diffèrent (voir section 6.2.3).

Conclusion sur la comparaison (o2) vs (o1) Au vu des valeurs de log-vraisemblance, il semble acceptable d'utiliser une fonction sigmoïde qui ne soit pas individualisée à chacun des sujets. En effet, les résultats obtenus à partir d'une optimisation à la condition d'écoute et à la métrique (optimisation o2) ne sont pas significativement moins bons que les résultats obtenus pour une optimisation complète, i.e. tenant compte du sujet. Cependant, les pourcentages de réussite se dégradent nettement lorsque le sujet n'est pas considéré dans le choix de la fonction sigmoïde optimale. Dans l'objectif de prédire la localisation de stimuli qui n'auraient pas été testés expérimentalement (e.g. pour prédire les directions perçues avec d'autres HRTFs non-individuelles), l'adaptation de la fonction sigmoïde à chaque auditeur semble donc nécessaire. Dans le cadre de la sélection d'un jeu de HRTFs pour un nouvel auditeur, l'optimisation de la fonction sigmoïde à l'auditeur n'est pas envisageable étant donné que ces HRTFs individuelles n'ont pas été mesurées. De plus, nous cherchons un modèle qui prédise au mieux les directions perçues, i.e. pour lequel la log-vraisemblance actuelle est minimisée. Le pourcentage de réussite n'est pas considéré puisque notre objectif n'est pas de prédire les directions perçues avec d'autres stimuli que ceux utilisés lors du test de localisation. Nous continuerons donc nos analyses en considérant uniquement l'optimisation de type (o2), i.e. spécifique au doublet [condition ; métrique].

7.3.2 Etude des poids relatifs des indices interauraux et spectraux

Cette seconde analyse se penche sur l'effet des poids relatifs de l'indice d'ITD et de l'indice spectral sur les résultats de prédiction. Pour rappel, le poids associé à la distance spectrale est de $(1 - w_{ITD})$. Nous réalisons, en parallèle des observations, une analyse ANOVA à mesures répétées sur la log-vraisemblance L_a avec les facteurs intra-groupe suivants : le poids de l'ITD (11 valeurs, de 0 à 1) et la métrique spectrale (10 catégories). L'analyse est menée séparément pour chaque condition d'écoute. Il est important de noter que ces résultats sont relatifs à des espaces de prédiction de largeur latérale 40° et que les résultats peuvent être amenés à varier pour des espaces plus larges ou plus restreints. Par exemple, pour une prédiction sur un plan sagittal, la contribution de l'indice interaural est négligée. Les paramètres de cette analyse sont donnés dans le tableau 7.2.

Correction du biais	correction du biais en azimut
Condition d'écoute	individuelle et non-individuelle, séparément
Espace de prédiction	intervalle latéral de largeur 40°
Fonction sigmoïde	optimisée pour la condition d'écoute et la métrique spectrale
Dispersion	sans dispersion
w_{ITD}	de 0 à 1 par pas de 0.1
Métrique spectrale	les 10 du tableau 6.1
Pondération binaurale	méthode de Baumgartner
Evaluation du modèle	log-vraisemblance actuelle, L_a

TABLE 7.2 – Paramètres associés à l'analyse sur le poids de l'ITD.

Tout d'abord l'ANOVA à mesures répétées présente un effet significatif du poids de l'ITD sur la log-vraisemblance aux réponses L_a en conditions individuelle et non-individuelle ($F(10, 110) = 44.93$; $p < 0.001$ et $F(1, 10) = 19.40$; $p < 0.001$, respectivement). L'interaction entre les facteurs poids de l'ITD et métrique spectrale est également significative dans les deux conditions d'écoute ($F(90, 990)=10.32$; $p < 0.001$ en condition individuelle et $F(90, 990) = 8.46$; $p < 0.001$ en condition non-individuelle). La figure 7.16(a) et 7.16(c) présente cet effet, pour chaque métrique spectrale. On observe que la log-vraisemblance est optimisée pour des poids d'ITD autour de 0.4 et qu'elle augmente pour des poids inférieurs ou supérieurs. La log-vraisemblance converge vers le seuil de chance ($L_a = 1$) à mesure que le poids w_{ITD} tend vers $w_{ITD} = 1$ (pas de prise en compte de l'indice spectral). On note également que la tendance en fonction de w_{ITD} est plus ou moins marquée selon les métriques.

Une partie des résultats du test post-hoc associé à l'interaction entres les deux facteurs sont visibles dans le tableau 7.3. Ce tableau présente plus spécifiquement la significativité statistique entre les résultats de prédiction associés aux poids optimaux (i.e. w_{ITD} qui minimisent L_a) et les autres poids testés. Les signes "+" dans les cases vertes indiquent les poids d'ITD qui optimisent la prédiction pour chaque métrique spectrale et chaque condition d'écoute. Les cases blanches indiquent les cas non significatifs par rapport aux poids optimaux. Les astérisques dans les cases rouges et les croix dans les cases jaunes indiquent respectivement les cas offrant des résultats significatifs et marginalement significatifs des valeurs de L_a associées aux poids optimaux.

D'après le tableau 7.3, on note que le paramètre w_{ITD} optimal varie selon la métrique. Par exemple, les métriques *posGrad-Mean* et *mfcc-MSE* montrent des tendances bien distinctes. Pour la métrique *posGrad-Mean*, les résultats de log-vraisemblance associés aux poids $w_{ITD} = 0.4$ et $w_{ITD} = 0.5$ montrent des résultats significativement meilleurs que les cas $w_{ITD} < 0.2$ et $w_{ITD} > 0.8$. Pour la métrique *mfcc-MSE*, les résultats pour $0 \leq w_{ITD} \leq 0.5$ présentent des résultats significativement meilleurs que les cas $w_{ITD} \geq 0.6$ et ne sont pas significatifs entre eux ($w_{ITD} = 0$ offre des résultats équivalents à $w_{ITD} = 0.5$) .

Ces observations peuvent être mises en parallèle avec la quantité d'information interaurale intrinsèquement encodée par la métrique spectrale (voir discussion en section 6.4.5.3). En effet, nous avons vu que la métrique *mfcc-MSE* est porteuse d'information interaurale. Les cartes de similarité spectrale associées à cette métrique sont corrélées à 0.83 avec les cartes de similarité associée à l'ITD. Lorsque cette métrique spectrale est combinée à la métrique d'ITD, l'information interaurale est redondante et le poids de l'information spectrale dans la prédiction est réduit. Cela explique pourquoi les résultats de prédiction sont optimisés pour des poids d'ITD faibles et significativement dégradés lorsque l'ITD est considéré à plus de 50% de la métrique spectrale. Au contraire, les cartes de similarité spectrale associées à la métrique *posGrad-Mean* présentait une corrélation nulle (voire négative) avec les cartes de similarité associés à la métrique d'ITD. Cela explique que les résultats de prédiction soient significativement moins bons lorsque $w_{ITD} < 0.2$. Cette métrique semble nécessiter davantage d'information interaurale pour optimiser la log-vraisemblance actuelle.

Le poids de l'ITD optimal dépendrait alors de la quantité d'information interaurale délivrée par la métrique spectrale. Pour tester cette hypothèse, nous avons mis en parallèle le coefficient de corrélation entre les cartes de similarité spectrale et les cartes de similarité d'ITD obtenus en section 6.4.5.3 pour chaque métrique spectrale, avec la différence de log-vraisemblance observée entre les cas $w_{ITD} = 0$ et $w_{ITD} = 0.5$. La dépendance entre ces deux quantités est quantifiée en terme du coefficient de corrélation et est présenté

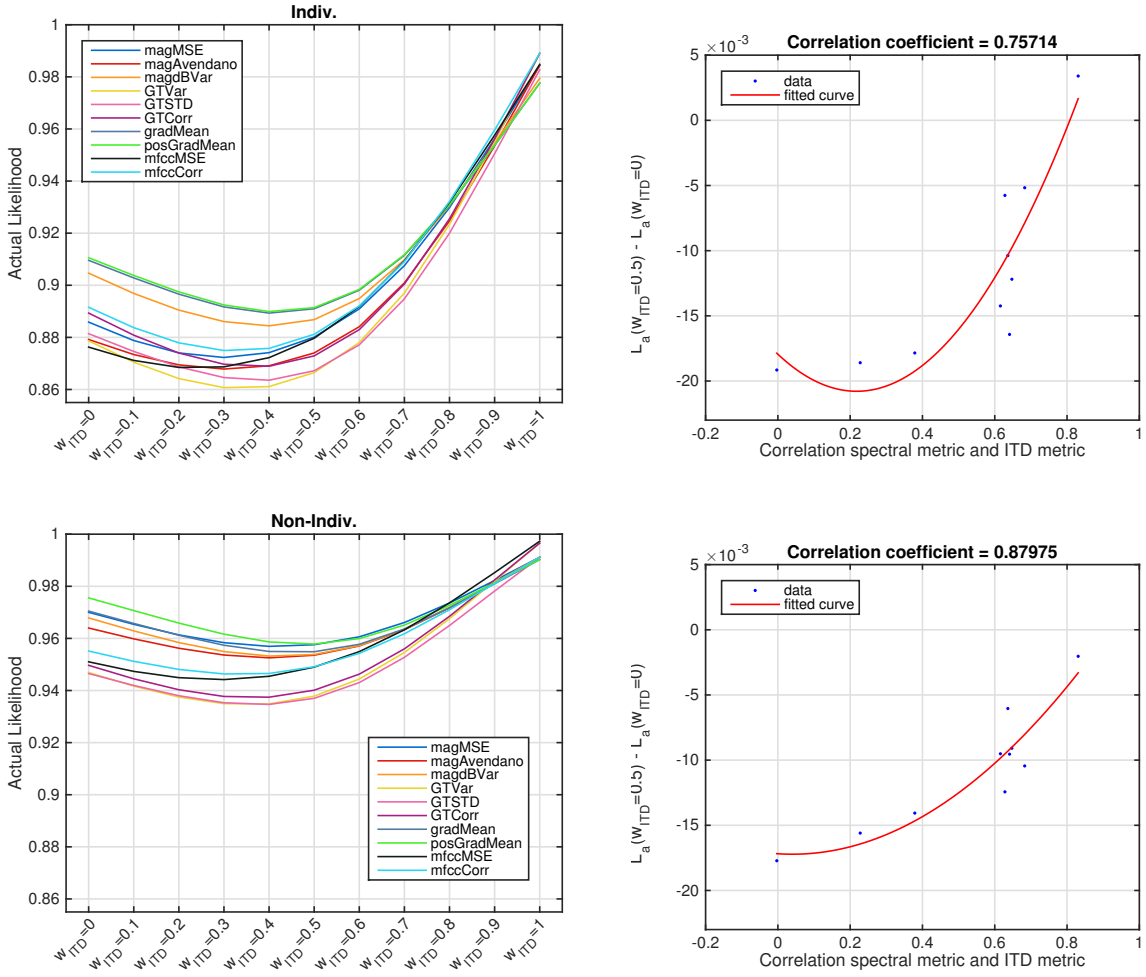


FIGURE 7.16 – (a et c) Evolution de la log-vraisemblance actuelle moyennée sur les 12 sujets en fonction du poids de l'ITD pour chacune des métriques spectrales (traits de couleur). (b et d) Corrélation entre la différence de log-vraisemblance actuelle qui distingue les cas $w_{ITD} = 0.5$ et $w_{ITD} = 0$ et les coefficients de corrélation entre cartes de similarités spectrale et d'ITD. Les résultats sont basés sur les données en condition d'écoute individuelle (a et b) et non-individuelle (c et d).

dans les figures 7.16(b) et 7.16(d). On observe que ces deux quantités sont bien corrélées à hauteur de 0.76 et 0.88 : plus la métrique spectrale encode l'information interaurale, moins la prise en compte de l'ITD améliore la prédiction. Notre hypothèse est donc vérifiée.

Pour conclure, cette étude a mis en évidence que l'importance de la prise en compte de l'indice d'ITD dans la prédiction dépend de la métrique spectrale utilisée. Elle est d'autant plus importante que l'information interaurale encodée intrinsèquement par la métrique spectrale est pauvre. Cette caractéristique liée à chaque métrique spectrale peut être quantifiée au moyen d'un calcul de corrélation entre les cartes de similarité spectrale et les cartes de similarité d'ITD. De manière générale, les résultats de log-vraisemblance sont optimisés pour des poids compris entre $w_{ITD} = 0.2$ et 0.5, et sont non significatifs entre eux. Nous continuons notre analyse avec le paramètre $w_{ITD} = 0.5$, tel que fixé *a priori*.

Métrique	condition d'écoute	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
mag-MSE	individuelle	*			+			*	*	*	*	*
	non-individuelle	*			+	+	+			*	*	*
mag-Avendano	individuelle	×			+			*	*	*	*	*
	non-individuelle	×				+			×	*	*	*
magdB-Var	individuelle	*	×			+			*	*	*	*
	non-individuelle	*				+			×	*	*	*
GT-Var	individuelle	*		+	+	+		*	*	*	*	*
	non-individuelle	*			+	+		*	*	*	*	*
GT-STD	individuelle	*				+		*	*	*	*	*
	non-individuelle	*				+			*	*	*	*
GT-Corr	individuelle	*	×			+		*	*	*	*	*
	non-individuelle	*			+	+			*	*	*	*
grad-Mean	individuelle	*	*			+			*	*	*	*
	non-individuelle	*	×			+	+			*	*	*
posGrad-Mean	individuelle	*	*			+			*	*	*	*
	non-individuelle	*	*			+	+			*	*	*
mfcc-MSE	individuelle	+	+	+	+	+		*	*	*	*	*
	non-individuelle				+			×	*	*	*	*
mfcc-Corr	individuelle	*		+	+	+		*	*	*	*	*
	non-individuelle			+	+	+	+	*	*	*	*	*

TABLE 7.3 – Tableau de significativité statistique entre la log-vraisemblance L_a associée aux poids w_{ITD} optimaux (pour lesquels L_a est minimisée) et les autres poids testés $\in [0, 1]$. Les signes "+" dans les cases vertes indiquent les poids de l'ITD qui optimisent la prédiction pour chaque métrique et condition d'écoute. Les astérisques dans les cases rouges indiquent par rapport à quels cas les résultats associés aux poids optimaux représentés en vert sont significativement différents. Les cases blanches et cases jaunes indiquent les cas non significatifs et marginalement significatifs respectivement, par rapport aux résultats optimaux.

7.3.3 Etude de la dispersion

Nous analysons ici l'impact du modèle de dispersion sur les résultats de prédiction. Jusqu'ici nous n'avons étudié que le cas où il n'était pas appliqué. Comme présenté en section 6.4.7.2, la dispersion de pointage a été estimée à partir des données du test sur haut-parleurs. La concentration $\kappa_{[R]}$ ainsi obtenue est en moyenne égale à $\bar{\kappa}_{[R]} = 144$ mais varie dans l'espace, avec une concentration plus élevée à l'avant qu'à l'arrière. Nous nous intéressons ici de savoir si le fait de considérer une concentration variable dans l'espace a un avantage par rapport à une concentration fixe, c'est-à-dire pour $\kappa = 144$ sur tout l'espace. De plus, nous tentons de diminuer progressivement le paramètre de concentration κ, i.e. d'augmenter la dispersion, pour en évaluer l'impact les résultats de log-vraisemblance actuelle. Une analyse ANOVA à mesures répétées est réalisée en parallèle pour évaluer l'effet de ce facteur et de son interaction avec le facteur métrique spectrale, pour chaque condition d'écoute. Les paramètres de cette analyse sont visibles dans le tableau 7.4.

Correction du biais	correction du biais en azimut
Condition d'écoute	individuelle et non-individuelle, séparément
Espace de prédiction	intervalle latéral de largeur 40°
Fonction sigmoïde	optimisée selon la condition d'écoute et la métrique spectrale
Dispersion	sans, $\kappa_{[R]}$, $\kappa = 144, 60, 20, 10, 5, 0.001$
w_{ITD}	0.5
Métrique spectrale	les 10 du tableau 6.1
Pondération binaurale	méthode de Baumgartner
Evaluation du modèle	log-vraisemblance actuelle, L_a

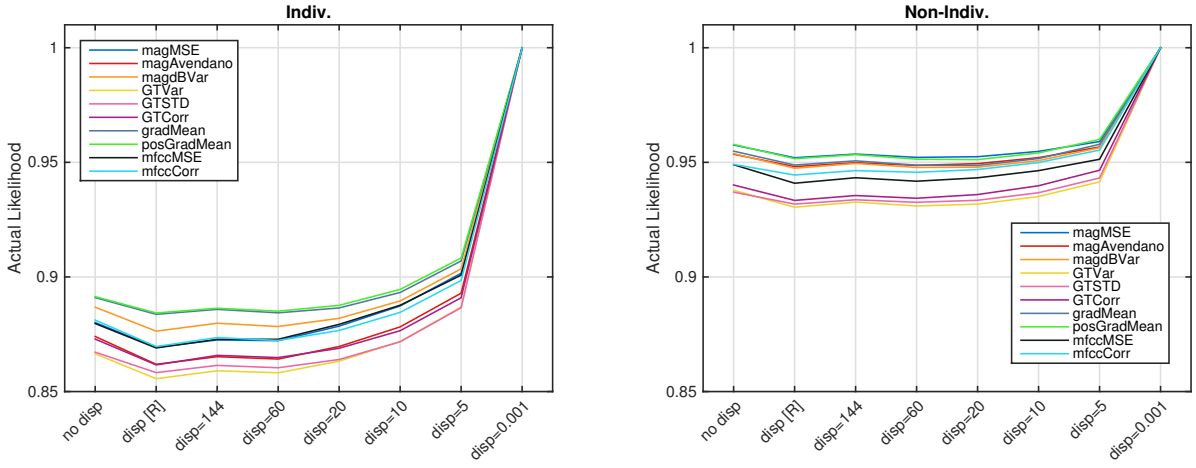
TABLE 7.4 – Paramètres associés à l'analyse sur la dispersion.

Tout d'abord, l'effet de la dispersion est significatif sur les résultats de prédiction, que ce soit en condition individuelle ou non-individuelle ($F(7, 77) = 145.39$; $p < 0.001$ et $F(7, 77) = 49.62$; $p < 0.001$, respectivement). L'interaction entre les facteurs dispersion et métrique spectrale est également significative ($F(63, 693) = 4.95$; $p < 0.001$ en condition individuelle et $F(63, 693) = 6.33$; $p < 0.001$ en condition non-individuelle). La figure 7.17 présente l'évolution des valeurs de log-vraisemblance actuelle pour chaque métrique spectrale en fonction de la valeur de concentration κ considérée dans le modèle de dispersion. L'abscisse correspond à des valeurs de κ ordonnées globalement de manière décroissante (i.e. pour une dispersion croissante). On observe que la log-vraisemblance est minimisée lorsque la dispersion est appliquée avec $\kappa_{[R]}$, soit une concentration estimée à partir des données expérimentales du test sur haut-parleurs et variable dans l'espace. Les résultats varient peu du cas sans dispersion à $\kappa = 10$ puis se dégradent de manière marquée à partir de $\kappa < 5$. Une partie des résultats relatifs aux tests post-hoc associés à l'interaction entre les facteurs dispersion et métrique pour chaque condition d'écoute sont présentés dans le tableau 7.5. Le tableau répertorie la significativité statistique des différences entre les résultats associés au cas optimal (indiqué par les cases vertes) et les autres cas de dispersion testés.

L'application du modèle de dispersion n'améliore pas toujours les résultats de manière significative. En condition individuelle, lorsque le modèle est appliqué avec $\kappa_{[R]}$, les résultats de prédiction sont significativement meilleurs du cas sans dispersion pour la plupart des métriques excepté *GT-STD*, *grad-Mean* et *posGrad-Mean*. Cependant, lorsque $\kappa = 144$, ne diffèrent pas significativement du cas sans dispersion. Cela signifie que le modèle de dispersion présente un avantage uniquement lorsque la concentration est estimée à partir de données réelles et variable dans l'espace. En condition non-individuelle, l'application du modèle de dispersion avec des concentrations supérieures à 5 améliore les résultats de manière marginalement significative pour les métriques *GT-Corr* et *GT-Var*, significative pour la métrique *mfcc-MSE* mais non significative pour les autres métriques. Il semblerait que le modèle de dispersion ait davantage d'impact sur les résultats de prédiction en condition individuelle.

On observe dans le tableau que les cas $\kappa_{[R]}$, $\kappa = 144$ et $\kappa = 60$ ne sont jamais significatifs entre eux, de même pour $\kappa = 20$ sauf dans la condition individuelle pour les métriques *mag-MSE* et *mfcc-MSE*. Lorsque le modèle de dispersion est appliqué avec une concentration $\kappa = 10$, les résultats se dégradent significativement dans la condition individuelle mais ne diffèrent pas significativement des concentrations $\kappa_{[R]}$, $\kappa = 144, 60$ et 20 dans la condition non-individuelle. Cette observation peut être mise en parallèle avec la dispersion des réponses, au sein des répétitions, calculée dans les conditions individuelle et non-individuelle. La dispersion moyenne κ^{-1} au sein des répétitions (pour tous les sujets) dans la condition individuelle est de 0.063 (soit $\kappa = 16$). Pour les HRTFs non-individuelles, la dispersion est plus élevée et de 0.096 en moyenne (soit $\kappa = 10.4$). Cela pourrait expliquer que lorsque le modèle de dispersion est appliqué avec une concentration de $\kappa = 10$, la log-vraisemblance ne soit pas altérée de manière significative en condition non-individuelle alors qu'elle se dégrade significativement en condition individuelle.

Pour $\kappa = 5$ et $\kappa = 0.001$, les résultats de prédiction sont significativement moins bons quel que soit la condition d'écoute. On note d'ailleurs dans la figure 7.17 que l'application d'un modèle de dispersion avec une concentration proche de zéro correspond à $L_a = 1$ soit une log-vraisemblance actuelle égale au seuil de chance. En effet, lorsque $\kappa = 0.001$ la distribution de probabilité devient uniforme sur l'espace.



(a) Condition individuelle

(b) Condition non-individuelle

FIGURE 7.17 – Log-vraisemblance actuelle moyennée sur les 12 sujets en fonction de la valeur du paramètre de concentration κ utilisé dans le modèle de dispersion. Les traits de couleur indiquent les résultats pour chaque métrique spectral et les figures (a) et (b) correspondent respectivement aux conditions individuelle et non-individuelle.

Métrique	condition d'écoute	sans disp.	$\kappa_{[R]}$	$\kappa = 144$	$\kappa = 60$	$\kappa = 20$	$\kappa = 10$	$\kappa = 5$	$\kappa = 0.001$
mag-MSE	individuelle	*	+			×	*	*	*
	non-individuelle		+		+			×	*
mag-Avendano	individuelle	*	+				*	*	*
	non-individuelle		+		+			*	*
magdB-Var	individuelle	*	+				*	*	*
	non-individuelle		+		+	+		*	*
GT-Var	individuelle	*	+				*	*	*
	non-individuelle	×	+		+			*	*
GT-STD	individuelle		+	+	+		*	*	*
	non-individuelle		+	+	+	+		*	*
GT-Corr	individuelle	*	+				*	*	*
	non-individuelle	×	+					*	*
grad-Mean	individuelle		+				×	*	*
	non-individuelle		+		+	+		*	*
posGrad-Mean	individuelle		+				*	*	*
	non-individuelle		+		+	+		*	*
mfcc-MSE	individuelle	*	+			*	*	*	*
	non-individuelle	*	+					*	*
mfcc-Corr	individuelle	*	+				*	*	*
	non-individuelle		+	+	+	+		*	*

TABLE 7.5 – Tableau de significativité statistique entre la log-vraisemblance actuelle associée aux cas où la valeur de concentration κ prise en compte dans modèle de dispersion optimise les résultats (cases vertes) et les autres cas de dispersion testés. Les codes couleurs sont identiques au tableau 7.3 (c.f. légende).

En conclusion, l'application du modèle de dispersion sur les cartes de similarité n'améliore significativement les résultats de prédiction qu'en condition individuelle et pour des concentrations estimées à partir de données expérimentales et variables dans l'espace ($\kappa_{[R]}$). De plus, cela ne concerne que certaines métriques : les métriques *GT-STD*, *grad-Mean* et *posGrad-Mean* ne voient pas leurs résultats s'améliorer significativement avec ce modèle de dispersion. En condition non-individuelle, le modèle de dispersion n'a pas d'impact significatif sauf pour la métrique *mfcc-MSE* pour laquelle le modèle basé sur $\kappa_{[R]}$ améliore significativement les résultats. Etant donné que le modèle de dispersion n'améliore globalement pas les résultats en condition non-individuelle, il ne sera pas appliqué pour la suite de l'analyse.

7.3.4 Etude de la méthode de pondération binaurale

Nous nous intéressons ici à l'effet de la méthode de pondération binaurale. En effet, nous n'avons jusqu'ici utilisé que la méthode de Baumgartner qui relève plutôt d'un modèle mais nous souhaitons la comparer à la méthode de Middlebrooks basée sur la mesure (voir section 6.4.6). Une ANOVA à mesures répétées est réalisée en parallèle avec les facteurs suivants : méthode de pondération binaurale et métrique spectrale, pour les conditions d'écoute individuelle et non-individuelle séparément. Les paramètres associés à l'analyse sont présentés dans le tableau 7.6.

Correction du biais	correction du biais en azimut
Condition d'écoute	individuelle et non-individuelle, séparément
Espace de prédiction	intervalle latéral de largeur 40°
Fonction sigmoïde	optimisée selon la condition d'écoute et la métrique spectrale
Dispersion	sans dispersion
w_{ITD}	0.5
Métrique spectrale	les 10 du tableau 6.1
Pondération binaurale	méthodes de Baumgartner et de Middlebrooks
Evaluation du modèle	log-vraisemblance actuelle, L_a

TABLE 7.6 – Paramètres associés à l'analyse sur la méthode de pondération binaurale.

En condition individuelle comme en condition non-individuelle, la méthode de pondération binaurale n'a pas d'effet significatif sur la log-vraisemblance actuelle ($F(1, 11) = 1.22$; $p = 0.29$ et $F(1, 11) = 0.74$; $p = 0.41$, respectivement) et son interaction avec le facteur métrique n'est pas significative ($F(9, 99) = 1.93$; $p = 0.06$ et $F(9, 99) = 0.63$; $p = 0.77$, respectivement). Pour conclure, la méthode de pondération binaurale n'est pas un paramètre déterminant les résultats du modèle de prédiction.

7.3.5 Etude de la métrique spectrale

Enfin, nous comparons le mérite des différentes métriques spectrales. Une ANOVA à mesures répétées est réalisée pour identifier les différences significatives entre les métriques au sein de chaque condition d'écoute. Les paramètres de l'analyse sont donnés dans le tableau 7.7.

Correction du biais	correction du biais en azimut
Condition d'écoute	individuelle et non-individuelle, séparément
Espace de prédiction	intervalle latéral de largeur 40°
Fonction sigmoïde	optimisée selon la condition d'écoute et la métrique spectrale
Dispersion	sans dispersion
w_{ITD}	0.5
Métrique spectrale	les 10 du tableau 6.1
Pondération binaurale	méthode de Baumgartner
Evaluation du modèle	log-vraisemblance actuelle, L_a

TABLE 7.7 – Paramètres associés à l'analyse des métriques spectrales.

Les métriques spectrales ont un impact significatif sur la prédiction ($F(9, 99) = 4.95$; $p < 0.001$ en condition individuelle et $F(9, 99) = 6.43$; $p < 0.001$ en condition non-individuelle). La figure 7.18 permet de visualiser les tendances. Un premier tableau, le tableau 7.8, illustre l'ensemble des résultats du test post-hoc, soit la significativité statistique entre chaque paire de métriques spectrales. Un second tableau, le tableau 7.9, offre une visualisation plus concise en présentant la significativité statistique des différences entre les métriques optimales et les autres métriques, pour chaque condition d'écoute, de la même façon qu'effectué précédemment.

D'après le tableau 7.8, les métriques basées sur les mêmes modes de représentation ne sont pas significatives entre elles. On a donc 4 groupes de métriques formés par les représentations suivantes : spectre d'amplitude (linéaire ou en dB), profils spectraux (en dB), gradient des profils spectraux (dont gradient positif) et MFCCs. Les groupes de métriques basées sur le spectre d'amplitude, les gradients et les MFCCs ne sont pas significatifs entre eux. Les seules différences significatives apparaissent pour les métriques *GT-Var* et *GT-STD* qui offrent de meilleurs résultats par rapport aux métriques basées sur le spectre d'amplitude et le gradient.

D'après le tableau 7.9, en condition individuelle, les métriques *GT-Var* et *GT-STD* se démarquent de manière significative des métriques *grad-Mean* et *posGrad-Mean* et marginalement significative de la métrique *magdB-Var*. En condition non-individuelle, les résultats associés aux métriques *GT-Var* et *GT-STD* sont significativement meilleurs que les métriques *posGrad-Mean* et *mag-MSE* et marginalement significatifs des résultats associés aux métriques *mag-Avendano*, *magdB-Var* et *grad-Mean*.

Pour conclure, on peut dire que les métriques *GT-STD* et *GT-Var* offrent toutes deux des résultats significativement meilleurs que les métriques basées sur la représentation par le gradient ou le spectre d'amplitude. Plus largement, les probabilités associées aux directions pointées sont maximisées dans le cas des métriques basées sur les profils spectraux et les MFCCs.

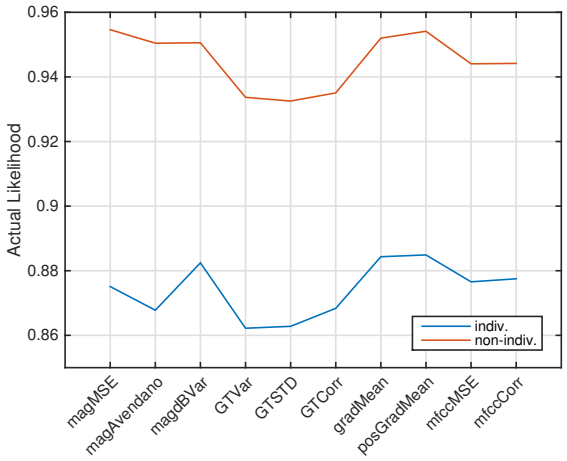


FIGURE 7.18 – Log-vraisemblance moyennée sur les sujets en fonction des métriques spectrales et pour les conditions individuelle et non-individuelle.

Métrique	condition d'écoute	*mag-MSE*	*mag-Avendano*	*magdB-Var*	*GT-Var*	*GT-STD*	*GT-Corr*	*grad-Mean*	*posGrad-Mean*	*mfcc-MSE*	*mfcc-Corr*
mag-MSE	individuelle	\									
	non-individuelle	\			$p = 0.001$	$p < 0.001$	$p = 0.006$				
mag-Avendano	individuelle		\								
	non-individuelle		\		$p = 0.02$	$p = 0.01$					
magdB-Var	individuelle			\	$p = 0.02$	$p = 0.03$					
	non-individuelle			\	$p = 0.02$	$p = 0.01$					
GT-Var	individuelle			$p = 0.02$	\			$p = 0.002$	$p = 0.001$		
	non-individuelle	$p = 0.001$	$p = 0.02$	$p = 0.02$	\			$p = 0.009$	$p < 0.001$		
GT-STD	individuelle			$p = 0.03$		\		$p = 0.003$	$p = 0.002$		
	non-individuelle	$p < 0.001$	$p = 0.01$	$p = 0.01$		\		$p = 0.005$	$p < 0.001$		
GT-Corr	individuelle						\				
	non-individuelle	$p = 0.006$					\	$p = 0.04$	$p = 0.005$		
grad-Mean	individuelle				$p = 0.002$	$p = 0.003$		\			
	non-individuelle				$p = 0.009$	$p = 0.005$	$p = 0.04$	\			
posGrad-Mean	individuelle				$p = 0.001$	$p = 0.002$			\		
	non-individuelle				$p < 0.001$	$p < 0.001$	$p = 0.005$		\		
mfcc-MSE	individuelle									\	
	non-individuelle									\	
mfcc-MSE	individuelle										\
	non-individuelle										\

TABLE 7.8 – Résultats des tests post-hoc associés au facteur métrique spectrale, pour chaque condition d'écoute. Les cases rouges indiquent les différences significatives et les cases jaunes, les différences marginalement significatives. Les cases blanches signifient que les couples de métriques n'offrent pas de résultats significativement différents.

condition d'écoute	*mag-MSE*	*mag-Avendano*	*magdB-Var*	*GT-Var*	*GT-STD*	*GT-Corr*	*grad-Mean*	*posGrad-Mean*	*mfcc-MSE*	*mfcc-Corr*
individuelle			×	+	+		*	*		
non-individuelle	*	×	×	+	+		×	*		

TABLE 7.9 – Tableau présentant les différences significatives en termes de log-vraisemblance actuelle vis-à-vis des métriques optimales *GT-Var* et *GT-STD*, pour chaque condition d'écoute.

7.4 Discussion

Le test de localisation a permis jusqu'ici de collecter des données pour optimiser le modèle de prédiction de la localisation avec des HRTFs individuelles et non-individuelles. Dans la suite de ce travail, ces données expérimentales permettront la réalisation d'une étape qui simule la mise à l'épreuve du modèle dans un cas réel pour proposer automatiquement un jeu de HRTFs adapté à l'auditeur, sur la base de ses réponses à un test de localisation en binaural dans lequel il doit désigner la direction perçue de signaux filtrés avec des HRTFs tirées d'une base de données.

Après avoir discuté des observations réalisées dans la section précédente, les résultats de prédiction sont comparés aux études de la littérature ayant développé des modèles de prédiction de la localisation.

7.4.1 Paramètres optimaux

Optimisation des paramètres de la fonction sigmoïde La première analyse s'est intéressée au couple de paramètres (Γ,S) associés à la fonction sigmoïde qui transforme des distances normalisées en indices de similarité. L'optimisation des paramètres (Γ,S) a pour but d'adapter la sélectivité spatiale des distributions de probabilité associées à chaque couple métrique spectrale-condition d'écoute. L'analyse a mis en évidence que, bien qu'une adaptation supplémentaire de la fonction sigmoïde à chaque individu améliore la log-vraisemblance actuelle, la tendance n'est pas significative. Une adaptation spécifique à la condition d'écoute plutôt qu'au sujet s'est avérée davantage nécessaire.

Des disparités en termes de méthodologie apparaissent clairement sur ce point entre le modèle de Baumgartner et al. et le présent modèle, et sont directement liés à la finalité de chacune des études. Pour rappel, la méthode de Baumgartner et al. consiste à individualiser le paramètre S de la fonction sigmoïde au sein d'une optimisation des résultats de prédiction en condition individuelle, et à utiliser le modèle ainsi calibré pour la prédiction des performances de localisation dans d'autres conditions d'écoute. La calibration du modèle au sein de chaque condition y est rejetée dans ce cas, puisqu'il s'agit de palier la mise en place de tests pour ces autres conditions d'écoute. La méthode d'évaluation adoptée par Baumgartner et al. consiste à estimer la capacité du modèle à prédire les directions pointées ou les performances de localisation. Le pourcentage de réussite est un critère adapté au type d'évaluation réalisé par ces auteurs. Celui-ci a été analysé dans notre étude et les observations effectuées suggèrent que les données prédites sont significativement améliorées lorsque le modèle est calibré de manière individuelle, i.e. pour chaque individu. Dans le cadre du présent modèle, une individualisation systématique des paramètres de la fonction sigmoïde à chaque auditeur n'est pas envisageable étant donné qu'il vise à être étendu à des auditeurs dont les HRTFs individuelles ne sont pas disponibles. Cependant, la calibration du modèle au sein de chaque condition d'écoute peut être réalisée en amont de l'application du modèle pour un nouvel auditeur. Nous avons utilisé le critère de la log-vraisemblance actuelle pour calibrer notre modèle puisque la procédure d'optimisation consiste à identifier les paramètres avec lesquels la probabilité aux directions pointées est maximisée. Pour ce critère d'évaluation, nous avons observé que les résultats n'étaient pas significativement améliorés avec une individualisation de la fonction sigmoïde. Pour terminer, bien que la fonction sigmoïde ait également été utilisée par Baumgartner et al., les paramètres qui la définissent sont difficilement comparables étant donné que les valeurs de distance ne sont pas du même ordre de grandeur.

Les conditions d'écoute étudiées ici concernent la localisation de stimuli filtrés par des HRTFs individuelles ou non-individuelles. Il serait intéressant d'étudier si ces mêmes observations sont valables pour d'autres types de stimuli comme des stimuli à bande étroite ou filtrés par des HRTFs ayant préalablement subi des modifications spectrales.

Importance de l'indice d'ITD pour la prédiction de localisation Le présent modèle est appliqué dans le cadre de la prédiction de la localisation sur un espace non réduit à un plan sagittal. La seconde analyse a permis d'évaluer l'importance de la prise en compte à la fois des indices interauraux et spectraux. L'amélioration des résultats observée avec la prise en compte de l'indice interaural dépend de la métrique spectrale. Par exemple, dans le cas particulier des métriques basées sur une représentation des HRTFs par les MFCCs, les résultats de prédiction ne sont pas significativement améliorés. L'analyse a mis en évidence un lien avec la quantité d'information interaurale portée par la métrique spectrale. Cette caractéristique est associée à la sensibilité de la métrique envers les différences de gain global. Cette observation est en accord avec le chapitre précédent qui montre que les premiers coefficients MFCCs encodent l'information de gain global. Enfin, nous avons identifié qu'une prédiction basée sur des cartes de probabilité issues de la combinaison pondérée des indices d'ITD et spectraux était optimisée pour l'ensemble des métriques spectrales testées, moyennant un poids de l'indice d'ITD à hauteur de 20% à 50% du poids de l'indice spectral.

Il est important de noter que ces poids optimaux varient selon l'espace de prédiction considéré. Dans le cadre d'une prédiction dans le plan sagittal uniquement, telle que celle réalisée par Baumgartner et al. ou Langendijk et Bronkhorst, il est justifié de ne pas prendre en compte l'indice interaural étant donné qu'il ne contribue pas (ou très peu) à la localisation en élévation et à la discrimination avant-arrière (en synthèse statique). Il se peut qu'une métrique spectrale 100% porteuse d'information spectrale, i.e. non corrélée à

l'indice interaural, soit mieux adaptée à ce type de prédiction comme par exemple la métrique basée sur les gradients positifs, tel qu'utilisé dans l'étude de Baumgartner et al. Dans le cadre d'une prédiction sur toute la sphère, il est vraisemblable que l'indice d'ITD soit davantage considéré dans la combinaison pondérée des informations interaurales et spectrales.

Modélisation de la dispersion La troisième analyse s'est intéressée à l'apport de la modélisation de la dispersion de pointage. Les observations montrent qu'une amélioration significative des résultats peut être obtenue en condition individuelle si le paramètre de concentration κ est déterminé de manière spécifique à la méthode de report utilisée dans le test expérimental et défini de manière variable dans l'espace (soit en utilisant le paramètre $\kappa_{[R]}$). L'amélioration globale des résultats de prédiction obtenue avec la prise en compte de la dispersion de pointage est en accord avec les résultats de Baumgartner et al. [BML14]. Cependant, le paramètre de dispersion ϵ utilisé n'avait pas été défini de manière variable dans l'espace et aucune analyse statistique n'avait été menée dans cette étude. Nos observations indiquent que dans ce cas (soit pour $\kappa = 144$ constant sur tout l'espace), le modèle de dispersion n'améliore pas significativement les résultats de prédiction. Par ailleurs, les analyses statistiques révèlent que l'amélioration des résultats avec l'application du modèle de dispersion est bien souvent non significative. Les résultats de prédiction associés aux métriques spectrales *GT-STD*, *grad-Mean* et *posGrad-Mean* ne sont pas significativement améliorés avec la prise en compte de la dispersion, quelle que soit la condition d'écoute. Pour les autres métriques, le modèle de dispersion n'améliore les résultats de manière significative qu'en condition individuelle. L'avantage du modèle de dispersion peut être masqué par l'optimisation de la fonction sigmoïde. En effet, celle-ci a pour but d'adapter la sélectivité spatiale de la distribution de probabilité de sorte à minimiser la log-vraisemblance aux réponses. Il se peut que la dispersion liée aux réponses soit d'ores et déjà considérée à ce stade. Les fonctions sigmoïdes optimales en condition individuelle sont d'ailleurs plus sélectives que les fonctions sigmoïdes en condition non-individuelle, ce qui est cohérent avec une dispersion des réponses plus faible en condition individuelle que non-individuelle.

Le paramètre de dispersion ϵ appliqué par Baumgartner et al. a été déterminé suivant une optimisation des résultats de prédiction en condition individuelle et s'apparente à une valeur de déviation standard des réponses en élévation. En effet, ces auteurs étudient la prédiction des performances de localisation dans une seule dimension, la dimension polaire. Dans notre étude, la prédiction est effectuée en 3 dimensions. Le modèle de dispersion est basé sur les distributions de Von Mises Fisher pour une sphère de dimension 2 (c.f. section 6.4.7.2). Le paramètre de concentration κ utilisé ne peut être comparé directement au paramètre ϵ de l'étude de Baumgarter et al. (exprimé en degrés). Cependant, nous pouvons estimer la déviation standard en élévation sur nos données de localisation. Celle-ci est en moyenne de 9° (sujets et cibles confondues, en condition individuelle) contre $\epsilon = 17°$ pour Baumgartner et al. La même différence est observée avec les résultats de Langendijk et Bronkhorst [LB02] pour qui la déviation standard moyenne en élévation est de 17°, le nombre de répétitions étant pourtant identique (5 répétitions). Ces disparités peuvent s'expliquer par des différences dans la procédure expérimentale, notamment en termes de méthode de report. Dans [LB02], les participants indiquent la direction perçue à l'aide d'un pointeur virtuel acoustique émettant un son harmonique complexe filtré par les HRTFs individuelles de l'auditeur. Baumgartner et al. se basent sur des données expérimentales de tests de localisation utilisant la méthode de pointage au pistolet, qui implique les mouvements du corps entier. D'après les résultats de l'étude comparative sur les méthodes de pointage (chapitre 4), bien que les 3 méthodes étudiées ne se distinguent pas significativement en termes de dispersion des réponses, la méthode de pointage au pistolet présente une dispersion significativement plus élevée à l'arrière, ce qui n'est pas le cas de la méthode proximale. Nous n'avons pas évalué la méthode de pointage acoustique utilisée par Langendijk et Bronkhorst. Cependant, étant donné que la localisation de stimuli à bande fréquentielle étroite filtrés par les HRTFs individuelles de l'auditeur entraîne des erreurs en élévation [Mid92], on peut se poser la question de la fiabilité de cette méthode de pointage. Les différences de dispersion entre la présente étude et l'étude de Langendijk et Bronkhorst ne peuvent s'expliquer par le niveau d'expertise des participants ou l'égalisation du casque audio car ils sont comparables d'une étude à l'autre.

Pondération binaurale La quatrième analyse, très courte, a montré que la méthode de pondération binaurale n'a pas d'effet significatif sur les résultats de prédiction. En parallèle à cette observation, Baumgartner et al. avaient étudié comparativement l'influence de la prise en compte de l'information spectrale ipsilatérale seulement ($w_{ipsi} = 1$, $w_{contro} = 0$), contralatérale seulement ($w_{ipsi} = 0$, $w_{contro} = 1$) ou combinée (poids déterminés selon la position latérale de la cible). Leurs résultats montrent que l'information spatiale aux oreilles ispilatérale et controlatérale sont semblables et la seule dégradation observée concerne la condition où seule l'information de l'oreille controlatérale est considérée. Dans notre cas, deux méthodes d'obtention des poids associés à la combinaison des informations spectrales ispsilatérale et contralatérale ont été comparées. L'observation selon laquelle les différences sont très faibles est cohérente avec les remarques de la précédente étude.

Métriques spectrale Le principal paramètre d'intérêt dans cette étude concerne la métrique spectrale. Nous avons noté un avantage des métriques basées sur le spectre d'amplitude moyenné par bandes fréquentielles suivant l'application d'un banc de filtres auditifs (profils spectraux). Ce résultat confirme la nécessité d'utiliser une représentation des HRTFs proche de la modélisation du système auditif. Une représentation logarithmique du spectre en amplitude ne suffit pas, la prise en compte de la résolution fréquentielle apparaît nécessaire pour prédire les directions perçues (*magdB-Var* et *GT-Var* marginalement significatives). L'étude a par ailleurs mis en évidence que le choix du mode de représentation était plus important que le critère de distance. En effet, les métriques basées sur le même mode de représentation n'ont pas montré de différences signficatives. Dans le cas des profils spectraux par exemple, bien que l'inter-corrélation offre des résultats moins bons que la variance ou la déviation standard, le résultat n'est pas significatif. On remarquera que l'avantage de la déviation standard par rapport à l'inter-corrélation est en accord avec Langendijk et Bronkhorst [LB02]. Par définition, les métriques *GT-Var* et *GT-STD* sont identiques à une racine carrée près. Elles se distinguent uniquement par leur caractère de sélectivité spatiale (comme présenté section 6.4.3). Par conséquent, suite à l'adaptation de la fonction sigmoïde à chacune d'entre elles, il est vraisemblable qu'elles offrent des résultats de prédiction similaires. Les résultats obtenus avec ces deux métriques sont significativement meilleurs par comparaison aux métriques basées sur le gradient ou le spectre d'amplitude. La faiblesse des métriques basées sur le gradient peut venir du fait que nous travaillons avec des stimuli large bande à spectre plat. En effet, Baumgartner et al. [BML14] ont montré que la métrique *posGrad-Mean* était mieux adaptée à des stimuli à bande fréquentielle étroite ou à spectre chahuté et que la métrique *GT-STD* présentait de meilleurs résultats dans le cas de stimuli large bande à spectre plat. De plus, la dégradation des résultats avec la prise en compte du gradient est en accord avec Langendijk et Bronkhorst [LB02]. Cependant, elle va à l'encontre de l'hypothèse de Zakarauskas et Cynader [ZC93] selon laquelle les dérivées spectrales offriraient une meilleure représentation de l'information spectrale pertinente. Enfin, la représentation des HRTFs par les MFCCs, non significative des métriques spectrales optimales basées sur les profils spectraux, s'est avérée également bien adaptée. Ce résultat est en accord avec l'étude de Lee et Lee [LpL11].

7.4.2 Résultats de prédiction

Afin d'évaluer le mérite de notre modèle, nous effectuons ici quelques comparaisons avec les résultats de la littérature.

7.4.2.1 Comparaison avec Middlebrooks (1992)

La méthode d'estimation du score de prédiction adoptée par Middlebrooks [Mid92], puis reprise par Bremen et al. [BvWvO10], consiste à comparer l'histogramme des valeurs des indices de similarité (SI) à l'endroit des réponses par rapport aux valeurs prises sur tout l'espace de prédiction. La normalisation de l'histogramme par la somme permet de s'abstraire du nombre de points utilisé pour construire l'histogramme et de rendre possible la comparaison entre l'histogramme des SI sur l'espace de prédiction par rapport à l'histogramme des SI à l'endroit des réponses (basés sur un nombre de points différent). Il faut s'assurer que l'échantillonnage sphérique à partir duquel on construit l'histogramme est régulier pour ne pas risquer de multiplier les valeurs de SI localisés dans une région spatiale plus dense.

De la même façon que Middlebrooks, nous comparons ici l'histogrammes des SI interpolés aux réponses par rapport l'histogramme des SI interpolés sur la grille complète d'échantillonnage régulière définie par 1488 points (voir section 6.4.2). Avant le calcul de l'histogramme, les indices de similarité issus de l'équation 6.10 sont divisés par le maximum pour obtenir une distribution entre 0 et 1. L'histogramme est ensuite calculé sur un vecteur de valeurs espacées de 0.01.

Etant donné qu'il existe une carte d'indices de similarité unique pour chaque cible et chaque participant, la moyenne des $U = 108$ histogrammes normalisés associés aux cartes uniques est calculée ainsi :

$$\frac{1}{U}\sum_{u=1}^{U}\frac{\text{hist}(\text{SI}_{j=1\ldots N_d})}{N_d} \tag{7.1}$$

avec $N_d = 1488$ le nombre de points de la grille échantillonnage sphérique et j l'indice de la direction. De plus, l'histogramme des indices de similarité interpolés à l'endroit des réponses et normalisé par le nombre de valeurs est donné par :

$$\frac{\text{hist}(\text{SI}_{r=1\ldots N_R})}{N_R} \tag{7.2}$$

où $N_R = 5 \times 108 = 540$ et r l'indice de la direction associée à chaque réponse. La fonction de répartition (*Cumulative Distribution Function*) est ensuite calculée pour l'histogramme normalisé des SI aux réponses et pour l'histogramme moyen normalisé des SI sur l'espace. La courbe ROC (*Receiver Operating Characteristic Curve*) est déterminée à partir des fonctions de répartition. L'aire sous la courbe ROC (*Area Under Curve*, AUC) définit alors le score de prédiction en termes du pourcentage correct ou du paramètre d' :

$$\text{percent correct} = 100 \cdot AUC \tag{7.3}$$

$$d' = \sqrt{2} \cdot \text{norminv}(AUC) \tag{7.4}$$

Ces deux critères permettent de quantifier la séparation entre les histogrammes et relèvent de la théorie de détection du signal. L'ensemble des ces étapes peut être visualisé grâce à l'exemple de la figure 7.19.

Notons qu'étant donné que les SI aux réponses sont interpolés à partir de la distribution sur l'espace de prédiction, la séparation parfaite des deux distributions n'est pas possible : le pourcentage correct ne peut donc atteindre 100%.

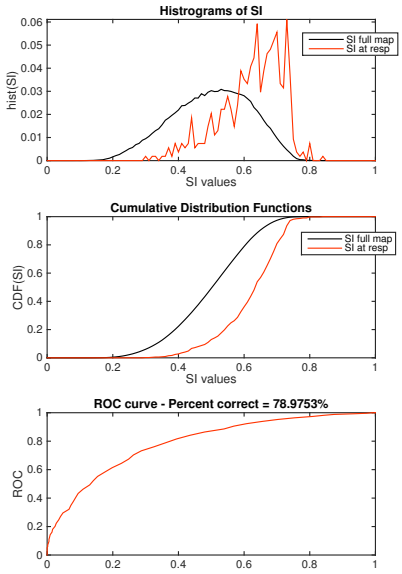


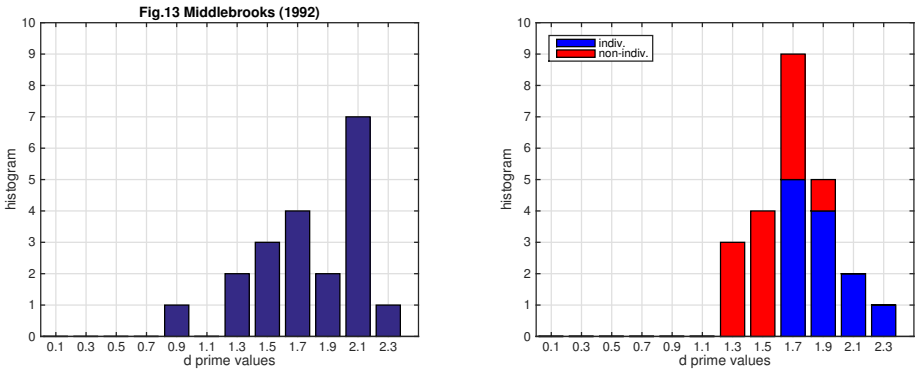
FIGURE 7.19 – Exemple d'analyse du score de prédiction en termes de pourcentage correct. Dans l'ordre de haut en bas : histogramme normalisé moyen des SI sur toute la sphère (courbe noire, c.f. équation 7.1) et histogramme des SI à l'endroit des réponses (courbe rouge, c.f. équation 7.2) ; courbes de répartition associée à chaque histogramme ; courbe ROC. Le pourcentage correct obtenu est indiqué dans le titre de la dernière figure et atteint ici 79%.

De façon similaire à la figure 13 de l'article de Middlebrooks présentée en figure 7.20(a), la figure 7.20(b) présente l'histogramme des valeurs du paramètre d' obtenu avec notre modèle, dans les conditions similaires au modèle de Middlebrooks, à savoir : une prédiction sur toute la sphère, des poids associés aux indices interauraux et spectraux égaux (soit $w_{ITD} = 0.5$), une méthode de pondération binaurale identique et une métrique spectrale basée sur l'inter-corrélation des profils spectraux (*GT-Corr*). L'histogramme de nos résultats présente 2×12 valeurs (12 sujets, 2 conditions d'écoute) soit 24 valeurs, contre 20 pour Middlebrooks (5 sujets, 4 conditions d'écoute). Les valeurs de d' obtenues avec notre modèle sont comprises entre 1.3 et 2.3 contre 0.8 et 2.3 pour Middlebrooks. Les résultats sont donc comparables. Par ailleurs, on note une distinction nette entre les résultats associés à la condition d'écoute individuelle (en bleu sur l'histogramme), en comparaison de la condition d'écoute non-individuelle (en rouge). Comme indiqué précédemment, les résultats de prédiction sont meilleurs en condition individuelle.

Concernant le pourcentage correct, Middlebrooks obtient des valeurs comprises entre 75%et 95% avec une médiane de 90%. En comparaison, nous obtenons des valeurs de pourcentage correct comprises entre 81% et 95% avec une médiane de 89% (résultats associés aux condition individuelle et non-individuelle confondus). Nos résultats sont donc comparables à ceux de l'étude de Middlebrooks associée à des sources virtuelles large bande ou à bande étroite, synthétisées avec les HRTFs individuelles de l'auditeur.

7.4.2.2 Comparaison avec Langendijk et Bronkhorst (2002)

L'étude de Langendijk et Bronkhorst [LB02] concerne la mise au point d'un modèle de prédiction de la localisation de stimuli large bande filtrés avec les HRTFs individuelles de l'auditeur brutes ou modifiées (dont les indices spectraux de certaines bandes fréquentielles ont été supprimés). La métrique spectrale utilisée dans leur modèle est la déviation standard des spectres d'amplitude moyennés par bande fréquentielle (soit *GT-STD*) à l'aide d'un banc de filtres de largeur de bande égale à $\frac{1}{6}$ d'octave, proches des largeurs de bande ERB. Contrairement à notre étude, la fonction qui transforme les distances en indices de similarité correspond à une fonction gaussienne, dont l'écart-type optimal est déterminé de sorte à minimiser la log-vraisemblance actuelle de manière globale, i.e. pour tous les sujets (pas d'individualisation de l'écart-type). Afin d'évaluer la validité du modèle, les auteurs comparent la log-vraisemblance actuelle L_a à la log-vraisemblance escomptée L_e. La méthode de calcul utilisée dans notre étude est basée sur celle définie dans ce papier mais considère une étape de division des valeurs de log-vraisemblance L_a et L_e par la log-vraisemblance L_{chance} associée à une probabilité constante sur l'espace de prédiction. Cela permet



(a) Résultats de Middlebrooks (1992)

(b) Résultats de notre étude

FIGURE 7.20 – (a) Histogramme des 20 valeurs de d' obtenus par Middlebrooks (c.f. figure 13 de l'article). (b) Histogramme des 24 valeurs de d' obtenues avec notre modèle pour les conditions individuelle (en bleu) et non-individuelle (en rouge) sous les même conditions que le modèle de Middlebrooks, à savoir : une prédiction sur toute la sphère avec $w_{ITD} = 0.5$, la métrique spectrale *GTCorr*, et sans appliquer le modèle de dispersion.

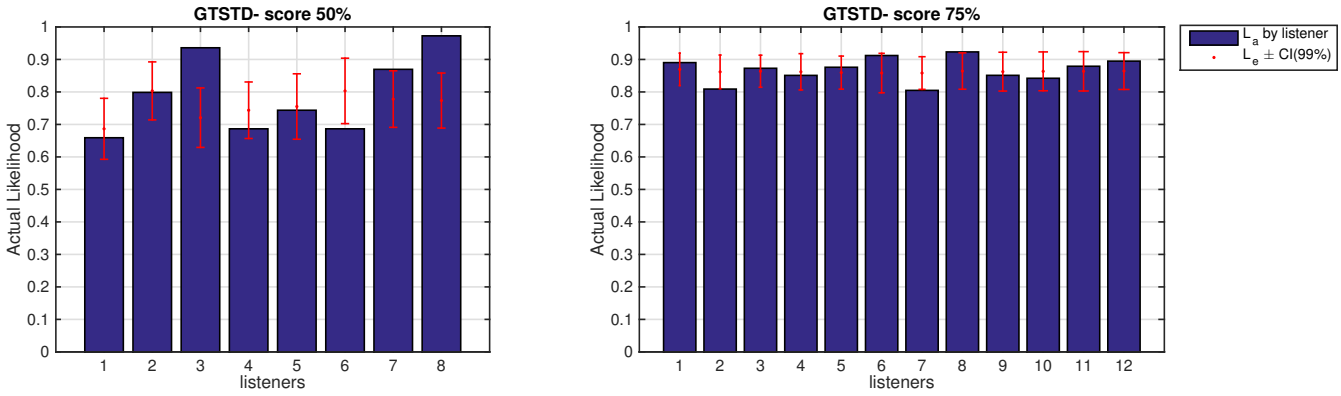
d'obtenir des valeurs qui ne tiennent pas compte du nombre de réponses considérées. Afin de rendre la comparaison avec les résultats de Langendijk et Bronkhorst possible, nous avons normalisé leurs valeurs de log-vraisemblance par la log-vraisemblance associée à une probabilité constante sur le plan médian (composé de 53 points) et un total de $11 \times 5 = 55$ réponses, soit $L_{chance} = -2 \cdot 55 \cdot \ln(\frac{1}{53}) = 437$, suivant l'équation 6.14.

Les résultats de Langendijk et Bronkhorst pour les stimuli filtrés par les HRTFs individuelles non modifiées sont donnés figure 7.21(a). On note que la log-vraisemblance actuelle n'est comprise dans l'intervalle de confiance de la log-vraisemblance escomptée que pour 50% des sujets (4 sur 8) : 3 sujets présentent le cas $L_a > L_e + CI_{99\%}(L_e)$ ce qui signifie que ces sujets sont moins performants que le modèle, et un sujet présente le cas $L_a < L_e - CI_{99\%}(L_e)$, ce qui signifie que ses réponses sont plus précises que ce que prévoit le modèle. En moyenne, la log-vraisemblance actuelle est de 0.79. Les résultats relatifs à notre étude, pour la condition individuelle et avec la métrique *GT-STD*, sont donnés dans la figure 7.22(b). On observe que la log-vraisemblance actuelle se situe dans l'intervalle de confiance escomptée pour 75% des sujets : deux d'entre eux (sujets 2 et 7) présentent une performance de localisation supérieure à celle prévue par le modèle et un d'entre eux (sujet 8), une performance moindre. En moyenne, la log-vraisemblance actuelle est de 0.87, soit légèrement supérieure à l'étude de Langendijk et Bronkhorst. Cette observation doit être principalement due aux différences dans le nombre de points considérés sur l'espace de prédiction. Comme présenté section 6.4.3 (voir aussi figure 6.6), les valeurs de log-vraisemblance sont dépendantes du nombre de points considérés sur l'espace de prédiction et ce, malgré la normalisation par L_{chance}. Les probabilités sont obtenues suite à la division des indices de similarité par la somme des indices sur l'espace. Ainsi, plus le nombre de points est important, plus les valeurs de probabilité sont faibles et plus la log-vraisemblance augmente. Langendijk et Bronkhorst réalisent une prédiction sur le plan médian uniquement, composé de 53 points, alors que nous utilisons des intervalle latéraux de largeur 40°, dont le nombre de points varie entre 150 et 450 points environ.

7.4.2.3 Comparaison avec Baumgartner et al. (2014)

Une comparaison du même ordre peut être réalisée avec l'étude de Baumgartner et al. [BML14] dans laquelle les résultats de log-vraisemblance actuelle (et de log-vraisemblance escomptée) sont donnés pour la condition d'écoute individuelle. Dans cette étude, la fonction sigmoïde est optimisée pour chacun des sujets dans la condition individuelle et avec la métrique *posGrad-Mean*. Les résultats de ces auteurs sont présentés figure 7.22(a). On observe que la log-vraisemblance actuelle se situe dans l'intervalle de confiance de la log-vraisemblance escomptée pour 61% des sujets (soit 14 sur 23). Dans le cadre de notre étude, pour une optimisation de la fonction sigmoïde à chacun des sujets dans la condition individuelle, avec la métrique *posGrad-Mean*, le pourcentage de réussite s'élève à 100%. L'individualisation à la fois des paramètres Γ et S de la fonction sigmoïde peut justifier l'amélioration observée. Cependant, il faut noter que nous avons un effectif deux fois plus faible.

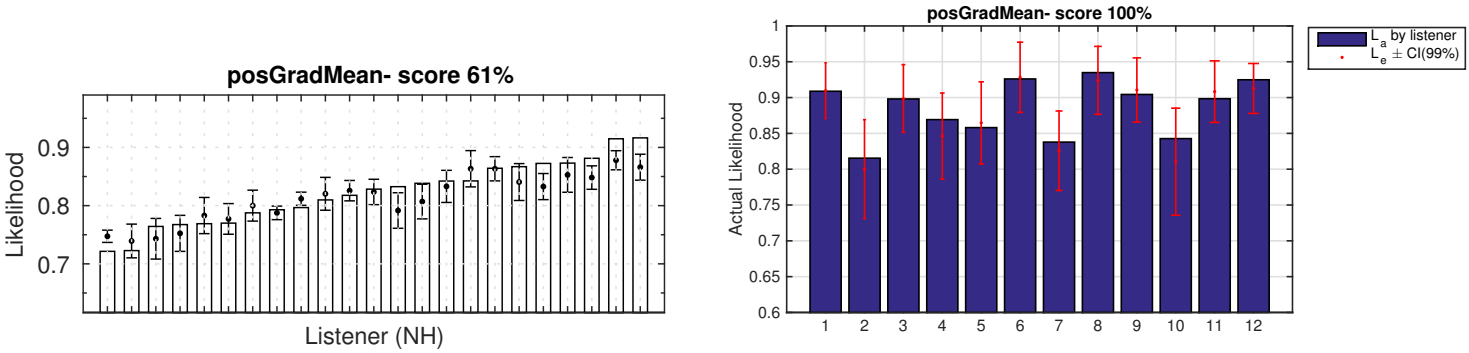
En moyenne, la log-vraisemblance actuelle est de 0.88 dans le cadre de notre étude (min.=0.82 ; max.=0.93 ; médiane=0.90) contre 0.82 pour l'étude de Baumgartner et al. (min.=0.72 ; max.=0.92 ; médiane=0.83). Cette différence peut à nouveau s'expliquer par le nombre de points considéré lors de la prédiction. En effet, Baumgartner et al. ont effectués la prédiction sur des plans sagittaux constitués d'environ 40 points.



(a) Résultats Langendijk et Bronkhorst

(b) Résultats de la présente étude

FIGURE 7.21 – Log-vraisemblance actuelle (barres bleues) et intervalle de confiance de la log-vraisemblance escomptée (barres d'erreur rouges), en condition d'écoute individuelle, pour (a) l'étude de Langendijk et Bronkhorst et (b) notre étude. Les valeurs ont été normalisées par la log-vraisemblance L_{chance}. Les résultats sont relatifs à la métrique *GT-STD* et à une optimisation de la fonction gaussienne ou sigmoïde non individualisée. Le score indiqué dans le titre indique le pourcentage de réussite du modèle, soit le nombre de fois sur les 8 ou 12 sujets où $L_a \in CI_{99\%}(L_e)$.



(a) Résultats Baumgartner et al. (figure 12 de l'article [BML14])

(b) Résultats de la présente étude

FIGURE 7.22 – Log-vraisemblance actuelle et intervalle de confiance de la log-vraisemblance escomptée (barres d'erreur), en condition d'écoute individuelle, pour (a) l'étude de Baumgartner et al. [BML14] et (b) notre étude. Les valeurs ont été normalisées par la log-vraisemblance L_{chance}. Les résultats sont relatifs à la métrique *posGrad-Mean* et à une optimisation de la fonction sigmoïde individualisée à chacun des sujets. Le score indiqué dans le titre indique le pourcentage de réussite du modèle, soit le nombre de fois sur les 23 ou 12 sujets où $L_a \in CI_{99\%}(L_e)$.

7.4.2.4 Différences de résultats entre les conditions d'écoute

Bien que les résultats de la prédiction en condition individuelle soient relativement satisfaisants en comparaison des autres études, la condition non-individuelle présente une dégradation marquée des résultats de prédiction. Les premières analyses des données expérimentales ont dévoilé une plus forte dispersion dans les jugements à l'écoute de stimuli synthétisés avec des HRTFs non-individuelles. Cela traduit une réelle difficulté à localiser dans cette condition d'écoute, qui peut être liée à des sources sonores plus confuses et moins externalisées, comme c'est souvent le cas dans le cadre de la localisation de stimuli synthétisés avec des HRTFs non-individuelles, mesurées en environnement anéchoïque et présentés de manière statique.

Une conséquence directe de la dispersion plus marquée des jugements en condition d'écoute non-individuelle est que les fonctions sigmoïdes optimales sont moins sélectives. Or, nous avons vu section 6.4.3 que plus les distributions de probabilités sont dispersées (i.e. moins elles sont sélectives), plus la log-vraisemblance actuelle se rapproche du seuil de chance ($L_a = 1$). De plus, lorsque les HRTFs cibles sont non-individuelles, nous avons vu que modèle prévoyait des régions de forte probabilité de réponse plus éparses. Cela est notamment visible sur la figure 6.15 de la section 6.4.5.

Cette difficulté à localiser en condition non-individuelle est plus ou moins marquée selon les individus et peut être liée à leur niveau d'expertise. La figure 7.23 présente l'amplitude des différences de log-vraisemblance actuelle pour chaque sujet entre les conditions individuelle et non-individuelle. La taille

des barres verticales quantifie cette différence, avec des minima et maxima correspondants aux valeurs de log-vraisemblance actuelle en condition individuelle et non-individuelle respectivement, et moyennées sur les 10 métriques spectrales testées. Les résultats sont relatifs à une optimisation de la fonction sigmoïde non-individualisée. On voit que ces différences dépendent beaucoup du sujet. Par exemple, le sujet 10 présente une forte différence entre les conditions d'écoute : la log-vraisemblance actuelle associée à ses réponses en condition individuelle est en-dessous de la moyenne et dépasse le seuil de chance pour la condition non-individuelle. Notons que les sujets 2 et 10 pour lesquels la différence de score de prédiction est importante ici, présentent une dispersion dans leur jugements qui diffère également de façon notable entre les conditions individuelles et non-individuelles (c.f. figure 7.9(b)). Cela suggère un lien entre les résultats de prédiction et la capacité du sujet à localiser les sources virtuelles. Par ailleurs, il peut y avoir un lien avec les HRTFs non-individuelles sélectionnées pour chacun des sujets.

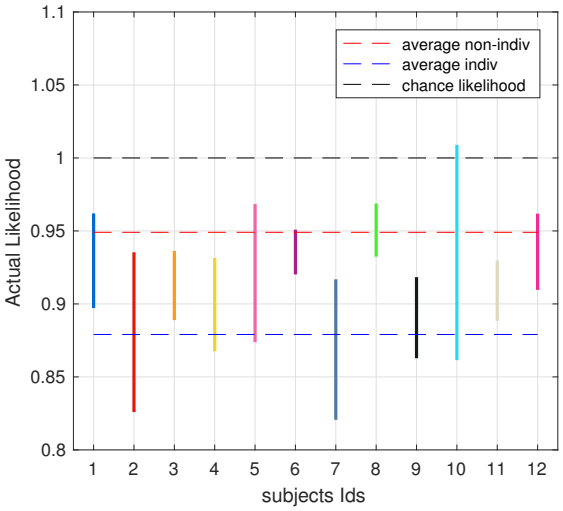


FIGURE 7.23 – Différence de log-vraisemblance actuelle en condition individuelle et non-individuelle pour chaque sujet (barres verticales de même code couleur que les figures 7.12 et 7.13), toutes métriques confondues. La courbe en pointillés bleue indique la valeur moyenne sur les sujets en condition individuelle, même chose en rouge pour la condition non-individuelle. La courbe en pointillés noire indique le seuil de chance.

7.4.2.5 Choix de l'indice d'ITD

Nous analysons ici la pertinence de l'indice d'ITD et de la méthode d'estimation de l'ITD pour prédire la position latérale des réponses. Pour ce faire, nous calculons le coefficient de corrélation entre l'ITD de l'auditeur à la direction pointée et l'ITD cible. Les ITDs sont estimés à partir des HRIRs de l'auditeur et une interpolation par les plus proches voisins permet d'obtenir l'ITD à la direction pointée. Pour rappel, les directions pointées ont été corrigées par le biais de pointage en azimut.

Tout d'abord, nos résultats peuvent être comparés à ceux de Middlebrooks (1992). Cet auteur a réalisé une étude similaire au paragraphe II.b. de l'article [Mid92] mais qui concerne l'indice d'ILD. En effet, leurs stimuli sont à bandes fréquentielles étroites, centrées sur des fréquences relativement élevées (égales ou supérieure à 6 kHz), et l'ILD est l'indice interaural prédominant dans cette situation [WK92]. Cependant, dans le cas de stimuli à bande fréquentielle large et plus généralement, contenant des basses fréquences, ce qui est la cas de notre expérience, l'ITD est l'indice interaural prédominant.

Les résultats de l'analyse de Middlebrooks présentent des coefficients de corrélation compris entre 0.881 et 0.957 avec une médiane de 0.945. Nos résultats sont semblables voire meilleurs : les coefficients de corrélation entre l'ITD aux directions pointées et l'ITD cible varient entre 0.939 et 0.977 avec une médiane de 0.957. Si on considère l'indice d'ILD, on remarque que les coefficients de corrélation diminuent avec un minimum de 0.875, un maximum de 0.945 et une médiane de 0.913, ce qui confirme que l'indice interaural d'ITD prédomine dans notre cas. Les coefficients de corrélation sont alors plus faibles que ceux mis en évidence par Middlebrooks. Notons par ailleurs que la méthode de calcul de l'ILD diffère quelques peu. En effet, Middlebrooks estime l'ILD dans la bande fréquentielle étroite du stimulus (de largeur $\frac{1}{6}$ d'octave) alors que nous travaillons avec des stimuli à bande fréquentielle large où l'estimation de l'ILD est réalisée dans la bande $[1, 5]$ kHz, en référence à [Lar01].

Pour aller un peu plus loin dans l'étude de l'indice interaural, on s'intéresse ici aux coefficients de corrélation obtenus avec différentes méthodes d'estimation de l'ITD : les méthodes "MaxIACC", "MaxIACCr",

"CenIACC", "CenIACCr", "Threshold", et "DELMOD", présentées en section 2.3.1.2. Pour rappel, la méthode "DELMOD" consiste à calculer la différence entre les retards purs gauche et droit estimés à partir de la pente de régression linéaire des spectres de phase de la composante à excès de phase. Ici, le calcul de la pente est restreint aux fréquences inférieures à 8 kHz, ce qui correspond à la zone fréquentielle où la pente est approximativement linéaire.

Pour rappel, la méthode utilisée dans l'étude est la méthode du maximum de l'inter-corrélation des réponses impulsionnelles gauche et droite soit "MaxIACCr". Sur la figure 7.24, on observe un désavantage marqué de la méthode "DELMOD". La méthode "CenIACC" présente également un coefficient de corrélation moins élevé que les méthodes "MaxIACC", "MaxIACCr", "CenIACCr" et "Threshold", qui présentent les meilleurs résultats. Ces observations nous permettent de valider le choix de la méthode d'estimation de l'ITD utilisé lors de la prédiction de la localisation latérale.

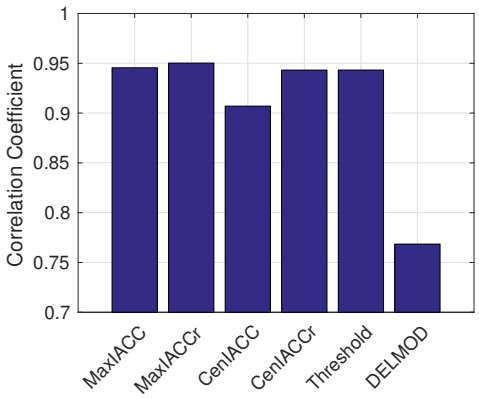


FIGURE 7.24 – Coefficient de corrélation moyenné sur les 12 sujets entre les ITDs aux directions pointées et les ITDs cibles, pour différentes méthodes d'estimation de l'ITD.

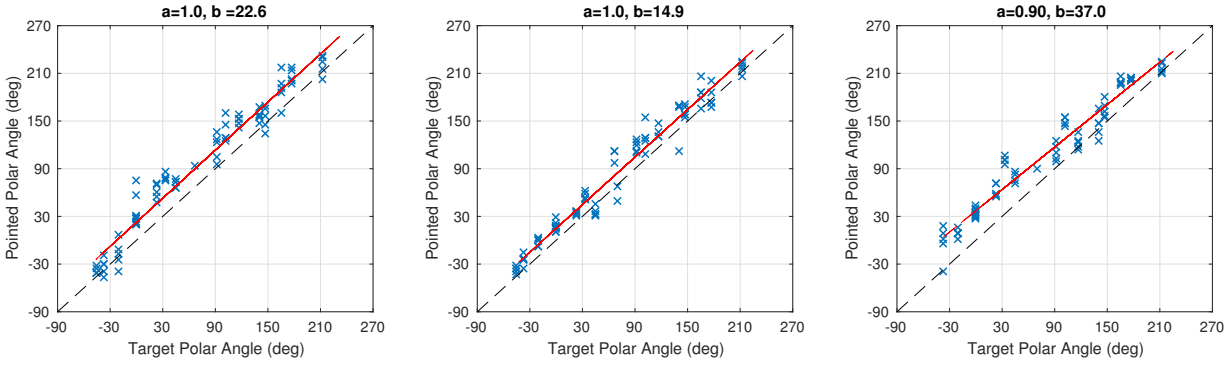
7.4.3 Décalage angulaire entre directions pointées et mesures de HRTFs

Un point critique n'a pas été mentionné jusqu'ici et mérite d'être abordé. Suivant la calibration réalisée lors du test de localisation, le plan horizontal a été défini par le plan d'intersection entre l'entrée des canaux auditifs gauche et droit et le bout du nez. Les directions pointées ont été calculées sur la base de cette définition du système de coordonnées relatif à la tête de l'auditeur. Or, elle n'est pas exactement en adéquation avec le système de coordonnées défini lors des mesures de HRTFs. Lors des mesures, l'axe interaural (repéré par l'entrée des canaux auditifs) a été aligné avec l'axe x du système de mesure. Cependant, aucune contrainte n'est imposée concernant l'inclinaison de la tête. Les sujets étaient invités à se positionner de manière "naturelle". Cela ne permet pas d'assurer que le bout du nez coïncide avec le plan horizontal. Au contraire, il arrive souvent qu'il soit plus bas que ce plan.

Ces éventuelles disparités dans la définition du système de coordonnées entre les mesures de HRTFs et la mesure des directions pointées peuvent être à l'origine d'une sur-estimation des erreurs polaires dans le test de localisation. Prenons un exemple : si une source virtuelle est synthétisée avec les HRTFs individuelles de l'auditeur à la direction nominative $(0°, 0°)$ et que l'auditeur la perçoit dans la direction exacte de synthèse, le sujet va pointer dans la direction où était placée la source sonore au moment de la mesure de la HRTF $(0, 0)$. Si celle-ci est décalée vers le haut par rapport au plan défini par l'axe interaural et le bout du nez, alors nous allons mesurer automatiquement une erreur polaire positive. Ce décalage d'angle polaire (ou *offset* polaire) sera alors valable pour toute direction pointée dans la condition individuelle. La figure 7.25 présente l'angle polaire des directions pointées en fonction de l'angle polaire cible (direction de synthèse) pour 3 sujets du test de localisation, en condition individuelle. On observe ici un *offset* polaire net pour ces 3 sujets qui varie de 15° à 37° dans ces exemples. Bien qu'il soit peu probable qu'il existe réellement un *offset* de l'ordre de 40°, ces figures mettent en évidence des réponses décalées de manière relativement consistante sur toute la dimension polaire, ce qui tend à évacuer l'hypothèse d'un biais dû à la méthode de pointage (de plus, cette caractéristique n'avait pas été observée au chapitre 4).

La correction des réponses par cet *offset* polaire devrait améliorer les résultats de prédiction dans la condition individuelle étant donné que le maximum de probabilité de réponse est localisé à la direction test. Il serait intéressant de voir si une rotation du système de coordonnées à partir duquel sont définies les directions nominatives des HRTFs *templates* (HRTFs de l'auditeur), autour de l'axe interaural et d'amplitude égal à l'*offset* mesuré en condition individuelle, permettrait également d'obtenir un modèle plus performant pour la prédiction des directions pointées en condition non-individuelle. Cette correction n'est cependant pas envisageable dans le cas pratique de la sélection guidée puisque les HRTFs individuelles du sujet n'ont pas été mesurées. Toutefois et de manière similaire, l'évaluation de chacun des jeux de HRTFs *templates* de la base de données testé lors de la procédure de sélection guidée (en termes de sa

capacité à prédire les réponses du sujet non mesuré), pourrait être évalué à une rotation près autour de l'axe interaural.



(a) (b) (c)

FIGURE 7.25 – Angle polaire des directions pointées en fonction de l'angle polaire des directions test en condition d'écoute individuelle. Les croix représentent les données expérimentales. La droite de régression linéaire est représentée en rouge et ses coefficients a (pente) et b (ordonnée à l'origine) sont indiqués dans le titre de chaque figure. La droite en pointillés noir représente le cas idéal ($y = x$).

7.5 Evaluation de la méthode d'individualisation proposée

L'objectif de cette étude concerne l'individualisation des HRTFs sous forme de sélection dans une base de données guidée par les résultats à un test de localisation. La première étape a consisté à mettre au point un modèle de prédiction de la localisation dans lequel plusieurs paramètres ont été optimisés de manière à maximiser la probabilité de réponse aux directions pointées. L'étape suivante consiste à évaluer si le modèle de prédiction peut être utilisé pour la sélection guidée. En d'autres termes il s'agit de déterminer si nous serions capables, à travers l'observation des réponses d'un sujet dans un test de localisation de sources virtuelles, de retrouver dans une base de données les HRTFs qui lui conviennent, ou du moins lui proposer un ensemble restreint de jeux de HRTFs parmi lesquels se trouvent celui ou ceux les plus susceptibles de le satisfaire. Nous tentons de modéliser les réponses du sujet tour à tour à partir de chacun des N jeux de HRTFs *templates* de la base de données. Ces N jeux peuvent ainsi être classés du plus probable au moins probable. On peut alors proposer à l'auditeur un nombre restreint de M jeux les plus susceptibles de le satisfaire.

L'ensemble N de HRTFs testées correspond à toute la base de données BiLi regroupant les HRTFs mesurées à l'IRCAM et à Orange Labs, soit un total de $N = 82$ jeux de HRTFs. Les HRTFs *templates* (H^{α}_{temp}) utilisées dans le modèle pour calculer les distributions de probabilité de réponse, et définies jusqu'ici par les HRTFs de l'auditeur, sont remplacées successivement par chacune des N HRTFs de la base de données (correspondant aux sujets $\alpha_n = \alpha_1, \ldots, \alpha_N$). La log-vraisemblance actuelle est calculée pour chaque jeu de HRTFs puis classée par ordre croissant (i.e. pour une probabilité aux réponses décroissante). Optimalement, les HRTFs de l'auditeur devraient se trouver en première position ou dans les M premiers jeux offrant le minimum de log-vraisemblance actuelle. La procédure est réalisée pour chaque métrique spectrale, de manière à identifier si l'une d'entre elles permet de classer le jeu de HRTFs de l'auditeur de manière plus systématique aux premières positions du classement et ainsi de réduire la taille du sous-ensemble optimal. Notons que la fonction sigmoïde utilisée correspond à la fonction sigmoïde optimisée pour chaque métrique spectrale dans la condition d'écoute non-individuelle (c.f. section 7.3.1). Ce travail consiste en effet à simuler la sélection d'un jeu de HRTFs permettant d'expliquer au mieux les directions pointées par un nouvel individu dans un test de localisation de sources virtuelles synthétisées par des HRTFs non-individuelles. Il s'agit donc de la condition d'écoute non-individuelle et d'une situation où l'optimisation de la fonction sigmoïde ne peut être individualisée étant donné que les HRTFs individuelles de l'individu ne sont pas disponibles.

La figure 7.26 présente pour chaque sous-ensemble des m premiers jeux de HRTFs du classement, le pourcentage de présence du jeu de HRTFs de l'auditeur (sur les 12 individus de test). Les figures 7.26(a) et 7.26(b) présentent respectivement les scores obtenus à partir d'une prédiction réalisée sur les réponses à des sources synthétisées avec les HRTFs de l'auditeur ou avec des HRTFs non-individuelles. Il peut sembler trivial de s'intéresser au cas individuel, étant donné qu'il ne peut faire partie du test de localisation mené dans un objectif de sélection guidée. Cependant, il nous permet de vérifier que la méthode est consistante. D'après la figure 7.26(a), on note que si le sujet est confronté à ses HRTFs (ou à des HRTFs qui objectivement lui conviennent), la méthode va les lui proposer (et ce quelle que soit la métrique adoptée). Cette remarque est par exemple valable pour 80% à 100% des sujets, selon la métrique utilisée, pour un sous-ensemble de $M = 20$ jeux de HRTFs proposés.

En général, les métriques offrent toutes des résultats significativement supérieurs au tirage aléatoire symbolisé par la diagonale (en noir pointillés) et ce aussi bien à partir des résultats observés en conditions non-individuelle et individuelle. Autrement dit, le participant n'a pas fait le test pour rien, puisque la méthode a permis de lui proposer un ensemble restreint de M jeux de HRTFs parmi lesquels celui qui est susceptible de le satisfaire a plus de chance d'y apparaître que si on avait tiré au hasard M jeux dans la base de données. Toutefois, on note que lorsque l'ensemble est très restreint, par exemple pour $M = 10$, certaines métriques, comme celles basées sur la représentation MFCC, restent dans ce cas proches du tirage aléatoire.

L'observation du cas basé sur les seules HRTFs non-individuelles est intéressante puisque cela signifie que, même si le sujet n'a pas été confronté dans le test à un jeu de HRTFs identiques aux siennes, les méthodes vont lui proposer "les siennes" (sous-entendu celles de la base de données qui lui conviennent) avec une probabilité significativement supérieure au tirage aléatoire, alors mêmes qu'elles ne faisaient pas partie des HRTFs testées.

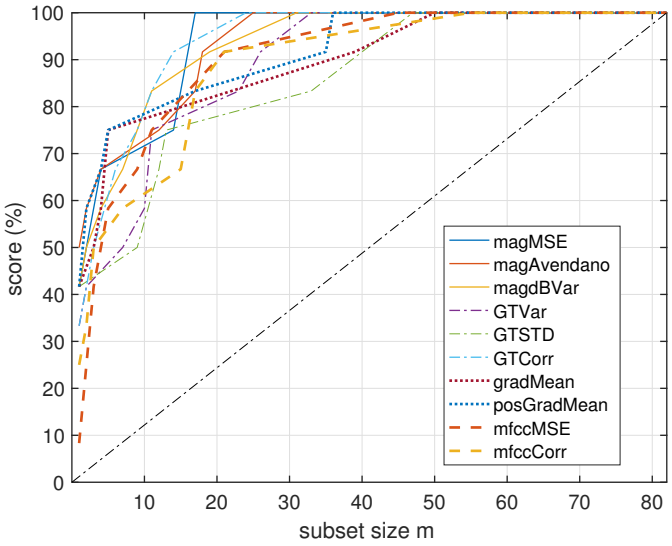
Les résultats ne permettent pas de conclure sur une métrique qui serait systématiquement supérieure aux autres. Cependant, les métriques basées sur les MFCCs présentent tout de même de moins bons résultats. Au contraire, le comportement des métriques basées sur le gradient spectral se montrent meilleures pour de très faibles nombres ($M < 10$) aussi bien dans le cas non-individuel, qu'individuel. Ensuite, l'avantage de ces métriques s'estompe. Il semblerait que ces métriques permettent d'identifier, au travers des réponses, un trait extrêmement caractéristique aux HRTFs individuelles. Pour 50% des sujets dans la condition non-individuelle, la métrique *gradMean* permet de proposer à l'auditeur un sous-ensemble de 7 jeux de HRTFs parmi lesquels se trouvent le jeu de HRTF optimal (individuel).

Ces observations peuvent être mises en parallèle avec les premières observations réalisées en section 6.4.5.3 sur la corrélation entre les métriques spectrales et la métrique interaurale. En effet, nous avons vu que les métriques basées sur les MFCCs encodent beaucoup plus l'information interaurale que les

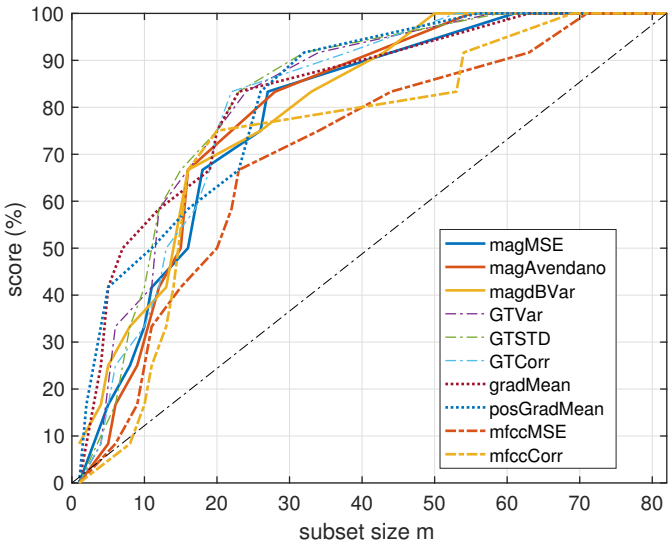
métriques basées sur les gradients spectraux. Or, nous savons que les indices spectraux présentent une variabilité inter-individuelle forte. Par conséquent, il est vraisemblable que ces indices soient davantage porteurs d'information individuelle. Cela expliquerait que les métriques basées sur le gradient spectral permettent plus facilement d'identifier les HRTFs de l'auditeur (ou du moins de les classer aux toutes premières positions du classement) autrement dit de caractériser sa carte auditive spatiale à partir de ses données expérimentales de localisation.

Ce type de représentation par le gradient avait également offert de bons résultats dans l'étude de Baumgartner et al. [BML14], concernant la prédiction des performances de localisation de stimuli contenant des enveloppes spectrales fortement ondulées. Au contraire, nous travaillions avec des stimuli large bande et à spectre plat, et visiblement ces métriques présentent un comportement qui peut être intéressant dans le cadre d'une sélection restreinte de jeux de HRTFs.

Les résultats offerts par la méthode de sélection guidée, à partir des réponses relevées pour 90 HRTFs (18 directions cibles $\times 5$ HRTFs non-individuelles), sont prometteurs. Bien que la méthode ne permette pas de sélectionner le jeu de HRTFs optimal, elle offre la possibilité de proposer à l'auditeur un sous-ensemble de jeux de HRTFs contenant le jeu de HRTFs optimal. Il serait intéressant de tester si, en augmentant le nombre de directions (exploration spatiale plus fine) ou le nombre de HRTFs non-individuelles pour chaque direction, le sous-ensemble optimal puisse être restreint. Il apparaît à présent nécessaire de valider par un test d'écoute le sous-ensemble de jeux de HRTFs sélectionné. On pourrait par exemple imaginer que, suite au test de localisation, l'auditeur soit invité à juger du rendu spatial de trajectoires horizontales et verticales synthétisées avec chacune des HRTFs non-individuelles du sous-ensemble pré-sélectionné (comme dans [KP12]) à l'issu du test de localisation.



(a) Scores établis sur la base des réponses en condition individuelle



(b) Scores établis sur la base des réponses en condition non-individuelle

FIGURE 7.26 – Pourcentage de présence du jeu de HRTFs individuelles de l'auditeur parmi les m premiers jeux HRTFs sélectionnés comme les plus probables par chaque métrique spectrale. La diagonale en pointillés noirs symbolise un tirage aléatoire dans la base de données.

Conclusion

Dans ce chapitre, les données d'un test de localisation de sources virtuelles synthétisées avec des HRTFs individuelles et non-individuelles a permis d'optimiser les paramètres d'un modèle visant à prédire les réponses des sujets. L'étude révèle que, pour être capable de prédire les observations expérimentales, la sélectivité spatiale des distributions de probabilité à la base du modèle doit s'adapter à la précision des auditeurs à localiser et à reporter la direction perçue dans chacune des conditions d'écoute. L'ajustement de la sélectivité spatiale peut être effectué à travers la paramétrisation de la fonction qui traduit les distances objectives sous forme de probabilité de réponse. Cette fonction modélise la dépendance de la probabilité que l'auditeur localise dans une direction donnée en fonction de l'amplitude des différences entre les indices acoustiques de la cible et ses indices propres. Une calibration du modèle dans chaque condition d'écoute (individuelle et non-individuelle) s'est avérée indispensable étant donné que la précision des auditeurs à identifier les directions les plus vraisemblables à l'écoute d'une source virtuelle diffère de manière significative entre une synthèse individuelle ou non-individuelle. La dispersion des jugements mesurée dans le test est non seulement liée à cette imprécision de localisation mais également à une imprécision dans le report des jugements spécifique de la méthode de pointage utilisée. Ces deux facteurs sont indissociables. Par conséquent, l'apport d'un modèle de dispersion de pointage *a posteriori* de l'adaptation de la sélectivité spatiale des distributions de probabilité n'a pas permis d'améliorer les résultats de manière significative.

L'analyse du comportement du modèle a permis de mettre en évidence les paramètres optimaux, en particulier les métriques spectrales les mieux adaptées à la prédiction des directions pointées. Tout d'abord, les métriques spectrales basées sur une représentation des HRTFs par un spectre d'amplitude lissé suivant des considérations physiologiques (banc de filtres auditifs) a montré offrir les meilleurs résultats. De plus, nous avons identifié que le critère de distance basé sur un calcul de variance de la différence spectrale permettait de s'approcher au mieux des observations expérimentales. Il serait intéressant d'approfondir la modélisation interne des HRTFs en tenant compte d'autres mécanismes du système auditif périphérique, tels que l'effet de la transduction des cellules ciliées [BML13], afin d'évaluer si cela améliore les résultats de prédiction. Une pondération des différences spectrales par bande fréquentielle suivant la résolution fréquentielle du système auditif pourrait également être testée [HZK99, NLBB06]. De plus, nous n'avons pas fait varier la métrique associée à l'indice interaural. Des modèles plus poussés de modélisation de l'information interaurale suivant par exemple le modèle d'égalisation-annulation, tel qu'utilisé par Park et al. [PNK08], pourraient être mis en œuvre pour prédire la localisation dans la dimension latérale. Par ailleurs, l'analyse comparative d'un ensemble de méthodes d'estimation de l'ITD a permis de souligner la pertinence perceptive de certaines méthodes d'estimation, telles que le maximum d'inter-corrélation des HRIRs gauche et droite.

Nous avons également montré que l'information traitée par les métriques spectrales présentait une corrélation plus ou moins marquée avec l'information interaurale. Ces caractéristiques, propres à chacune, doivent être prises en compte lors du choix de la métrique spectrale. Par exemple, pour caractériser les HRTFs d'un individu, les métriques porteuses d'information interaurale sont moins performantes puisque l'information spectrale discrimine davantage les individus entre eux. Dans ce cas, les métriques basées sur le gradient spectral paraissent mieux adaptées.

Le modèle s'est révélé bien moins performant pour prédire les directions pointées en condition d'écoute non-individuelle. Au vu de la dispersion des réponses dans cette condition, il semble que la localisation des sources virtuelles synthétisées avec des HTFs non-individuelles est perçue de manière diffuse et entraîne un difficulté à en indiquer une direction. Etant donné qu'en situation d'écoute réelle, l'auditeur utilise les différentes modalités sensorielles pour localiser, il serait intéressant de mener un test de localisation en synthèse dynamique, où la position des sources virtuelles serait rafraîchie en fonction des mouvements de la tête, de manière à évaluer si la localisation se précise en condition non-individuelle et si la prédiction des directions pointées est améliorée. Le modèle que nous avons développé s'appuie uniquement sur l'information acoustique. Dans le cadre d'un test de localisation autorisant les mouvements de la tête en synthèse binaurale dynamique, le modèle devra alors prendre en compte l'information spatiale acquise de l'interaction entre les actions motrices et les autres modalités sensorielles afin de prédire les directions perçues. La problématique associée à un tel type de test est que le processus de suivi de la tête va favoriser un apprentissage (plasticité cérébrale) qui risque de minimiser l'importance de ce que l'on cherche à observer, à savoir quel jeu de HRTFs permettrait le mieux d'expliquer les réponses, autrement dit les erreurs de localisation. Une idée pourrait être d'ausculter le voisinage de la direction pointée par le sujet en réponse à une HRTF statique, en testant les variations locales de différents jeux de HRTFs et en invitant le sujet à suivre de la main la direction qu'il perçoit.

La méthode de sélection guidée proposée a été évaluée en fin de chapitre. Sans garantir une identification exacte des HRTFs de la personne, la méthode s'est néanmoins avérée capable de la proposer parmi un jeu restreint, de manière significativement supérieure à un tirage aléatoire. Du point de vue du classement, aucune métrique ne s'est réellement détachée des autres en termes de capacité d'identification du jeu de HRTFs individuel. Cependant, le comportement des métriques de type MFCC a manifesté une certaine difficulté à se détacher de celui du tirage aléatoire, contrairement aux métriques de type gradient, qui ont présenté une certaine capacité à proposer le jeu de HRTFs individuel au sein d'un ensemble très restreint.

Tel qu'appliqué ici, la méthode recherche simultanément le jeu de HRTFs qui satisfasse simultanément

l'ITD et les indices spectraux. On pourrait dissocier les deux de sorte à retrouver un degré de liberté dans la sélection guidée. Aussi, les directions test du protocole expérimental ont été définies de manière à baliser le mieux possible l'espace et à collecter des données pertinentes pour l'optimisation du modèle. Dans le cadre de la recherche d'un protocole visant la sélection guidée, le critère de choix des directions pourrait être différent et insister plutôt sur les directions qui différencient le mieux les individus (directions discriminantes).

L'une des questions à l'issue de cette première tentative de mise en application de la méthode de sélection proposée, est évidemment de savoir si les HRTFs non-individuelles élues seraient effectivement jugées acceptables au travers d'un test d'écoute subjectif. Il serait également intéressant de déterminer si le classement obtenu est très discriminant i.e. si les HRTFs des premières positions apparaissent perceptivement déjà très différentes ou au contraire très proches. De plus, la robustesse du critère de décision des jeux de HRTFs sélectionnés mérite d'être évaluée. On pourrait imaginer mettre en œuvre un test de type "jackknife" où les réponses à chacune des HRTFs non-individuelles, pour lesquelles les données de localisation ont été récoltées, seraient retirées successivement afin d'observer l'impact sur le classement obtenu et ainsi de tirer des conclusions sur sa stabilité vis-à-vis du bruit de mesure.

Conclusion et perspectives

Le potentiel extraordinaire de la synthèse binaurale réside dans sa capacité *a priori* à restituer une information spatiale complexe à partir d'une technique et d'un dispositif de restitution très simples. A l'instar de la synthèse d'instruments de musique par échantillonnage, il n'est *a priori* nul besoin de connaissance sur la physique ou la perception auditive humaine. Dans le cadre de la synthèse instrumentale, on enregistre un instrument note par note et on rejoue celles-ci avec une garantie d'authenticité immédiate, hormis l'influence du haut-parleur (réponse en fréquence et directivité). Dans le cadre de la technique binaurale, on enregistre une scène sonore ou on engrange des réponses impulsionnelles directionnelles de la tête au plus près des conduits auditifs. Moyennant quelques précautions élémentaires de compensation de la chaîne de mesure et de reproduction, le résultat est également garanti en donnant accès de manière immédiate à l'ensemble de la sphère auditive.

Cependant, dès lors que l'identité entre la personne qui écoute et celle sur laquelle les échantillons de fonction de directivité ont été mesurés n'est pas respectée, il devient alors délicat de comprendre quelles en seront les conséquences. Il faut alors de nouveau convoquer différentes disciplines (acoustique, traitement numérique audio, perception et cognition auditive) pour tenter de modéliser les artefacts possibles et les méthodes permettant de les compenser.

Le déploiement de solutions de diffusion de contenus audio 3D (radio, télé) grand public, aisément accessibles sur dispositifs portables (tablette, téléphone, etc.) impose de trouver des solutions pour rétablir cette identité en sorte de fournir à l'auditeur la meilleure qualité d'expérience possible. Le projet BiLi, cadre dans lequel s'est déroulée cette thèse, entendait développer en parallèle différentes approches d'individualisation sur la base de mesures acoustiques simplifiées, de modélisation numérique de la diffraction des ondes sur la tête ou encore la sélection guidée d'un jeu de HRTFs dans une base de données d'HRTFs de référence.

La méthode d'individualisation proposée dans cette thèse repose sur la modélisation de la localisation auditive par une approche psycho-physique et de traitement du signal. Elle vise à sélectionner un jeu de HRTFs dans une base de données pour un individu à partir de l'observation de ses réponses à un test de localisation de sources sonores virtuelles. L'hypothèse à l'origine de cette initiative suggère que la direction perçue par un individu à l'écoute d'un stimulus filtré par une paire d'HRTFs non-individuelle peut être prédite par la mesure des similarités entre les indices acoustiques cibles et les indices acoustiques de localisation propres à cet individu.

Cette hypothèse permet d'expliquer pourquoi l'utilisation de filtres HRTFs non-individuels à la synthèse donne lieu à des défauts de spatialisation à la restitution, en particulier des ambiguïtés avant-arrière, des erreurs de localisation en élévation et un manque d'externalisation des sources sonores virtuelles. A l'instar des techniques de réalité virtuelle, ces défauts pourraient être réduits par l'utilisation d'un suivi de tête conjointement à une synthèse dynamique de la spatialisation des sources. Ces dispositifs permettent de rétablir ainsi la situation d'écoute naturelle dans laquelle notre perception de l'environnement s'appuie sur la boucle auditivo-motrice (apprentissage de l'évolution des informations acoustiques en fonction de nos propres mouvements). Cependant, l'accès à des dispositifs de suivi de la tête n'est pas encore généralisé. Même si des solutions sont en cours de commercialisation, les contenus adaptés sont encore très peu répandus. De plus, il ne garantit pas la qualité du rendu d'autres critères comme le rendu timbral des sources. Aussi, reste-t-il nécessaire de trouver des solutions d'individualisation pour lesquelles une compréhension des artefacts de localisation entraînées par l'usage de HRTFs non-individuelles peut être un guide précieux.

La première contribution de la thèse concerne l'acquisition d'une base de données de HRTFs individuelles d'une cinquantaine d'individus. Cette étude a fait l'objet d'un article de conférence. Les données collectées ont pu être exploitées tout au long du travail de thèse pour mener à bien nos objectifs. Notamment, les individus ayant été mesurés ont participé au test de localisation de sources sonores virtuelles. Cela a permis d'étudier la localisation auditive en synthèse binaurale non-individuelle pour des auditeurs dont nous possédions les HRTFs individuelles et de répondre ainsi à la contrainte imposée par les objectifs de recherche ayant suscité la mise en place de ce test.

La seconde contribution concerne la comparaison de trois méthodes de pointage égocentriques pour le report de la localisation auditive dans le cadre d'un test de localisation de sources réelles. Cette étude a fait l'objet d'un article de revue. Elle inclut deux méthodes de pointage très répandues (pointage avec la tête ou pointage manuel bras tendu) ainsi qu'une méthode peu documentée jusqu'à présent, la méthode de pointage proximale. Nous avons relevé certaines limitations pour chacune d'entre elles qui peuvent

être associées à la tâche motrice. Les résultats exposés permettent de guider le choix d'une méthode de pointage pour les tests de localisation auditive. Certains avantages, notamment pratiques, observés avec la méthode de pointage proximale, nous ont encouragé à utiliser cette méthode pour le report des directions perçues dans le test de localisation de sources sonores virtuelles. Une perspective intéressante à la mise en application de la méthode proximale serait de l'utiliser pour collecter des jugements sur la distance perçue de sources sonores virtuelles synthétisées au casque dans des directions de l'hémi-champ frontal. Elle permettrait ainsi l'évaluation de l'externalisation des sources virtuelles, qui constitue une caractéristique importante et pourtant souvent délaissée dans les méthodes mises œuvre pour évaluer la qualité du rendu spatial. Par ailleurs, avec la démocratisation des dispositifs inertiels (embarqués par exemple dans les téléphones ou manettes de jeu), il n'est pas exclu de pouvoir transposer le principe de ce test dans un contexte domestique pour collecter des performances de localisation auditive pouvant être exploitée pour guider l'individualisation.

Le développement du modèle de prédiction de la localisation auditive fait l'objet de la contribution principale de cette thèse. L'ensemble des étapes et éléments clé du calcul de la distribution de probabilité de réponse ont été étudiés en détails. Un intérêt particulier a été porté à la question de la métrique spectrale, utilisée pour évaluer la similarité entre HRTFs. Les directions indiqués par les participants dans un test de localisation auditive à l'écoute de sources virtuelles synthétisées avec des HRTFs non-individuelles ont servi de base pour la calibration du modèle. L'observation des performances de prédiction obtenues selon les paramètres utilisés a notamment permis de mettre en évidence deux métriques spectrales particulièrement adaptées à la prédiction des directions perçues. Ces deux métriques s'appuient sur une représentation lissée des spectres d'amplitude des HRTFs par bandes fréquentielles suivant un modèle de filtres auditifs. On peut envisager enrichir la modélisation des traitements appliqués par les différents éléments qui composent le système auditif afin de déterminer si cela permet d'offrir de meilleures performances de prédiction. Concernant la métrique interaurale, les méthodes d'estimation de l'ITD basées sur l'inter-corrélation des HRIRs gauche et droite ont montré pouvoir prédire de près les réponses latérales. L'utilisation d'une méthode de type *Equalization-Cancellation* à la place d'une métrique d'ITD est à tester.

Enfin, le modèle optimisé a été appliqué pour examiner la méthode d'individualisation par sélection guidée proposée. Sur la base du test de localisation, les métriques basées sur le gradient spectral des HRTFs se sont montrées susceptibles d'identifier de manière plus efficace les HRTFs individuelles des participants dans la base de données à partir de leurs données de localisation, comparativement aux autres métriques testées. Ce résultat suggère que la représentation des HRTFs par le gradient spectral permet de mieux caractériser les HRTFs d'un individu et ainsi de mieux discriminer les individus entre eux. Ce résultat devrait cependant être confirmé par des tests à plus grande échelle ou avec des stimuli plus complexes. Globalement, nous avons observé que la plupart des métriques permettaient de proposer au participant un ensemble restreint de jeux de HRTFs incluant le jeu optimal (ici symbolisé par le jeu individuel). La méthode doit être à présent validée par un test perceptif pour vérifier que les jeux de HRTFs non-individuels qui permettent d'expliquer au mieux les jugements de localisation offrent un bon compromis à la mesure de HRTFs individuelles.

Annexe A

Article sur la mesure de la base de données d'HRTFs du projet BiLi

7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Measurement of a head-related transfer function database with high spatial resolution

Thibaut Carpentier, Hélène Bahu, Markus Noisternig, Olivier Warusfel
UMR STMS IRCAM-CNRS-UPMC, 1 place Igor Stravinsky, 75004 Paris, France.

Summary
This paper describes a database of high spatial resolution head-related transfer functions (HRTF) measurements for 54 subjects (42 males, 12 females) and 3 dummy heads. The head-related impulse responses (HRIR) have been measured in IRCAM's anechoic chamber using the exponentional sweep sine technique and a sampling rate of 96 kHz. Microphones were positionned at the entrance of the blocked ear canal. The spatial sampling scheme is based on a Gaussian grid and includes 1680 directions with full azimuth range (0° to 360°), and elevation ranging from -51° to +86°. The angular step size is approximately 6 degrees in both dimensions. The subject's head position and orientation are tracked with an infrared optical motion capture system. The HRIR are publicly available in the standardized SOFA file format.

PACS no. 43.66.Pn, 43.60.Ek

1. Introduction

Head Related Transfer Functions (HRTFs) describe the linear filtering of a free-field sound from a given direction caused by the physical propagation and diffraction around the head, body and ears of a listener. When represented in time domain these functions are typically referred to as Head Related Impulse Responses (HRIR). They comprise the different sound localization cues and are thus essential for the design and the evaluation of spatial audio systems (e.g. binaural synthesis, room auralization, virtual reality applications, etc.).

Several research groups measured HRTF data and made them available for other researchers (see e.g. [1–8]). This work presents a new HRIR database measured on human subjects and three different dummy heads that provides:

- a large number of human subjects;
- a high resolution sampling grid;
- a high sampling rate;
- low harmonic distortions and signal-to-noise ratio well suited for binaural sound reproduction.

The article is organized as follows: Section 2 presents the measurement apparatus and the data acquisition procedure. Section 3 discusses the obtained HRIR data and the corresponding database.

2. Measurement setup

2.1. Infrastructure

All measurements were performed in IRCAM's full anechoic chamber. This sound isolated cuboid-shaped room (room-within-a-room structure supported on neoprene mounts; all interior surfaces and mechanical structures are covered with anechoic wedges) provides a useable volume of 103 m^3 ($5.7 \times 4.3 \times 4.2$ m^3), a lower limiting frequency of 75 Hz, and approximately 18.5 dB SPL (A Weighting) background noise. The anechoic chamber is equipped with a pivoting arc positioning system and a turntable that allow for positioning a sound source at any arbitrary position on a sphere around the head (see Fig. 1). The head is centered and aligned by means of three coincident laser beams pointing at the entrance of the ear canals and the rotation axis of the turntable, respectively. Head movements are tracked during measurements using an infrared camera motion capture system.

2.1.1. Turntable

Grating floor sections were installed in the center of the anechoic chamber to support a Brüel & Kjaer 9640 turntable system (see Fig. 1). This stepper motor turntable provides an angular resolution of 1° and is connected to a ball-bearing mounted rotary plate that supports the relatively high axial load. An adjustable-height stool is mounted on the plate, which is equipped with a slim back and neck rest. The rotation axis of the turntable is aligned by means of a ceiling-mounted vertical pointing laser beam.

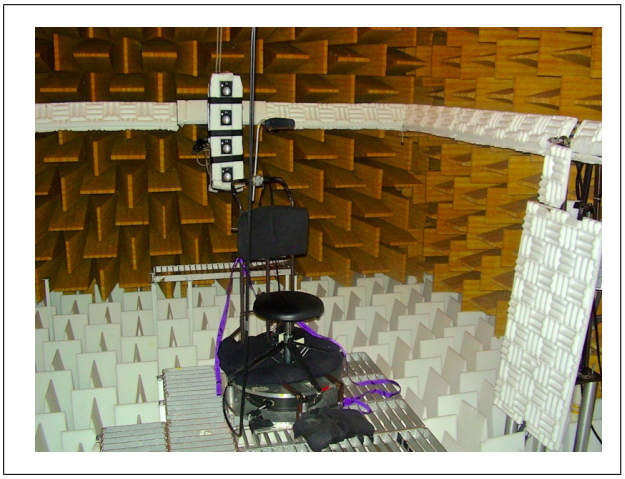


Figure 1: View of the anechoic chamber. Turntable with stool in the foreground; 4 loudspeakers mounted on the pivoting arc in the background.

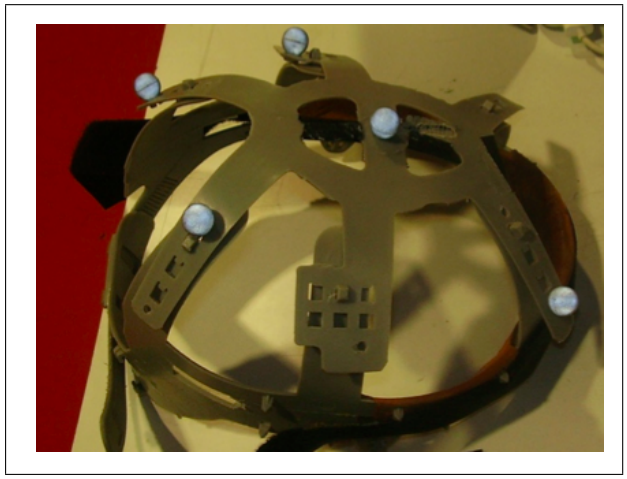


Figure 2: Safety helmet frame with reflective markers for head motion capture.

2.1.2. Pivoting arc

Measurement loudspeakers are mounted on a stepper-motor driven pivoting arc that is controlled by a 10 bit PIC microcontroller. It provides an angular resolution of $360/2^{10} \approx 0.35^{\circ}$. During measurements elevation angles are monitored with two high-precision electronic angular position sensors. The axis of rotation is indicated by two laser beams, which are aligned with the left and right stepper motor axes, respectively. The position of the subject – sitting on the stool – is then adjusted such that his/her interaural axis is aligned with the arc's rotation axis. To damp the vibrations of the arc a tension belt is stretched along the mechanical structure. The structure itself is further covered with absorbent material to minimize acoustic reflections. The arc supports up to four loudspeakers at a distance of 2.06 m from the center of the measurement sphere (see Fig. 1).

2.1.3. Motion tracking system

Head movements were monitored during measurements using six OptiTrack V100:R2 infrared motion capture cameras. The motion tracking software can identify rigid bodies defined by a set of reflective markers. We mounted five reflective markers onto the inner frame of a safety helmet (see Fig. 2) that can be easily adjusted to a subject's head. The center marker was used as a reference point and was aligned with both the system's vertical rotation axis and the center of the subject's head. Prior to the actual measurement session the tracking system was re-calibrated each time when a rigid body was adjusted to a subject's head. An additional webcam was installed so that the experimenter could monitor the subject during measurements from outside the anechoic chamber.

2.2. Audio equipment

2.2.1. Loudspeakers

Non-coaxial two-way ELAC 301 loudspeakers were used for the HRIR measurements. The distance between tweeter and woofer is approximately 6 cm; the cross-over frequency is 3.2 kHz. The tweeter-woofer distance corresponds to an elevation-angle bias of less than 2° that will be neglected for the remainder of this paper. The loudspeakers are driven by a 4-channel Yamaha P2040 amplifier that delivers up to 20 watts into 8 Ω loads.

2.2.2. Ear molds and microphones

For each subject individualized silicon ear molds were fabricated by a hearing aid specialist. They were cut to fit to the external auditory meatus [3], varnished for rigidification, and drilled to hold the Knowles FG26107 C34 miniature microphones and cable connectors (see Fig. 3). The microphones are connected to a custom made low-noise preamplifier and signal conditioner with a stabilized voltage regulated power supply to reduce total harmonic distortions. Audio signals are then transmitted to a RME Fireface 800 digital audio interface, which is also used for measurement signal playback.

2.3. Software components

2.3.1. Architecture overview

The measurement environment is divided into three main functional units: (i) automatic sequence control and monitoring of the measurement process with Matlab, (ii) real-time audio processing and communication with peripheral audio devices with Max/MSP, and (iii) real-time head motion capture using Natural Point's TrackingTools software.

The software units communicate trough the following internet protocols: Matlab sends instructions and

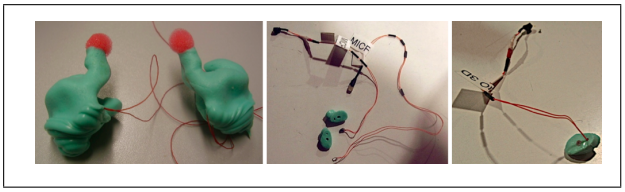


Figure 3: Different stages of the crafting of the ear molds: (left) silicon ear mold; (middle) drilled holes for microphone and cable connector; (right) ear mold with mounted microphone.

receives notifications ("error" or "success") to/from Max/MSP applying the Open Sound Control (OSC) protocol; the Virtual Reality Peripheral Network (VRPN) protocol is used for streaming the tracking data from Natural Point's interface to Max/MSP. The interaction in between the functional units is completely defined by the data exchange protocols and allows for the substitution of any element by a suitable replacement unit, which adheres to the defined protocol. The hardware and software architecture is illustrated in Fig. 4.

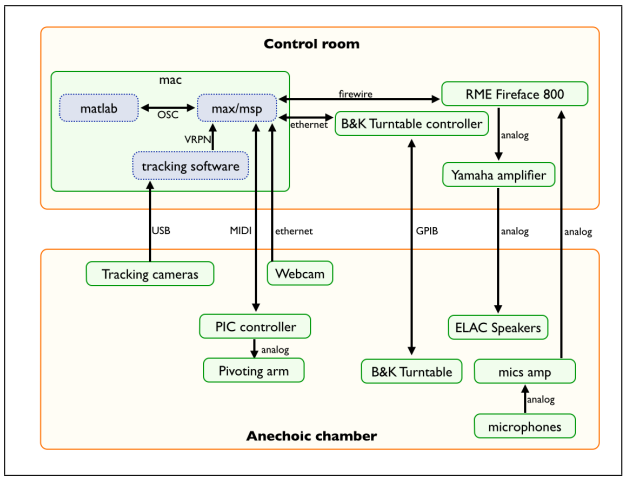


Figure 4: Overview of the measurement setup: hardware devices (green solid line); software components (blue dashed line).

2.3.2. Audio acquisition

HRIRs are measured with the exponential sweep sine method [9] at 96 kHz sampling rate. The benefits of sweep-based techniques over other acoustical measurements have been extensively discussed in literature (see e.g. [9, 10]). Basically, they allow for high signal-to-noise ratios (SNR), suppression of the harmonic distortion artifacts, and they are less vulnerable to the effects of time variance.

The sweep length is chosen to be 2^{16} samples (i.e. 683 ms) for a frequency range from 0 to 48 kHz. A 50 samples fade-out is applied to guarantee for a zero-crossing at the end of the signal. The length of the sweep is chosen as a trade-off between a good signal-to-noise ratio and the shortest possible duration of the overall measurement procedure. Short sweeps further reduce the risk of head movements during measurements so that the time-invariance hypothesis holds. A pause of 150 msec is set after the emission of each sweep; this pause is sufficiently long given (i) the typical length of HRIRs (about a dozen of milliseconds), and (ii) the low latency of the system.

To obtain the HRIR from sweep measurements the recorded signal has to be deconvolved with the excitation signal. The measurement system performs the deconvolution on the fly and allows for in-situ monitoring of the results. All computations are performed with double-precision floating point arithmetic; the numerical noise resulting from spectral inversion of the sweep is less than −110 dB. After deconvolution, the SNR is estimated through backward integration of the energy. Measurements with an estimated SNR below a threshold of 35 dB SNR are repeated. On average, SNRs are approximately 75 and 55 dB for ipsilateral and contralateral sides respectively (for 0° elevation). All data (raw recordings, sweep signal, HRIR data, etc.) are saved as 64-bit Matlab files along with miscellaneous textual metadata.

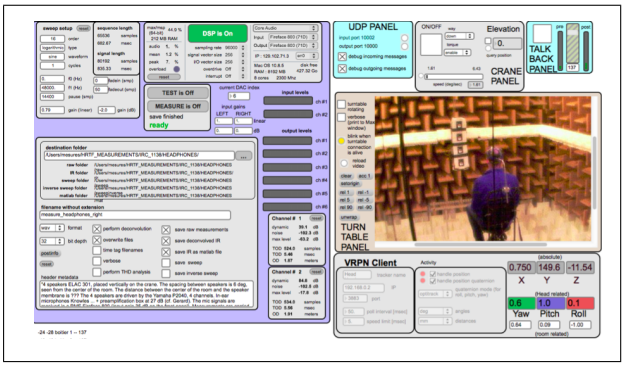


Figure 5: Experimenter interface in Max/MSP: (left) sweep settings, input/output signal vu-meters, real-time SNR estimations, etc.; (right) robotic arc elevation control, webcam monitoring, head-tracking data.

2.3.3. Control of mechanical devices

The B&K turntable system (type 9640) consists of a controllable turntable (type 5960) and a turntable controller (type 5997). A software plugin was developed to control the turntable from Max/MSP applying the GPIB (IEEE-488) protocol. The plugin receives acknowledgments from the B&K controller and unblocks HRIR measurements only when the turntable has reached the indicated position.

Max/MSP further controls the pivoting arc by sending MIDI commands to the PIC microcontroller. The actual angular position is monitored with the high-precision electronic angular position sensors that are connected to a mass balance. A control loop ensures that the arc reaches the target elevation.

3. Description of the database

3.1. Measurement grid

The Fourier-Bessel Expansion (FBE), or also Spherical Harmonics Expansion (SHE), of the HRTF data has been proven useful for many applications, such as spatial interpolation and range extrapolation (see e.g. [11] for further details). To accurately evaluate the discrete spherical wave spectrum of the measured HRTF data proper sampling of the sphere is indispensable. The number and distribution of sampling points limits the useable angular bandwidth, i.e. the maximum harmonic order N of the spherical data. Various spherical sampling schemes have been compared in [12]. Rafaely [13] shows that a Gaussian sampling grid with $2\,(N+1)^2$ sampling points allows for the exact computation of the spherical wave spectral coefficients. Although other sampling schemes offer the same property with less sampling points, they are often difficult to implement and/or it takes too much time to use them with a scanning array due to the non-regular angular distribution of points on (or even inside) the sphere.

When decomposing HRTFs onto a basis of discrete spherical harmonics, spatial aliasing limits the upper frequency for analysis to $kr < N$ (see e.g. [13]), where k is the wavenumber and r the radius of the listener's head. With a typical head radius of $r \simeq 10$ cm and an upper frequency limit of $f \simeq 16$ kHz the SHE order should be chosen $N \geq 29$. For the measured database, a nearby Gaussian grid with $N = 29$ was used. A Gaussian grid is uniformly distributed in azimuthal direction and close to a uniform angular distribution in elevation. The constant azimuthal step for the chosen order N is exactly 6°, which is compatible with the angular resolution of the B&K turntable. The elevation step is almost constant to approximately 5.9°. Therefore, the four available loudspeakers have been mounted with a vertical offset of 6°. This configuration allows for a good approximation of the theoretical positions (see Tab. I) while minimizing the number of displacements of the pivoting arc (which is rather slow). Due to practical constraints, such as the presence of a grating floor and the anechoic wedges on the ground floor, the lowest measured elevation is limited to −50.5°. This results in a polar gap of 360 non-measured directions (20%) with respect to the theoretical 29^{th} order Gaussian grid.

It is important to note that a 29^{th}-order Gaussian grid does not include data on the equator of the sphere. However, it is essential for many applications to provide HRTF data in the horizontal plane. Therefore, we performed an additional measurement with the pivoting arc at 0° elevation (measuring the four loudspeakers at -12, -6, 0 and 6° elevation). In summary the used measurement grid provides 1680 directions (see Fig. 6). With the described measurement setup, the data acquisition takes about 90 minutes per subject.

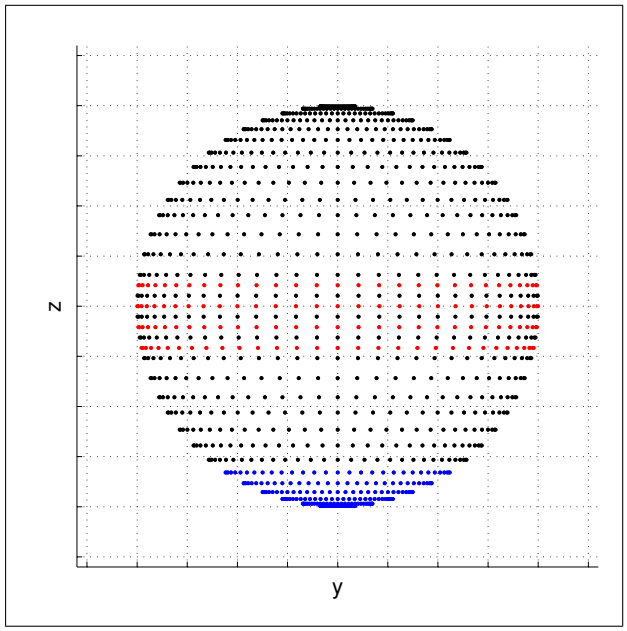


Figure 6: Measurement grid: (black dots) upper part of the Gaussian grid; (blue dots) lower part of the Gaussian grid that was not measured; (red dots) additional measurement points (not Gaussian).

3.2. Content of the database

This first release of the database includes head-related impulse responses for 54 subjects (42 men and 12 women), as well as 3 dummy heads (a Neumann KU 100, and a Brüel & Kjaer type 4100D with and without pinna). Free-field measurements of the loudspeakers and microphones responses are also available. Individual loudspeaker impulse responses were measured with a high-precision pre-polarized Brüel & Kjaer 1/2" free-field microphone (type 4189-L-001) and can be used for both free-field and diffuse-field equalization of HRTFs.

The downloadable package includes additional technical documentation and Matlab programs for screening the data. The database will be upgraded in the near future adding (i) HRIR data for additional subjects, (ii) anthropometric measurements, and (iii) high-precision 3D scans of the head and pinna for some of the subjects.

3.3. Data format

HRIR data are stored in the standardized AES-X212/SOFA "Spatial Acoustic Data File Format" (see e.g. [14] and http://www.sofaconventions.org for details). SOFA builds on netCDF (Network Common Data Form, http://www.unidata.ucar.edu/) and provides data compression, network transfer, file hierarchy, and partial data access over networks via OPeNDAP (Open-Source Project for a Network Data Access Protocol, http://www.opendap.org).

Data are made public-domain (for research and educational purposes) and are available at http://www.hrtf.ircam.fr.

4. Conclusions

We presented a new HRIR database measured for 54 human subjects that further provides tracking data for the actual head position during the measurements. All HRTF measurements were performed at the entrance of the blocked ear canals. This uniform database aims at enabling studies on interpersonal differences and facilitating research on individualized HRTFs.

Future work will focus on correcting the data according to the actual sound source directions, and completing the database with 3D meshes and anthropometric data. It is also planned to further investigate the impact of incomplete data (missing directions in the lowest part of the auditory sphere) on spherical harmonics analysis and modeling.

Acknowledgement

This work was funded by the French FUI project BiLi ("Binaural Listening", www.bili-project.org, FUI- AAP14) with support from "Cap Digital Paris Region".

References

[1] F.L. Wightman, D.J. Kistler: Headphone simulation of free-field listening (I): Stimulus synthesis. J. Acoust. Soc. Am. **85** (1989) 858 – 867.

[2] P. Majdak, M. J. Goupell, B. Laback: 3-D localization of virtual sound sources: effects of visual environment, pointing method, and training. Attention, Perception, and Psychophysics, **72** (2010) 454 – 469.

[3] H. Møller, M.F. Sørensen, D. Hammershøi, C.B. Jensen: Head-Related Transfer Functions of Human Subjects. J. Audio Eng. Soc, **43** (1995) 300 – 321.

[4] J. Blauert, C. Brueggen, A.W. Bronkhorst et al: The AUDIS catalog of human HRTFs. J. Acoust. Soc. Am. **103** (1998) 2901 – 2902.

[5] V.R. Algazi, R.O. Duda, D.M. Thompson, C. Avendano: The CIPIC HRTF database. Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 99 – 102.

[6] B. Xie, X. Zhong, D. Rao, Z. Liang: Head-related transfer function database and its analyses. Science in China Series G: Physics, Mechanics and Astronomy **50** (2007) 267 – 280.

[7] O. Warusfel: Listen HRTF Database, 2003. recherche.ircam.fr/equipes/salles/listen

[8] J.G. Bolaños, V. Pulkki: HRIR database with measured actual source direction data. Proc. 133rd Convention of the Audio Engineering Society, 2012.

[9] A. Farina: Simultaneous measurement of impulse response and distorsion with a swept-sine technique. Proc. 108th Convention of the Audio Engineering Society, 2013.

[10] S. Müller, P. Massarani: Transfer-function measurement with sweeps. J. Audio Eng. Soc. **49** (2001) 443 – 471.

[11] M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, M. Noisternig: Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition. Acta Acustica united with Acustica, **98** (2012) 72 – 82.

[12] M. Noisternig, F. Zotter, B. F. G. Katz: Reconstructing sound source directivity in virtual acoustic environments. In : Principles and Applications of Spatial Hearing. Y. Suzuki, D. S. Brungart, H. Kato (eds.). World Scientific Publishing, 2011.

[13] B. Rafaely: Analysis and design of spherical microphone arrays. IEEE Transactions on Speech and Audio Processing, **13** (2005) 135 – 143.

[14] P. Majdak, R. Nicol, T. Carpentier, Y. Suzuki, H. Wierstorf, H. Ziegelwanger, M. Noisternig: Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions. Proc. 134th Convention of the Audio Engineering Society, 2013.

Table I: Theoretical elevations (for 29th order Gaussian grid) versus actually measured elevations.
Black font: upper part of the Gaussian grid; blue font: lower part of the Gaussian grid that was not measured; red font: additional measurement points (not Gaussian). All values are expressed in degrees.

theoretical	-85.4826	-79.6307	-73.7443	-67.8500	-61.9528	-56.0542	-50.1547	-44.2548	-38.3546
actual	-	-	-	-	-	-	-50.5	-44.5	-38.5
deviation	-	-	-	-	-	-	0.3453	0.2452	0.1454
theoretical	-32.4541	-26.5535	-20.6528	-14.7521	-8.8513	-2.9504	2.9504	8.8513	14.7521
actual	-32.5	-26.5	-20.5	-14.5	-8.5	-3.0	3.0	9.0	15.0
deviation	0.0459	0.0535	0.1528	0.2521	0.3513	0.0496	0.0496	0.1487	0.2479
theoretical	20.6528	26.5535	32.4541	38.3546	44.2548	50.1547	56.0542	61.9528	67.8500
actual	20.5	26.5	32.5	38.5	44.0	50.0	56.0	62.0	67.5
deviation	0.1528	0.0535	0.0459	0.1454	0.2548	0.1547	0.0542	0.0472	0.3500
theoretical	73.7443	79.6307	85.4826	-	-	-	-		
actual	73.5	79.5	85.5	-12.0	-6.0	0.0	6.0		
deviation	0.2443	0.1307	0.0174	-	-	-	-		

Annexe B

Article sur la comparaison des méthodes de pointage

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 102 (2016) 107 – 118

DOI 10.3813/AAA.918928

Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments

Hélène Bahu, Thibaut Carpentier, Markus Noisternig, Olivier Warusfel
IRCAM, CNRS, Sorbonne Universités, UPMC Univ Paris 06, UMR 9912 STMS, 75004 Paris, France.
helene.bahu@hotmail.fr, markus.noisternig@ircam.fr

Summary
This study evaluates several methods for reporting the perceived location of real sound sources. It is well known that the method used for collecting judgments in auditory-localization experiments has a strong influence on the accuracy of a subject's response. Previous works on auditory-localization tasks revealed that egocentric pointing methods (which are based on a body-centered coordinate system) allow for more accurate judgments than verbal reporting or exocentric pointing techniques (which are based on a 2D or 3D reporting device). Three different egocentric methods are compared: the most commonly applied "manual pointing" and "head pointing" methods, and the "proximal pointing" method, which forces the participants to indicate the apparent direction by pointing in the proximal region of the head with a marker held at the fingertips. The two first methods involve a rotation of the body of the participant, whereas the third method only involves movements of the arm(s) and hand(s) with a fixed head. Sound stimuli were presented randomly over 24 loudspeakers that were uniformly distributed on the upper hemisphere around the subject. The merits of the different methods are compared and discussed with regard to localization errors and to practical considerations. Although they show similar trends, each of the different methods affects the pointing accuracy in a specific way. The proximal pointing method, for example, is more accurate for sources located at high elevation angles. However, at rear locations close to the median plane an increased bias appears due to difficulties in performing the motor task to reach these positions. The proximal pointing method shows faster response times, which may be advantageous when planning 3D sound localization experiments.

PACS no. 43.66.Qp

1. Introduction

The evaluation of human sound localization performances typically relies on measuring the bias and variability between the perceived and the actual sound location. The precision and reproducibility of such an evaluation not only depend on the subject's ability to localize sound sources, but also on the accuracy of the method used for collecting the responses.

Several reporting methods were proposed and discussed in literature. Although frequently used, the paradigm of verbal report (in which the subject indicates the apparent spatial position in terms of azimuth and elevation angles [1, 2]) was found to be unintuitive for unexperienced subjects, and thus yielding inaccurate judgments [3].

Exocentric reporting techniques were also investigated. In [4, 5, 6, 7], subjects reported the perceived sound location on a 2D/3D graphical interface, which schematically represents different views of the environment. Another reporting technique, the "God's Eye Localization Pointing" method (GELP, see *e.g.* [8, 9]), uses a rigid sphere that represents the auditory space and the subject indicates the perceived source direction by pointing with a stick onto the sphere. This method allows for fast response times. However, a systematic judgment error was observed for such exocentric methods, as subjects typically have difficulties to "project" their own auditory space onto the reporting device, *i.e.* the rigid sphere or the graphical user interface [5, 9].

Egocentric pointing methods have been shown to be more precise than exocentric methods [5, 8, 9]. The most common egocentric methods use manual pointing (*e.g.*, with the finger or with a tracked object held in the hand) [5, 9, 10, 11] or head pointing [9, 10, 12, 13]. As has been shown in Frank *et al.* [14], the advantage of the manual pointing method (implemented with a toy gun in their study) is that it is very intuitive and technically reliable. The different methods are sometimes used along with some visual feedback that indicates the pointed direction (*e.g.*, a laser pointer mounted on the head or handheld pointing device) [14, 15, 16, 17, 18, 19, 20], as it has been shown that a visual feedback improves the pointing accuracy [12, 18]. Seeber *et al.* [21, 22] introduced a reporting method, where the subject indicates the perceived

Received 26 February 2015,
accepted 11 October 2015.

direction with a trackball-controlled laser pointer, while the head is maintained in a fixed position. This method reduces the complex interaction effects between the different sensorial modalities (*i.e.*, vestibular, visual, and auditory), but is restricted to the field of vision (*i.e.* frontal targets). Other methods use an acoustic pointer that has to be aligned with a reference sound [23, 24, 25]. Those methods have also been omitted, as they are not very well adapted to 3D localization tests with real sound sources.

This study focuses on 3D localization tests with real sound sources (which are regularly distributed around the listeners) and only involves the auditory and proprioceptive modality. It compares three different egocentric pointing methods: (i) The "head pointing" method, where the participant is asked to turn his body towards the perceived sound source, and then to point to its direction with the head. When applied in previous studies (see *e.g.* [9, 10, 12]), this method showed many advantages over other methods, despite a difficulty to reach highly elevated targets. (ii) The "manual pointing" method is investigated using a toy-gun held in one hand. This method is also referred to as "gun pointing". The participant has to rotate on a swivel chair before aiming at the perceived sound direction with the arm stretched out. Majdak *et al.* [10] compared this method to the head pointing method. Both methods showed similar results, except for a higher accuracy of the manual pointing method for sources at high elevations. (iii) The present study investigates the benefits of a particular manual pointing method, which does not involve any head or torso movements. The participants have to indicate the perceived source direction by placing a marker held at the fingertips in the proximal region of the head. This method will be referred to as "proximal pointing" method throughout the remainder of this article.

The proximal pointing method is closely related to the finger pointing method reported in Djelani *et al.* [9], but allows for the use of both hands. It thus overcomes the systematic pointing error that is associated to movement restrictions when the subject points to sound sources in the hemispace opposite to the hand he is using for pointing. Contrary to [9], in this experiment the subject remained still during the stimulus playback in order to avoid that dynamic localization cues influence the test results. In addition, more source directions were evaluated than in Djelani *et al.*'s study, and real sound sources (*i.e.* sound stimuli played back over loudspeakers) were used instead of virtual sound sources (*i.e.* sound stimuli played back over headphones).

The motivation for studying the proximal pointing method is fourfold: (i) when a subject reports the perceived source direction with the fingertips in the proximal region of her head, being allowed to freely choose either the left or the right hand for pointing, she should easily point to any direction in 3D space; (ii) the proximal pointing method does not require any head or torso movements (not even for rear directions) and thus allows for fast response times; (iii) the measurement apparatus can be used for localization tests with both stimuli presentations over loudspeakers and headphones; (iv) this method may also be used for collecting judgments on the perceived distance of virtual sound stimuli played back over headphones in binaural sound reproduction systems. We however expect several limitations linked to limits of the human motor control. For instance, source directions close to the median plane in the rear hemifield may be difficult to reach with both the left and the right hand. One may also expect that a subject's pointing performance varies greatly depending on if he either responds with the dominant or the non-dominant hand.

This paper is organized as follows: Section 2 presents the experimental apparatus and the different methods. Section 3 examines the results in terms of localization accuracy, dispersion of responses and response time. Section 4 compares the results with reference studies in literature, discusses the respective merits of the different methods, and the limitations and possible future improvements of the proximal method.

2. Methods

2.1. Participants

39 participants (15 female and 24 male), ranging in age from 21 to 55 years, served as paid volunteers. 19 participants were audio professionals and 4 of them had already participated in auditory localization experiments. The experiment consisted of three different conditions (which are detailed in sections 2.4 and 2.5). 13 participants (8 male and 5 female) took part in each condition (in other words, each participant tested only one condition). All participants reported normal hearing. They were informed that the sound could emanate from any direction in 3D space, but had no prior knowledge of the loudspeakers positions. All tests were performed in the same acoustic environment and in total darkness, so that the loudspeakers were invisible for the subjects.

2.2. Experimental Setup

The experiment was conducted in an acoustically damped room with dimensions 7.8 m × 6 m × 4.2 m (experimental studio with all walls and the ceiling covered with porous broadband sound absorption panels, $RT_{60} = 150$ ms at 1 kHz; the floor was covered with additional sound absorbing fabric material) and a background noise level of 27 dBA. The room was equipped with a dome of twenty-four coaxial Amadeus PMX-5 loudspeakers, with an effective frequency range from 80 Hz to 22 kHz. The loudspeakers were uniformly distributed on a hemisphere with a radius of 2.65 m. The spatial distribution was symmetric about the median plane and composed of 3 horizontal rings at elevations of −5°, 26° and 57°, respectively, plus one additional loudspeaker at the zenith of the sphere. The loudspeaker positions are detailed in Table I. In the remainder of this paper, azimuth angles are measured clockwise from the median plane (front to rear), *i.e.* negative and positive azimuth angles are in the left and right hemispace, respectively. Elevation angles are measured from the

horizontal plane, with positive values in the upper hemisphere. All loudspeakers were oriented towards the center of the hemisphere, which also determines the origin of the reference coordinate system.

All participants were seated on a height-adjustable swivel chair and their heads were positioned in the center of the sphere. The position and the orientation of both the participant's head and the pointing devices were captured in real-time (with an update rate of 60 Hz) by the means of four infrared motion capture cameras (A.R.T. Track3). With the motion capture software, several tracking objects were defined, each consisting of a set of reflective markers mounted on a rigid body. For the gun pointing and proximal pointing methods, a toy-gun and small ad hoc objects were used, respectively. For head tracking, several markers were mounted on the inner frame of a safety helmet, which could be easily adjusted to the participant's head in such a way that it did not move or change its shape during the experiment. Prior to each experiment, and after aligning the participant's head with the reference coordinate system, the motion capture system was calibrated. A set of four coincident self-leveling cross-line laser beams was used to center and align the head (the interaural axis, the median plane, and the horizontal plane). This defines the initial position of the head. Then the head was tracked in real-time during the repositioning phase and for calculating the pointed direction (see section 2.5).

A 17-inch computer monitor was placed below the frontal loudspeaker to display an interactive graphical user interface (GUI). The GUI was solely switched on during the repositioning phase, in order to guide the participants to retrieve the initial head position, *i.e.* $(x, y, z) = (0, 0, 0)$ and orientation of 0° in azimuth and elevation. The GUI depicted a top view of the listener's head (left side of the screen) for adjusting the (x, y) position, and the orientation of the listener's head (right side of the screen). A vertical slider guided the participants to adjust their vertical head position. The GUI was refreshed in real time according to the head tracking data and the background color was switched from red to green to indicate that the subject was well positioned.

In order to make sure that none of the participants could see the loudspeakers, they (i) entered the darkened room blindfolded, and (ii) had to wear opaque sunglasses during the entire experiment. A preliminary test showed that, although the computer monitor was switched on during the repositioning phase, none of the loudspeakers was visible. Prior to each trial, the GUI was switched off to avoid visual anchors during the localization and pointing tasks.

2.3. Stimuli

The sound stimuli should be short enough in order to avoid that subjects receive dynamic auditory cues resulting from head movements while the sound is being played. According to Blauert [26], it takes at least 200 ms to initiate head movements. In other words, stimuli with a duration of less than 200 ms will not be affected. Other studies, which investigated the influence of the stimulus duration on the sound localization (see *e.g.* [12, 27, 28]), revealed that a stimulus of at least 80 ms is required for a stable estimation of the target sound elevation. Katz and Parseihian [29] further showed an improved source localization accuracy in azimuthal direction when using successive short bursts instead of a single noise burst. In this work, the target sound stimulus was a train of four 50 ms Gaussian noise bursts with 10 ms raise and fall times ($\cos^2$ slopes), and 10 ms pauses in between successive bursts. Therefore, the total duration of the target sound stimulus was 230 ms. The sound pressure level of each burst was set to 65 dBA, with a maximum deviation of 2 dBA over all loudspeakers.

2.4. Conditions 1 and 2

Condition 1 – head pointing: Participants were instructed to point their nose towards the direction of the perceived sound source. The reported direction is computed from the intersection of the head orientation with a virtual spherical surface, that has the same radius and center as the loudspeaker hemisphere. The computed intersection point is then converted into polar coordinates. Participants sat on a swiveling chair to facilitate body rotations in azimuthal direction. Once facing the perceived direction, they had to validate their response by double-clicking a button on a hand-held device. Participants had the free choice of which hand to use for the response button. In this experiment, one of the participants showed noticeable differences in performance from the overall average. Therefore, we decided to remove her results from the analysis, resulting in a total of 12 participants, thereof 8 male and 4 female.

Condition 2 – gun pointing: Participants were instructed to aim at the perceived location of the sound with a toy-gun (equipped with reflective markers for motion tracking) held in the hand of their choice. As before, the reported direction is computed from the intersection of the toy-gun orientation with a virtual spherical surface, that has the same radius and center as the loudspeakers hemisphere. The computed intersection point is then converted into polar coordinates. Here again, rotating the body on the swiveling chair was allowed. Participants were instructed to aim at the target with a stretched out arm and to position the gun such that it aligns with the arm (*i.e.* to avoid wrist movements). They had to validate their response by pressing twice the trigger of the gun.

In both conditions, participants entered the room blindfolded, guided by the experimenter. Once seated, their head was positioned in the center of the sphere using a set of coincident laser beams. After calibration, all laser beams and disturbing lights were turned off and the participants remained in total darkness. Then the blindfold was removed so that participants could see the graphical user interface. One trial consisted of the following steps: (1) the subject had to return to the initial position using the GUI; (2) once the position was reached (within a tolerance range of ±1° in azimuth and elevation, and ±2 cm in x, y and z) and maintained for two seconds, the participant was instructed not to move and the GUI was switched off; (3) two seconds later, the sound stimulus was played back

Table I. Loudspeaker coordinates (azimuth and elevation in degrees).

Elevation (°)	Azimuth (°)								
-5	-160	-120	-80	-40	0	40	80	120	160
26	-20	-60	-100	-140	180	140	100	60	20
57	-144		-72		0		72		144
90					0				

over a randomly selected loudspeaker; (4) the subject had to turn his body and head or the gun towards the location of the sound, and then to validate his reported direction by pressing twice the button or gun trigger.

2.5. Condition 3

Condition 3 – proximal pointing: Participants had to place their head on a neck-rest during the experiment in order to avoid head movements. In each hand, they held a rigid body with the reference marker at the fingertips. Participants were instructed to indicate the direction of the sound by placing the fingertips in the perceived direction within the proximal region of the head (see Figure 1). They were free to use the hand of their choice according to the most comfortable way to indicate the corresponding direction. No constraints about the distance of pointing from the head were imposed. Although the participants were instructed not to move their heads, the tracking data showed unconscious head movements during the pointing phase (average position and angular displacements in the order of 1 cm and 1°, respectively). Hence, the reported direction was evaluated between the position of the fingertips and the actual position of the head.

Here again, participants entered the room blindfolded, guided by the experimenter. They were seated on a non-swiveling height-adjustable chair. The position of the neck-rest was adjusted in order to position the participant's head at the center of the loudspeaker hemisphere. As for the other conditions, participants remained in total darkness, and the repositioning GUI was used to ensure the exact positioning of the head just before the stimulus playback. One trial consisted of the following steps: (1) the subject had to adjust the position and orientation of her head according to the GUI; (2) once well placed during 2 seconds, the subject was asked to put the arms onto the armrests; (3) once the arms remained well positioned for at least 1 second, the GUI was switched off; (4) two seconds later, the sound stimulus was played back over a randomly selected loudspeaker; (5) the participant had to indicate the perceived source direction by placing one of the two rigid bodies close to the head, while keeping the other hand on the arm rest, and had to confirm the response by pressing a foot-pedal.

2.6. Training and experiment

The experiment consisted of 192 trials (8 repetitions for each of the 24 source locations), with a total duration of about 40 minutes. No feedback was given with regard to the localization performance, neither during the training nor during the experiment.

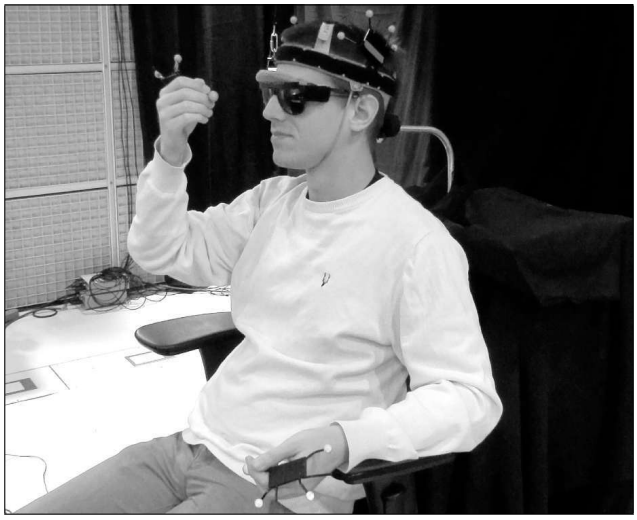


Figure 1. Photo of the proximal pointing task. The participant holds a rigid body in each hand, aligning the reference marker with his fingertips. The head is supported by a neck-rest to prevent head movements, and the head position and orientation were monitored in real-time using the motion capture system (inner frame of a helmet with reflective markers). The subjects were free to use the hand of their choice according to the most comfortable way to indicate the direction of the sound source, which is computed with regard to the current head position.

The subjects were trained to get familiar with the respective pointing method using the same stimuli as for the experiment. A training session consisted of 10 different target sound directions that were identical for all subjects and conditions. The different training directions were selected such that at least one target sound direction per quadrant was tested.

3. Data analysis and results

3.1. Definition of errors

In order to analyze the data, measures used to calculate the localization error and the coordinate system need to be defined. The choice may be determined from practical considerations (such as the spatial distribution of the sound sources), the hypothesized auditory localization model, or the reporting modalities. Wightman and Kistler [1] used a single pole coordinate system, where the azimuth and elevation angles correspond to the longitude and latitude, respectively. Data were analyzed in terms of the average angle of error (*i.e.* the mean of the unsigned angles between

the judgment vector and the target vector), the direction of the judgment centroid, and the associated dispersion. Morimoto and Aokata [30] and Majdak *et al.* [10] used a lateral/polar coordinate system. In lateral directions (*i.e.* the arc length between the judgment and the vertical median plane), errors were analyzed in terms of a lateral bias and a lateral precision. In polar directions (*i.e.* the elevation angle within a circle of constant lateral angle), errors were analyzed in terms of quadrant errors [10]. The use of the lateral angle is in good accordance with auditory localization models that depend on the interaural time and level differences (ITD, ILD).

However, in order to assess the localization error, the sensory-motor constraints (auditory and proprioceptive) of a pointing task should be considered [31, 32]. For instance, the use of the lateral angle is less suitable for head pointing and gun pointing tasks, as they require body rotations (given by the rotation axis of the swiveling chair) and a head and/or arm movement. For this reason, a single pole coordinate system is used in the present study and the localization errors are analyzed as in Makous and Middlebrooks [12] and Gilkey *et al.* [8]. The vertical error is given by the angular distance in between the elevation of the reported direction and the target sound source. As illustrated in Figure 2, the horizontal error is given by the angle between the target vector and a vector from the center of the sphere to a point on the surface of the sphere, whose longitude and latitude are equal to the judgment longitude and target latitude, respectively [8]. This error definition takes the azimuth compression at the zenith into account. For a source at the zenith, the horizontal error becomes zero. Consequently, this target source will not be considered when analyzing the horizontal localization performance.

3.2. Data analysis

Both the signed and unsigned localization errors are computed: unsigned errors only consider the magnitude of the error and thus reflect the localization accuracy, whereas signed errors provide information about the direction of the bias with reference to the target source position. A positive elevation error corresponds to a pointed direction higher than the target source. A positive horizontal error indicates that the subject rotated his body farther than the target position, the zero reference being the initial position, *i.e.* the body and head orientated towards (az, el)=(0°,0°).

No head movements were allowed during the stimulus playback, which results in an increase of front-back confusions [33]. During data analysis, front-back confusions were detected as described in Wightman and Kistler [1]: if the absolute angle between the target direction and the judgment gets smaller when the judgment is mirrored about the frontal plane, it is considered as a front-back confusion. Sources which are close to or on the frontal plane are excluded from the detection of confusions [34]. This applies to targets located in the following areas: (az, el) = (±80°, −5°); (az, el) = (±100°, 26°); (az, el) = (±72°, 57°), and (az, el) = (0°, 90°).

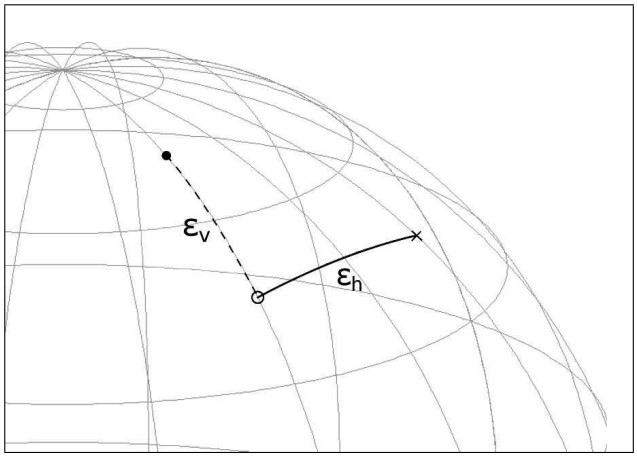


Figure 2. Illustration of the horizontal error angle (ϵ_h) and vertical error angle (ϵ_v) measured on the unit sphere subtended at the center of the head as defined in Makous and Middlebrooks [12] and Gilkey *et al.* [8]. The grey lines represent a part of the sphere's latitude and longitude lines, drawn with 20° increments; the filled circle corresponds to the response direction and the cross represents the stimulus direction. The point represented by the open circle is defined by the longitude of the response direction and the latitude of the stimulus direction.

As the proximal pointing method involves both hands, we examined the differences in the responses given with the dominant and non-dominant hands. This effect will be referred to as lateral dominance with items "dominant hand" and "non-dominant hand". It is worth noting that the participants always used the hand corresponding to the hemispace where the target source was located (*e.g.*, the right hand to indicate a source in the right hemispace); for sources located on the median plane, the hand differed over repetitions, for a given source and the same participant.

A repeated measures ANOVA (analysis of variance) was performed with the pointing condition (head pointing, H; gun pointing, G; proximal pointing, P) as between-subjects factor, and with several within-subject factors, such as (i) the repetition index ranging from 1 to 8; (ii) the hemifield of the target, with the categories "Front" associated to targets ranging clockwise from azimuth −80° to +80°, and "Back" associated to targets ranging clockwise from azimuth +100° to −100° (*i.e.* all targets on the frontal plane are excluded); (iii) the target elevations (−5°, 26°, 57°, and 90°); (iv) the elevation range, with items "Mid" combining the elevations −5° and 26°, and "High" for elevations 57° and 90°.

Following the recommendations given in Johnson [35], a tested effect was considered significant when the p-value was below 0.005, and only marginally significant for values $0.005 < p < 0.05$. Post-hoc tests were performed using the Tukey-Kramer method [10].

3.3. Results

3.3.1. Front-back confusions

Most of the time, subjects were rather consistent in their judgments. We found that the judgments were distributed

as follows: for a given target, the subjects either pointed in the wrong hemifield for all trials (see Figure 3, left subfigure) or just for a minor part of the repeated trials (see Figure 3, right subfigure). This behavior shows a very consistent perception of the target direction (possibly on the wrong hemifield), rather than a dispersion of pointing directions. Consequently, response directions for which confusions were detected were resolved prior to further analysis by mirroring the judgments about the frontal plane.

A repeated measures ANOVA was performed on the front-back reversal rates, with the pointing condition (H, G, P) as between-subjects factor, and the repetition (1-8), the hemifield (Front, Back), and the target elevation (−5°, 26°, 57°) as within-subjects factors. No significant trend was noticed over repetitions ($F(7, 245) = 1.53$; $p = 0.16$) and no significant effect of the condition was found ($F(2, 35) = 0.46$; $p = 0.63$). This last observation may be related to the fact that front-back confusions reflect perceptual ambiguities, and thus should not be affected by the pointing method itself. It further ensures that the applied reversal treatments do not affect the comparison between the different pointing methods. Confusion rates appeared to be significantly higher in the rear hemifield than in the frontal hemifield ($F(1, 35) = 63.54$; $p < 0.001$), which means that most of the confusions were back-to-front confusions. Furthermore, the target elevation has a significant effect on the reversal rate ($F(2, 70) = 49.75$; $p < 0.001$). The interaction effect between target elevation and hemifield was found to be significant too ($F(2, 70) = 52.07$; $p < 0.001$). This is illustrated in Figure 4. A post-hoc test showed that, only for an elevation of 57°, the reversal rates are significantly higher in the rear hemifield than in the front hemifield ($p < 0.001$).

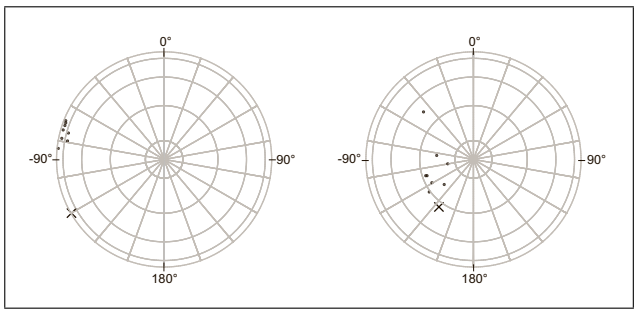


Figure 3. Top views of a schematic representation of the upper sphere (on which the loudspeakers are located). The black crosses represent the target locations and the grey points represent the pointed directions. The shown examples highlight the two different kinds of judgment distributions that reflect front-back confusions. Judgments presented here are raw data (*i.e.* the front-back confusions are not yet treated).

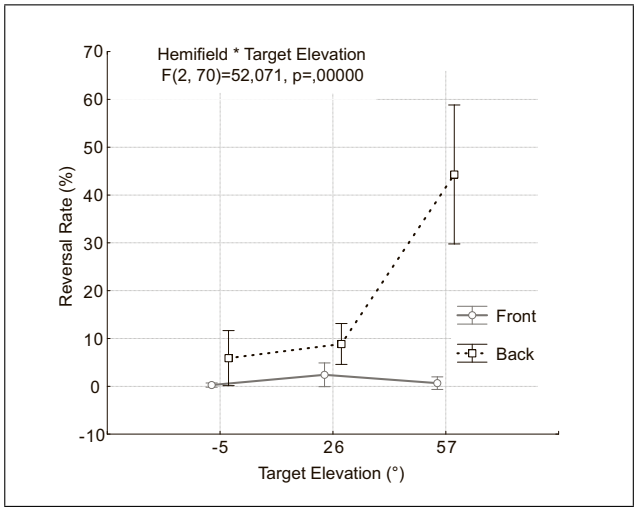


Figure 4. Reversal rate (%) observed on all subjects and all targets located in the front (solid grey line) and back (dashed black line) hemifields. Significance of the interaction effect between source elevation and hemifield is indicated on top of the figure by F and p values. Vertical bars represent standard errors.

3.3.2. Horizontal errors

A repeated measures ANOVA was performed on the unsigned horizontal errors, with the pointing condition (H, G, P) as between-subjects factor, and the repetition (1-8) and the hemifield (Front, Back) as within-subjects factors. Firstly, the effects of the repetition and of the interaction between condition and repetition were not significant ($F(7, 245) = 1.40$; $p = 0.21$ and $F(14, 245) = 1.08$; $p = 0.37$, respectively). Secondly, the condition appeared to be significant ($F(2, 35) = 11.63$; $p < 0.001$). A post-hoc test highlighted significant differences between the proximal pointing method and the other methods (P vs. H: $p = 0.004$; P vs. G: $p < 0.001$), whereas no significant differences could be observed between the head pointing and the gun pointing ($p = 0.58$). The proximal pointing condition showed, on average, larger horizontal errors than the other conditions.

In Figure 5 (left subfigure), a progressive increase of the unsigned horizontal errors from frontal to rear target source locations can be observed for all conditions. According to the ANOVA, the effect of the hemifield was shown to be significant ($F(1, 35) = 167.95$; $p < 0.001$). A significant interaction effect between condition and hemifield was also observed ($F(2, 35) = 6.37$; $p = 0.004$) and a post-hoc test revealed that the significantly greater rate of errors for the proximal pointing condition only appears for targets located in the back hemifield.

The right subfigure in Figure 5 depicts the systematic bias between the target directions and the reported directions. A repeated measures ANOVA was performed on the horizontal bias (*i.e.* the signed horizontal errors), using similar between-subjects and within-subjects factors. A significant effect of the repetition was noticed ($F(7, 245) = 2.93$; $p = 0.006$), which is linked to a slight reduction of the horizontal bias over repetitions of about 2°. The interaction effect between condition and repetition was found not to be significant ($F(14, 245) = 0.69$; $p = 0.78$). Here again, the ANOVA revealed a significant effect of the condition ($F(2, 35) = 9.91$; $p < 0.001$), and a post-hoc test showed that the results from the proximal method significantly differ from the results of other methods (P vs. H: $p = 0.005$, and P vs. G: $p < 0.001$). The hemi-

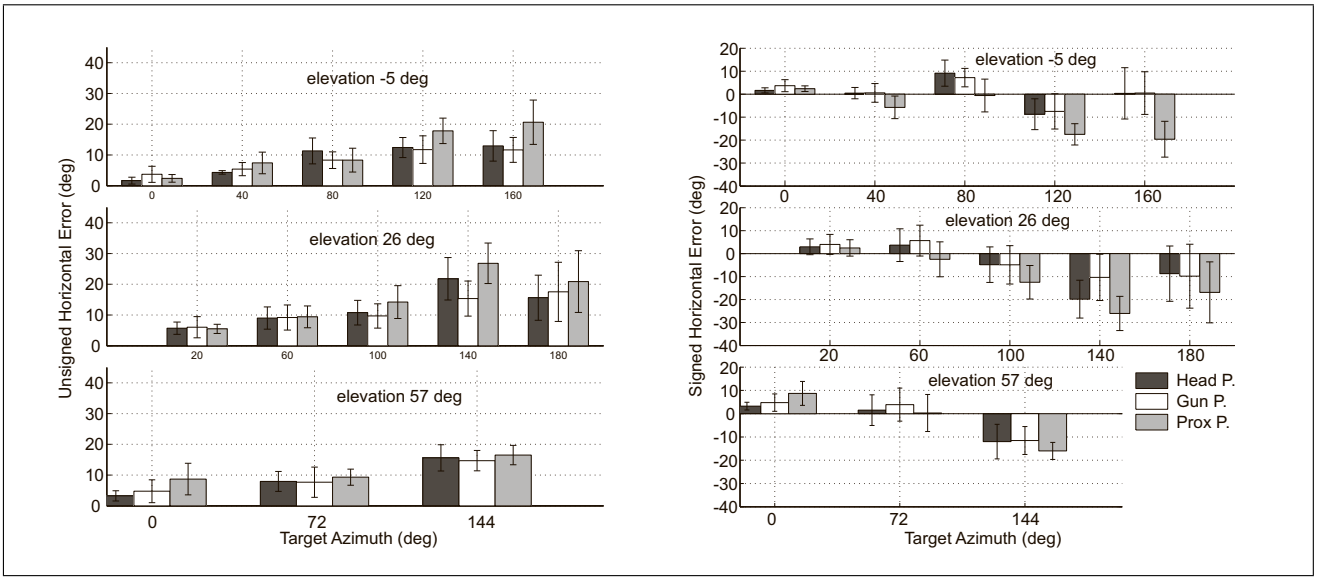


Figure 5. Unsigned (left) and signed (right) horizontal localization errors for each target elevation as a function of the target's azimuth position. Data were averaged over repetitions and subjects for each condition. For improved readability, data are combined for the left and right hemispaces. Vertical black lines show the standard deviation, and thus the dispersion of subject responses about the mean errors. It is important to note that this figure only gives a global overview over the distribution and magnitude of errors as the different averages were not taken over the same number of observations (*e.g.*, the average over the left and right hemispaces is not applied to loudspeakers located on the median plane).

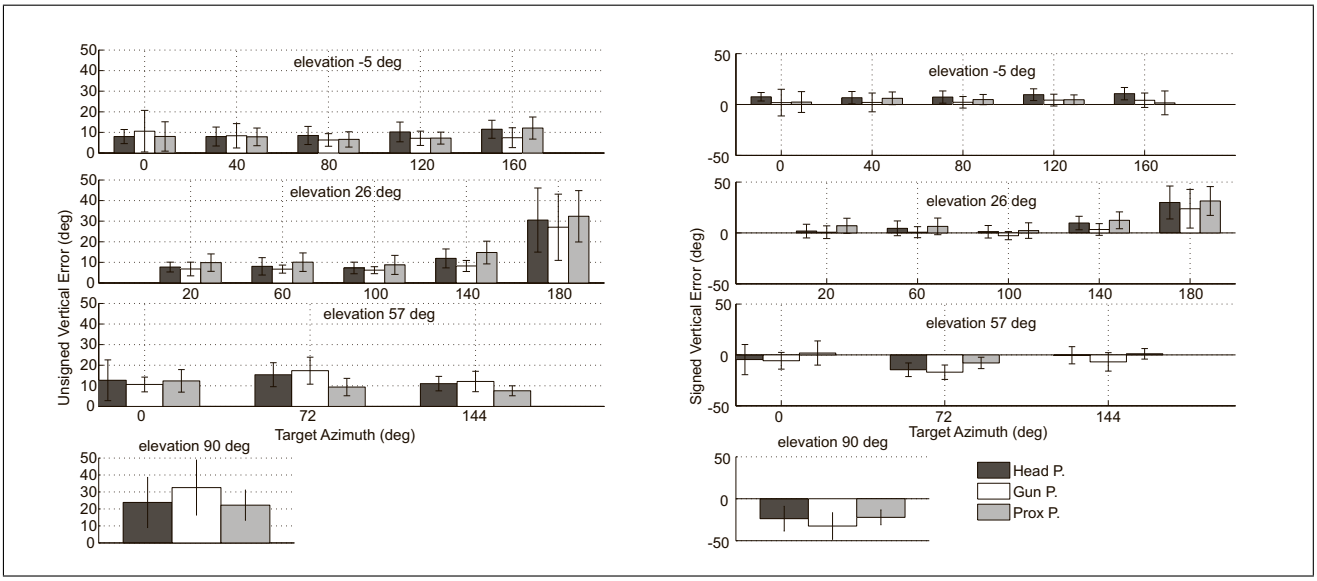


Figure 6. Unsigned (left) and signed (right) vertical localization errors for each target elevation as a function of the target's azimuth position. See legend of Figure 5 for more details.

field has a significant effect on the signed horizontal errors ($F(1, 35) = 281.48$; $p < 0.001$), as well as the interaction effect between condition and hemifield ($F(2, 35) = 6.90$; $p = 0.003$). According to a post-hoc test, the larger systematic horizontal bias of the proximal pointing method compared to the other methods only holds for targets in the back hemifield. Figure 7 shows almost no systematic bias for the proximal pointing method for frontal targets, whereas a positive systematic bias can be observed for the other methods. For rear targets, a negative systematic bias was found for all conditions, which is significantly higher for the proximal pointing method.

3.3.3. Vertical errors

A repeated measures ANOVA was performed on the unsigned vertical errors, with the pointing condition (H, G, P) as between-subjects factor, and the repetition (1-8) and the elevation range (Mid, High) as within-subjects factors. Neither the effect of the repetition ($F(7, 245) = 1.67$; $p = 0.12$) nor the interaction between condition and repetition ($F(14, 245) = 1.50$; $p = 0.11$) were found to be significant. Unsigned vertical errors do not differ significantly between the conditions ($F(2, 35) = 1.35$; $p = 0.27$), but they increase significantly from mid to high eleva-

tions ($F(1, 35) = 17.37$; $p < 0.001$). The interaction effect between condition and elevation range appeared to be marginally significant ($F(2, 35) = 4.68$; $p = 0.02$). According to the post-hoc test, and as shown in Figure 8 (left subfigure), the significant increase of the unsigned vertical errors between mid and high elevations only holds for the gun pointing method ($p < 0.001$). The proximal pointing method shows almost no difference in vertical pointing accuracy. Moreover, the proximal pointing method is more accurate in vertical pointing at high elevations than the gun pointing method ($p = 0.02$).

The same analysis was performed on the signed vertical errors. No significant effect of the repetition could be found ($F(7, 245) = 1.55$; $p = 0.15$); however, the interaction effect between repetition and condition appeared to be significant ($F(14, 245) = 4.69$; $p < 0.001$). This is linked to the fact that we observed symptoms of fatigue for the gun pointing condition. Subjects gradually decreased the pointing height with ongoing repetitions (about 6° in total). The effect of the condition was found to be marginally significant ($F(2, 35) = 3.95$; $p = 0.03$): a global negative vertical bias was noticed for the gun pointing method, whereas the average bias remained close to zero for the head pointing and proximal pointing. The dependence of the signed vertical errors on the elevation range was found to be significant ($F(1, 35) = 322.93$; $p < 0.001$): an average positive bias of +6° was found for mid elevations as well as an averaged negative bias of −10° for high elevations. In other terms, the vertical range of the pointed elevation is compressed compared to the actual elevation range of the target sources. The interaction effect between condition and elevation range is also marginally significant ($F(2, 35) = 3.64$; $p = 0.037$): a marginally significant difference in vertical pointing accuracy is noticed at the high elevation range between the proximal and gun pointing methods ($p = 0.015$). As shown in Figure 8 (right subfigure), the compression effect is lower for the proximal pointing than for the two other conditions.

3.3.4. Dispersion of responses

Following [1], the dispersion parameter κ^{-1} was investigated in order to summarize the consistency of the subjects' responses, regardless of the horizontal and vertical dimensions. The parameter κ is estimated from the length of the sum of all unit-length vectors corresponding to the directions pointed by a given subject for a given target.

$$\kappa = (N-1)^2 / N(N-R),$$

where N is the number of repetitions and R is the length of the resultant vector. The dispersion parameter κ^{-1} varies from 0 (no dispersion) to 1 (full dispersion, *i.e.* the reported directions are randomly distributed on the sphere).

A repeated measures ANOVA was performed on the dispersion, with the pointing condition (H, G, P) as between-subjects factor, and the target elevation (−5°, 26° and 57°) and the hemifield (Front, Back) as within-subjects factors. The results are depicted in Figure 9. The

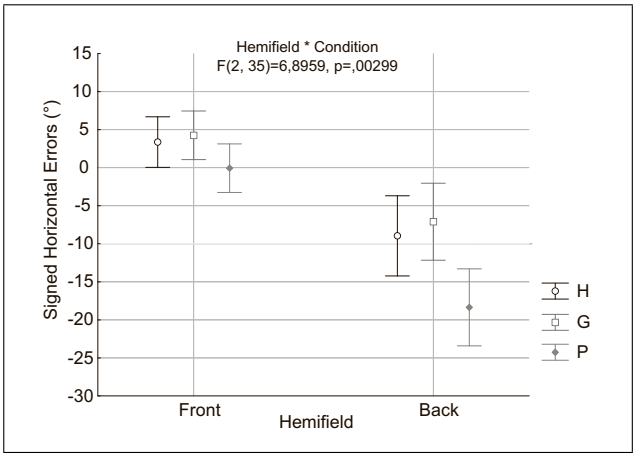


Figure 7. Signed horizontal errors averaged over subjects and repetitions. Results are presented separately for front and back hemifields and for each experimental condition (H:head, G:gun, P:proximal).

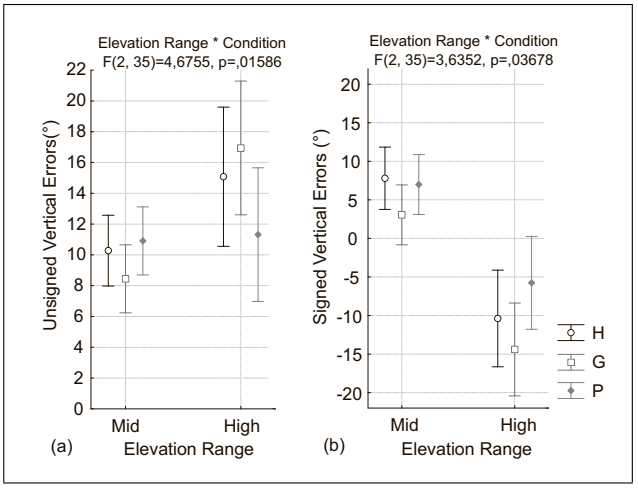


Figure 8. Unsigned (a) and signed (b) vertical errors as a function of the target elevation range. "Mid" includes the target elevations −5° and +26°, and "High" is related to 57° and 90°. Data are displayed separately for each condition (H:Head, G:Gun, P:Proximal).

condition effect was found not to be significant ($F(2, 35) = 0.75$; $p = 0.48$). The dispersion appeared to be significantly larger for rear targets than for frontal targets ($F(1, 35) = 40.05$; $p < 0.001$), and to significantly increase from low to high elevations ($F(2, 70) = 16.94$; $p < 0.001$). However, a more detailed analysis of the interactions reveals that the results of the proximal pointing are more homogeneous. A marginally significant interaction effect between condition and target elevation was found ($F(4, 70) = 2.72$; $p = 0.04$). The associated post-hoc test revealed, that the significant increase of the dispersion between frontal and rear targets only holds for the head and gun pointing methods. Moreover, the interaction effect between condition, hemifield and target elevation appeared to be significant ($F(4, 70) = 9.43$; $p < 0.001$). According to a post-hoc test, the proximal pointing method shows a significantly lower dispersion at an elevation of 57° in

the rear hemifield compared to the gun pointing method ($p = 0.002$). Note that the target elevation of 90° was excluded from the current statistical analysis because it doesn't belong to any hemifield. Then, in order to evaluate the dispersion effect for the different pointing methods at the highest elevation, an ANOVA was performed with condition as between-subjects factor and target elevation as within-subjects factor (including elevation 90°). However, no significant difference between pointing methods emerged from the associated post-hoc test at this elevation.

3.3.5. Response time

We further compared the different pointing methods with respect to the response time, *i.e.* the time delay between the playback of the stimulus and the (validated) response of a participant. In a localization test with open loop conditions (*i.e.* where the stimulus playback is stopped before the participant points at the target), the response time cannot be considered as a relevant indicator of the localization performance. The observed time differences will be solely determined by the motor task itself. However, the response time is an important parameter that may be taken into account when planning and designing a localization experiment. For this reason, this parameter was studied in the present work. A repeated measures ANOVA was performed on the response time, with the pointing condition (H, G, P) as between-subjects factor, and the repetition (1-8) and the absolute target angle as within-subjects factors. The absolute target angle refers to the magnitude of the angular distance between the frontal direction (*i.e.* the resting position) and the target position on the sphere. Firstly, the analysis results revealed a marginally significant effect of the pointing condition on the response time ($F(2, 35) = 5.46$; $p = 0.009$): we obtained average response times of 3.5 s, 3.1 s, and 2.5 s for the head pointing, gun pointing, and proximal pointing methods, respectively. Secondly, the response time decreases significantly over the repetitions ($F(7; 245) = 15.22$; $p < 0.001$). However, according to the interaction effect between condition and repetition ($F(14, 245) = 2.80$; $p < 0.001$), this was only observed for the proximal and the head pointing methods (see Figure 10). No improvement on the reaction time can be noticed for the gun pointing method.

As expected, the absolute target angle has a significant effect on the response time ($F(12, 420) = 43.54$; $p < 0.001$): there is an increase in the response time as a function of the absolute target angle. According to the interaction effect between condition and absolute target angle ($F(24, 420) = 6.40$; $p < 0.001$), this effect is well pronounced for the gun pointing method and, in particular, for the head pointing method, whereas no clear trend could be observed for the proximal condition.

3.3.6. Effect of the hand used to point

For the proximal pointing method, we examined the responses given with the participant's dominant or non-dominant hands. Note that the analysis considers only one pointing condition, which eliminates the between-subjects factor. A two-way repeated measures ANOVA

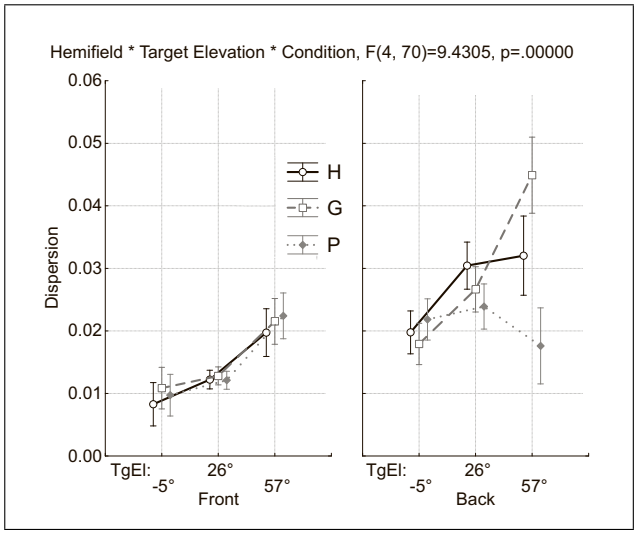


Figure 9. Response dispersion κ^{-1} as a function of the experimental condition (H:Head, G:Gun, P:Proximal), hemifield (Front, Back), and target elevation (−5°, 26°, 57°).

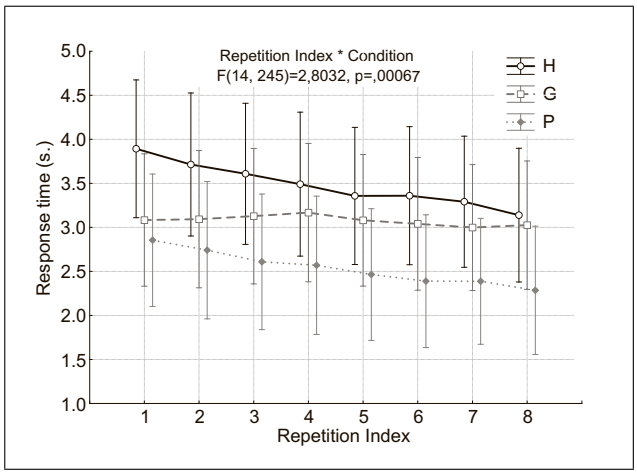


Figure 10. Evolution of the response time ($s.$) for the same target source over the repetitions. The data are displayed separately for each experimental condition (H:Head, G:Gun, P:Proximal).

was performed with the lateral dominance (Dominant, Non-dominant) and the hemifield (Front, Back) as within-subjects factors. The effect of the lateral dominance on the signed and unsigned vertical errors is not significant ($F(1, 12) = 0.24$; $p = 0.64$ and $F(1, 12) = 0.37$; $p = 0.55$, respectively). Therefore, the analysis was focused on the horizontal dimension and showed a significant effect of the lateral dominance on the unsigned horizontal errors ($F(1, 12) = 14.78$; $p = 0.002$) as well as of the hemifield ($F(1, 12) = 103.4$; $p < 0.001$). The interaction effect between those factors was found not to be significant ($F(1, 12) = 2.06$; $p = 0.18$). Regarding the signed horizontal errors, the effect of the hemifield is significant ($F(1, 12) = 188.6$; $p < 0.001$), but the lateral dominance effect ($F(1, 12) = 4.34$; $p = 0.06$) and the interaction effect between the two factors ($F(1, 12) = 0.81$; $p = 0.39$) are not significant. To summarize, the choice of the hand affects the horizontal pointing accuracy only, regardless of

the hemifield of the sound source. The averaged unsigned horizontal error is 12.3° for the dominant hand compared to 14.8° for the non-dominant hand.

4. Discussion

The primary goal of this study is to compare the proximal pointing method to more conventional pointing methods, such as the head and gun pointing. The localization accuracy was analyzed on a relatively small number of spatial directions (mostly in the upper hemisphere). It is thus essential to verify the results by comparing them to the results of previously published reference studies.

The study of Makous and Middlebrooks [12] has been selected for several reasons: Firstly, because it is based on a head pointing task for reporting the localization of real sound sources distributed over a sphere. Secondly, some of the tested positions are very close to those used in this study. Thirdly, the definition of localization errors is identical. Finally, the results of Makous and Middlebrooks' study were grounded on observations gathered over a relatively dense grid of tested directions and after a long training of the participants. For similar reasons, the results from Gilkey *et al.* [8] were used for comparisons, although they used an exocentric reporting method (GELP) in their study.

Figure 11 presents the unsigned horizontal and vertical errors observed in the two reference studies compared to the different conditions of the present experiment. A very similar trend is observed for all studies with increasing errors in the back hemifield, especially from azimuth 110° to 160° for the unsigned horizontal errors as already pointed out in [11]. Unsigned errors found in the study of Gilkey *et al.* are systematically greater than those of Makous and Middlebrooks, which confirms the advantage of egocentric pointing methods compared to exocentric ones [8]. On average, the results gathered in the different conditions of the present experiment are lying in between the two reference methods. In particular, the unsigned horizontal errors observed for the gun pointing and the head pointing conditions perfectly match the corresponding curve of the study of Makous and Middlebrooks [12], except for the head pointing condition at 140°. This result provides confidence to the present study, especially when considering the very limited training of our participants (10 directions, completed in approximately 2 minutes) compared to the study of Makous and Middlebrooks (1000 directions, with a total duration of the training period up to two hours).

The main exception to this behavior are the higher unsigned horizontal errors for targets located in the back hemifield (starting from azimuth 120°) observed for the proximal method, as previously noticed.

With regard to signed horizontal errors, an average positive bias was found in the frontal hemifield as well as an average negative bias in the back hemifield (see section 3.3.2). This general extra-lateralization of the reported directions has already been underlined in [11, 34, 10]. Moreover, the particular positive vertical bias observed for

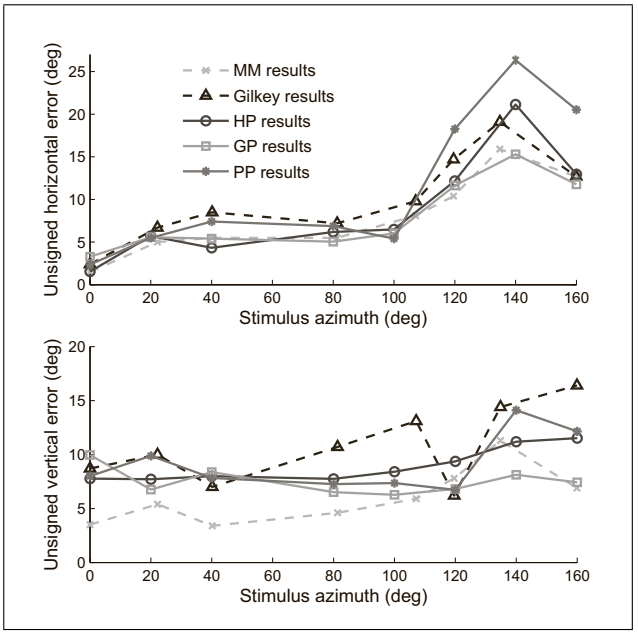


Figure 11. Unsigned horizontal errors (top) and unsigned vertical errors (bottom) as a function of target azimuth (in degrees). Data from the present study are compared to the results from Makous and Middlebrooks' open-loop experiment [12] (based on the head pointing method) and the results from Gilkey *et al.* [8] (using the GELP technique). All data were averaged over the left and right hemispaces and confusions were treated in the same way (*i.e.* removed from the data).

highly elevated sources located in the rear has been also pointed out by Oldfield and Parker [11].

Regarding the confusions, the important reversal rate observed for targets located at high elevations in the rear hemifield, is consistent with literature [12, 34]. More generally, in the present study higher confusion rates were noticed in the rear hemifield, which corresponds to a prevailing occurrence of back-to-front confusions. This observation is in accordance with a localization test conducted by Bronkhorst [36], in which he was using a reporting method similar to that one used in this study (*i.e.* head pointing and blindfolded subjects). Bronkhorst shows a predominant occurrence of front-to-back and diagonal confusions for virtual sound sources (*i.e.* the playback of binaural signals on headphones), but a predominant occurrence of back-to-front confusions for real sound sources.

Another comparison can be done with the results reported in Djelani *et al.* [9], who compared a finger pointing method with both a head pointing and a GELP pointing method, although they only provide global values of the absolute angle error and the dispersion coefficient κ^{-1} averaged over 12 directions and eight subjects. The higher accuracy provided by the egocentric methods (finger and head pointing) against the exocentric GELP technique is in good agreement with the observations reported in this article. However, some differences can also be noticed. The averaged dispersion coefficient κ^{-1} measured in their study is much higher (between 0.04 and 0.08 depending on the method) than what we observed in the present study

(between 0.01 and 0.04 according to the pointed directions and methods). A possible explanation is that their study was based on virtual sound sources rendered over headphones, whereas real sources were used in this study. Finally, we found larger horizontal errors in the back hemifield for the proximal pointing method compared to the head pointing method, which was apparently not the case in the study of Djelani *et al.* [9]. However, this may originate from an important difference in the pointing task compared to the present study. They used a closed-loop condition (*i.e.* continuous stimulus during the pointing task) for the GELP technique and the finger pointing method, whereas an open-loop condition (*i.e.* stimulus stopped before pointing) was used for the head condition. In contrast, we used an open-loop condition for all methods in order to allow for a more consistent comparison of the data. The choice of a continuous stimulus was excluded in our experiment since it would mean that under the head pointing and gun pointing conditions, the localization task would always end up with the sound target located in front of the participant [9]. Nevertheless, results of Djelani *et al.* give an indication that the relative weakness of the proximal method observed in the back hemifield could probably be overcome or at least reduced in a closed-loop condition as expected from other auditory pointing tasks [37]. It could as well benefit from a longer training.

According to the observations made in Section 3.3, the proximal method shows a lack of accuracy in the horizontal dimension for rear locations close to the median plane. The negative horizontal bias observed for these locations suggests a motor task difficulty to indicate directions in the back hemifield (task described in [9] as "uncomfortable"). In contrast, for the head and gun pointing methods, the horizontal error increase observed at rear locations may rather come from the inaccuracy induced by the large body rotations.

In comparison with the proximal pointing method, the vertical pointing accuracy of the gun pointing method is lower at high elevations, and the dispersion of judgments for rear targets located at high elevations is higher. This observation reflects the difficult motor task that combines large arm movements with large body rotations. There is also a tendency that the subjects point too low when reporting the perceived sound position with the head or a gun. This is probably caused by the physical strain experienced when pointing to high elevations with these pointing methods [9].

It was shown that the response time of the proximal pointing method is 30% and 20% shorter than that of the head and gun pointing methods, respectively. Although this advantage was partly compensated in the present study by the additional time needed for the complex repositioning procedure (note that the percentage is divided in half when considering the repositioning time), it can be taken into account when designing localization experiments. Shorter response times help to reduce the total duration of experiments, which limits the fatigue of the participants. With shorter response times, more responses can be collected with the same test duration. This strengthens the statistical results.

5. Conclusion

Egocentric pointing tasks for reporting a perceived sound source position involve moving different body segments, and the motor control may affect the localization performance. This study shows similar trends for the spatial distribution of localization errors for the three methods under investigation: a progressive increase of the horizontal errors with the azimuth, a negative horizontal bias for rear locations, and a negative vertical bias for sources located at high elevations. It further highlights some limitations specific to each pointing method. For sources located at high elevations, the head and gun pointing methods (which both involve a rotation of the whole body) exhibit a larger dispersion and a larger vertical bias than the other method. For sources located in the rear hemifield close to the median plane, the proximal pointing method presents a larger horizontal bias. However, this method has several practical advantages and interesting improvement perspectives for future 3D sound localization experiments. The proximal pointing method's shorter response times allow for collecting a larger number of judgments for a given test duration. In contrast to other pointing methods, it is also well suited for being used in a closed-loop condition, i.e., where the sound stimuli remain switched on during the pointing task. Exploiting the auditory-motor loop, combined with an extensive training, may help to improve the pointing accuracy, especially for rear locations and for responses collected with the non-dominant hand. The proximal pointing method may also be extended to assess the perceived distance of virtual sound sources played back over headphones, *e.g.*, to investigate the externalization phenomenon in binaural synthesis.

Acknowledgement

This work was funded by the French FUI project BiLi ("Binaural Listening", www.bili-project.org, FUI-AAP14) with support from Cap Digital Paris Region.

References

[1] F. L. Wightman, D. J. Kistler: Headphone simulation of free-field listening. ii: Psychophysical validation. The Journal of the Acoustical Society of America **85** (1989) 868–878.

[2] E. Wenzel, M. Arruda, D. Kistler, F. Wightman: Localization using nonindividualized head-related transfer functions. The Journal of the Acoustical Society of America **94** (1993) 111–123.

[3] M. J. Evans: Obtaining accurate responses in directional listening tests. 104th Audio Engineering Society Convention, May 1998.

[4] V. Larcher: Techniques de spatialisation des sons pour la réalité virtuelle. Dissertation. Université de Pierre et Marie Curie, Paris VI, 2001.

[5] J.-M. Pernaux, M. Emerit, R. Nicol: Perceptual evaluation of binaural sound synthesis: the problem of reporting local-

ization judgments. 114th Audio Engineering Society Convention, Amsterdam, March 2003.

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

[7] M. Schoeffler, S. Westphal, A. Adami, H. Bayerlein, J. Herre: Comparison of a 2d- and 3d-based graphical user interface for localization listening tests. Proc. of the EAA Joint Symposium on Auralization and Ambisonics, April 2014.

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

[9] T. Djelani, C. Pörschmann, J. Sahrhage, J. Blauert: An interactive virtual-environment generator for psychoacoustic research ii: Collection of head-related impulse responses and evaluation of auditory localization. Acta Acustica united with Acustica **86** (November 2000) 1046–1053.

[10] P. Majdak, M. Goupell, B. Laback: 3-d localization of virtual sound sources: effects of visual environment, pointing method, and training. Attention, Perception & Psychophysics **72** (2010) 454–469.

[11] S. R. Oldfield, S. P. A. Parker: Acuity of sound localisation: a topography of auditory space. ii. pinna cues absent. Perception **13** (1984) 601–617.

[12] J. C. Makous, J. C. Middlebrooks: Two-dimensional sound localization by human listeners. The Journal of the Acoustical Society of America **87** (1990) 2188–2200.

[13] V. Best, S. Carlile, C. Jin, A. van Schaik: The role of high frequencies in speech localization. The Journal of the Acoustical Society of America **118** (2005) 353–363.

[14] M. Frank, L. Mohr, A. Sontacchi, F. Zotter: Flexible and intuitive pointing method for 3-d auditory localization experiments. Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation, June 2010.

[15] S. Perrett, W. Noble: The contribution of head motion cues to localization of low-pass noise. Perception & Psychophysics **59** (1997) 1018–1026.

[16] R. L. Martin, K. I. McAnally, M. A. Senova: Free-field equivalent localization of virtual audio. J. Audio Eng. Soc **49** (2001) 14–22.

[17] H. Wierstorf, A. Raake, S. Spors: Localization of a virtual point source within the listening area for wave field synthesis. 133th Audio Engineering Society Convention, Oct 2012.

[18] V. Tabry, R. J. Zatorre, P. Voss: The influence of vision on sound localization abilities in both the horizontal and vertical planes. Frontiers in Psychology **4** (2013).

[19] T. Ashby, R. Mason, T. Brookes: Head movements in three-dimensional localization. 134th Audio Engineering Society Convention, May 2013.

[20] P. Paukner, M. Rothbucher, K. Diepold: Sound localization performance comparison of different hrtf-individualization methods. Dissertation. Technische Universität München, Lehrstuhl für Datenverarbeitung, April 2014.

[21] B. Seeber: A new method for localization studies. Acta Acustica united with Acustica **83** (1997) 1–2.

[22] B. Seeber: Untersuchung der auditiven lokalisation mit einer lichtzeigermethode (development and test of a new method to study auditory localization and application of it to virtual acoustics). Dissertation. Technical University of Munich, Munich, 2003.

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

[24] V. Pulkki, T. Hirvonen: Localization of virtual sources in multichannel audio reproduction. IEEE Transactions on Speech and Audio Processing **13** (2005) 105–119.

[25] E. H. A. Langendijk, A. W. Bronkhorst: Contribution of spectral cues to human sound localization. The Journal of the Acoustical Society of America **112** (2002) 1583–1596.

[26] J. Blauert: Spatial hearing: The psychophysics of human sound localization. MIT Press, 1997.

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

[28] P. M. Hofman, A. J. Van Opstal: Spectro-temporal factors in two-dimensional human sound localization. The Journal of the Acoustical Society of America **103** (1998) 2634–2648.

[29] B. F. G. Katz, G. Parseihian: Perceptually based head-related transfer function database optimization. The Journal of the Acoustical Society of America **131** (2012) EL99–EL105.

[30] M. Morimoto, H. Aokata: Localization cues of sound sources in the upper hemisphere. The Journal of the Acoustical Society of Japan (E) **5** (1984) 165–173.

[31] M. Aytekin, C. F. Moss, J. Z. Simon: A sensorimotor approach to sound localization. Neural Comput. **20** (2008) 603–635.

[32] I. Viaud-Delmon, O. Warusfel: From ear to body: the auditory-motor loop in spatial cognition. Frontiers in Neuroscience **8** (2014) 283.

[33] F. L. Wightman, D. J. Kistler: Resolution of front–back ambiguity in spatial hearing by listener and source movement. The Journal of the Acoustical Society of America **105** (1999) 2841–2853.

[34] S. Carlile, P. Leong, S. Hyams: The nature and distribution of errors in sound localization by human listeners. Hearing Research **114** (1997).

[35] V. E. Johnson: Revised standards for statistical evidence. Proceedings of the National Academy of Sciences **110** (2013) 19313–19317.

[36] A. W. Bronkhorst: Localization of real and virtual sound sources. The Journal of the Acoustical Society of America **98** (1995) 2542–2553.

[37] E. O. Boyer, B. M. Babayan, F. Bevilacqua, M. Noisternig, O. Warusfel, A. Roby-Brami, S. Hanneton, I. Viaud-Delmon: From ear to hand: the role of the auditory-motor loop in pointing to an auditory source. Frontiers in Computational Neuroscience **7** (2013).

Bibliographie

[AAD01] Algazi, V. R., Avendano, C., and Duda, R. O. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3) :1110–1122, 2001.

[AAK12] Aussal, M., Alouges, F., and Katz, B. Hrtf interpolation and itd personalization for binaural synthesis using spherical harmonics. In *Audio Engineering Society Conference : UK 25th Conference : Spatial Audio in Today's 3D World*, Mar 2012.

[ABK15] Andreopoulou, A., Begault, D. R., and Katz, B. F. G. Inter-laboratory round robin hrtf measurement comparison. *IEEE Journal of Selected Topics in Signal Processing*, 9(5) :895–906, 2015.

[ADA99] Avendano, C., Duda, R. O., and Algazi, V. R. Modeling the contralateral hrtf. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, Mar 1999.

[ADTA01] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102, 2001.

[AK15] Andreopoulou, A. and Katz, B. F. G. On the use of subjective hrtf evaluations for creating global perceptual similarity metrics of assessors and assessees. In *21st International Conference on Auditory Display*, July 2015.

[AK16] Andreopoulou, A. and Katz, B. F. G. Subjective hrtf evaluations for obtaining global similarity metrics of assessors and assessees. *Journal on Multimodal User Interfaces*, pages 1–13, 2016.

[AMB13] Ashby, T., Mason, R., and Brookes, T. Head movements in three-dimensional localization. In *134th Audio Engineering Society Convention*, May 2013.

[AMS08] Aytekin, M., Moss, C. F., and Simon, J. Z. A sensorimotor approach to sound localization. *Neural Comput.*, 20(3) :603–635, 2008.

[AMS13] Andéol, G., Macpherson, E. A., and Sabin, A. T. Sound localization in noise and sensitivity to spectral shape. *Hearing Research*, 304 :20–27, 10 2013.

[AR14] Andreopoulou, A. and Roginska, A. Evaluating hrtf similarity through subjective assessments : Factors that can affect judgment. In *Internaional Conference on Audio Display*, 2014.

[ARB13] Andreopoulou, A., Roginska, A., and Bello, J. P. Reduced representations of hrtf datasets : A discriminant analysis approach. In *135th Audio Engineering Society Convention*, Oct 2013.

[ASG15] Andéol, G., Savel, S., and Guillaume, A. Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources. *Frontiers in Neuroscience*, 8 :451, 2015.

[Bas03] Baskind, A. *Modèles et méthodes de description spatiale de scènes sonores, application aux enregistrement binauraux*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2003.

[BB77] Butler, R. A. and Belendiuk, K. Spectral cues utilized in the localization of sound in the median sagittal plane. *The Journal of the Acoustical Society of America*, 61(5) :1264–1269, 1977.

[BBB+13] Boyer, E. O., Babayan, B. M., Bevilacqua, F., Noisternig, M., Warusfel, O., Roby-Brami, A., Hanneton, S., and Viaud-Delmon, I. From ear to hand : the role of the auditory-motor loop in pointing to an auditory source. *Frontiers in Computational Neuroscience*, 7(26), 2013.

[BCJvS05] Best, V., Carlile, S., Jin, C., and van Schaik, A. The role of high frequencies in speech localization. *The Journal of the Acoustical Society of America*, 118(1) :353–363, 2005.

[BCNW16] Bahu, H., Carpentier, T., Noisternig, M., and Warusfel, O. Comparison of different egocentric pointing methods for 3d sound localization experiments. *Acta acustica united with Acustica*, 102 :107–118, 2016.

[BDPW13] Bertet, S., Daniel, J., Parizet, E., and Warusfel, O. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta acustica united with Acustica*, 4-99 :642–657, 2013.

[Ber14] Bernard, B., Mathieu. *Active audition and sensorimotor integration for a bioinspired autonomous robot*. PhD thesis, Université Pierre et Marie Curie - Paris VI, May 2014.

[BK01] Breebaart, J. and Kohlrausch, A. Perceptual (ir)relevance of hrtf magnitude and phase spectra. In *Audio Engineering Society Convention 110*, May 2001.

[Bla97] Blauert, J. *Spatial Hearing : The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[BML13] Baumgartner, R., Majdak, P., and Laback, B. Assessment of sagittal-plane sound localization performance in spatial-audio applications. In Blauert, J., editor, *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pages 93–119. Springer Berlin Heidelberg, 2013.

[BML14] Baumgartner, R., Majdak, P., and Laback, B. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2) :791–802, 2014.

[BN03] Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6) :1373–1396, June 2003.

[BR99] Brungart, D. S. and Rabinowitz, W. M. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3) :1465–1479, 1999.

[BR00] Berg, J. and Rumsey, F. Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. In *109th AES Convention*, September 2000.

[Bre13] Breebaart, J. Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis. *The Journal of the Acoustical Society of America*, 133(1) :EL1–EL6, 2013.

[Bro95] Bronkhorst, A. W. Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, 98(5) :2542–2553, 1995.

[Bus06] Busson, S. *Individualization of acoustic cues for binaural synthesis*. PhD thesis, Université de la Méditerranée - Aix-Marseille II, Jan 2006.

[BvWvO10] Bremen, P., van Wanrooij, M. M., and van Opstal, A. J. Pinna cues determine orienting response modes to synchronous sounds in elevation. *The Journal of Neuroscience*, 30(1) :194–204, 2010.

[BW97] Blommer, M. A. and Wakefield, G. H. Ieee transactions on speech and audio processing. *Pole-zero approximations for head-related transfer functions using a logarithmic error criterion*, 5(3) :278–287, 1997.

[BWA01] Begault, D. R., Wenzel, E. M., and Anderson, M. R. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc*, 49(10) :904–916, 2001.

[CB13] Carlile, S. and Blackman, T. Relearning auditory spectral cues for locations inside and outside the visual field. *JARO : Journal of the Association for Research in Otolaryngology*, 15(2) :249–263, 04 2013.

[CBK14] Carlile, S., Balachandar, K., and Kelly, H. Accommodating to new ears : The effects of sensory and sensory-motor feedback. *The Journal of the Acoustical Society of America*, 135(4) :2002–2011, 2014.

[CBNW14] Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O. Measurement of a head-related transfer function database with high spatial resolution. In *Forum Acusticum*, 2014.

[CJvR00] Carlile, S., Jin, C., and van Raad, V. Continuous virtual auditory space using hrtf interpolation : acoustic and psychophysical errors. pages 220–223, 2000.

[CK05] Colburn, H. S. and Kulkarni, A. *Models of Sound Localization*. Springer New York, 2005.

[CLH97] Carlile, S., Leong, P., and Hyams, S. The nature and distribution of errors in sound localization by human listeners. *Hearing Research*, 114(1-2), 1997.

[CP94] Carlile, S. and Pralong, D. The location,Äêdependent nature of perceptually salient features of the human head,Äêrelated transfer functions. *The Journal of the Acoustical Society of America*, 95(6) :3445–3459, 1994.

[CSL03] C.-S., F. and Lo, Y.-C. On the clustering of head-related transfer functions used for 3-d sound localization. *J. Inf. Sci. Eng.*, 19(1) :141–157, 2003.

[DAA99] Duda, R. O., Avendano, C., and Algazi, V. R. An adaptable ellipsoidal head model for the interaural time difference. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 965–968 vol.2, 1999.

[DC77] Domnitz, R. H. and Colburn, H. S. Lateral position and interaural discrimination. *The Journal of the Acoustical Society of America*, 61(6) :1586–1598, 1977.

[DPK96] Dau, T., Pschel, D., and Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, 99(6) :3615–3622, 1996.

[DPSB00] Djelani, T., Pörschmann, C., Sahrhage, J., and Blauert, J. An interactive virtual-environment generator for psychoacoustic research ii : Collection of head-related impulse responses and evaluation of auditory localization. *Acta Acustica united with Acustica*, 86(6) :1046–1053, November 2000.

[DR05] Duraiswami, R. and Raykar, V. C. The manifolds of spatial hearing. In *In Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 285–288, 2005.

[DRP+92] Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M. On the externalization of auditory images. *Presence : Teleoper. Virtual Environ.*, 1(2) :251–257, May 1992.

[DS03] Dhillon, I. S. and Sra, S. Modeling data using directional distributions. Technical report, 2003.

[DW02] Durant, E. A. and Wakefield, G. H. Efficient model fitting using a genetic algorithm : pole-zero approximations of hrtfs. *IEEE Transactions on Speech and Audio Processing*, 10(1) :18–27, 2002.

[Eva98] Evans, M. J. Obtaining accurate responses in directional listening tests. In *104th Audio Engineering Society Convention*, May 1998.

[Far00] Farina, A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th Audio Engineering Society Convention*, Feb 2000.

[FCP08] Fischer, B. J., Christianson, G., and Peña, J. Cross-correlation in the auditory coincidence detectors of owls. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(32) :8107–8115, 08 2008.

[FMB14] F., R., M., E., and B., K. Evaluation objective et subjective de différentes méthodes de lissage des hrtf. In *Congres Francais d'Acoustique*, pages 2215–2221, June 2014.

[FMSZ10] Frank, M., Mohr, L., Sontacchi, A., and Zotter, F. Flexible and intuitive pointing method for 3-d auditory localization experiments. In *Audio Engineering Society Conference : 38th International Conference : Sound Quality Evaluation*, June 2010.

[FR15] F. Rugeles, J. D., M. Emerit. A fast measurement of high spatial resolution head related transfer functions for the bili project. In *International Conference on Spatial Audio 2015*, 2015.

[FZ06] Fastl, H. and Zwicker, E. *Psychoacoustics : Facts and Models*. Springer-Verlag New York, Inc., 2006.

[GGE+95] Gilkey, R., Good, M., Ericson, M., Brinkman, J., and Stewart, J. A pointing technique for rapidly collecting localization responses in auditory research. *Behavior Research Methods, Instruments, & Computers*, 27(1) :1–11, 1995.

[GM90] Glasberg, B. R. and Moore, B. C. J. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47 :103–138, 1990.

[GMGF14] Grijalva, F., Martini, L., Goldenstein, S., and Florencio, D. Anthropometric-based customization of head-related transfer functions using isomap in the horizontal plane. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, pages 4473–4477, May 2014.

[GML10] Goupell, M. J., Majdak, P., and Laback, B. Median-plane sound localization as a function of the number of spectral channels using a channel vocoder. *The Journal of the Acoustical Society of America*, 127(2) :990–1001, 02 2010.

[Grö02] Gröhn, M. Localization of a moving virtual sound source in a virtual room : The effect of a distracting auditory stimulus. In *Stimulus, in International Conference on Auditory Display*, 2002.

[Gui09] Guillon, P. *Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*. PhD thesis, Université du Maine, Juin 2009.

[HG10] Hugeng, W. W. and Gunawan, D. Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements. *Journal of Telecommunications*, 2 :31–41, May 2010.

[HOR99] Hofman, P. M., Opstal, A. J. V., and Riswick, J. G. A. V. Relearning sound localization with new ears. *The Journal of the Acoustical Society of America*, 105, 1999.

[HPP08] Hwang, S., Park, Y., and Park, Y.-s. Modeling and customization of head-related impulse responses based on general basis functions in time domain. *Acta Acustica united with Acustica*, 94(6), 2008.

[HVO98] Hofman, P. M. and Van Opstal, A. J. Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103(5) :2634–2648, 1998.

[HVO03] Hofman, P. and Van Opstal, A. Binaural weighting of pinna cues in human sound localization. *Experimental Brain Research*, 148(4) :458–470, 2003.

[HW74a] Hebrank, J. and Wright, D. Are two ears necessary for localization of sound sources on the median plane ? *The Journal of the Acoustical Society of America*, 56(3) :935–938, 1974.

[HW74b] Hebrank, J. and Wright, D. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6) :1829–1834, 1974.

[HZK99] Huopaniemi, J., Zacharov, N., and Karjalainen, M. Objective and subjective evaluation of head-related transfer function filter design. *J. Audio Eng. Soc*, 47(4) :218–239, 1999.

[HZMW08] Hu, H., Zhou, L., Ma, H., and Wu, Z. Hrtf personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics*, 69(2) :163 – 172, 2008.

[HZZ+06] Hu, H., Zhou, L., Zhang, J., Ma, H., and Wu, Z. Head related transfer function personalization based on multiple regression analysis. In *2006 International Conference on Computational Intelligence and Security*, volume 2, pages 1829–1832, 2006.

[Iwa06] Iwaya, Y. Individualization of head-related transfer functions with individualization of head-related transfer functions with tournament-style listening test : Listening with other's ears. *Acoustical Society of Japan*, 27(6) :340–343, 2006.

[JCCvS04] Jin, C., Corderoy, A., Carlile, S., and van Schaik, A. Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America*, 115(6) :3124–3141, 2004.

[Jef48] Jeffress, L. A. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1) :35–39, 1948.

[JLL+00] Jin, C., Leong, P., Leung, J., Corderoy, A., and Carlile, S. Enabling individualized virtual auditory space using morphological measurements. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing*, pages 235–238, 2000.

[Joh13] Johnson, V. E. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48) :19313–19317, 2013.

[JSC00] Jin, C., Schenkel, M., and Carlile, S. Neural system identification model of human sound localization. *The Journal of the Acoustical Society of America*, 108(3) :1215–1235, 2000.

[Kah00] Kahana, Y. *Numerical modelling of the head-related transfer function.* Original typescript, 2000.

[Kas15] Kasarapu, P. Modelling of directional data using kent distributions. *CoRR*, abs/1506.08105, 2015.

[Kat98] Katz, B. F. G. *Measurement and calculation of individual Head-Related Transfer Functions using a boundary element model including the measurement and effect of skin and hair impedance.* PhD thesis, 1998.

[KC98] Kulkarni, A. and Colburn, H. S. Role of spectral detail in sound-source localization. *Nature*, 396(6713) :747–749, 12 1998.

[KC00] Kulkarni, A. and Colburn, H. S. Variability in the characterization of the headphone transfer-function. *The Journal of the Acoustical Society of America*, 107(2) :1071–1074, 2000.

[KC05] Kim, S.-M. and Choi, W. On the externalization of virtual sound images in headphone reproduction : A wiener filter approach. *The Journal of the Acoustical Society of America*, 117(6) :3657–3665, 2005.

[KD08] Keyrouz, F. and Diepold, K. A new hrtf interpolation approach for fast synthesis of dynamic environmental interaction. *J. Audio Eng. Soc*, 56(1/2) :28–35, 2008.

[KIC99] Kulkarni, A., Isabelle, S. K., and Colburn, H. S. Sensitivity of human subjects to head-related transfer-function phase spectra. *The Journal of the Acoustical Society of America*, 105(5) :2821–2840, 1999.

[KMKK08] Kapralos, B., Mekuz, N., Kopinska, A., and Khattak, S. Dimensionality reduced hrtfs : a comparative study, 2008.

[KN14] Katz, B. F. and Noisternig, M. A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, 135(6) :3530–3540, 2014.

[KP12] Katz, B. F. G. and Parseihian, G. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, 131(2) :EL99–EL105, 2012.

[KW92] Kistler, D. J. and Wightman, F. L. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3) :1637–1647, 1992.

[Lar01] Larcher, V. *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Université de Pierre et Marie Curie, Paris VI, 2001.

[LB00] Langendijk, E. H. A. and Bronkhorst, A. W. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America*, 107(1) :528–537, 2000.

[LB02] Langendijk, E. H. A. and Bronkhorst, A. W. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4) :1583–1596, 2002.

[LBCP10] Le Bagousse, S., Colomes, C., and Paquier, M. State of the art on subjective assessment of spatial sound quality. In *Audio Engineering Society Conference : 38th International Conference : Sound Quality Evaluation*, Jun 2010.

[LCB+05] Lemaire, V., Clerot, F., Busson, S., Nicol, R., and Choqueuse, V. Individualized hrtfs from few measurements : a statistical learning approach. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2041–2046, July 2005.

[LH13] Li, L. and Huang, Q. Hrtf personalization modeling based on rbf neural network. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3707–3710, 2013.

[LJV98] Larcher, V., Jot, J.-M., and Vandernoot, G. Equalization methods in binaural technology. In *Audio Engineering Society Convention 105*, Sept 1998.

[LKW01] Langendijk, E. H. A., Kistler, D. J., and Wightman, F. L. Sound localization in the presence of one or two distracters. *The Journal of the Acoustical Society of America*, 109(5) :2123–2134, 2001.

[LpL11] Lee, K.-S. and pil Lee, S. A relevant distance criterion for interpolation of head-related transfer functions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6) :1780–1790, 2011.

[MA84] Morimoto, M. and Aokata, H. Localization cues of sound sources in the upper hemisphere. *The Journal of the Acoustical Society of Japan (E)*, 5(3) :165–173, 1984.

[Maa16] Maazaoui, O., M. et Warusfel. Estimation des hrtfs individuelles sur la base d'enregistrements binauraux en conditions non-contrôllées. In *CFA*, 2016.

[MBL14] Majdak, P., Baumgartner, R., and Laback, B. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Frontiers in Psychology*, 5(319), 2014.

[MCD+12] Mendonça, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P., and Santos, J. A. On the improvement of localization accuracy with non-individualized hrtf-based sounds. *J. Audio Eng. Soc*, 60(10) :821–830, 2012.

[MF05] Maki, K. and Furukawa, S. Reducing individual differences in the external-ear transfer functions of the mongolian gerbil. *The Journal of the Acoustical Society of America*, 118(4) :2392–2404, 2005.

[MG96] Moore, B. C. J. and Glasberg, B. R. A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2) :335–345, 1996.

[MGL10] Majdak, P., Goupell, M., and Laback, B. 3-d localization of virtual sound sources : effects of visual environment, pointing method, and training. *Attention, Perception & Psychophysics*, 72(2) :454–469, 2010.

[MHJS95] Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. Transfer characteristics of headphones measured on human ears. *J. Audio Eng. Soc*, 43(4) :203–217, 1995.

[MIC+13] Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H., and Noisternig, M. Spatially oriented format for acoustics : A data exchange format representing head-related transfer functions. In *134th Audio Engineering Society Convention*, May 2013.

[Mid92] Middlebrooks, J. C. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5) :2607–2624, 1992.

[Mid99a] Middlebrooks, J. C. Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of The Acoustical Society of America*, 106, 1999.

[Mid99b] Middlebrooks, J. C. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3) :1493–1510, 1999.

[Mil58] Mills, A. W. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4) :237–246, 1958.

[MJHS96] Møller, H., Jensen, C. B., Hammershøi, D., and Sørensen, M. F. Using a typical human subject for binaural recording. In *Audio Engineering Society Convention 100*, May 1996.

[MM90] Makous, J. C. and Middlebrooks, J. C. Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5) :2188–2200, 1990.

[MM02a] Macpherson, E. A. and Middlebrooks, J. C. Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5) :2219–2236, 2002.

[MM02b] McAnally, K. I. and Martin, R. L. Variability in the headphone-to-ear-canal transfer function. *J. Audio Eng. Soc*, 50(4) :263–266, 2002.

[MMG89] Middlebrooks, J. C., Makous, J. C., and Green, D. M. Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1) :89–108, 1989.

[MMO00] Middlebrooks, J. C., Macpherson, E. A., and Onsan, Z. A. Psychophysical customization of directional transfer functions for virtual sound localization. *The Journal of the Acoustical Society of America*, 108(6) :3088–3091, 2000.

[MMS01] Martin, R. L., McAnally, K. I., and Senova, M. A. Free-field equivalent localization of virtual audio. *J. Audio Eng. Soc*, 49(1/2) :14–22, 2001.

[MN82] Morimoto, M. and Nomachi, K. Binaural disparity cues in median-plane localization. *Journal of the Acoustical Society of Japan (E)*, 3(2) :99–103, 1982.

[MNFC14] Mattes, S., Nelson, P. A., Fazi, F. M., and Capp, M. Exploration of a biologically inspired model for sound source localization in 3d space. In *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, pages 1–7, April 2014.

[Mor01] Morimoto, M. The contribution of two ears to the perception of vertical angle in sagittal planes. *The Journal of the Acoustical Society of America*, 109(4) :1596–1603, 2001.

[MPC05] Minnaar, P., Plogsties, J., and Christensen, F. Directional resolution of head-related transfer functions required in binaural synthesis. *J. Audio Eng. Soc*, 53(10) :919–929, 2005.

[MS07] Macpherson, E. A. and Sabin, A. T. Binaural weighting of monaural spectral cues for sound localization. *The Journal of the Acoustical Society of America*, 121(6) :3677–3688, 2007.

[MSJH96] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. Binaural technique : Do we need individual recordings ? *J. Audio Eng. Soc*, 44(6) :451–469, 1996.

[MWL13] Majdak, P., Walder, T., and Laback, B. Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *The Journal of the Acoustical Society of America*, 134(3) :2148–2159, 2013.

[NLBB06] Nicol, R., Lemaire, V., Bondu, A., and Busson, S. Looking for a relevant similarity criterion for hrtf clustering : A comparative study. In *Audio Engineering Society Convention 120*, May 2006.

[NYS92] Neti, C., Young, E. D., and Schneider, M. H. Neural network models of sound localization based on directional filtering by the pinna. *Journal of the Acoustical Society of America*, 92(6) :3140–3156, 1992.

[Ols72] Olson, H. F. The measurement of loudness. *Audio*, pages 18–22, 1972.

[OP84] Oldfield, S. R. and Parker, S. P. A. Acuity of sound localisation : a topography of auditory space. ii. pinna cues absent. *Perception*, 13(5) :601–617, 1984.

[PC96] Pralong, D. and Carlile, S. The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America*, 100, 1996.

[PEN03] Pernaux, J.-M., Emerit, M., and Nicol, R. Perceptual evaluation of binaural sound synthesis : the problem of reporting localization judgments. In *114th Audio Engineering Society Convention*, Amsterdam, March 2003.

[PH05] Pulkki, V. and Hirvonen, T. Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, 13(1) :105–119, 2005.

[PK12] Parseihian, G. and Katz, B. F. G. Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America*, 131(4) :2948–2957, 2012.

[PN97a] Perrett, S. and Noble, W. The contribution of head motion cues to localization of low-pass noise. *Perception & Psychophysics*, 59(7) :1018–1026, 1997.

[PN97b] Perrett, S. and Noble, W. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4) :2325–2332, 1997.

[PNK08] Park, M., Nelson, P. A., and Kang, K. A model of sound localisation applied to the evaluation of systems for stereophony. *Acta Acustica united with Acustica*, 94(6) :825–839, 2008.

[PNW+12] Pollow, M., Nguyen, K.-V., Warusfel, O., Carpentier, T., Müller-Trapet, M., Vorländer, M., and Noisternig, M. Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition. *Acta Acustica united with Acustica*, 98(1) :72–82, 2012.

[PRD14] Paukner, P., Rothbucher, M., and Diepold, K. *Sound Localization Performance Comparison of Different HRTF-Individualization Methods*. PhD thesis, Technische Universität München, Lehrstuhl für Datenverarbeitung, April 2014.

[PRH+92] Patterson, R. D., Robinson, K., Holdsworth, J., Mckeown, D., Zhang, C., and Allerhand, M. Complex sounds and auditory images. In *Proceedings of the 9th International Symp Hearing Audit., Physiol. Perception9th International Symp Hearing Audit., Physiol. Perception*, 1992.

[Raf05] Rafaely, B. Analysis and design of spherical microphone arrays. *Speech and Audio Processing, IEEE Transactions on*, 13(1) :135–143, 2005.

[RBA+10] RH., S., B., N., A., H., J., B., J., B., and KL., L. Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound : cluster analysis and an experimental study. *Ergonomics*, 53(6) :767–81, June 2010.

[RDSD10] Rothbucher, M., Durkovic, M., Shen, H., and Diepold, K. Hrtf customization using multiway array analysis. *18th European Signal Processing Conference*, pages 229–233, 2010.

[RE02] Rio E., W. O. Optimization of multi-channel binaural formats based on statistical analysis. In *Forum Acusticum*, 2002.

[ROEK15] Rugeles Ospina, F., Emerit, M., and Katz, B. F. The three-dimensional morphological database for spatial hearing research of the bili project. *Proceedings of Meetings on Acoustics*, 23(1), 2015.

[RS00] Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290(2323–2326), 2000.

[RY05] Reiss, L. A. J. and Young, E. D. Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. *The Journal of Neuroscience*, 25(14) :3680–3691, 2005.

[RYW00] Runkle, P., Yendiki, A., and Wakefield, G. H. Active sensory tuning for immersive spatialized audio. In *Proc. ICAD*, 2000.

[SBCD75] Searle, C. L., Braida, L. D., Cuddy, D. R., and Davis, M. F. Binaural pinna disparity : another auditory localization cue. *The Journal of the Acoustical Society of America*, 57(2) :448–455, 1975.

[SC05] Scarpaci, J. W. and Colburn, H. S. Principal components analysis interpolation of head related transfer functions using locally,Äêchosen basis functions. *The Journal of the Acoustical Society of America*, 117(4) :2561–2562, 2005.

[See97] Seeber, B. A new method for localization studies. *Acta Acustica united with Acustica*, 83 :1–2, 1997.

[See03] Seeber, H., Bernhard U ; Fastl. Subjective selection of non-individual head-related transfer functions. In *In Proceedings of the 2003 International Conference on Auditory Display*, pages 1–4, July 2003.

[SFK08] Schonstein, D., Ferr,àö©, L., and Katz, B. F. Comparison of headphones and equalization for virtual auditory source localization. *The Journal of the Acoustical Society of America*, 123(5) :3724–3724, 2008.

[SHH94] Shimada, S., Hayashi, N., and Hayashi, S. A clustering method for sound localization transfer functions. *J. Audio Eng. Soc*, 42(7/8) :577–584, 1994.

[SK12] Schönstein, D. and Katz, B. F. Variability in perceptual evaluation of hrtfs. *J. Audio Eng. Soc*, 60(10) :783–793, 2012.

[SP14] Smith, R. C. G. and Price, S. R. Modelling of human low frequency sound localization acuity demonstrates dominance of spatial variation of interaural time difference and suggests uniform just-noticeable differences in interaural time difference. *PLoS ONE*, 9(2) :e89033, 2014.

[SR03] Saul, L. K. and Roweis, S. T. Think globally, fit locally : Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4 :119–155, Dec. 2003.

[SWA+14] Schoeffler, M., Westphal, S., Adami, A., Bayerlein, H., and Herre, J. Comparison of a 2d- and 3d-based graphical user interface for localization listening tests. In *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, April 2014.

[TBK12] Tame, R. P., Barchiesi, D., and Klapuri, A. Headphone virtualisation : Improved localisation and externalisation of nonindividualised hrtfs by cluster analysis. In *Audio Engineering Society Convention 133*, Oct 2012.

[TdSL00] Tenenbaum, J. B., de Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *American Association for the Advancement of Science*, 290 :2319–2323, 2000.

[TG98] Tan, C.-J. and Gan, W.-S. User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems. *Electronics Letters*, 34, 1998.

[TUN99] Toyama, M., Uchiyama, M., and Nomura, H. Head related transfer function representation of directional sound for spatial acoustic events modeling. In *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on*, pages 221–226, 1999.

[TZV13] Tabry, V., Zatorre, R. J., and Voss, P. The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in Psychology*, 4(932), 2013.

[vdMPvdH08] van der Maaten, L., Postma, E. O., and van den Herik, H. J. Dimensionality reduction : A comparative review, 2008.

[VHF08] Völk, F., Heinemann, F., and Fastl, H. Externalization in binaural synthesis : effects of recording environment and measurement procedure. *The Journal of the Acoustical Society of America*, 123(5) :3935–3935, 2008.

[VVO04] Vliegen, J. and Van Opstal, A. J. The influence of duration and level on human sound localization. *The Journal of the Acoustical Society of America*, 115(4) :1705–1713, 2004.

[WAKW93] Wenzel, E., Arruda, M., Kistler, D., and Wightman, F. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1) :111–123, 1993.

[Wal40] Wallach, H. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4) :339, 1940.

[Wal10] Walder, T. Sound-source localization through warped head-related transfer functions. Master's thesis, University of Music and Performing Arts, Graz, Austria, 2010.

[WK89] Wightman, F. L. and Kistler, D. J. Headphone simulation of free-field listening. ii : Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2) :868–878, 1989.

[WK92] Wightman, F. L. and Kistler, D. J. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3) :1648–1661, 1992.

[WK99] Wightman, F. L. and Kistler, D. J. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5) :2841–2853, 1999.

[WK05] Wightman, F. and Kistler, D. Measurement and validation of human hrtfs for use in hearing research. *Acta Acustica united with Acustica*, 91(3), 2005.

[WRS12] Wierstorf, H., Raake, A., and Spors, S. Localization of a virtual point source within the listening area for wave field synthesis. In *133th Audio Engineering Society Convention*, Oct 2012.

[WWKF88] Wenzel, E., Wightman, F., Kistler, D., and Foster, S. Acoustic origins of individual differences in sound localization behavior. *The Journal of the Acoustical Society of America*, 84(S1) :S79–S79, 1988.

[XDNXB13] Xie, B., Dr. Ning Xiang, R., and Blauert, J. . . *Head-Related Transfer Function and Virtual Auditory Display : Second Edition*. J Ross Publishing, 2013.

[XZH15] Xie, B., Zhong, X., and He, N. Typical data and cluster analysis on head-related transfer functions from chinese subjects. *Applied Acoustics*, 94 :1 – 13, 2015.

[XZZ13] Xie, B., Zhang, C., and Zhong, X. A cluster and subjective selection-based hrtf customization scheme for improving binaural reproduction of 5.1 channel surround sound. In *Audio Engineering Society Convention 134*, May 2013.

[Zah02] Zahorik, P. Auditory display of sound source distance. In *Proceedings of the 2002 International Conference on Auditory Displays*, pages 326–332, 2002.

[ZBS+06] Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K., and Tam, C. Perceptual recalibration in human sound localization : Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America*, 120(1) :343–359, 2006.

[ZC93] Zakarauskas, P. and Cynader, M. S. A computational theory of spectral cue localization. *The Journal of the Acoustical Society of America*, 94(3) :1323–1331, 1993.

[ZDD02] Zotkin, D. N., Duraiswami, R., and Davis, L. S. Customizable auditory displays. *Proc. 2002 International Conference on Auditory Displays*, pages 167–176, 2002.

[ZH10] Zhang, P. X. and Hartmann, W. M. On the ability of human listeners to distinguish between front and back. *Hearing research*, 260(1-2), 2010.

[ZKM15] Ziegelwanger, H., Kreuzer, W., and Majdak, P. Mesh2hrtf : An open-source software package for the numerical calculation of head-related transfer functions. In *Proceedings of the 22nd International Congress on Sound and Vibration (ICSV)*, pages 1–8. ICSV 2015, 2015.

[ZM14] Ziegelwanger, H. and Majdak, P. Modeling the direction-continuous time-of-arrival in head-related transfer functions. *The Journal of the Acoustical Society of America*, 135(3) :1278–1293, 2014.

[ZS15] Zhong, X. and Shi, B. Reliability of headphone equalization in virtual sound reproduction. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on*, volume 2, pages 87–90, 2015.

[Zwi61] Zwicker, E. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33(2) :248–248, 1961.

[ZX09] Zhong, X.-L. and Xie, B.-S. Maximal azimuthal resolution needed in measurements of head-related transfer functions. *The Journal of the Acoustical Society of America*, 125(4) :2209–2220, 2009.