Université du Maine

THESE

pour obtenir le grade de

Docteur de l'Université du Maine

Discipline : ACOUSTIQUE

présentée et soutenue publiquement

par

Pierre GUILLON

le 11 Juin 2009

Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF

Jury :

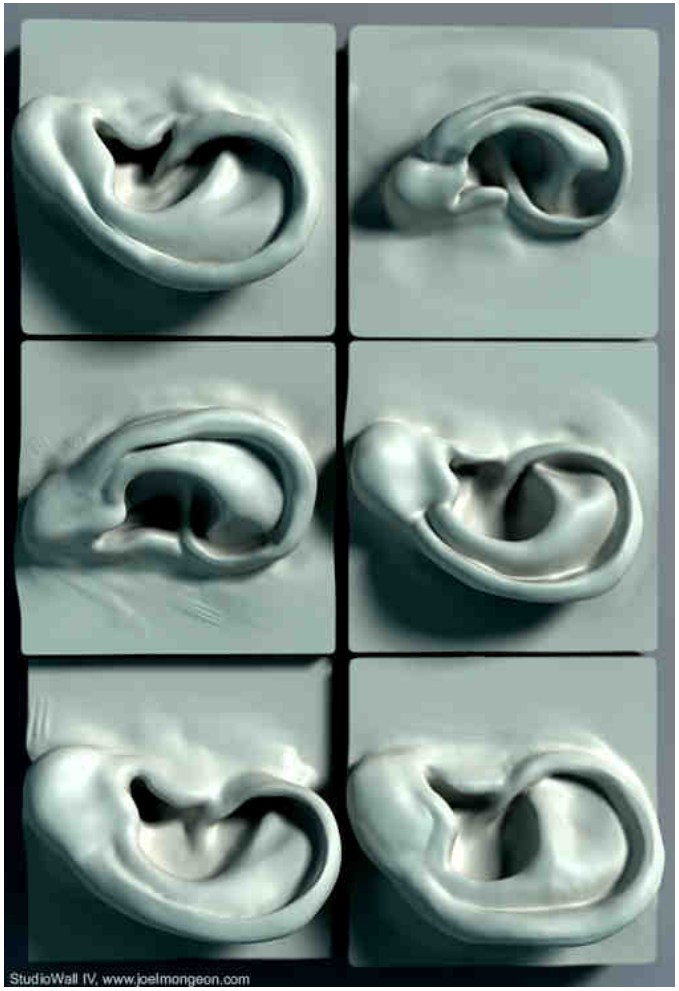
MM.	Enrique LOPEZ-POVEDA	Rapporteur
	Ewan MACPHERSON	Rapporteur
	Laurent SIMON	Directeur de thèse
	Olivier WARUSFEL	Examinateur
	Mathieu PAQUIER	Examinateur
Mlle	Rozenn NICOL	Encadrante

Sommaire

Abréviations

ACP	Analyse en Composantes Principales
ARMA	Auto-Regressive Moving Average
BEM	Boundary Element Method
BRIR	Binaural Room Impulse Response
CPA	Covert Peak Area
DCN	Dorsal Cochlear Nucleus
DTF	Directional Transfer Function
FIR	Finite Impulse Response
HPTF	HeadPhone Transfer Function
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
IBEM	Inverse Boundary Element Method
IC	Inferior Colliculus
ICP	Iterative Closest Point
IIR	Infinite Impulse Response
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
IS	Indices Spectraux
ISD	Interaural Spectral Difference
ISSD	Inter-Subject Spectral Difference
ITD	Interaural Time Difference
JND	Just Noticeable Difference
LSO	Lateral Superior Olive
MLE	Maximum Likelihood Estimation
MSO	Medial Superior Olive
OCR	Optimal Coordinate Rotation
OSF	Optimal Scaling Factor
RANSAC	RANdom SAmple Consensus
RMSE	Round Mean Sqaure Error
SC	Superior Colliculus

SFRS	Spatial Frequency Response Surface
SO(3)	Groupe des rotations
SRF	Spatial Receptive Field
STPS	Spherical Thin Plate Splines
SVD	Singular Value Decomposition
VAS	Virtual Auditory Space


StudioWall IV, www.joelmongeon.com

Introduction

Les systèmes de réalité virtuelle ont pour objectif de créer des environnements immersifs convaincants, en combinant des représentations visuelle, sonore et haptique. Historiquement, ce sont les aspects graphiques, et les systèmes de rendu stéréoscopiques qui ont initialement mobilisé les principaux efforts de recherche. Il s'avère néanmoins que la modalité auditive est tout aussi importante que le canal visuel pour fournir à l'utilisateur la sensation d'être réellement plongé dans le monde virtuel. On nomme *Virtual Auditory Space*, ou VAS, la partie acoustique des environnements virtuels, dont le but est de créer l'illusion d'un espace sonore naturel entourant l'auditeur. Au-delà des questions classiques de fidélité, relatives à toute diffusion de signaux audio, la composante majeure dans la génération de VAS est l'aspect spatial : il s'agit notamment d'assurer une localisation correcte des objets sonores virtuels. On peut ainsi considérer la spatialisation sonore comme le pendant de la production d'images en relief dans le canal visuel des environnements virtuels.

Techniquement, plusieurs stratégies sont disponibles pour générer une scène sonore spatialisée. On peut choisir de travailler sur la pression acoustique à générer dans un volume plus ou moins étendu de l'espace d'écoute, dans lequel l'auditeur viendra simplement se plonger. C'est le but de l'holophonie, équivalent acoustique de l'holographie, qui repose sur le principe de Huygens. On parle classiquement de *Wave Field Synthesis* ou WFS, qui n'est qu'un cas particulier de l'holophonie, et qui nécessite l'utilisation d'un réseau de haut-parleurs. L'approche ambisonique s'apparente à l'holophonie : basée sur une décomposition de la pression acoustique en une série de Fourier-Bessel, son objectif est aussi d'atteindre une reconstruction physique fidèle. Ses faibles performances de reconstruction aux hautes fréquences sont compensées grâce à des stratégies basées sur des aspects psychoacoustiques. Cette discipline est aujourd'hui connue au travers de HOA (*Higher Order Ambisonics*), qui constitue sa mise en œuvre la plus prometteuse. Enfin on distingue les technologies dites "binaurales", basées sur le constat que seuls deux signaux acoustiques captés aux deux oreilles suffisent pour percevoir l'espace au quotidien. La synthèse binaurale, qui est le cadre de notre étude, vise à sculpter les signaux à présenter au niveau des tympans, en leur associant tous les indices de localisation présents en situation

d'écoute naturelle. La diffusion sur casque ou écouteurs est particulièrement adaptée pour un contrôle fin de ces signaux. Si la mise en œuvre de la synthèse binaurale est réalisée avec soin, l'auditeur perçoit les sons comme provenant de sources nettement éloignées de sa tête, dans des directions bien définies. Cette perception illusoire permet d'approcher au mieux les sensations d'une écoute naturelle.

Du fait de leur capacité à simuler de façon efficace des situations variées, les environnements virtuels s'étendent à de nombreux champs d'application. La synthèse binaurale est particulièrement appropriée pour la mise en œuvre des VAS dans les exemples suivants. Dans le domaine des jeux vidéos, tous les progrès techniques visent à favoriser l'immersion du joueur dans l'univers synthétique qui lui est proposé. Le son spatialisé y contribue grandement, et c'est actuellement un des axes majeurs de recherche. Les environnements virtuels sont aussi utilisés dans une optique d'apprentissage ou d'entraînement à une tâche qui serait coûteuse ou dangereuse à mettre en place en réalité, comme la simulation du pilotage d'un avion, ou d'un sous-marin. Dans un but thérapeutique, on emploie désormais des systèmes de réalité virtuelle pour aider des patients phobiques à dépasser leurs angoisses. Il s'agit de désensibiliser les patients en les exposant aux situations anxiogènes. L'immersion dans un monde virtuel est une solution contrôlable pour le thérapeute, et ainsi plus confortable pour le patient. Enfin, en architecture, les techniques de réalité virtuelle sont d'un intérêt capital pour guider la conception d'un bâtiment, dès lors qu'on y inclut des techniques d'auralisation, c'est-à-dire des outils numériques permettant de simuler la réponse acoustique d'une salle en cours de conception.

Les applications de la synthèse binaurale dépassent le cadre strict des environnements immersifs. Dans le domaine des télécommunications, les conférences audio peuvent aisément bénéficier de cette technique. La spatialisation sonore permet de percevoir les voix des participants comme s'ils étaient répartis autour de l'auditeur, et ainsi d'assurer une bonne identification des locuteurs, et un meilleur confort dans la communication. On peut aussi envisager la spatialisation comme une dimension supplémentaire pour diffuser des informations dans un environnement qui en est déjà surchargé. Ainsi, dans un cockpit d'avion, la diffusion d'un signal sonore spatialisé peut accompagner l'apparition d'un signal d'alerte lumineux, et permettre d'améliorer le temps de réaction du pilote. Enfin, si sa mise au point se base sur une certaine connaissance des mécanismes et des performances de la localisation auditive, la synthèse binaurale constitue elle-même un outil d'investigation flexible et puissant pour étudier les fondements de la perception auditive de l'espace.

Selon l'application considérée, le réalisme de l'espace sonore n'est pas forcément le but à atteindre, mais des informations directionnelles assez élémentaires peuvent suffire. Dans notre étude, on se place dans le cas le plus contraignant : on cherche à assurer à un auditeur l'illusion parfaite qu'il se trouve immergé dans un espace réaliste.

En synthèse binaurale, quarante ans de recherches ont permis de cerner l'essentiel des conditions requises pour permettre cette illusion. Il s'agit de produire fidèlement, au niveau des tympans de l'auditeur, la pression sonore qui y aurait été engendrée dans la réalité par la source sonore à spatialiser. On utilise pour cela des filtres linéaires qui intégrent toutes les transformations subies par une onde sonore entre la source et les tympans : diffraction par la tête, réflexions sur le buste, résonances des pavillons. Les HRTF (*Head-Related Transfer Function*) sont les fonctions de transfert acoustiques qui portent en elles tous ces phénomènes, et c'est leur connaissance, pour chaque individu, qui permet de générer les filtres binauraux appropriés.

Un obstacle subsiste cependant pour le déploiement au grand public de cette technique : il faudrait idéalement mesurer les HRTF pour chaque utilisateur, car les phénomènes acoustiques à l'origine de leur formation sont très liés à la morphologie. Malheureusement, la mesure acoustique des HRTF est longue, coûteuse, et inconfortable, car elle doit être réalisée pour de nombreuses directions de l'espace. Nos travaux de recherche visent à proposer des solutions alternatives d'individualisation pour la synthèse binaurale afin d'offrir à chaque auditeur des filtres adaptés à sa perception. Plus précisément, ce sont les colorations spectrales des HRTF, appelées indices spectraux, que nous cherchons à reproduire : principalement générées par les résonances du pavillon, elles permettent la localisation auditive des sources sonores en élévation, et varient largement d'un individu à l'autre.

Malgré les différences apparentes, d'un individu à l'autre, entre les indices spectraux, il existe des comportements communs. Ces similarités sont toutefois susceptibles d'être masquées par deux sources morphologiques de variabilité : d'une part des différences de taille des pavillons d'oreille, et d'autre part des différences d'orientation spatiale de ces pavillons. C'est l'idée majeure qui a guidé nos travaux de recherche. Nous proposons des outils permettant de dépasser ces différences apparentes, afin de se focaliser sur ce qui est vraiment spécifique à chaque individu. Deux solutions techniques d'individualisation des HRTF sont développées en utilisant avantageusement la diversité des comportements offerte par les HRTF d'une base de données.

La première solution, développée au chapitre 5, consiste à adapter pour un nouvel auditeur un jeu de HRTF issues d'une base de données. C'est la comparaison morphologique entre ce nouvel auditeur et le propriétaire des HRTF qui permet de guider les transformations à appliquer aux HRTF pour qu'elles lui conviennent au mieux. La seconde solution, décrite au chapitre 6, permet de reconstruire les HRTF d'un auditeur dans une direction quelconque, à partir d'un nombre réduit de mesures individuelles. Afin de dépasser les performances des techniques classiques d'interpolation, la méthode proposée utilise des informations *a priori* issues de l'analyse d'une large base de données. La validation subjective de cette technique est finalement l'occasion de proposer une nouvelle méthode d'évaluation, basée sur l'utilisation de la

synthèse binaurale dynamique.

Les principes et les réflexions qui ont mené à l'émergence de ces solutions reposent sur les nombreux résultats physiques et psychophysiques disponibles dans la littérature. Ces résultats sont présentés dans les trois premiers chapitres. Le chapitre 1 rappelle les principes fondamentaux de la localisation auditive. Le principe de la synthèse binaurale est décrit au chapitre 2. L'état des connaissances sur les indices spectraux, qui sont au cœur de l'individualisation, est exposé au chapitre 3. Le lecteur dispose alors de l'ensemble des informations nécessaires pour aborder le problème de l'individualisation des indices spectraux, qui est posé au chapitre 4. On y montre en quoi il est nécessaire d'individualiser les HRTF, avant d'exposer les différentes solutions de l'état de l'art.

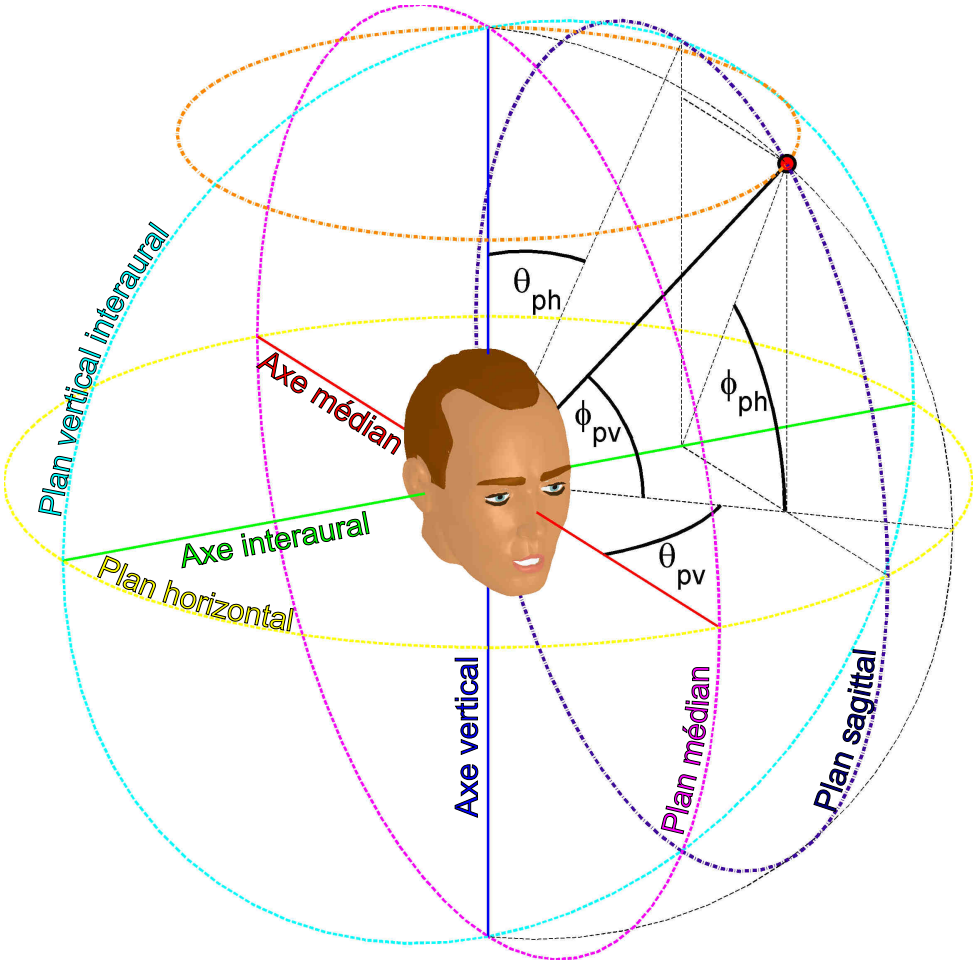


Figure 1 – Systèmes de coordonnées utilisés. Dans le système *Polaire Vertical*, les lieux d'élévations ϕ_{pv} constantes correspondent à des cercles centrés autour de l'axe vertical, inscrits dans des plans parallèles au plan horizontal. L'élévation ϕ_{pv} décrit l'intervalle $[-\pi/2; \pi/2]$ de bas en haut. L'azimut θ_{pv} décrit l'intervalle $[0; 2\pi]$: 0 correspond à la direction frontale, $\pi/2$ à gauche, π à l'arrière, $3\pi/2$ à droite (référentiel égocentrique). Dans le système *Polaire Horizontal*, les lieux d'azimuts θ_{ph} constants correspondent à des cercles centrés autour de l'axe interaural, inscrits dans des plans sagittaux, i.e. parallèles au plan médian. L'azimut θ_{ph} décrit l'intervalle $[-\pi/2; \pi/2]$: $\pm\pi/2$ correspondent à l'axe interaural, 0 au plan médian. L'élévation ϕ_{ph} décrit l'intervalle $[0; 2\pi]$, en parcourant les hémisphères avant et arrière : 0 correspond au plan horizontal avant, $\pi/2$ au plan vertical interaural supérieur, π au plan horizontal arrière.

"*[...] Mon corps, en fait, il est toujours ailleurs. Il est lié à tous les ailleurs du monde, et à vrai dire il est ailleurs que dans le monde. Car c'est autour de lui que les choses sont disposées, c'est par rapport à lui, et par rapport à lui comme par rapport à un souverain, qu'il y a un dessus, un dessous, une droite, une gauche, un avant, un arrière, un proche, un lointain. Le corps est le point zéro du monde. Là où les chemins et les espaces viennent se croiser. Le corps, il n'est nulle part. Il est au coeur du monde, ce petit noyau utopique à partir duquel je rêve, je parle, j'avance, j'imagine, je perçois les choses en leur place et je les nie aussi par le pouvoir indéfini des utopies que j'imagine. Mon corps, il est comme la cité du soleil, il n'a pas de lieu. Mais c'est de lui que sortent et que rayonnent tous les lieux possibles réels ou utopiques.*" Michel Foucault, Utopies et hétérotopies, 1966 [71].

Chapitre 1

Fondements de la localisation auditive

Ce premier chapitre est consacré à la présentation des fondements de la localisation auditive, dont la connaissance est essentielle pour la mise au point de systèmes de spatialisation sonore. Si la localisation auditive est possible, c'est que les tympans gauche et droit captent un même évènement sonore à des positions différentes de l'espace, et que le corps de l'auditeur, plongé dans le champ sonore, transforme celui-ci. Dès lors que ces transformations dépendent de la position de la source sonore, il en résulte l'apparition d'indices acoustiques de localisation, c'est-à-dire des quantités objectives dont le système auditif peut extraire l'information spatiale. La formation du percept de localisation est cependant le résultat d'une intégration multisensorielle : l'environnement visuel, ainsi que les connaissances et les attentes de l'auditeur contribuent également. L'effet ventriloque est un exemple probant de cette intégration [106] : il résulte d'un conflit entre la localisation de l'évènement auditif et des signaux visuels. La localisation de l'événement général - par exemple une marionnette qui parle - est alors déterminée préférentiellement par la localisation du stimulus visuel, c'est-à-dire la bouche de la marionnette bouge, et pas celle du ventriloque. Conscient de cette complexité, nous détaillons ici uniquement les indices acoustiques de localisation, ceux que les expériences physiques et psychophysiques peuvent quantifier et/ou contrôler, afin d'en décrire les variations et les conditions d'exploitation par le système auditif.

1.1 Indices interauraux

La théorie dite *duplex*, introduite par Lord Rayleigh [216], se fonde sur le constat suivant : "la différence principale entre les deux oreilles est le fait qu'elles ne sont pas au même endroit". Ainsi la captation par les deux oreilles d'une même onde sonore donne lieu à des indices dits "interauraux" : l'ITD (*Interaural Time Difference*) et l'ILD (*Interaural Level Difference*).

1.1.1 La différence interaurale de temps (ITD)

Pour toute source située en dehors du plan médian, il existe une différence de trajet acoustique entre la source et chaque oreille, et donc un retard, résumé par le concept unique d'ITD. Cet indice recouvre cependant deux mécanismes du système auditif, agissant sur deux plages fréquentielles distinctes.

Si l'on considère une onde plane sinusoïdale de pulsation ω, il existe une différence de phase entre les signaux captés à chaque oreille, appelée IPD (*Interaural Phase Difference*). Kuhn a étudié l'IPD d'un modèle de tête sphérique en tenant compte des phénomènes de diffraction combinant ondes incidentes et réfléchies à la surface d'une sphère [125] (cf. Fig. 1.1). Si l'on note a le rayon de la sphère équivalente à

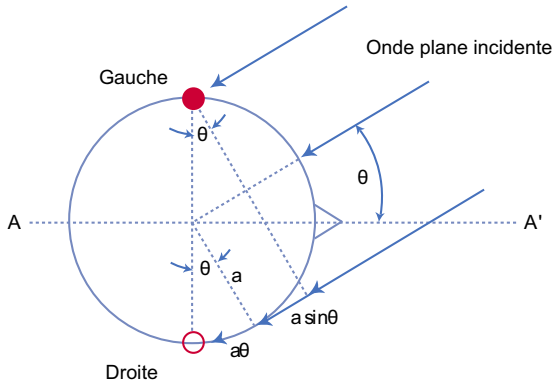


Figure 1.1 – Modèle de tête sphérique : a est le rayon de la sphère et θ l'azimut, i.e. l'angle d'incidence de l'onde plane (d'après [233]).

la tête, λ la longueur d'onde, c la célérité du son dans l'air, $k = 2\pi/\lambda$ le nombre d'onde, et θ l'azimut de la source, on peut approcher l'IPD aux basses fréquences ($(ka)^2 << 1$) par la relation :

$$IPD_{BF} = 3ka.\sin\theta \tag{1.1}$$

Cette différence de phase mène à l'expression suivante de l'ITD pour les basses fréquences :

$$ITD_{BF} = \frac{IPD_{BF}}{\omega} = 3\frac{a}{c}.\sin\theta \tag{1.2}$$

Ce modèle, corroboré par la mesure, prédit une ITD constante avec la fréquence jusqu'à environ 500 Hz à 600 Hz, ce qui constitue un codage simple de l'azimut d'une source. Un tel indice peut donc être facilement identifié par le système auditif. Cependant au delà de 600 Hz, plusieurs aspects suggèrent que l'ITD basée sur une différence de phase est difficilement exploitable [225]. D'abord, elle varie de façon non monotone avec la fréquence : un minimum est observé entre 1000 Hz et 2000 Hz, et selon l'azimut il évolue de façon assez variable aux plus hautes fréquences (cf. Fig. 1.2). Par ailleurs, une même différence de phase peut être engendrée par un grand nombre de valeurs de retards interauraux. Cette ambiguïté ne peut être levée que pour des signaux de période supérieure au double de l'ITD maximal, c'est à dire en dessous de 1500 Hz environ [177]. Enfin, les capacités du système nerveux central à encoder des différences de phase pour des sons purs sont limitées aux basses fréquences [192, 289]. Aux plus hautes fréquences, c'est le retard d'enveloppe entre les signaux captés aux oreilles gauche et droite qui constitue l'indice temporel le plus plausible. L'ITD est alors bien approchée par une modélisation sphérique de la tête d'un auditeur ne tenant compte que de la différence de marche entre les oreilles de

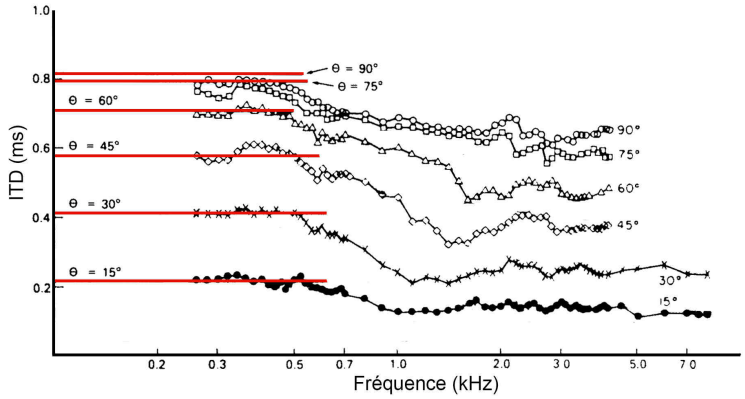


Figure 1.2 – ITD liée à un retard interaural de phase, d'après les mesures de Kuhn sur une tête artificielle [125]. Les lignes rouges correspondent aux valeurs asymptotiques du modèle de tête sphérique aux basses fréquences (cf. équation 1.2).(d'après [125]).

l'auditeur (modèle de Woodworth [278]).

$$ITD_{HF} = \frac{a}{c}(\sin\theta + \theta) \tag{1.3}$$

Plusieurs études, utilisant des stimuli bien spécifiques appelés "stimuli hautes fréquences transposés", et dont la distribution d'énergie est centrée autour de 4 kHz, montrent que l'information de la latéralisation d'une source peut être efficacement décodée à partir de l'enveloppe de tels signaux [13–15]. La sensibilité du système auditif à des retards d'enveloppe décroît cependant pour des fréquences supérieures à 4 kHz [171], et la pertinence perceptive de cet indice est discutée, dans la mesure où il n'apparaît suffisamment saillant que pour des fluctuations fortes de l'enveloppe [144]. La coexistence de ces deux types d'ITD asymptotiques suggère néanmoins que le système auditif utilise deux mécanismes distincts pour l'exploitation des indices temporels de localisation.

Selon chacun des modèles - basses et hautes fréquences - l'ITD ne dépend que de l'azimut de la source, ce qui met en évidence l'existence de loci particuliers : ce sont les cônes de confusion, introduits par Woodworth, sur lesquels l'ITD serait constante. Mathématiquement, ce sont des hyperboloïdes de révolution, dont les oreilles constituent les foyers. Pour une distance donnée de la source par rapport au centre de la tête, les loci d'ITD constante sont des cercles inscrits dans des plans sagittaux (cf. Fig. 1.3). Cependant la forme de la tête est plus ovale que sphérique, ce qui a un impact pour les positions de source les plus latéralisées, c'est-à-dire pour des directions qui s'approchent de l'axe interaural : l'ITD n'est donc pas constante pour un azimut donné, mais varie légèrement avec l'élévation. Ces variations sont aussi liées au fait que les oreilles ne sont pas diamétralement opposées de part et d'autre de la tête [38]. On représente figure 1.4 les lignes de niveau iso-ITD sur la sphère.

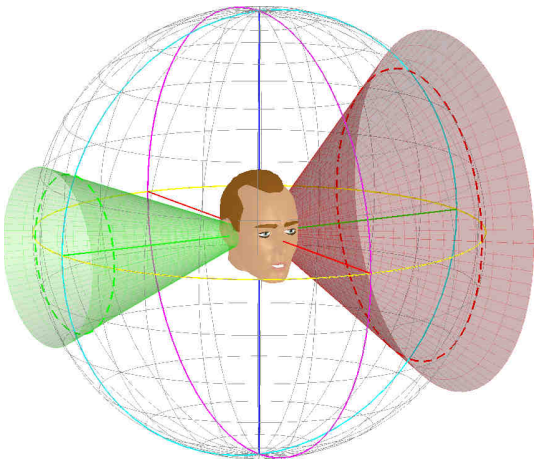


Figure 1.3 – Illustration de la notion de cône de confusion. Les hyperboloïdes correspondent chacune aux positions de source qui génèrent une ITD constante selon le modèle de Woodworth (une valeur d'ITD pour l'hyperboloïde rouge, une autre pour la verte). A mesure que l'on s'éloigne de l'auditeur, ces surfaces ressemblent de plus en plus à des cônes. Pour une distance donnée de la source par rapport à l'auditeur, les lignes iso-ITD sont des cercles centrés sur l'axe interaural, inscrits dans des plans sagittaux (cercles en pointillés).

On observe effectivement que les lignes iso-ITD sont excentrées par rapport à l'axe interaural, et qu'elles ne sont pas parfaitement circulaires. Cependant l'ITD reste un indice largement ambigu selon l'élévation, et par ailleurs, il demeure indépendant de la distance de la source par rapport à l'auditeur, même pour des sources proches [32]. L'ITD porte donc essentiellement l'information du degré de latéralisation de la source sonore (azimut dans le système polaire horizontal, cf. Fig. 1.3).

1.1.2 La différence interaurale d'intensité (ILD)

La tête d'un auditeur agit comme un obstacle face à une onde acoustique incidente. Il en résulte des différences d'intensité entre les sons captés à chaque oreille, qui dépendent de la position de la source. L'ILD est l'indice interaural de localisation qui traduit ces différences. Une modélisation simple de la tête comme une sphère rigide permet d'appréhender les différents phénomènes physiques à l'origine de l'ILD (cf. Fig. 1.5). Aux basses fréquences ($\lambda \simeq a$), l'ILD est très faible, car la tête diffracte peu l'onde incidente. Pour des fréquences croissantes ($\lambda < a$), la présence de la tête induit une combinaison complexe de perturbations réflectives et diffractives, dont résulte une distribution de pression acoustique dépendant à la fois de l'angle d'incidence et de la fréquence. Enfin pour des longueurs d'onde très inférieures à la distance interaurale ($\lambda << a$), la surface de la sphère réfléchit parfaitement l'onde incidente, et il en résulte un gain asymptotique de 6 dB pour

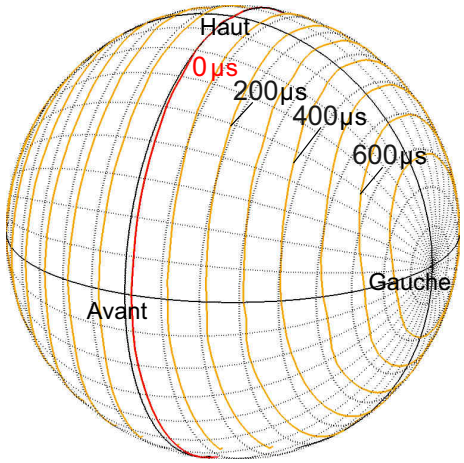


Figure 1.4 – Evolution de l'ITD aux basses fréquences sur la sphère. Les lignes relient les directions correspondant à une même valeur d'ITD. La ligne rouge représente l'iso-ITD $0\mu s$, les lignes sont espacées de 100 μs. (Estimation par régression linéaire de la phase aux basses fréquences des HRTF du sujet n°1 de la base privée de Orange Labs. L'ITD est la différence des retards monauraux gauche et droit, qui sont les pentes des droites de régression des phases des HRTF rsp. gauche et droite pour une même direction).

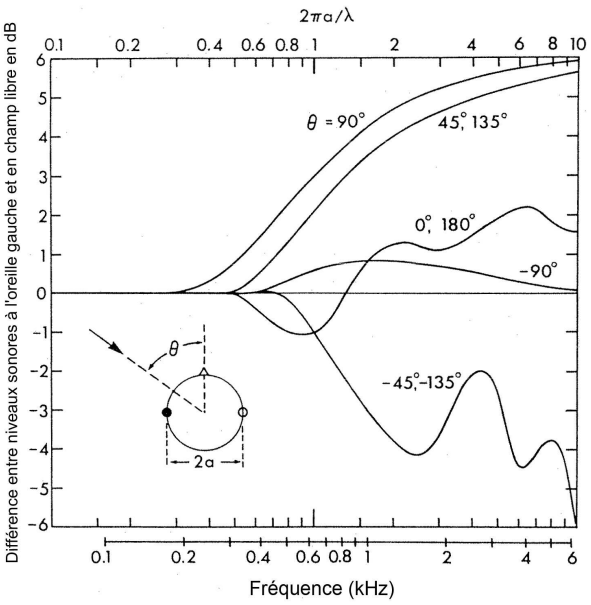


Figure 1.5 – Différence entre le niveau perçu à une oreille et le niveau mesuré au centre de la tête en l'absence de celle-ci, en fonction de l'angle d'incidence de acoustique (modèle de tête sphérique, d'après [233]).

une source située dans l'axe de l'oreille ipsilatérale. A ces considérations simples viennent s'ajouter aux hautes fréquences les phénomènes spécifiques induits par les pavillons. Les différences interaurales de niveau dépendent donc de façon complexe de la fréquence, et ces variations prennent des valeurs significatives aux hautes fréquences. Cependant, comme le suggère Blauert [19], le système auditif n'évalue pas chaque détail des dissimilarités interaurales, mais tire les informations nécessaires à partir d'attributs nets et facilement reconnaissables [1]. Ainsi, comme le montrent plusieurs expériences psychophysiques (cf. 3.2.8), l'analyse de l'ILD consiste plutôt en l'intégration fréquentielle des différences de niveau observées dans des bandes de fréquences discrètes [144]. On résume donc avantageusement les indices interauraux de niveau par une ILD indépendante de la fréquence, selon l'expression proposée par Larcher [132] :

$$ILD = 10\,log_{10}\frac{\int_{\omega_1}^{\omega_2}|\Phi_L(j\omega)|^2 d\omega}{\int_{\omega_1}^{\omega_2}|\Phi_R(j\omega)|^2 d\omega} \tag{1.4}$$

où Φ_L et Φ_R sont respectivement les pressions acoustiques aux tympans gauche et droit, pour une direction donnée, et ω_1 et ω_2 sont les bornes fréquentielles de la bande utile. On représente figure 1.6 les variations de l'ILD ainsi définie selon la position de la source (intégration sur la bande de fréquence [1.5 kHz - 10 kHz]). Les lignes de niveau iso-ILD montrent une dépendance marquée de l'ILD selon l'azimut de la source, mais révèlent une ambiguïté en fonction de l'élévation, bien que l'évolution de l'ILD sur un plan sagittal soit plus complexe que celle observée pour l'ITD. L'ILD apparaît comme un indice potentiel de latéralisation, précisément dans la zone de fréquence où l'ITD devient moins saillante. Des expériences psychophysiques révèlent néanmoins que l'ILD ne serait utilisée par le système auditif que pour la localisation de signaux dépourvus d'énergie aux basses fréquences, et que son poids perceptif est limité dans le cas de signaux large bande [47]. Ces résultats correspondent à des positions de sources en champ lointain ($r > 10a$). En champ proche la nature sphérique des ondes sonores doit être considérée. En particulier, pour des distances inférieures à 1 m, des différences de niveau conséquentes apparaissent aux basses fréquences (jusqu'à 20 dB à 0.12 m) [34], constituant des indices potentiels de localisation en distance pour des sources proches [32] (cf. Fig. 1.7).

1.1.3 Limites de la théorie duplex

On représente figure 1.8 la superposition des lignes iso-ITD et iso-ILD. Il apparaît nettement que l'utilisation exclusive de ces deux indices ne permet pas au

1. *"...the system does not evaluate every detail of the complicated interaural dissimilarities, but rather derives what information is needed from definite, easily recognizable attributes"* Blauert, 1997, p. 138 [19].

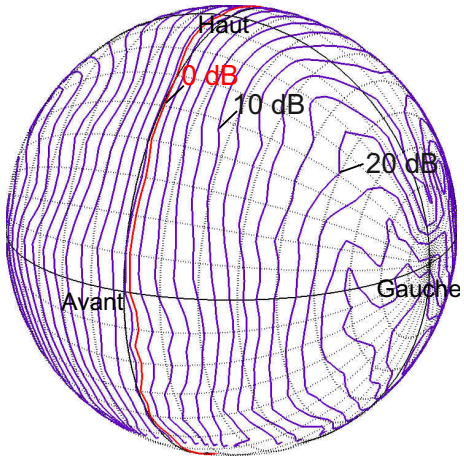


Figure 1.6 – Evolution de l'ILD sur la sphère (sujet n°1 de la base privée de HRTF de Orange Labs, intégration de l'énergie sur la bande de fréquence [1.5 kHz - 10 kHz]). Les lignes relient les directions correspondant à une même valeur d'ILD, et sont espacées de 2 dB. La ligne rouge représente l'iso-ILD 0 dB.

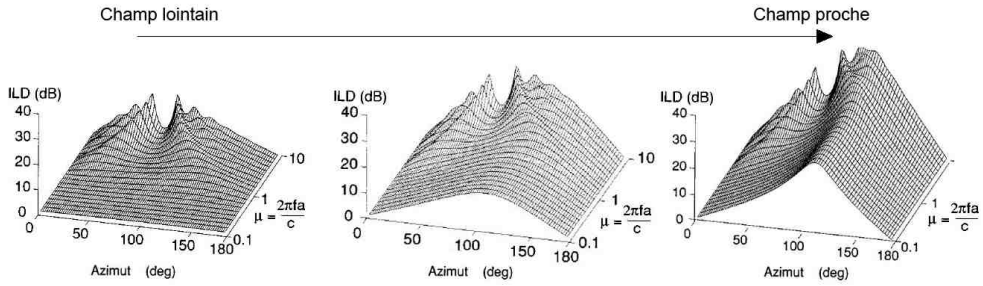


Figure 1.7 – ILD dans le plan horizontal en fonction de l'azimut et de la fréquence normalisée $\mu = 2\pi f a/c$, pour trois valeurs différentes du rapport $\rho = r/a$ entre la distance de la source r et le rayon a de la tête (modèle de tête sphérique, oreilles située aux azimuts $\pm 100^\circ$, d'après [63]). De gauche à droite : $\rho = 100$, $\rho = 2$, $\rho = 1.25$. Aux basses fréquences, on voit apparaître en champ proche des différences de niveau interaural absentes en champ lointain.

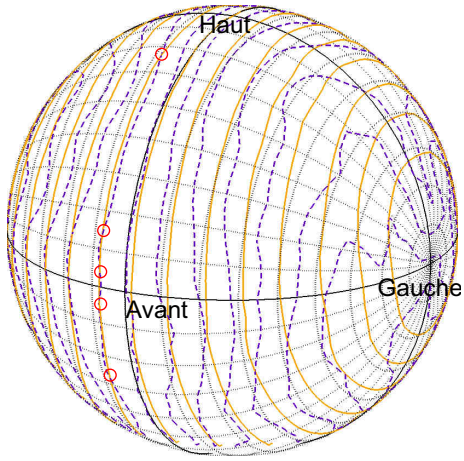


Figure 1.8 – Evolution de l'ITD basses fréquences et de l'ILD sur la sphère. Deux lignes iso-ITD et iso-ILD données présentent des intersections pour plusieurs directions, rendant ces indices ambigus pour la perception de l'élévation (par exemple, dans les directions cerclées de rouge). Les lignes iso-ITD (en orange) sont espacées de 100 μs, les lignes iso-ILD (en pointillés violets) sont espacés de 4 dB.

système auditif de localiser convenablement en élévation. En effet, il existe une infinité de positions qui engendrent un même couple de valeurs d'ITD et d'ILD. La théorie duplex ne parvient donc pas à expliquer la perception de l'élévation, ni la discrimination de positions de sources situées dans le plan médian, pour lesquelles les indices interauraux sont très faibles. C'est ce qui a mis en évidence l'existence d'autres indices, dits indices spectraux monauraux.

1.2 Indices spectraux monauraux et HRTF

Les phénomènes de réflexion et de diffraction subis par l'onde sonore lors de son interaction avec le torse, la tête et les pavillons, engendrent d'un point de vue signal un filtrage : des creux et des pics, dépendant de la direction de la source, viennent entacher le spectre du signal d'origine. Ce sont ces colorations spectrales que l'on nomme indices spectraux, et ils sont monauraux au sens où ils apparaissent aux tympans indépendamment pour chaque oreille.

1.2.1 Approche système du problème et définition des HRTF

Il convient d'adopter une approche système des phénomènes acoustiques, pour décrire d'un point de vue signal les phénomènes acoustiques observés entre la source et les tympans d'un auditeur. Considérons la figure 1.9. La source émet un signal acoustique x, et les signaux reçus aux oreilles gauche et droit sont x_L et x_R. Plus

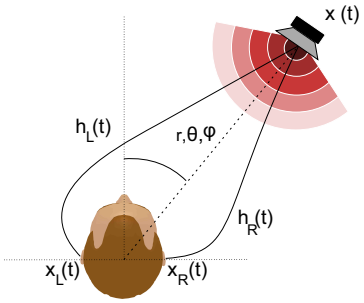


Figure 1.9 – Approche système du problème acoustique. Les phénomènes acoustiques modifiant le signal entre sa génération à la source, et sa captation aux oreilles sont modélisés par des filtres linéaires invariants dans le temps h_L et h_R, qui dépendent de la position de la source dans l'espace.

précisément, ce sont les signaux reçus à l'entrée des conduits auditifs qui nous intéressent, car il a été démontré que la fonction de transfert entre l'entrée du conduit auditif et le tympan est indépendante de la direction de la source [154]. Dans le cas d'une source fixe dans le repère centré sur l'auditeur, les phénomènes de propagation acoustique observés entre la source et l'entrée des conduits auditifs gauche et droit peuvent être modélisés comme deux systèmes linéaires invariants, caractérisables par leurs réponses impulsionnelles $h_L(t)$ et $h_R(t)$, telles que :

$$x_{L,R}(t) = h_{L,R} * x(t) \tag{1.5}$$

Sous sa forme fréquentielle, le problème est défini par les relations :

$$X_{L,R}(j\omega) = H_{L,R}(j\omega).X(j\omega) \tag{1.6}$$

où H_L et H_R sont les fonctions de transfert traduisant les phénomènes acoustiques subis par le signal x entre la source et l'entrée des conduits. Ces fonctions sont appelées *Head-Related Transfer Functions* ou HRTF, soit en français les fonctions de transfert relatives à la tête. Leurs versions temporelles h_L et h_R sont appelées *Head-Related Impulse Responses*, ou HRIR. Les HRTF dépendent à la fois de la position de la source par rapport à l'auditeur dans les trois dimensions de l'espace, et de la fréquence. De fortes différences sont observées d'un individu à l'autre : elles traduisent l'impact majeur des différences morphologiques sur les phénomènes acoustiques. Les HRTF offrent une approche globale, car elles contiennent toutes les informations dont dispose le système auditif pour localiser une source fixe dans une position donnée de l'espace.

1.2.2 Les colorations spectrales comme indices de localisation

L'observation du module spectral des HRTF permet de décrire les colorations subies par un signal du fait de l'interaction de l'onde sonore avec la morphologie de

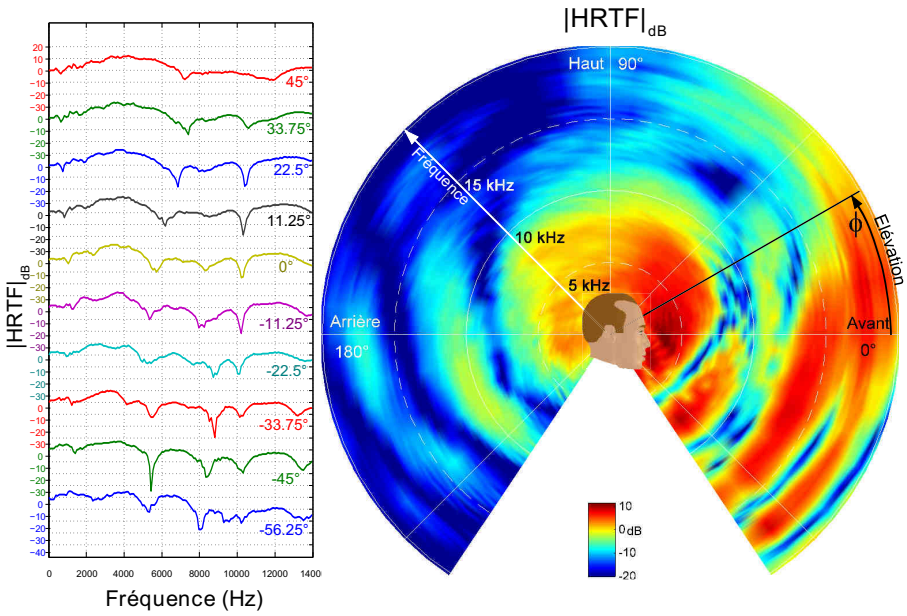


Figure 1.10 – Indices spectraux monauraux. A gauche : représentation du module du spectre des HRTF pour différentes élévations dans le plan médian (oreille droite). A droite : représentation polaire du module des HRTF en dB dans le plan médian, en fonction de la fréquence (rayon) et de l'élévation (angle). Il apparaît une dépendance forte des caractéristiques du spectre avec l'élévation, suggérant l'existence d'indices monauraux de localisation. (Sujet n°5 de la base privée de HRTF d'Orange Labs, oreille droite.)

l'auditeur (cf. Fig. 1.10). On remarque que les différents accidents du spectre - creux et pics - sont très dépendants de la direction de la source, en particulier aux hautes fréquences. C'est en cela que les colorations spectrales sont des indices potentiels de localisation. De plus, il a été démontré que pour bien localiser en élévation, le signal de la source doit contenir suffisamment d'énergie aux hautes fréquences [131]. De nombreux autres résultats viennent confirmer le fait que le système auditif utilise des données fréquentielles pour localiser une source (cf. 3.2.2). Grossièrement, la localisation sur la base des indices spectraux s'apparenterait à un processus d'identification entre les filtrages subis par le signal source et les indices spectraux "stockés" en mémoire. Le système auditif utiliserait en quelque sorte des facultés de reconnaissance de formes. La description des indices spectraux et la question de leur analyse par le système auditif sont traitées chapitre 3. Notons seulement qu'ils permettent d'expliquer la perception de l'élévation (haut/bas et avant/arrière).

1.3 Indices dynamiques de localisation

Pour localiser une source sonore dans l'espace, un auditeur cherche naturellement à orienter sa tête vers celle-ci de façon à lui faire face : c'est en effet dans cette position que les sons sont localisés le plus précisément. Cependant, Perrett et Noble [199] montrent qu'une amélioration de la précision de localisation en azimut est apportée par les indices dynamiques même quand le son est trop court pour que l'auditeur se tourne face à la source. Ce résultat montre que des indices de localisation dits dynamiques, induits par ces mouvements de la tête, contribuent en eux-mêmes à la formation du percept de localisation d'une source.

L'existence de cônes de confusion rend ambigüe la localisation, et malgré l'existence d'indices spectraux pouvant lever ces ambiguïtés, on observe parfois ce que l'on appelle des confusions avant-arrière : une source située à l'avant peut être perçue à l'arrière, à la position symétrique de sa position réelle par rapport au plan vertical interaural, ou inversement. Wightman et Kistler [276] ont montré que les indices apportés par des mouvements de la tête permettent de réduire ces confusions. Par ailleurs, les mêmes résultats sont observés quand l'auditeur garde la tête fixe, mais contrôle lui-même les mouvements de la source. Cela confirme l'hypothèse de Wallach [263] : les mouvements de tête ne sont pas absolument nécessaires pour générer des indices dynamiques, mais seule la connaissance des mouvements relatifs de la source par rapport à la tête suffisent. Des rotations de la tête de l'ordre de 5° (à $50°/s$) sont suffisantes pour générer des indices dynamiques exploitables [142, 143], c'est pourquoi leur effet bénéfique pour la discrimination avant/arrière est également observé pour des sons courts [199]. L'ambiguïté avant/arrière est levée par l'analyse dynamique des changements d'ITD et d'ILD de la façon suivante. Par exemple pour une position frontale de la source dans le plan horizontal, si l'auditeur tourne la tête vers la droite, la source apparaît plus proche de son oreille gauche, et s'il tourne la tête vers la gauche, c'est de l'oreille droite qu'elle se rapproche. Dans le cas où la source est positionnée directement à l'arrière de l'auditeur, c'est l'inverse qui se produit (cf. Fig. 1.11).

Perrett et Noble [199, 200] ont étudié l'effet des mouvements de tête pour la localisation en élévation. Ils semblent bénéfiques pour détecter des positions de sources très élevées ou bien très basses (au delà de ±30°) : dans ces conditions l'amplitude des variations dynamiques des indices interauraux induites par des rotations de la tête sont plus faibles que pour des sources proches du plan horizontal [262, 263]. Les expériences de Perrett et Noble, utilisant des signaux filtrés passe-bas, indiquent que ce sont les changements dynamiques de l'ITD aux basses fréquences qui sont exploités, plutôt que ceux de l'ILD [200]. Cependant, Faure [68] montre pour des sources large bande, que les indices dynamiques ne permettent pas de réduire les erreurs de

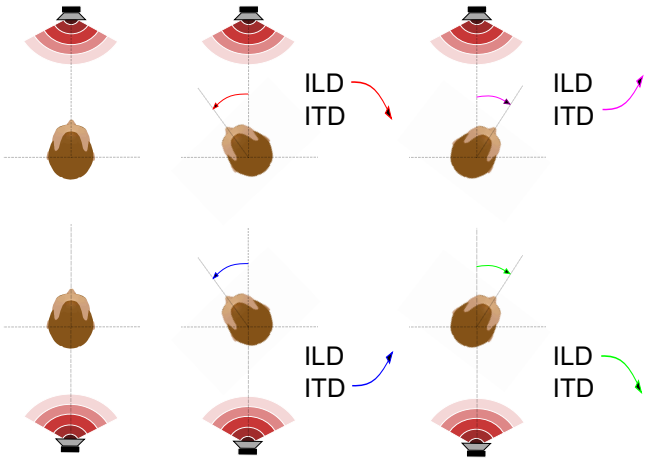


Figure 1.11 – Illustration de l'évolution dynamique des indices interauraux (ITD et ILD définies par rapport à l'oreille gauche).

perception en élévation induites par des indices spectraux altérés, ce qui suggère que les indices spectraux restent prépondérants pour la formation du percept d'élévation.

1.4 Perception de la distance

La variation de la distance d'une source sonore affecte souvent de multiples façons les propriétés acoustiques du son atteignant les tympans de l'auditeur. Ainsi, la formation du percept de distance est le fruit de l'analyse combinée de plusieurs indices acoustiques disponibles [281].

Le premier indice pouvant renseigner sur la distance d'une source est le niveau acoustique perçu. En champ libre, la décroissance du niveau sonore est de 6 dB pour un doublement de la distance, ce qui peut donc fournir un indice fiable si le niveau sonore de la source est invariant et si l'auditeur le connaît [165]. Dans un espace réverbérant, cette décroissance avec la distance est cependant moindre, mais un autre indice intervient : le niveau relatif d'énergie ν entre champs direct et réverbéré. Bien que la géométrie des lieux influence énormément le jeu de réflexions atteignant l'auditeur, ν dépend généralement de la distance. A proximité de la source, le champ direct est prépondérant dans le signal perçu, tandis qu'à distance croissante, la part relative du champ réverbéré augmente. La pertinence perceptive de cet effet a été démontrée par von Bekesy [259].

Le spectre perçu est aussi systématiquement affecté par la façon dont le signal source est réfléchi et réverbéré. Par exemple quand l'auditeur est proche d'un mur, les réflexions de l'onde acoustique sur celui-ci sont à l'origine d'un filtrage en peigne

caractéristique, constituant potentiellement un indice supplémentaire [31]. Notons enfin que les effets liés à l'environnement réverbérant sont non seulement fonction de la distance égocentrique de la source, mais aussi des positions absolues de la source et de l'auditeur dans la salle.

Pour des distances supérieures à 15 m, l'atténuation de l'énergie aux hautes fréquences lors de la propagation des ondes sonores dans l'air affecte l'équilibre spectral du signal sonore, et agit comme un filtre passe-bas, dont la fréquence de coupure est liée à la distance. Des expériences ont montré que la distance apparente d'une source dépend effectivement du rapport d'énergie de ses composantes hautes et basses fréquences [44]. Cependant, c'est un indice d'évolution très lente : environ quelques décibels pour 100 m [101], et il n'apparaît utilisable par le système auditif que comme un indice relatif [134].

A des distances proches ($<$ 1 m à 2 m), la nature sphérique des ondes acoustiques doit être considérée dans l'analyse des indices de localisation. Comme explicité en 1.1.2, la courbure du front d'onde affecte nettement l'ILD, et par la même occasion le spectre des HRTF. L'évolution de l'ILD avec la distance n'apparaît exploitable par le système auditif que pour des sources à moins de 1 m et hors du plan médian [240]. Il existe de plus un effet de parallaxe acoustique [33, 118] : à grande distance, les angles latéraux observés entre la source et chacun des plans sagittaux contenant l'entrée des conduits auditifs sont proches de l'azimut défini, lui, par rapport au centre de la tête (cf. Fig. 1.12). Par contre, à mesure que la distance diminue, les différences entre ces angles augmente (cf. Fig. 1.12). Ainsi le filtrage directionnel imposé par chacun des pavillons dépend aussi de la distance si l'on conserve le système de coordonnées attaché au centre de la tête. Cet effet est particulièrement important pour des positions de sources proches du plan médian, mais disparaît progressivement à mesure qu'elles se rapprochent de l'axe interaural.

1.5 Performances de localisation

L'estimation des performances de localisation du système auditif en champ libre a fait l'objet de nombreuses études psycho-expérimentales : historiquement ces recherches ont commencé avec Lord Rayleigh [216], et plus récemment Blauert [19], Oldfield et Parker [190], et enfin Carlile *et al.* [49]. On présente ici synthétiquement les résultats de cette dernière étude, concernant la perception de la direction de signaux courts et large bande présentés une seule fois, à distance constante de l'auditeur et en conditions de champ lointain.

Les auteurs proposent d'employer la distribution de Kent (cf. Annexe A), permettant d'analyser des nuages de points étirés de façon elliptique sur la sphère, et ainsi d'exprimer l'asymétrie des reports de localisation autour d'une direction cible

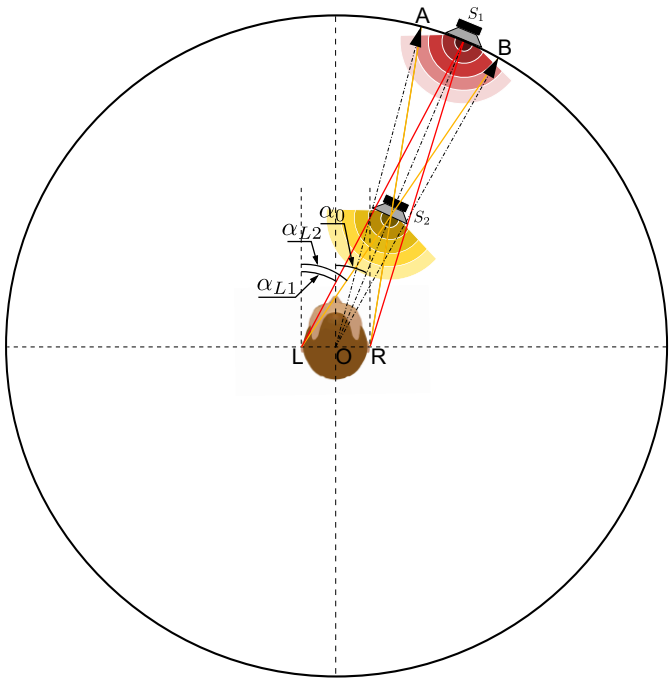


Figure 1.12 – Illustration du parallaxe acoustique. Pour une distance égocentrique décroissante de la source, l'angle d'incidence apparente pour chaque oreille s'écarte de plus en plus de l'azimut défini par rapport au centre de la tête ($\alpha_{L2} - \alpha_0 > \alpha_{L1} - \alpha_0$).

particulière. On représente figure 1.13 les résultats de cette étude. Les traits rouges résument le résultat de l'ajustement d'une distribution de Kent sur les reports de localisation recueillis pour 19 sujets. Ils sont centrés sur la direction moyenne des reports de localisation, et leur longueur exprime la dispersion des résultats : elle est égale à l'écart type dans la direction d'étirement maximal, celle du grand axe des contours d'équiprobabilité ellipsoïdaux. La dispersion est la plus faible juste au dessus du plan horizontal, près du plan médian (écart-type : 5°). La dispersion croît à mesure que les sources se rapprochent de l'axe interaural, et une légère augmentation est encore observée pour les sources situées à l'arrière, jusqu'à atteindre un écart-type d'environ 10°. Les dispersions les plus fortes sont observées aux élévations extrêmes (±40°) : l'écart-type atteint environ 10° à 12° quelque soit l'azimut. La moindre précision de localisation observée dans l'hémisphère arrière est peut-être due au protocole expérimental, d'après les auteurs. En effet, les sujets devaient pointer la tête dans la direction perçue, perdant ainsi en partie leurs repères spatiaux quand ils devaient atteindre les positions arrière. D'autres expériences montrent des performances similaires en termes de précision azimutale dans les hémisphères avant et arrière [25]. En termes de précision, on observe des erreurs de localisation en moyenne de 3° en azimut, et 4° en élévation (système de coordonnées polaire-vertical) entre les directions moyennes des nuages de points et les directions cibles. Les erreurs les plus faibles sont enregistrées pour l'hémisphère frontal, autour du plan horizontal. A l'avant, les positions de sources les plus hautes et les plus basses sont reportées plus près du plan horizontal. Il existe également un biais de localisation vers l'arrière pour les positions proches de l'axe interaural.

Les confusions avant/arrière sont faibles : environ 3.4% des réponses, correspondant en grande majorité à des positions de sources à moins de 30° du plan vertical contenant l'axe interaural. La moyenne des taux de confusion observés est d'environ 5% à 6% si l'on se réfère à d'autres études [29, 50, 273]. On représente figure 1.14 les positions concernées par ces confusions. Le rayon des cercles est proportionnel au nombre relatif de confusions pour chaque position.

Nos performances à percevoir la distance des sources sonores sont relativement médiocres. Zahorik [281] a montré un phénomène général de compression entre la distance réelle et la distance perçue : pour des distances inférieures à 1 m, la distance est plutôt surestimée, tandis qu'elle est sous-estimée au delà. La dispersion des reports de distance perçue est très dépendante des capacités de l'auditeur, mais elle est toujours assez élevée : autour de 20% à 60% de la distance réelle de la source. Ce résultat révèle plutôt le flou du percept de distance que la variabilité expérimentale du report de jugement [282]. La familiarité de l'auditeur avec le signal sonore joue un rôle important dans cette faculté. Par exemple pour la parole normale, Gardner [73] rapporte des performances excellentes entre 1 m et 9 m, mais les chuchotements

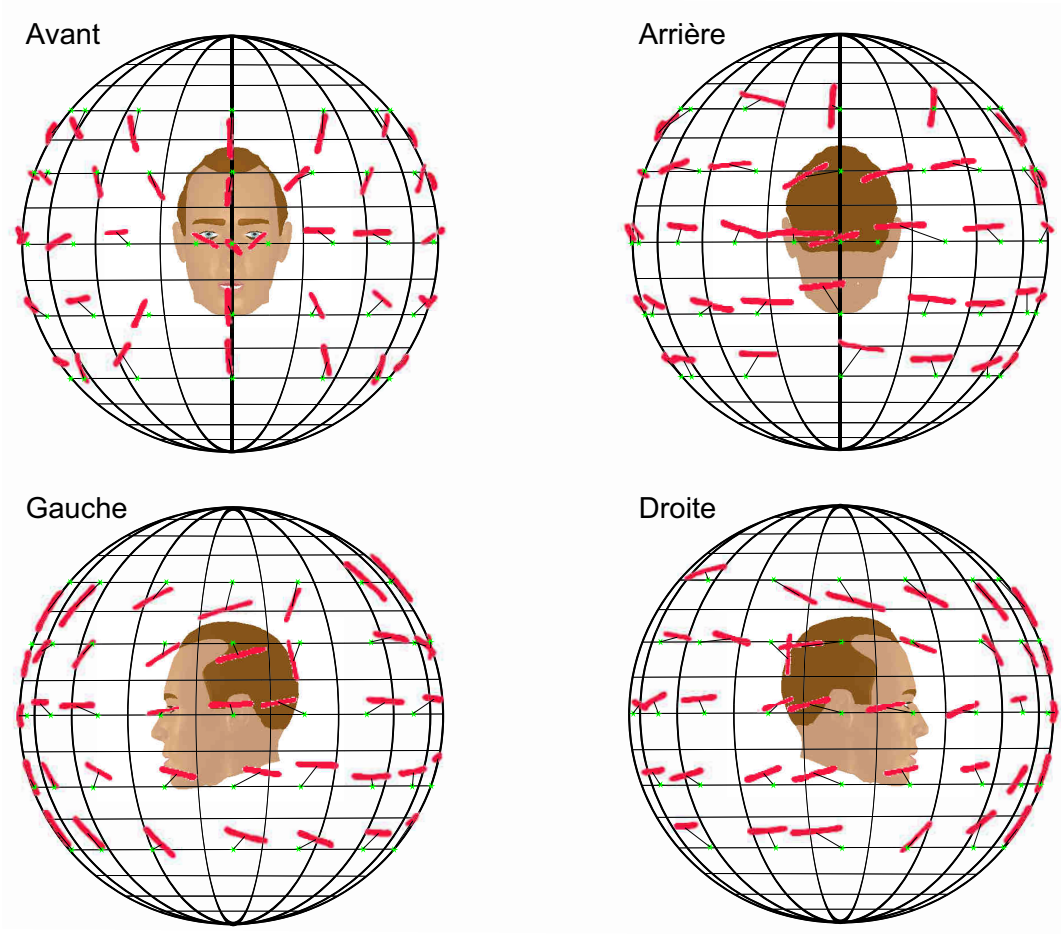


Figure 1.13 – Illustration des performances de localisation (d'après [49]). Les étoiles vertes matérialisent les directions des sources à localiser, tandis que les traits rouges résument le résultat de l'ajustement d'une distribution de Kent sur les reports de localisation recueillis pour 19 sujets. Ces traits sont centrés sur la direction moyenne des reports de localisation, et leur longueur exprime la dispersion des résultats : elle est égale à l'écart type dans la direction d'étirement maximal, celle du grand axe des contours d'équiprobabilité ellipsoïdaux.

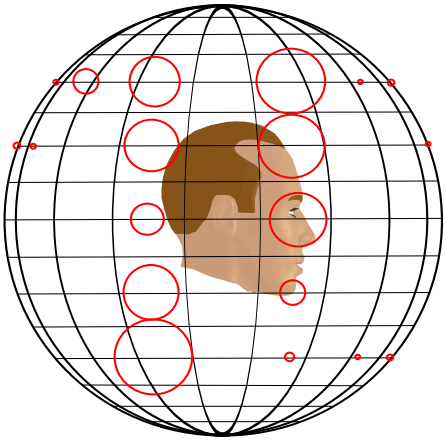
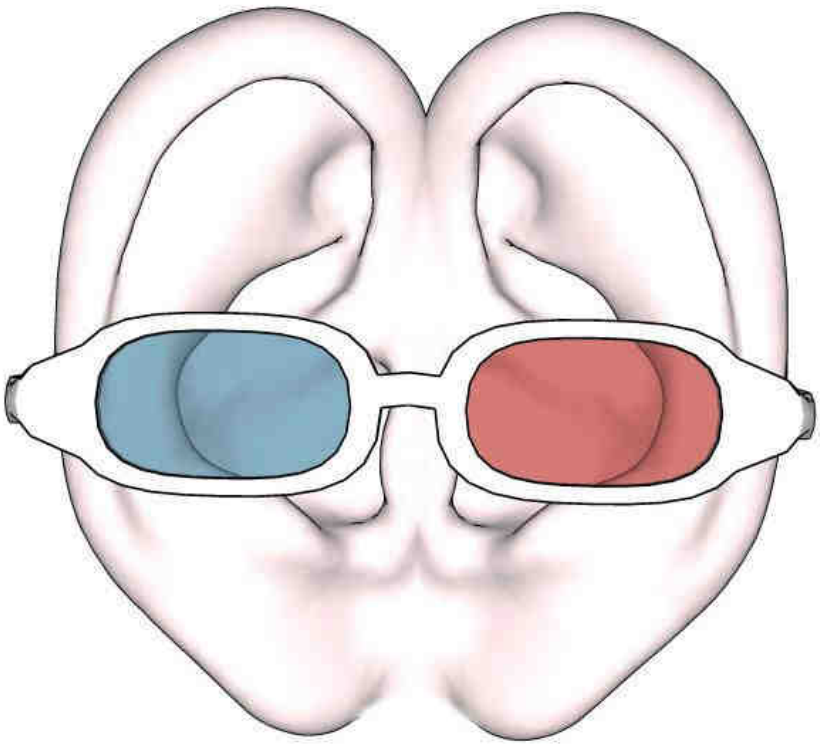


Figure 1.14 – Illustration des confusions avant/arrière observées pour 3.4% des reports de localisation (d'après [49]). Le rayon des cercles rouges est proportionnel au nombre de confusions.

et les cris mènent respectivement à une sous-estimation et à une surestimation de la distance. Pour des clics, les performances peuvent être similaires si l'auditeur est exposé au préalable au signal sonore.

Il existe enfin un phénomène appelé *proximity-image effect* [72] : la distance apparente d'une source est déterminée par la distance de l'objet à portée de vue le plus proche susceptible d'être à l'origine du son perçu. Cet effet peut être très gênant dans le cadre de la synthèse sonore spatialisée si l'auditeur a les yeux ouverts, et si aucun environnement visuel synthétique n'est présent pour matérialiser les sources sonores virtuelles [282].

Ces différents résultats permettent de dresser une cartographie des faiblesses du système auditif à localiser des sources. Comme il est peu probable que la résolution des processus neuronaux impliqués dans la localisation varie selon les positions de sources, ces résultats révèlent plutôt le fait que les indices physiques eux-mêmes offrent un encodage spatial plus ou moins net selon la position [49]. On peut noter que l'ouïe offre un champ de perception plus large que le champ visuel, mais avec une précision moindre. Les deux sont en fait très liés. En effet la localisation auditive est une faculté en partie acquise. Un apprentissage est réalisé dès le premier âge : la position d'une source sonore perçue auditivement est constamment reliée à sa position perçue visuellement. Même si cela n'explique pas tout, il semble donc naturel d'observer que les performances de localisation sont optimales pour des sources positionnées dans le champ visuel, et moindres à l'arrière et au dessus de la tête.



Chapitre 2

Synthèse binaurale

La synthèse binaurale est une technique de spatialisation dont le but ultime est de procurer à l'auditeur l'illusion parfaite qu'il est immergé dans une scène sonore. Pour ce faire, l'idée est de créer de façon synthétique les signaux qu'il aurait perçus lors d'une écoute naturelle, et de générer le champ acoustique correspondant au niveau de ses tympans. L'utilisation de filtres binauraux issus des HRTF permet de reproduire au mieux tous les indices de localisation nécessaires, et ainsi d'assurer une illusion satisfaisante. Nous abordons dans ce chapitre les principes et les différents aspects techniques de la synthèse binaurale.

2.1 Principe

La synthèse binaurale est basée sur l'utilisation de paires de filtres binauraux, qui découlent idéalement des HRTF. Pour chaque position de l'espace (r, θ, ϕ) il existe une paire de HRTF, qu'on obtient soit par un modèle, soit par la mesure, auquel cas elles ne sont connues rigoureusement que pour un échantillonnage discret de l'espace. Pour placer une source virtuelle à une position désirée de l'espace, on trouve directement la paire de HRIR correspondantes dans une base de données, si elle est disponible à cette position, ou bien on la calcule par interpolation dans le cas contraire, puis on en déduit une paire de filtres binauraux h_L et h_R sous une forme adaptée à l'implémentation choisie. Pour une diffusion sur casque d'écoute stéréophonique classique (Cf Fig. 2.1), on convolue chacun de ces filtres au signal source monophonique et anéchoïque x, pour obtenir les signaux x_L et x_R à présenter aux écouteurs. Cette version de la synthèse binaurale est la plus simple à mettre en œuvre, car elle permet de contrôler indépendamment les signaux présentés à chacune des oreilles. Une restitution sur deux haut-parleurs est également possible (Cf Fig. 2.2). On parle alors de synthèse binaurale sur haut-parleurs, dont un exemple connu de mise en œuvre est le système *Transaural*™[6, 9]. Une difficulté supplémentaire intervient alors, car le signal de chacun des haut-parleurs est perçu par les deux oreilles. Cela impose de développer une stratégie d'annulation des trajets croisés (*cross-talk cancellation*) : les HRTF correspondant aux positions physiques des haut-parleurs sont utilisées pour réaliser un matriçage, dont le but est de rendre transparent le système de restitution. Que la restitution soit réalisée au casque ou bien sur haut-parleurs, il convient de compenser la réponse des transducteurs pour pouvoir contrôler finement les signaux générés.

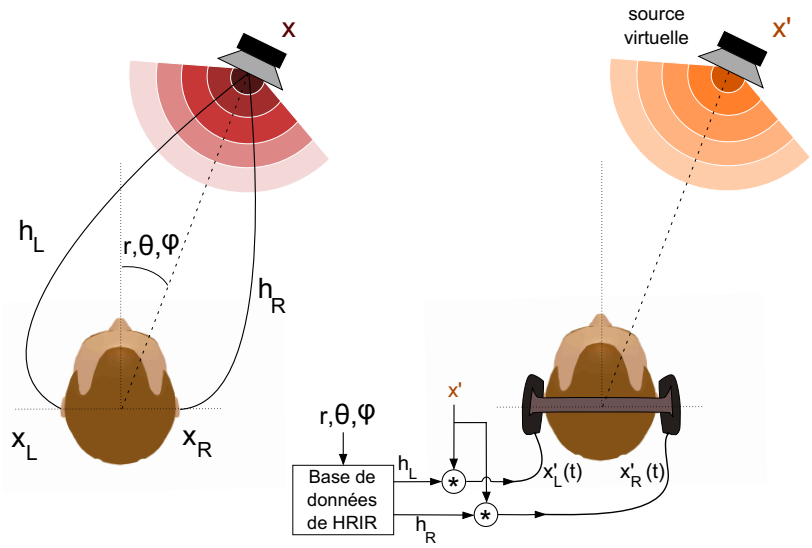


Figure 2.1 – Principe de la synthèse binaurale : restitution sur casque.

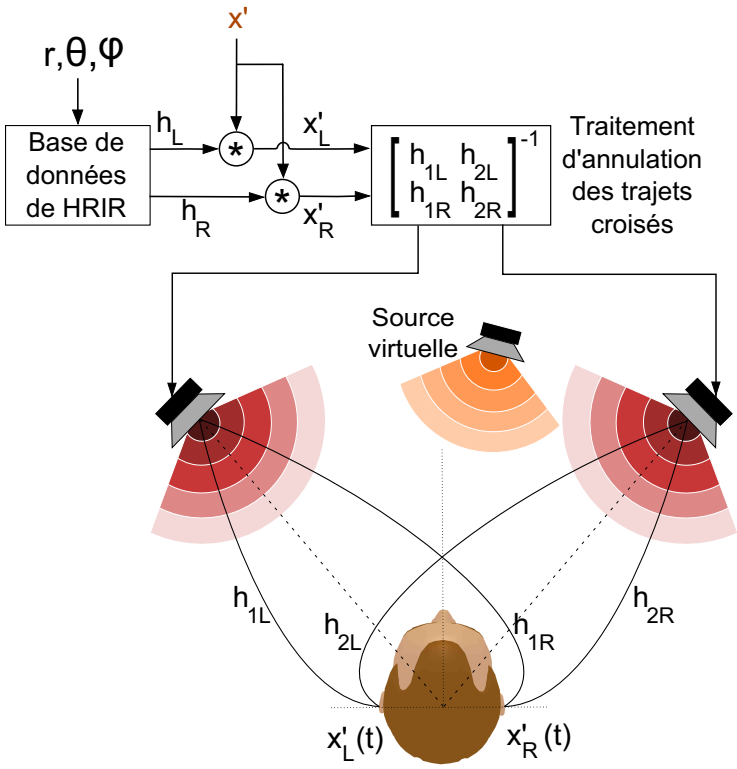


Figure 2.2 – Principe de la synthèse binaurale : restitution sur haut-parleurs

2.2 Mesure des HRTF

Une première façon d'obtenir les HRTF est la mesure acoustique. Basée sur des principes relativement simples, elle s'avère en fait lourde à mettre en œuvre, notamment parce qu'il faut se placer dans une chambre anéchoïque, afin de ne capter que les effets acoustiques propres au corps de l'auditeur. On dispose deux microphones miniatures au niveau des oreilles d'un sujet, puis on positionne précisément sa tête au centre d'un dispositif qui permet de déplacer un haut-parleur autour de ce point, et de décrire un échantillonnage spatial suffisamment fin. Pour chacune des positions du haut-parleur, un signal est émis et une mesure est effectuée (cf. Fig. 2.3). Deux approches sont possibles pour le positionnement des microphones. Soit on dispose des petites capsules microphoniques à l'entrée du conduit auditif, avec un moule qui bouche le conduit auditif (mesure dite "conduit bloqué"), ou bien on dispose une sonde au plus près du tympan, le conduit auditif restant ouvert (mesure dite "conduit ouvert"). Les deux solutions sont correctes dans la mesure où le comportement acoustique du canal auditif est indépendant de la position de la source [154]. Cependant l'existence de nœuds et de ventres de pression liés aux ondes stationnaires dans le conduit auditif rend délicat le positionnement d'une sonde microphonique et affecte le rapport signal à bruit lors de la mesure. C'est donc la mesure conduit bloqué qui est aujourd'hui préférée, comme illustré figure 2.3, pour sa facilité de mise en œuvre et parce que le positionnement du microphone est parfaitement reproductible.

En pratique, différents signaux d'excitation peuvent être utilisés : impulsions étirées dans le temps (*Time Stretched Pulse* ou TSP) [5, 29, 197], séquences de longueur maximale (*Maximum-Length Sequence* ou MLS) [160, 195, 243], codes de Golay [70, 172, 205], sinus glissants [147]. Quelle que soit la méthode choisie, la définition physique des HRTF gauche et droite H_L et H_R est commune [154] :

$$H_{L,R}(r,\theta,\phi)(j\omega) = \frac{\Phi_{L,R}(r,\theta,\phi)(j\omega)}{\Phi_0(j\omega)}$$

où (r,θ,ϕ) désigne la position en coordonnées sphériques du haut-parleur de mesure dans le référentiel auditeur, Φ_L et Φ_R sont les pressions sonores mesurées à l'entrée des conduits respectivement gauche et droit, et Φ_0 est la pression acoustique mesurée à la position du centre de la tête, le sujet étant absent. Il est à noter que si la mesure est réalisée sur une sphère, comme décrit précédemment, et ce qui est le plus courant, on choisit sciemment de négliger la dépendance radiale des HRTF. Le rayon de la sphère est alors choisi à une valeur supérieure à 1 m - 1.5 m de telle façon que le sujet se trouve en condition de champ lointain au delà de 500 Hz. Il est également possible d'intégrer aux HRTF les effets liés à la propagation acoustique dans une salle : la procédure est en tous points identique, à ceci près qu'elle est réalisée dans une salle non-anéchoïque, et que les signaux d'excitation devront être plus longs pour capter

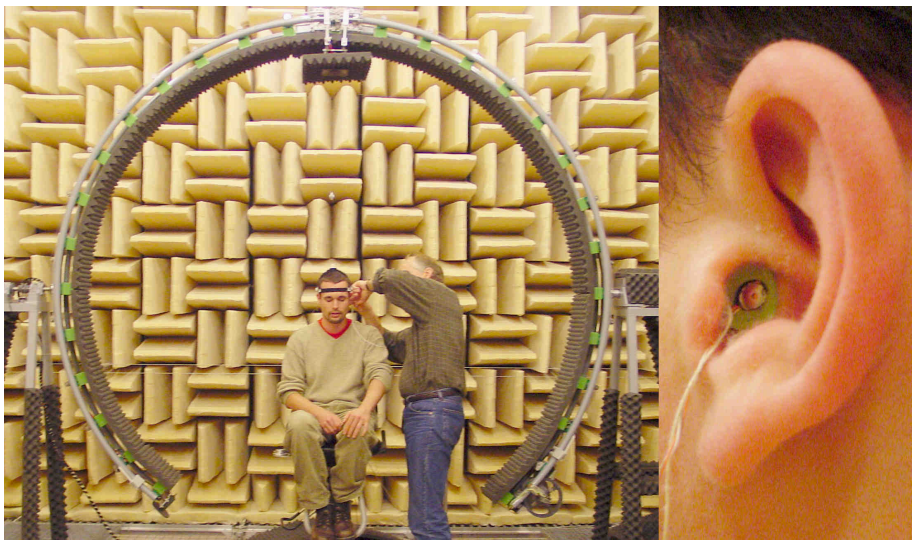


Figure 2.3 – Mesure de HRTF en chambre anéchoïque (d'après [197]).

l'effet de salle. Les réponses impulsionnelles associées à de telles HRTF sont appelées *Binaural Room Impulse Responses* ou BRIR.

2.3 Synthèse de la distance

Si la dimension radiale et l'effet de salle ne sont pas intégrés dans la mesure des HRTF en champ lointain, il existe des moyens indirects pour jouer sur la distance des sources virtuelles, en modélisant les phénomènes décrits en 1.4.

Pour des sources en champ lointain ($r > 1\,m$), les effets de décroissance du niveau, et l'atténuation des hautes fréquences peuvent être simulées de façon assez simple. La simulation de l'effet de salle peut présenter plus de difficultés. En effet le nombre de sources constituant une scène sonore peut rendre prohibitifs les calculs nécessaires pour simuler la spatialisation des multiples réflexions et du champ réverbéré. Les algorithmes classiques simplifient le problème en ne spatialisant que les premières réflexions [112], et en reproduisant la réverbération à l'aide d'un bruit décorrélé spatialisé comme un champ diffus. La décorrélation entre les signaux gauche et droit est en effet un indice concomitant à la diminution du ratio énergétique entre champ direct et champ réverbéré [30]. Il apparaît de plus que les premières réflexions peuvent être synthétisées avec une moindre fidélité sans que cela n'affecte leur perception : tant que la spatialisation du son direct est traitée fidèlement, on peut conserver grossièrement les indices interauraux générés par les premières réflexions, et relâcher la contrainte de fidélité des indices spectraux [283]. Il convient d'utiliser avec précaution la réverbération dans un VAS, car si elle améliore la perception de

la distance, elle peut dégrader la perception de la direction, et également dégrader l'intelligibilité de signaux de parole [239].

Romblom propose de reproduire conjointement les différents effets de spatialisation observés en champ proche [223]. Un modèle de tête sphérique peut être utilisé pour calculer les *filtres différences* reliant les HRTF en champ lointain et les HRTF à une distance donnée en champ proche, direction par direction. Ces filtres sont ensuite superposés aux HRTF de l'auditeur mesurées en champ lointain. De plus, la parallaxe acoustique peut être reproduite en choisissant pour chaque oreille la HRTF correspondant à la direction apparente de la source (définies pour les oreilles gauche et droite respectivement par les vecteurs **OB** et **OA**, cf. Fig. 1.12, le cercle décrivant la distance à laquelle les HRTF sont disponibles).

2.4 Perception du spectre de phase des HRTF

Plusieurs études ont évalué la pertinence perceptive des différents attributs des HRTF. Kulkarni *et al.* [128] ont montré que des transformations affectant la phase des HRTF ne sont pas perceptibles lors d'une écoute monaurale, mais le sont seulement lors d'une présentation binaurale. De plus, les auteurs démontrent que la dépendance fréquentielle du retard interaural de phase peut être négligée lors de l'implantation des HRTF, tant que la moyenne de ce retard aux basses fréquences est conservée par rapport aux HRTF originales. Ces conclusions sont en accord avec les résultats de Wightman et Kistler [274], qui ont montré que la phase des HRTF aux hautes fréquences peut être rendue totalement aléatoire sans que cela n'affecte les performances de localisation. Constan et Hartmann [58] ont également observé des résultats similaires. Les indices temporels de localisation sont donc contenus de façon globale dans la phase des HRTF, et les variations fréquentielles fines n'ont pas d'impact perceptif. Ces résultats apportent la preuve de la validité d'une décomposition des HRTF largement répandue en synthèse binaurale : la décomposition en un retard pur et un filtre à phase minimale, qui permet de conserver rigoureusement le spectre d'amplitude des HRTF, tout en offrant la réponse impulsionnelle la plus compacte possible. Soit H une HRTF correspondant à une position de l'espace. On définit la décomposition comme suit :

$$
\begin{aligned}
H &= |H|.e^{j\varphi_0} && (2.1)\\
&= H_{min}.H_{exc} && (2.2)\\
H_{min} &= |H|.e^{j\varphi_{min}} && (2.3)\\
H_{exc} &= e^{j\varphi_{exc}} = e^{j(\varphi_0-\varphi_{min})} && (2.4)\\
\varphi_{min} &\triangleq \Re(Hilbert(-log(|H_{min}|))) && (2.5)
\end{aligned}
$$

où H_{min} est le filtre à phase minimale de même spectre d'amplitude que H, dont la phase φ_{min} est reliée à son module selon la relation 2.5. Le filtre H_{exc}, appelé filtre excès de phase, est de module unitaire, c'est donc un filtre passe-tout, dont la phase est définie de façon à vérifier l'égalité 2.2. La décomposition devient une approximation intéressante pour l'implantation quand on approche H_{exc} par un retard pur τ, c'est à dire :

$$H_{exc} \simeq e^{-j\omega\tau} \tag{2.6}$$

$$\hat{H} = H_{min}.e^{-j\omega\tau} \tag{2.7}$$

Mehrgardt et Mellert [164] ont en effet observé que l'excès de phase φ_{exc} est en général quasi linéaire jusqu'à 10 kHz. Soient $H_{L,appr}$ et $H_{R,appr}$ les HRTF approchées gauche et droite pour une direction donnée, et soient τ_L et τ_R les retards purs, dits monauraux, définis selon cette approximation. Il suffit donc de régler τ_L et τ_R de façon à ce que le retard interaural de phase moyen aux basses fréquences soit le même entre les HRTF originales H_L et H_R et entre les HRTF approchées $H_{L,appr}$ et $H_{R,appr}$. Ces résultats sont valables pour des HRTF mesurées en environnement anéchoïque, mais la décomposition n'est pas valable si elles intègrent une réponse de salle [128].

2.5 Perception du spectre d'amplitude des HRTF

Le système auditif réalise une analyse fréquentielle du signal, modélisable par un banc de filtres passe-bande, appelés filtres auditifs, qui se recouvrent continûment le long du spectre audible. D'un point de vue physiologique, cette caractéristique est issue de l'organisation tonotopique de l'oreille interne : chaque région de la membrane basilaire est sensible à une certaine plage de fréquences, et peut être assimilée à un filtre passe-bande. Cela se traduit par une résolution fréquentielle limitée, rendant inutilisables par le système auditif certains détails fins du spectre d'amplitude des HRTF.

Glasberg et Moore [75] ont décrit le banc de filtres cochléaires comme une série de filtres gammatones G_{f_c} de fréquence centrale f_c (cf. Fig. 2.4), dont la fonction de transfert est donnée par les relations suivantes :

$$G_{f_c,n}(f) = \left(1 + j.\left(\frac{f - f_c}{b(f_c, n)}\right)^{-n}\right) \tag{2.8}$$

$$b(f_c, n) = \frac{24,7(0,00437.f_c + 1)}{2\sqrt{2^{1/n} - 1}} \tag{2.9}$$

où n est l'ordre du filtre, et b un paramètre calculé de façon à ce que la largeur de bande à 3 dB corresponde aux ERB [75]. La résolution spectrale de la cochlée est

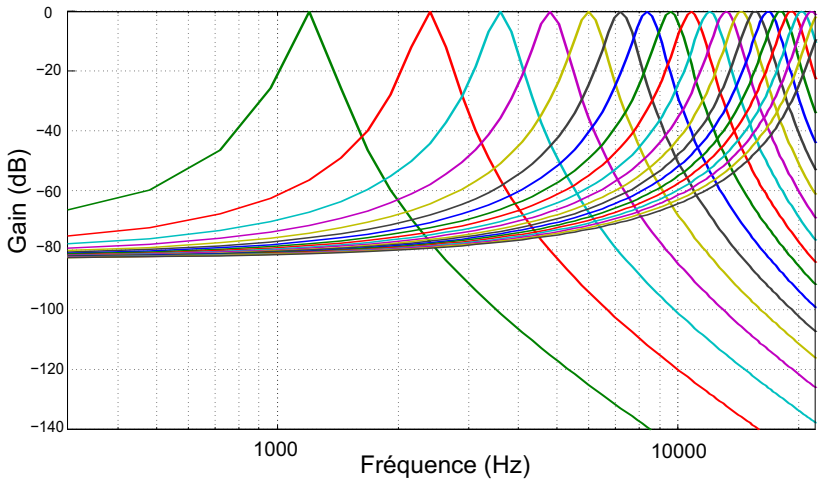


Figure 2.4 – Spectre d'amplitude des filtres de type gammatone sur le spectre audible.

bien décrite par ces filtres pour un ordre n = 4 [194]. On peut alors exprimer, pour une direction et une oreille données, le spectre d'amplitude perçu par le système central $|X(f)|$ pour un stimulus de type bruit blanc, à partir de la HRTF $H(f)$ correspondante [28] :

$$|X(f)| = \sqrt{\frac{\int_0^\infty |H(f')|^2.|G_{f,4}(f')|^2.df'}{\int_0^\infty |G_{f,4}(f')|^2.df'}} \tag{2.10}$$

On illustre figure 2.5 le spectre d'amplitude $|X(f)|$ pour une direction particulière : une diminution des accidents spectraux est bien observée, et elle est d'autant plus prononcée aux hautes fréquences. Cette prise en compte du filtrage auditif permet de se placer au plus près de l'information dont dispose le système nerveux pour établir un jugement de localisation. Si le système auditif présente une faible discrimination spectrale, alors la localisation doit nécessairement être insensible à des variations de la structure fine du spectre des HRTF, vecteur de l'information spatiale. C'est ce que les études suivantes se sont attachées à démontrer dans le cadre de la synthèse binaurale sur casque. Asano *et al.* [5] ont testé les capacités à localiser des sources virtuelles pour différents degrés de simplification des HRTF utilisées : le lissage spectral est réalisé dans leur étude par une modélisation ARMA plus ou moins fidèle. Les auteurs montrent que la localisation dans le plan médian reste peu dégradée jusqu'à l'utilisation d'un modèle à 20/20 coefficients (inclus). En-deçà d'un modèle à 10/10 coefficients, une augmentation des confusions avant/arrière apparaît nettement. Kulkarni et Colburn [127] effectuent un lissage cepstral des HRTF : une décomposition en série de Fourier du logarithme du spectre des HRTF est effectuée, et leur re-

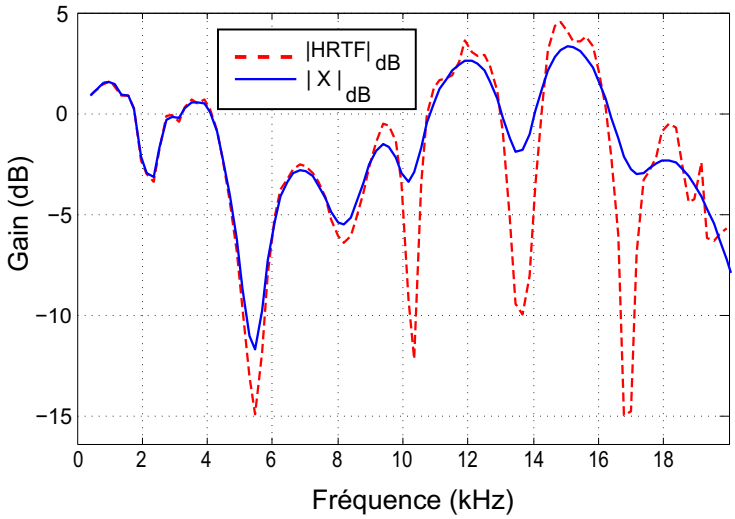


Figure 2.5 – Spectre d'amplitude dont dispose le système nerveux à une oreille pour un stimulus de type bruit blanc, dans une direction donnée. En pointillés rouges : le spectre d'amplitude de la HRTF correspondant à la direction considérée. En bleu : le même spectre passé par un banc de filtres de type gammatone.

construction est tronquée à un nombre réduit de coefficients. Aucune différence n'est perçue entre l'espace sonore réel et l'espace sonore virtuel qui le simule, jusqu'à une simplification cepstrale à 32 coefficients (incluse). Otten [191] montre que plus de 32 coefficients sont nécessaires dans certaines directions de l'espace, notamment en dessous du plan horizontal. Langendijk et Bronkhorst [131] montrent que le spectre des HRTF peut être moyenné par demi octaves sans dégradation significative des performances de localisation.

Bien que les lissages testés dans ces différentes expériences ne soient pas de même nature, et qu'ils affectent plus ou moins les détails fins du spectre, ces résultats indiquent que les indices de localisation contenus dans le spectre des HRTF sont robustes à une diminution de la résolution spectrale plus sévère encore que celle des filtres auditifs [28]. Des prétraitements appropriés, exploitant ces résultats, peuvent intervenir au moment du *design* des filtres binauraux [98]. Par ailleurs, puisqu'un lissage du spectre, donc une simplification des HRTF, est possible dans une certaine mesure sans induire de dégradation de la spatialisation, on peut envisager d'utiliser des techniques de réduction des données, se traduisant par une économie en espace mémoire. Une simplification efficace des filtres binauraux peut être obtenue par une décomposition des HRTF sur des bases bien choisies, suivie d'une reconstruction tronquée : on peut citer l'analyse en composantes principales [121] (ACP, cf. Annexe I), la transformée en ondelettes [78][250], les algorithmes génétiques [55].

2.6 Synthèse binaurale dynamique

Lors de la création d'une scène sonore spatialisée, chaque source peut être positionnée dans une direction désirée en lui associant, une fois pour toutes, la paire de filtres binauraux appropriée. Dans cette configuration, dite "statique", si l'auditeur tourne la tête, les sources virtuelles semblent suivre son mouvement. Dans le cas où l'environnement virtuel est généré avec des indices visuels, par exemple par le biais d'une représentation en trois dimensions de la scène sur un écran, cela induit une rupture entre les différentes modalités perceptives. De plus, en situation d'écoute réelle, on sait que l'aspect dynamique constitue un indice de localisation important : la possibilité de tourner la tête permet de confirmer et d'affiner la localisation d'une source sonore. Une solution consiste à placer chaque source virtuelle à une position fixe du référentiel absolu, et de mettre à jour les filtres binauraux en fonction de la position adoptée par la tête de l'auditeur : c'est le principe de la synthèse binaurale dynamique. En pratique, cela nécessite l'utilisation d'un système de suivi de la position de la tête, pilotant en temps réel la mise à jour des filtres. De plus, on doit disposer de filtres correspondant à n'importe quelle direction de l'espace et il se pose donc la question de l'interpolation, qui peut être commune à la synthèse binaurale statique. Par ailleurs, la synthèse binaurale dynamique soulève deux problèmes spécifiques : d'une part la mise à jour en temps réel des filtres doit être suffisamment rapide pour que la scène virtuelle s'adapte convenablement aux mouvements de la tête : c'est le problème de la latence du dispositif. Enfin, la commutation d'un filtre à un autre doit être réalisée sans artefact audible.

Systèmes de suivi de la position de la tête

Il existe différents systèmes de suivi de la tête ou *head-trackers*, permettant de mesurer à la fois la position et l'orientation de celle-ci (6 degrés de liberté). On recense principalement les systèmes suivants :

- **Acoustiques**
  La distance entre l'émetteur fixe et le récepteur solidaire de la tête est évaluée à partir d'ultrasons. Ces systèmes sont limités en portée, du fait de l'absorption de l'air.
- **Inertiels**
  Les capteurs inertiels utilisent deux technologies complémentaires : les gyroscopes qui permettent de calculer l'orientation de la tête après simple intégration, et les accéléromètres qui donnent accès à la position après double intégration. Précis à haute vitesse, ils le sont beaucoup moins à faible vitesse et souffrent de problèmes de stabilité.
- **Optiques**

Basés sur la mesure du positionnement de la tête à l'aide d'une ou plusieurs caméras et d'algorithmes appropriés, leur portée est limitée et ils demandent beaucoup de ressources informatiques.

– **Magnétiques**
Ils mesurent la position dans un champ magnétique créé par une source, ou bien par rapport au champ magnétique terrestre. Ils sont généralement très précis, mais souffrent d'une portée limitée et sont sensibles aux perturbations électromagnétiques.

Un bon système aura une latence minimale et une grande précision de mesure, pour offrir une résolution angulaire maximale. Pour des applications à la synthèse binaurale dynamique, Faure [67] a évalué ces différents systèmes et a opté pour un système électromagnétique avec source (*Polhemus Fastrack*®), qui permet d'obtenir une latence totale de l'ordre de 50 à 70 ms.

Effets de la latence

La latence du système est un paramètre important de la synthèse binaurale dynamique. Elle est définie comme le temps qui s'écoule entre l'instant où l'auditeur effectue un mouvement de tête et celui où les filtres correspondant à cette nouvelle position sont mis à jour. Sandvad [227] a montré que la latence était le premier facteur de dégradation des performances de localisation en synthèse binaurale dynamique. Le temps de réponse des sujets augmente et leur précision en localisation diminue pour un temps de latence croissant. La valeur de latence acceptable dépend en fait de la durée du son. Pour les sons longs, le seuil de 250 ms est avancé par Wenzel [268], tandis que pour les sons courts, il serait plutôt de 75 ms selon Brungart *et al.* [35]. Cette dernière valeur doit donc être retenue.

Interpolation des HRTF

La synthèse binaurale, sous sa forme statique comme dynamique, doit pouvoir simuler toutes les positions discernables par le système auditif. On choisit en général de réaliser une mesure des HRTF sur un maillage grossier de l'espace, puis de reconstruire les données manquantes : c'est un des objectifs de l'interpolation, opération qui peut être effectuée en temps différé (cf. chapitre 6). L'aspect supplémentaire, spécifique à l'implantation dynamique, est le calcul en temps réel des nouveaux filtres appropriés. Larcher [132] s'est intéressée à l'interpolation locale de filtres, c'est-à-dire entre les directions voisines de celle à synthétiser, pour lesquelles des HRTF sont disponibles en mémoire. L'étude a porté sur les divers choix possibles pour la structure d'implantation. Elle montre, dans le cas d'une décomposition en filtre à phase minimale et retard pur, que l'interpolation linéaire des coefficients FIR de la

composante à phase minimale s'avère efficace, conjointement à une interpolation linéaire des retards monauraux. En implantation IIR, l'interpolation optimale dépend de la structure choisie, transverse ou en treillis. On se rapportera à [132] pour plus de détails.

Commutation des HRTF

La transition d'une position de source à une autre est réalisée par la mise à jour, ou commutation, entre deux filtres. Une commutation brutale risque d'entraîner l'apparition d'artefacts audibles. Les clics que l'on peut alors entendre sont issus, dans le cas d'une implantation FIR, d'une discontinuité dans le signal, et dans le cas d'une implantation IIR, de l'instabilité des filtres liée à un problème de réinitialisation des mémoires. Larcher [132] propose une revue des techniques de commutation disponibles. La méthode la plus simple pour réaliser une commutation transparente est d'effectuer un fondu-enchaîné entre les signaux de synthèse correspondant aux positions successives, à mesure que l'on se déplace de la première à la seconde. En général, les techniques de commutation imposent un doublement du coût d'implantation par rapport à leur équivalent statique, et requièrent une résolution spatiale assez fine de la base des HRTF disponibles [197].

2.7 Externalisation

Un artefact de perception apparaît parfois en synthèse binaurale ou en écoute naturelle : les sources sonores ne sont pas localisées nettement à l'extérieur, mais très proches de la tête, ou même à l'intérieur de la tête. On parle alors de perception intra-crânienne, ou de défaut d'externalisation. Les conditions requises pour assurer l'externalisation ne sont pas encore tout à fait connues, mais Loomis *et al.* [135] détaillent quelques pistes pouvant expliquer les artefacts en synthèse binaurale.

Des informations d'ordre cognitif peuvent dégrader l'externalisation, comme le fait que l'auditeur sait que le son provient des écouteurs, et qu'il sent la pression du casque sur ses oreilles. L'absence d'indices visuels, ou bien la présence d'indices visuels incohérents par rapport aux indices auditifs, peuvent également jouer un rôle néfaste. On peut aussi incriminer les dégradations du signal acoustique généré au niveau des tympans causées par les distorsions des écouteurs, ou bien par le couplage acoustique non naturel entre les écouteurs et le conduit auditif. Il semble également nécessaire d'assurer une cohérence entre les différents indices de localisation portés par les filtres binauraux : s'il existe un conflit entre les indices interauraux notamment, la source peut être perçue comme dédoublée, floue, ou bien difficilement localisable, et la perception intra-crânienne accompagne souvent ces attributs am-

bigus. On sait enfin que les indices supplémentaires apportés par la présence d'effet de salle, ainsi que les indices dynamiques favorisent la perception extra-crânienne, même si l'externalisation est aussi possible en synthèse binaurale statique.

Quelques études portant spécifiquement sur la synthèse binaurale ont suggéré que les indices spectraux participent à une bonne externalisation des sources sonores, au sens où une dégradation de ces indices mène à une dégradation de cet attribut [81, 267, 272, 273]. Ces résultats sont discutés par Loomis *et al.* [136], qui affirment que leur rôle est en fait très limité par rapport à celui de l'ITD, de l'ILD, des indices dynamiques [64], et de la présence d'effet de salle. Les auteurs l'ont démontré grâce à l'expérience suivante [135]. Des sujets étaient équipés d'écouteurs intra-auriculaires, d'un casque anti-bruit posé par dessus, et d'une paire de microphones, disposés sur le casque, captant l'environnement sonore. Un système permettait aux sujets d'écouter ce que captaient les microphones via les écouteurs, et ainsi d'annuler les indices spectraux induits par les pavillons, tout en conservant les indices interauraux, les indices dynamiques, et les indices liés à la salle. Dans ces conditions, une bonne externalisation a été obtenue pour 96% des sujets.

Pour conclure, il semble en tout cas que l'externalisation est un attribut perceptif fragile, rapidement affecté quand la position de la source est ressentie comme ambigue [64].

2.8 Calibration du casque

Pour une diffusion sur un casque, la fonction de transfert de celui-ci doit être compensée, si l'on veut contrôler finement la pression acoustique aux tympans de l'auditeur. Plusieurs études se sont penchées sur la question difficile de la calibration individuelle du casque [119, 126, 158, 163, 206, 217]. La réponse acoustique des transducteurs d'un casque n'est en général pas plate, et de plus le couplage entre le casque et les pavillons est fonction du casque lui-même, de son positionnement sur l'oreille, et de la forme du pavillon, spécifique à chaque auditeur. On englobe tous ces phénomènes dans une fonction de transfert appelée HPTF (*HeadPhone Transfer Function*), par laquelle les signaux binauraux doivent être déconvolués pour une restitution transparente. Les HPTF peuvent être mesurées de la même manière que les HRTF, en utilisant le casque lui-même pour générer les signaux d'excitation. Le spectre des HPTF présente des résonances et antirésonances prononcées, de facteur de qualité élevé, qui ressemblent fortement aux caractéristiques spectrales des HRTF [126, 158, 163, 206, 217], traduisant entre autres l'effet de filtrage des pavillons sous le casque. Différents positionnements du casque sont possibles sur les oreilles de l'auditeur, ce qui affecte la reproductibilité des mesures de HPTF : l'impact est modéré aux basses fréquences (< 6 kHz), mais assez prononcé aux hautes

fréquences, où l'écart-type des HPTF, d'un positionnement du casque à un autre, peut atteindre 9 à 10 dB [163, 270]. La reproductibilité est meilleure pour un casque circum-auriculaire que pour un casque supra-auriculaire [217], ainsi que lorsque le sujet positionne lui-même le casque, avec la consigne d'obtenir le meilleur confort [158]. Pour les casques supra-auriculaires, les variations observées sont notamment liées au fait que les coussins déforment plus ou moins les pavillons d'oreilles, affectant donc de façon variable les phénomènes acoustiques qui s'y produisent.

Les HPTF présentant un caractère individuel, il conviendrait de les évaluer pour chaque couple casque/auditeur. Selon Kulkarni et Colburn [126], la difficulté à évaluer de manière fiable les HPTF reste un problème majeur. Les auteurs affirment que même la calibration par une HPTF moyennée à partir de plusieurs mesures reste insatisfaisante, car les colorations qui subsistent ressemblent fortement à celles des HRTF, et risquent donc d'affecter la spatialisation. Selon McAnally et Martin [163], ce problème n'est pas insurmontable, car les variations des mesures des HPTF présentent une variance généralement bien moindre que les colorations utiles des HRTF, qui portent l'information directionnelle. Pralong et Carlile [206] précisent tout de même que l'utilisation d'une calibration non-individuelle est susceptible d'engendrer une dégradation de la qualité des VAS en synthèse binaurale, d'une ampleur équivalente à celle que provoque l'utilisation de HRTF non-individuelles (cf. 4.1.3). Wightman et Kistler [277] montrent que sans calibration du casque, il apparaît une dégradation des performances de localisation en élévation et un augmentation des confusions avant/arrière. Pour Kim et Choi [119], la calibration individuelle du casque serait indispensable pour assurer une bonne externalisation.

Sans calibration individuelle, on pourra tout de même diffuser une scène sonore sur un casque dont la calibration, ou égalisation, est connue par construction. Deux types d'égalisation du casque sont rencontrés en pratique :

- l'égalisation champ libre : la réponse en fréquence du casque est rendue plate, donc neutre, pour des sources de direction frontale,

- l'égalisation champ diffus : la réponse en fréquence du casque est neutre, pour la restitution d'un champ diffus, c'est-à-dire pour des sons non corrélés provenant de toutes les directions.

L'égalisation champ diffus est la plus courante, car des tests psycho-acoustiques ont montré qu'elle offre l'écoute la plus fidèle pour des signaux stéréo [249]. Cependant, ces égalisations étant effectuées par construction pour un certain jeu de HRTF, elles ne peuvent pas convenir à tous les auditeurs. L'égalisation en champ diffus est celle qui apporte le moins de caractéristiques individuelles, et en ce sens, c'est la plus appropriée à la restitution de la synthèse binaurale. Si l'on fait ce choix, une égalisation en champ diffus des HRTF doit être effectuée en préalable à la génération des signaux binauraux [154]. Cela consiste pour chaque oreille à diviser spectralement

les HRTF de chaque direction par la moyenne des HRTF dans toutes les directions. Cette égalisation est aussi intéressante du point de vue du coût d'implémentation : les spectres des HRTF ainsi égalisées sont moins accidentés, et elles peuvent donc être matérialisées par des filtres d'ordre réduit. Retenons que cette solution est un pis-aller, et qu'elle ne permet pas de contrôler rigoureusement les signaux binauraux.

"*[...] En terminant, il serait bon de se demander à quoi sert l'oreille. Ce sera vite fait : à rien ou à presque rien. Chez beaucoup d'animaux - voyez le cheval - le pavillon de l'oreille est manifestement destiné à recueillir les sons extérieurs, à les concentrer en quelque sorte sur le tympan ; à cet effet il est souvent doué d'une assez grande mobilité lui permettant d'en diriger la cavité vers le lieu supposé d'où part le son. Chez l'homme, le pavillon est parfaitement immobile ; quelques personnes cependant peuvent lui imprimer de petits mouvements, mais ceux-ci sont tout au plus bons à "étonner la galerie". Un physiologiste s'est demandé si, néanmoins, les méandres de l'oreille ne pouvaient conduire les sons. En conséquence il a supprimé monts et vallées en remplissant ces dernières de cire : le résultat a été que l'audition n'était nullement changée. Tout au plus le "patient" remarqua-t-il qu'il se rendait un peu moins bien compte de l'endroit d'où partaient les ondes sonores. Le pavillon de l'oreille serait donc surtout destiné à nous faire connaître la direction des sons : il remplit bien mal son rôle et si un fabricant d'appareils acoustiques nous donnait un pareil instrument, nous le refuserions sans tarder. Ce qui ne veut pas dire qu'il faille se couper les oreilles ou les cacher tout simplement ; les oreilles ont peut être des vertus cachées que nous ignorons : qui se serait, par exemple, douté, il y a quelques années, qu'elles serviraient un jour à retrouver les escarpes et autres malandrins qui ne songent qu'à nous être nuisibles !*" Henri Coupin, 1901 [59].

Chapitre 3

Indices spectraux

Ce chapitre a pour but de présenter l'état des connaissances sur les indices spectraux de la localisation auditive. Il s'agit d'abord de comprendre comment ils apparaissent, c'est-à-dire, physiquement, comment l'interaction entre l'onde sonore et le corps de l'auditeur modifie spectralement le signal entre la source et le tympan. On introduit séparément les contributions des différentes parties du corps : la tête, le buste, le pavillon. Après une description des caractéristiques typiques des colorations observées, les différents aspects de leur utilisation par le système auditif sont détaillés, ainsi que les principaux modèles proposés pour décrire leur identification au niveau central.

3.1 Origine physique des colorations spectrales

On décrit ici l'origine physique des colorations spectrales en observant le spectre d'amplitude des HRTF. Les effets des différentes parties du corps sont observés séparément : la diffraction par la tête, les réflexions sur le buste, et enfin les effets des pavillons, interprétables en termes de réflexions/diffraction ou de résonances. Bien que ces phénomènes acoustiques ne soient pas totalement découplés, on les traitera comme superposables, pour la clarté de l'exposé.

3.1.1 Contribution de la tête

Considérons ici la tête de l'auditeur, sans buste, ni pavillon, en adoptant à nouveau sa modélisation sphérique. Rabinowitz *et al.* [209] ont proposé une solution analytique pour décrire la pression acoustique à la surface d'une sphère rigide. Pour une source ponctuelle située à la distance r du centre de la sphère, on définit la fonction de transfert H entre la pression sur la sphère et la pression en son centre en l'absence de celle-ci. On a :

$$H = -\frac{\rho}{\mu} e^{-j\mu\rho}\psi \tag{3.1}$$

$$\psi(\rho,\mu,\theta) = \sum_{m=0}^{\infty}(2m+1)P_m(cos\theta)\frac{h_m(\mu\rho)}{h'_m(\mu)} \tag{3.2}$$

où a est le rayon de la sphère, θ l'azimut, $\mu = 2\pi fa/c$ la fréquence normalisée, $\rho = r/a$ la distance normalisée ($\rho > 1$), P_m le polynôme de Legendre d'ordre m, h_m la fonction de Hankel sphérique d'ordre m. A un décalage angulaire de $\pi/2$ près, et si l'on considère les oreilles comme étant diamétralement opposées sur la sphère, H est un modèle de la HRTF. On représente figures 3.1 et 3.2 son module dans le plan horizontal, pour $\rho >> 1$, calculé d'après l'algorithme de Bauck et Cooper [10]. Cela

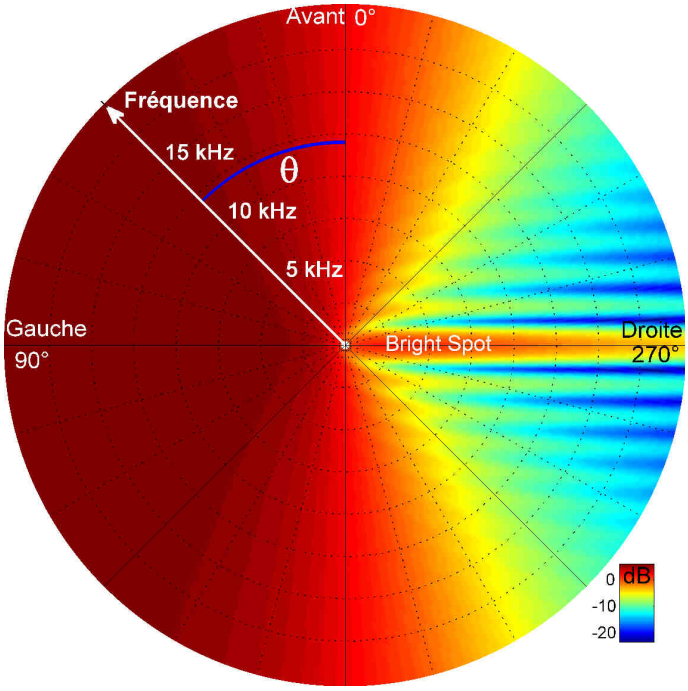


Figure 3.1 – Module en dB des HRTF gauches obtenues par le modèle simple de tête sphérique, sans buste ni pavillon, en représentation polaire. L'amplitude en dB est représentée en couleur, la fréquence est représentée radialement et croît à partir du centre, et l'azimut θ dans le système polaire-vertical est représenté en azimut.

correspond à une oreille gauche : on observe que le module est maximal en incidence normale, donc pour une source à 90°, et que la diffraction par la tête montre ses effets quand l'oreille devient controlatérale.

Du côté ipsilatéral, on observe une amplification des hautes fréquences correspondant à une réflexion spéculaire dans la direction de la source. Des interférences constructives entre onde incidente et onde réfléchie entraînent un gain de 6 dB au niveau de l'oreille. A mesure que la source s'éloigne de 90°, le phénomène s'estompe, et le gain diminue.

Du côté controlatéral, on note l'existence d'un autre phénomène, appelé *bright spot* : pour des sources situées dans l'ombre de la tête, une amplification est observée. On peut l'interpréter par l'existence de deux ondes se propageant d'un côté et de l'autre de la sphère, et qui s'ajoutent en phase quand les deux trajets sont de longueur égale, c'est-à-dire pour une incidence de 270° [63]. Autour de cette position, sur les intervalles [210°-270°] et [270°-330°], des interférences constructives et destructives créent des oscillations dans le spectre. Algazi *et al.* [1], et Brungart *et al.* [32] ont effectué des mesures de HRTF sur une tête artificielle, montrant une bonne adéquation avec le modèle. Cependant le spectre mesuré se révèle plus complexe au

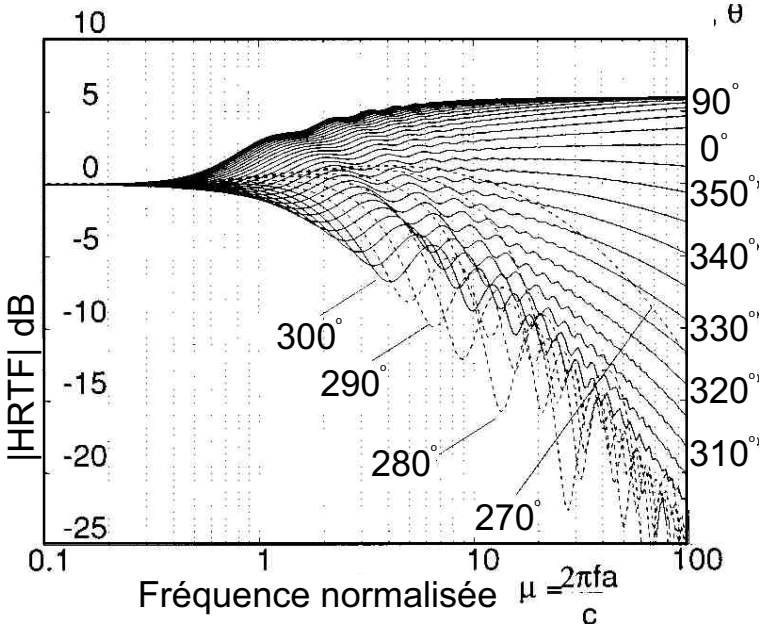


Figure 3.2 – Module en dB des HRTF gauches obtenues par le modèle simple de tête sphérique (d'après [63])

niveau des frontières de l'ombre créée par la tête. Par ailleurs, du fait de la position décentrée des oreilles vers l'arrière, le *bright spot* est observé à un azimut légèrement inférieur à 270°.

3.1.2 Contribution du buste

La principale contribution du buste réside dans l'existence d'une réflexion de l'onde incidente sur les épaules. Dans le domaine fréquentiel, cet écho se traduit classiquement par un filtrage en peigne : des oscillations viennent entacher le spectre d'amplitude, avec une période d'autant plus grande que le retard est petit. Pour décrire ces phénomènes, Algazi *et al.* proposent un modèle simple composé d'un buste sphérique surmonté d'une tête sphérique : c'est le modèle *Snowman* [2]. Dans le plan médian, le modèle donne les HRTF représentées figure 3.3. La différence de marche est maximale pour une élévation de 90°, entraînant des oscillations rapides du spectre. A mesure que les sources se décalent vers des positions plus basses, les oscillations du spectre sont de plus en plus lentes. Cette dépendance directionnelle de la période d'oscillation du spectre en fait un indice potentiel de localisation en élévation.

Un modèle de buste ellipsoïdal permet de mieux appréhender ces effets, avec comme paramètres les dimensions de la tête, du cou, et du buste, ainsi que le décalage de la position des oreilles par rapport à la tête, et celui de la tête par rapport au

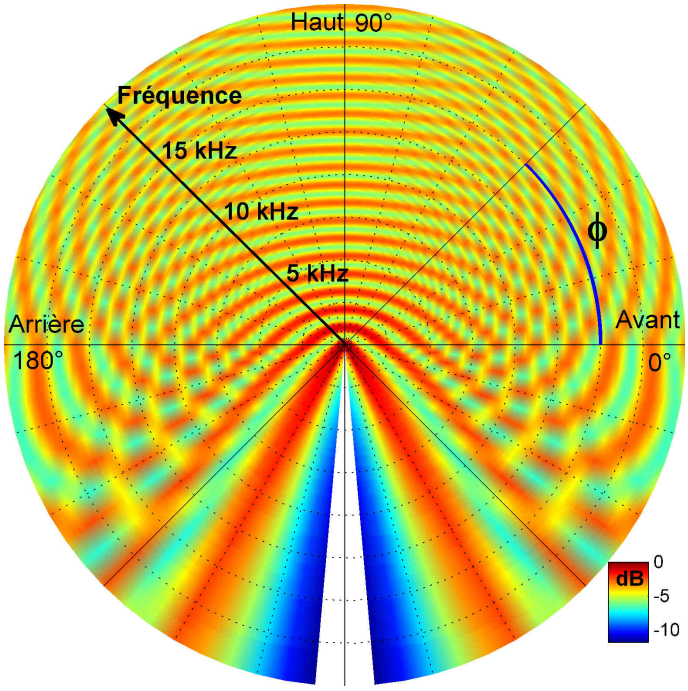


Figure 3.3 – Représentation polaire du module des HRTF dans le plan médian obtenues par le modèle *Snowman*[2]

buste [1].

3.1.3 Contribution des pavillons

Batteau [8] fut l'un des premiers à souligner l'importance des pavillons d'oreille dans la localisation auditive. L'auteur émit l'hypothèse que son rôle était de créer, grâce à ses multiples cavités, une série d'échos encodant des informations précises sur la localisation des sources. Des mesures ont été réalisées sur des répliques de pavillons humains, pour estimer l'évolution spatiale du retard entre l'onde directe, et l'onde réfléchie par la surface de la conque : il varierait de façon monotone de 10 à 100 μs avec l'azimut, et de 100 à 300 μs avec l'élévation. Les échos générés par le pavillon seraient détectés séparément dans le temps, et analysés par le système auditif dans le domaine temporel. La principale critique émise contre cette théorie temporelle est le fait que ces intervalles sont extrêmement courts, et qu'il est donc improbable que le système auditif soit en mesure d'en tirer parti [47]. Bien que cette question soit toujours controversée, il est plus probable que l'analyse des phénomènes induits par les pavillons soit exclusivement menée dans le domaine fréquentiel : à chaque direction de source est associé un profil spectral particulier, telle une signature, que le système auditif est capable de décoder pour localiser une source. C'est cette

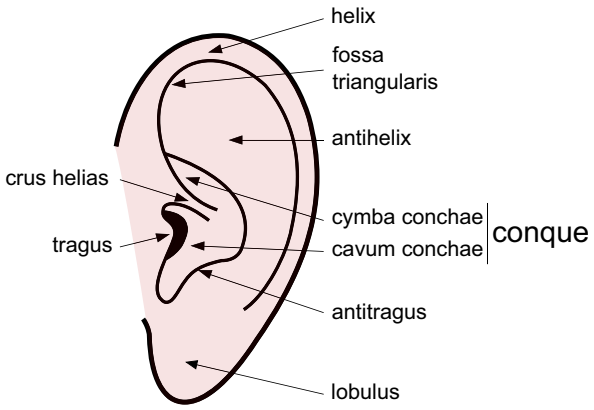


Figure 3.4 – Anatomie du pavillon

hypothèse, adoptée par la majorité des études postérieures à celle de Batteau, que nous retenons dans la suite de notre exposé.

Les pavillons jouent un rôle capital dans la création d'indices spectraux, et ils sont de premier intérêt pour la localisation car ils présentent une forte dépendance avec la direction. Cela est dû notamment aux fortes asymétries du pavillon : on peut observer figures 3.4 et 3.6 l'anatomie de la structure complexe du pavillon. On nommera conque l'ensemble des deux cavités nommées *cymba conchae* et *cavum conchae*, et on parlera de rebord du pavillon pour désigner l'ensemble constitué de *fossa triangularis*, *helix*, *antihelix* et *lobulus*. On illustre figure 3.5, les morphologies très différentes observées d'un individu à un autre. Après un descriptif des caractéristiques spectrales typiques induites par le pavillon, on détaillera leur origine physique selon différents modèles.

Description des colorations typiques induites par le pavillon

Sont décrites ici les modifications spectrales engendrées par le pavillon, d'après les observations communes à plusieurs études. Elles sont illustrées sur des HRTF issues de la base privée d'Orange Labs. Les azimuts et élévations indiqués dans ce qui suit sont relatifs au système polaire-vertical.

Shaw et Teranishi [236], et Shaw [234] ont montré que les effets du pavillon étaient observables dans le spectre d'amplitude des HRTF aux fréquences supérieures à 3.5 kHz à 4 kHz environ, ce qui est en accord avec ce que permettent de prédire les dimensions du pavillon. Les colorations peuvent être décrites en termes de pics et de creux dans le spectre, comme illustré figures 3.7, 3.8 et 3.9.

Pics spectraux

– Un pic lié à la résonance du conduit auditif a été identifié à environ 2.5 kHz

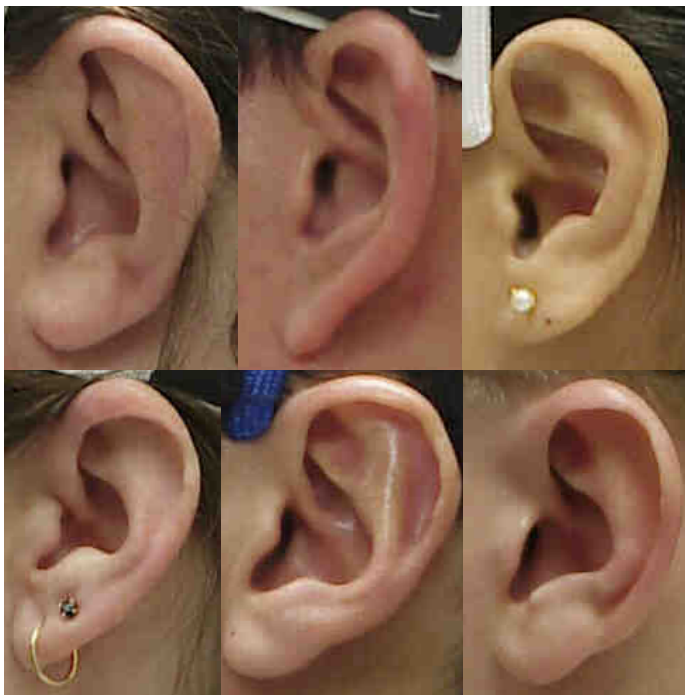


Figure 3.5 – Photographies de pavillons gauches de 6 individus (sujets de la base du CIPIC [3])

[236], mais sa présence dépend de la méthode employée pour la mesure des HRTF : il est présent pour une mesure conduit ouvert, absent pour une mesure conduit bloqué. Cette caractéristique n'est pas d'un grand intérêt pour la localisation, car ce pic ne présente pas de dépendance avec la direction de la source.

– Une asymétrie globale avant/arrière est observable dans le plan horizontal. Le module du côté ipsilatéral est plus élevé à l'avant qu'à l'arrière, ce qui est directement expliqué par l'effet d'ombre induit par le rebord de l'oreille [248].
– Un pic (P1) assez large est observé du côté ipsilatéral, et dans l'hémisphère avant entre 3.5 kHz et 5.5 kHz [51, 137]. Carlile *et al.* signalent un décalage progressif de la fréquence centrale de ce pic vers les hautes fréquences quand la source va du plan médian au plan vertical interaural [51].
– Autour de 8 à 10 kHz, un pic (P2) apparaît pour des directions proches du plan vertical interaural, côté ipsilatéral, et pour des élévations inférieures au plan horizontal [51, 137, 174].
– Autour de 11 kHz à 14 kHz, un pic (P3) apparaît seulement dans l'hémisphère avant. Son amplitude décroît pour une élévation croissante [5], et il apparaît parfois centré sur le plan médian, parfois un peu plus proche du plan vertical interaural [51].

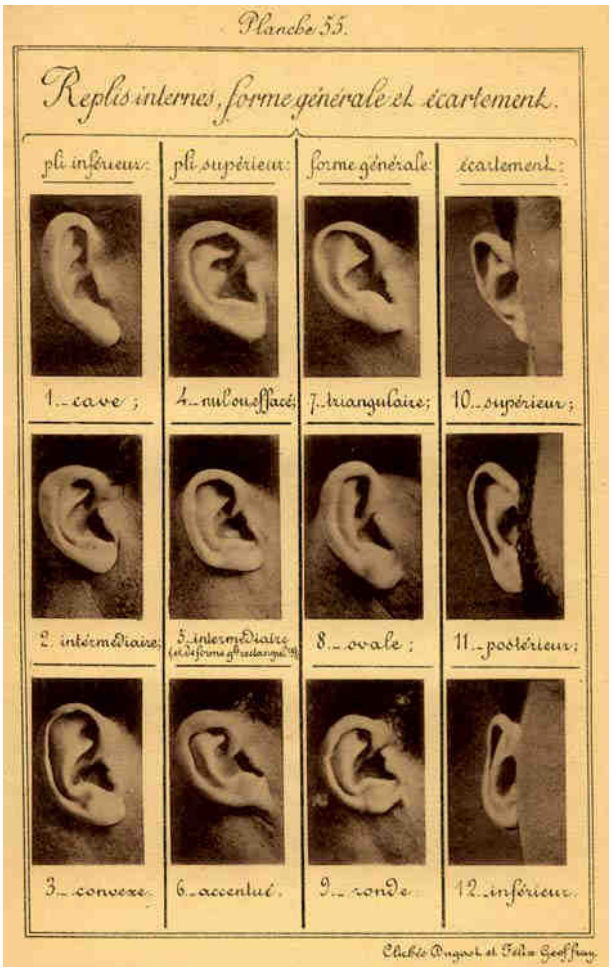


Figure 3.6 – Etude anthropométrique de l'oreille (d'après [16])

Creux spectraux

- Un premier creux (C1) a été révélé par de nombreuses études, dans l'hémisphère avant du côté ipsilatéral [5, 22, 40, 51, 85, 137, 236], et aussi côté controlatéral [137]. Quel que soit l'azimut, on observe une même dépendance avec l'élévation, mais avec un décalage progressif de la fréquence centrale vers les hautes fréquences pour des directions allant du plan médian au plan vertical interaural[51, 137]. La pente basses-fréquences de ce creux devient d'autant moins raide que l'élévation augmente [51].
- Un deuxième creux (C2) est mis en évidence. Il apparaît plutôt en dessous du plan horizontal, et sa fréquence centrale, quasi indépendante de l'élévation, évolue légèrement avec l'azimut : d'environ 10 kHz près du plan médian, à environ 12 kHz près du plan vertical interaural [137]. Ce creux est moins net que le premier, et se dédouble parfois.
- Un troisième et dernier creux (C3) est observé, dont les caractéristiques sont très dépendantes de l'individu [137]. Il apparaît plutôt pour des élévations inférieures au plan horizontal : sa fréquence centrale, d'environ 16 kHz à 17 kHz, augmente avec l'élévation. Ce creux présente la même dépendance en azimut que les deux premiers.

Evolution en termes de fonctions de directivité

Une représentation alternative des HRTF permet de cerner quelques comportements typiques. Il s'agit d'observer l'évolution spatiale du module des HRTF sous forme d'une série de fonctions de directivité pour une fréquence fixe (figures 3.10 et 3.11). Plus précisément, on peut s'intéresser aux principaux lobes de directivité, ainsi qu'à l'axe acoustique, défini comme la direction pour laquelle le module des HRTF est maximal à une fréquence donnée. Pour une fréquence croissante, on observe classiquement les phénomènes suivants [174](Cf Fig. 3.10 et 3.11) :

- Autour de 4 kHz, un lobe unique, ou monopôle, occupe l'essentiel de l'hémisphère ipsilatéral. L'axe acoustique est généralement situé à l'avant, près du plan horizontal.
- Entre 4 kHz et 7 kHz, Le lobe devient plus directif et se décale vers l'arrière et vers le haut, tandis qu'un autre lobe apparaît sous le plan horizontal.
- Autour de 7 kHz à 9 kHz, un comportement dipolaire est observé : deux lobes coexistent centrés près du plan vertical interaural, l'un dans l'hémisphère supérieur, l'autre dans l'hémisphère inférieur, séparés par une vallée proche du plan horizontal.
- On nomme *breakpoint* [37] la zone fréquentielle à laquelle le lobe inférieur devient prépondérant. L'évolution spatiale de l'axe acoustique connaît alors une discontinuité : il passe brutalement de haut en bas, éventuellement vers l'avant.

- Le lobe supérieur disparaît tandis que le lobe inférieur glisse progressivement vers l'arrière et vers le haut. Un autre lobe apparaît autour du plan horizontal, à l'avant.
- Autour de 12 kHz à 14 kHz, un autre comportement dipolaire apparaît : deux lobes centrés près du plan horizontal, l'un à l'avant, l'autre à l'arrière, sont séparés par une vallée proche du plan vertical interaural.
- Un second *breakpoint* est observé quand le lobe avant devient prépondérant. L'axe acoustique passe brutalement d'arrière en avant.
- Le lobe arrière disparaît et le lobe avant se décale de bas en haut, et d'avant en arrière.
- Aux plus hautes fréquences, le comportement est multipolaire, et surtout très variable d'un individu à l'autre.

Il en résulte une trajectoire typique de l'axe acoustique pour une fréquence croissante, prenant la forme d'un α (cf. Fig. 3.12).

Toutes les études, ainsi que nos investigations, montrent que les fréquences et les directions auxquelles ces caractéristiques sont observées varient d'un individu à un autre, et ce d'autant plus qu'ils apparaissent aux hautes fréquences ($>$ 12 kHz). On remarque que les caractéristiques spectrales induites par les pavillons sont prépondérantes aux hautes fréquences. Les colorations spectrales induites par le buste ne peuvent donc être perceptivement utiles qu'aux fréquences inférieures. L'effet d'ombre de la tête reste lui toujours visible du côté controlatéral, et on distingue encore le *bright spot* (cf. Fig. 3.7).

Origine physique des indices induits par le pavillon

Les effets de coloration décrits précédemment ont été attribués au comportement de résonateur multi-modes du pavillon. C'est Shaw le premier qui mit en évidence l'existence de modes de résonance dans les cavités du pavillon [232, 235, 236]. Il en a mesuré les caractéristiques, conduit auditif bloqué, sur des répliques de pavillons de dix sujets, positionnées sur un écran [234], et excités par une source en incidence rasante. Shaw observe que d'un pavillon à l'autre ces modes sont excités par des directions de source légèrement différentes, à des fréquences différentes, et des niveaux différents. Cependant, il identifie des comportements communs à tous les pavillons, et établit l'existence de six modes (cf. Fig. 3.13). Le mode 1 est omnidirectionnel, et correspond à une simple résonance quart d'onde, avec une pression uniforme dans la conque. Les cinq modes suivants sont essentiellement transverses, et on distingue les modes 2 et 3 qui sont dits "verticaux", des modes 4, 5 et 6 qui sont dits "horizontaux", simplement parce qu'ils sont excités de façon maximale par des sources situées près du zénith pour les premiers, et près du plan horizontal pour les suivants. Tous

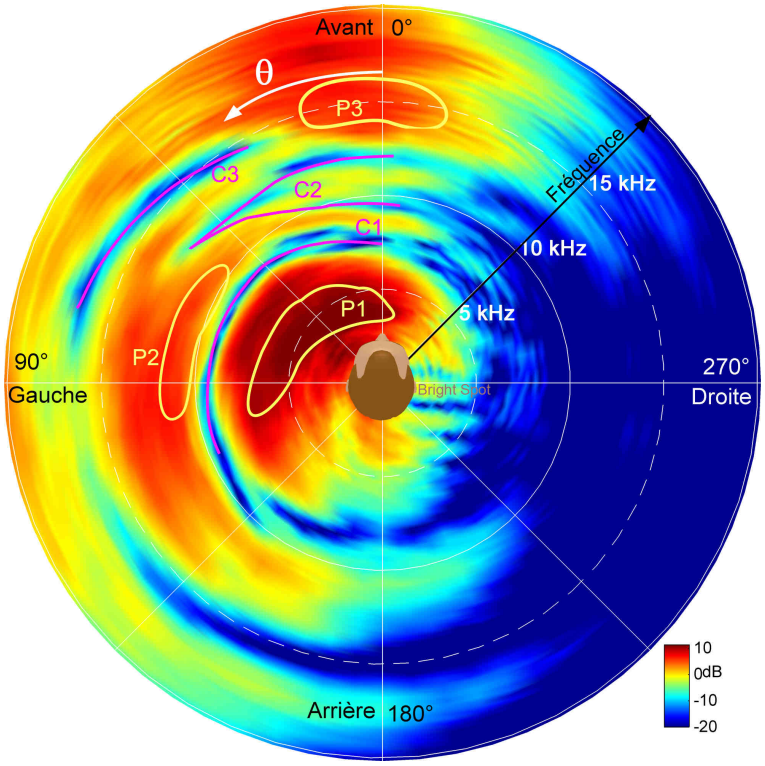


Figure 3.7 – Illustration des creux et pics caractéristiques observés dans le plan horizontal : représentation polaire du module des HRTF d'un sujet de la base privée d'Orange Labs (sujet n°5, oreille gauche, mesurées à l'entrée du conduit auditif, conduit bloqué), en fonction de l'azimut dans le système polaire-vertical (cf. Fig. 1).

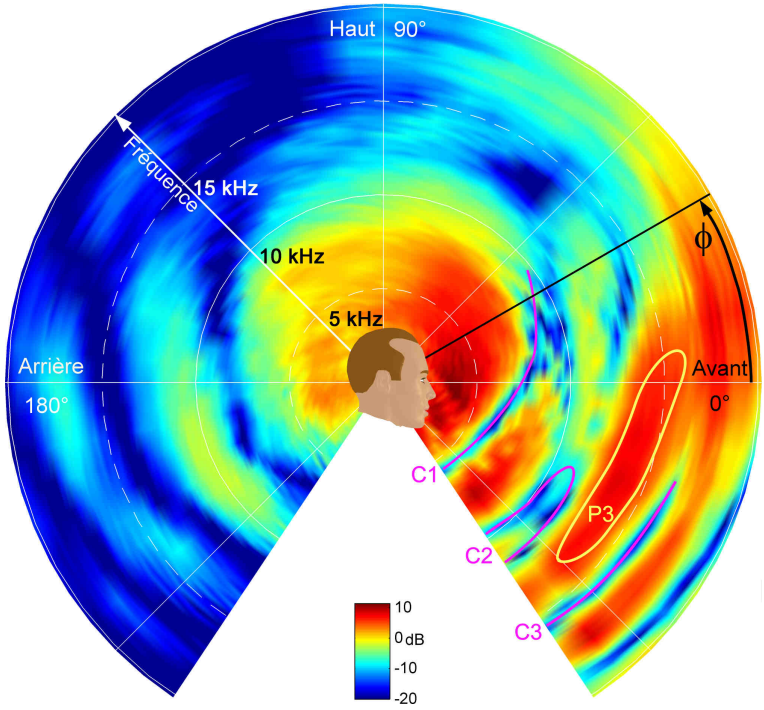


Figure 3.8 – Illustration des creux et pics caractéristiques observés dans le plan médian : représentation polaire du module des HRTF d'un sujet de la base privée d'Orange Labs (sujet n°5, oreille gauche, mesurées à l'entrée du conduit auditif, conduit bloqué), en fonction de l'élévation dans le système polaire-horizontal (cf. Fig. 1).

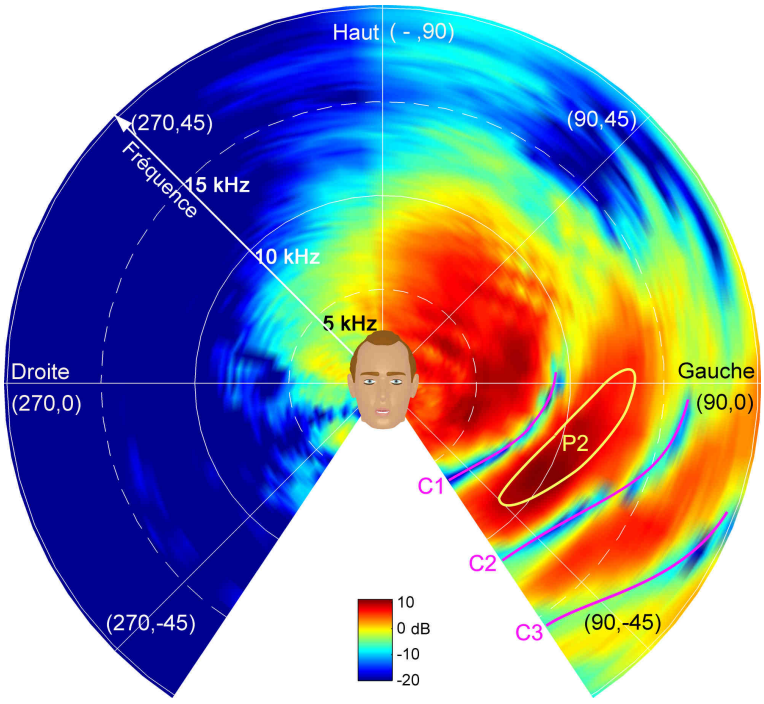


Figure 3.9 – Illustration des creux et pics caractéristiques observés dans plan vertical interaural : représentation polaire du module des HRTF d'un sujet de la base privée d'Orange Labs (sujet n°5, oreille gauche, mesurées à l'entrée du conduit auditif, conduit bloqué). On indique le couplet (azimut,élévation) en degrés dans le système polaire-vertical (cf. Fig. 1).

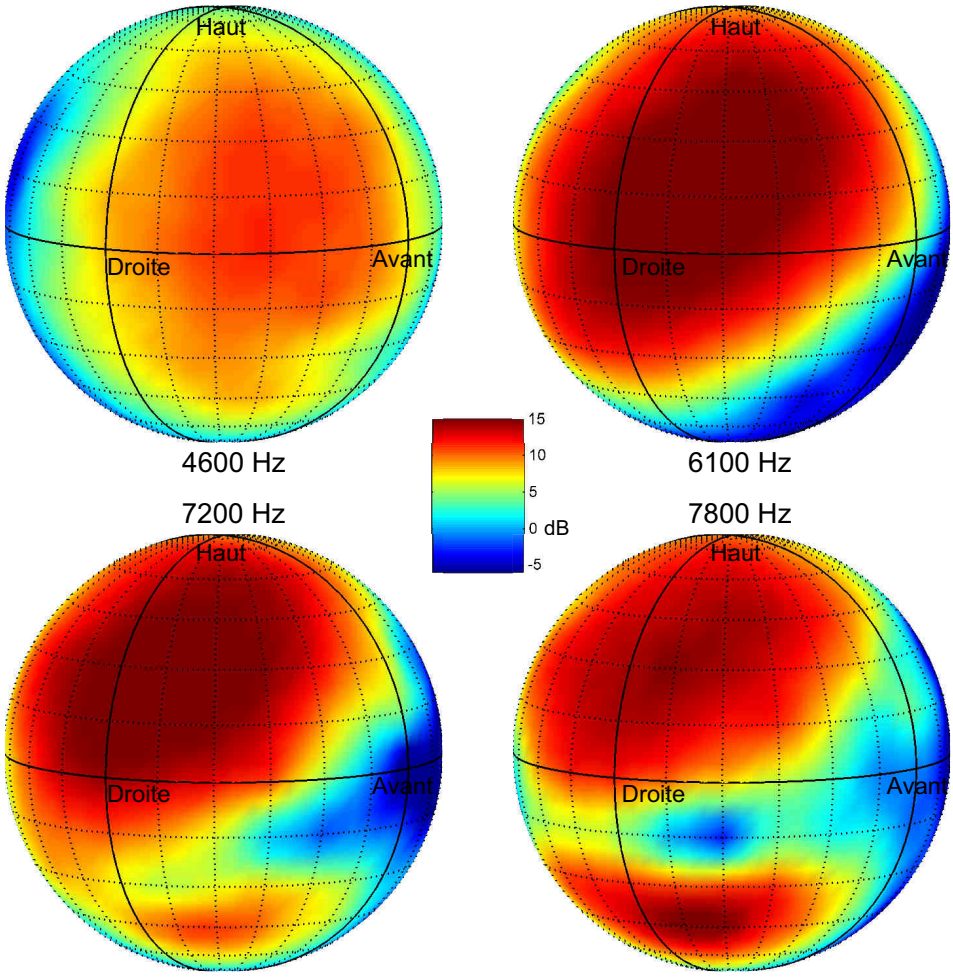


Figure 3.10 – Représentation sphérique du module des HRTF (en dB) pour une fréquence donnée (systèmes de coordonnées polaire-vertical (cf. Fig. 1), oreille droite du sujet n°59 de la base *Listen* de l'IRCAM, mesurées à l'entrée du conduit auditif, conduit bloqué). Pour une meilleure lisibilité, la moyenne spatiale du module des HRTF est retranchée, ce qui correspond à la définition des DTF (*Directional Transfer Functions*) (cf. *infra*).

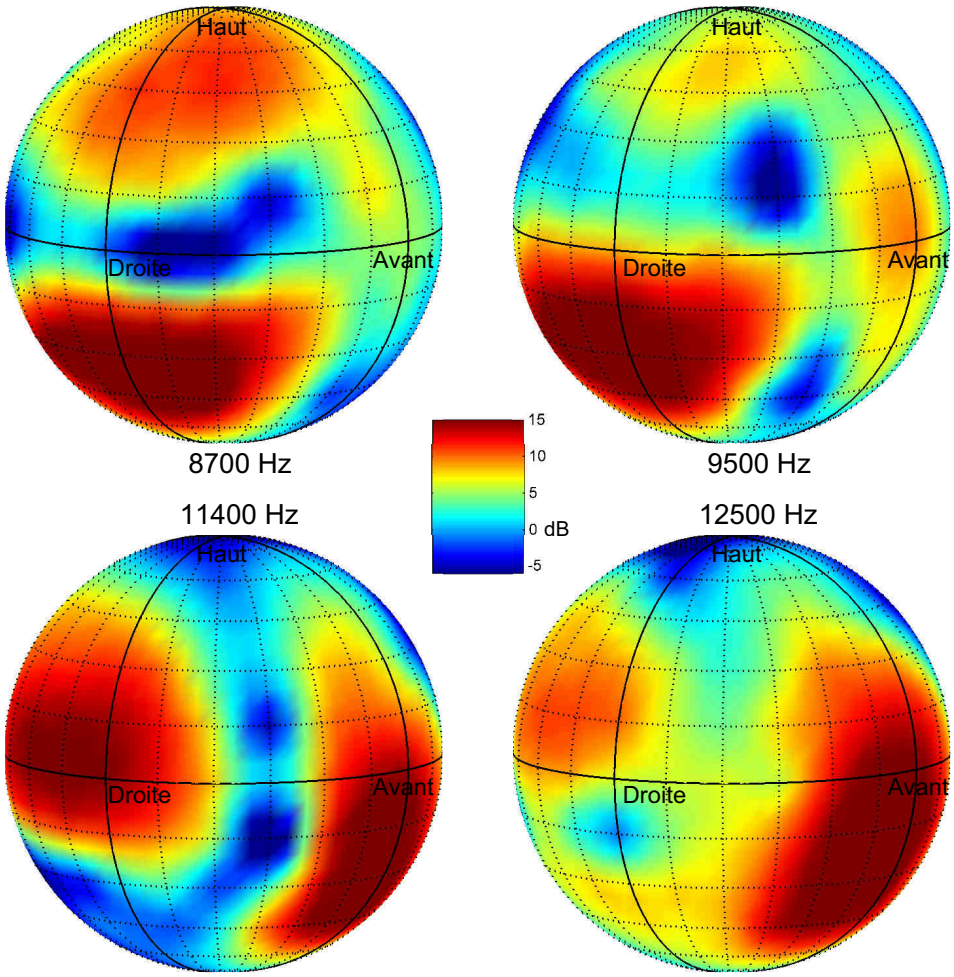


Figure 3.11 – Représentation sphérique du module des HRTF (en dB) pour une fréquence donnée (suite de la figure 3.10).

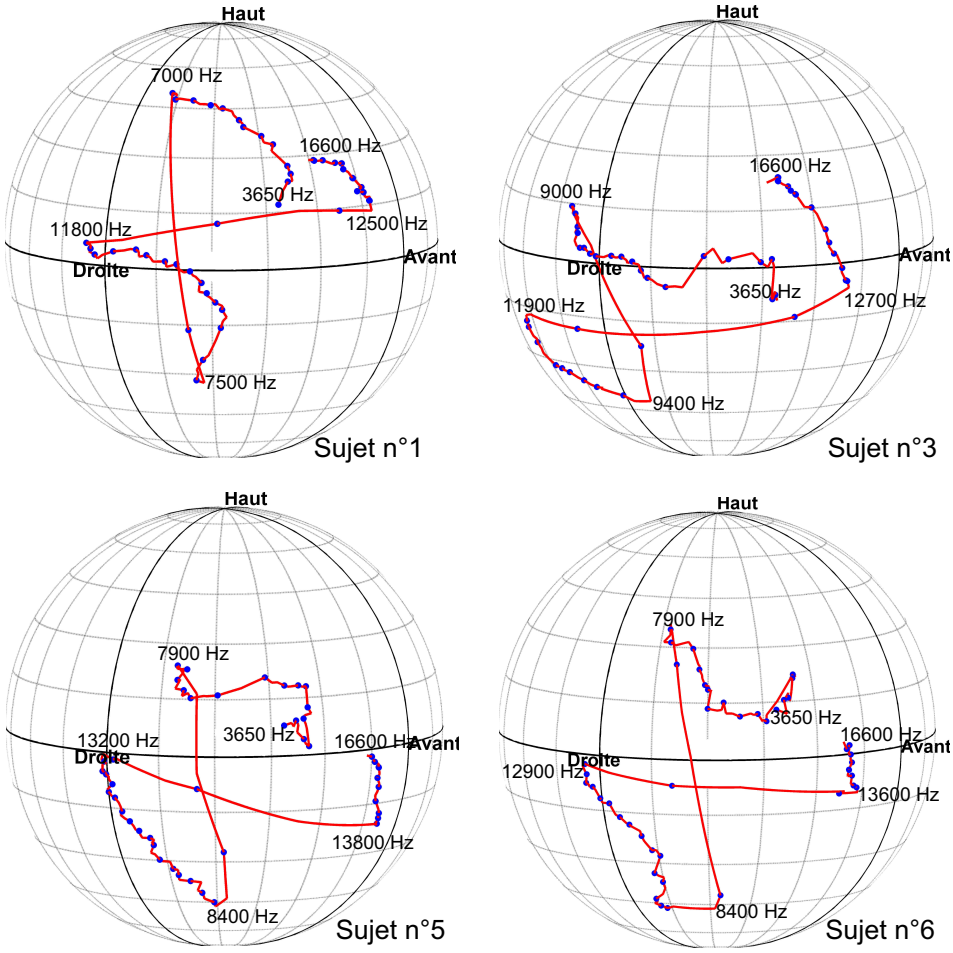


Figure 3.12 – Trajectoire de l'axe acoustique pour les oreilles droites de 4 sujets de la base privée de HRTF d'Orange Labs.

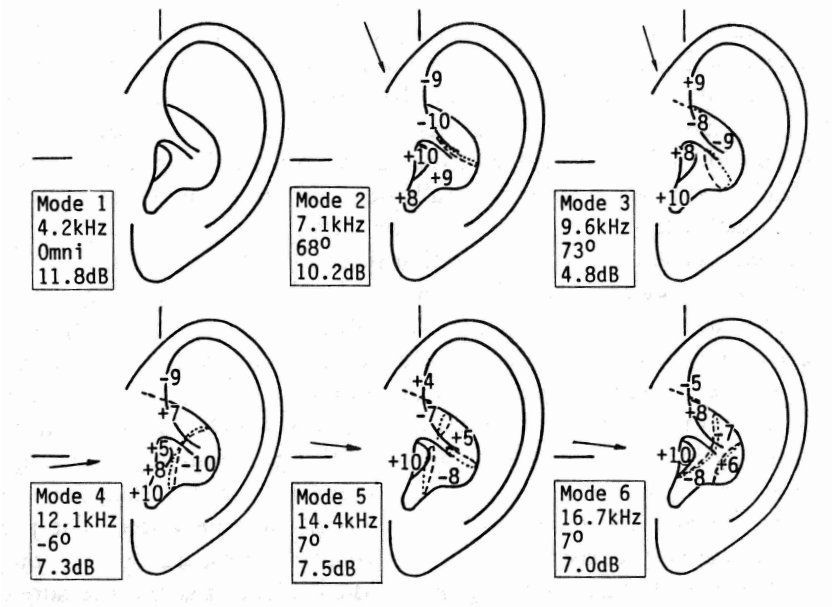


Figure 3.13 – Modes de résonance du pavillon, d'après [235]. Pour chacun des modes, on indique sa fréquence, l'élévation pour laquelle il est le plus excité (dans le plan médian), et le profil du mode dans le pavillon : maxima et minima de pression (en dB), et surfaces nodales. (Comportements moyens observés d'après les mesures pour 10 pavillons différents.)

ces modes transverses présentent un gain qui dépend de la direction d'excitation. Les techniques de modélisation numérique (BEM) ont permis à Kahana et Nelson [115] de calculer ces modes pour un pavillon en silicone équipant la tête artificielle KEMAR. Les auteurs observent une bonne adéquation avec les résultats de Shaw.

Shaw met en évidence le fait que la distribution de pression pour le mode 2 est celle d'un dipôle vertical, et que celle du mode 4 est celle d'un dipôle horizontal. Ces comportements peuvent être observés dans la représentation sphérique des HRTF : le mode 2 à 7600 Hz, figure 3.10, et le mode 4 à 12100 Hz, figure 3.11. L'analyse des modes de résonance du pavillon, initiée par Shaw, permet de comprendre la formation de pics dans le spectre des HRTF. Cependant l'apparition de creux, n'est pas clairement expliquée par le phénomène des modes de résonance. Shaw et Teranishi ont avancé l'hypothèse que le creux noté C1 serait le résultat d'une interférence entre les deux premiers modes du pavillon (mode 1 et mode 2) [236]. Musicant *et al.* [187] l'expliquent ainsi : ou bien à la fréquence du creux, l'amplitude des différents modes est particulèrement faible en comparaison avec les fréquences voisines, ou au contraire les modes adjacents sont excités fortement à cette fréquence, mais en opposition de phase, ce qui rejoint la thèse de Shaw et Teranishi. Ces hypothèses n'ont cependant été confirmées par aucun calcul, ni aucune expérience. Lopez-Poveda [138] propose un modèle considérant non seulement les nombreuses réflexions de l'onde incidente

dans la conque avant d'atteindre te tympan, mais aussi les effets de la diffraction par l'ouverture de la conque. En première approximation, il n'a considéré que les premières réflexions des ondes préalablement diffractées par l'ouverture. Suivant un modèle acoustique, l'auteur parvient à expliquer par le calcul l'apparition des creux C1 et C3, mais pas celle de C2.

Lopez-Poveda [138] propose une décomposition structurelle du pavillon : en retirant ou en conservant différentes parties de l'anatomie d'une réplique du pavillon, il cherche par la mesure à relier chaque caractéristique spectrale à la présence de telle ou telle partie du pavillon. Il conclut que cette correspondance simple n'existe pas : tous les éléments du pavillons contribuent conjointement à l'apparition des phénomènes observés. Néanmoins, l'auteur parvient à relier la formation du creux C2 à l'existence du *tragus*, et du *crus helias*, qui sépare la conque en deux cavités : *cymba conchae* et *cavum conchae* (Cf Fig. 3.4). Shaw et Teranishi avaient auparavant montré le rôle important du rebord du pavillon dans la dépendance directionnelle des modes 2 et 3 [248], et celui de la cavité nommée *fossa triangularis* et du *crus helias* dans la dépendance directionnelle des modes supérieurs [234].

3.2 Résultats psychophysiques d'intérêt

De nombreuses expériences ont été menées pour mieux cerner la nature des indices spectraux, et les conditions nécessaires pour leur bonne utilisation par le système auditif.

3.2.1 Preuve de l'utilité des IS induits par les pavillons

Les insuffisances de la théorie duplex ont mené à l'idée que les IS, introduits par le pavillon notamment, permettent de lever l'ambiguïté de localisation sur un cône de confusion. Il restait néanmoins à démontrer leur utilisation effective par le système auditif. Une preuve a été apportée par les expériences de Roffler et Butler [221] et Gardner [74]. Des tests de localisation dans le plan médian ont été réalisés sur des sujets dont on avait couvert les pavillons, ne laissant qu'une ouverture en face du conduit auditif [221], ou bien sur des sujets dont on avait progressivement comblé les différentes cavités du pavillon par des moulages [74]. Ces modifications des pavillons perturbent les phénomènes physiques décrits précédemment, et les indices spectraux sont donc altérés. Les deux études mènent aux mêmes résultats : les performances de localisation dans le plan médian sont d'autant plus dégradées que les pavillons sont modifiés. La précision des sujets à reporter l'élévation de la source est nettement amoindrie, et les confusions avant/arrière augmentent.

3.2.2 Bande fréquentielle des IS

La tête et le torse introduisent également des colorations spectrales qui présentent une dépendance avec la direction de la source aux basses fréquences. Algazi *et al.* [1] ont évalué les caractéristiques spectrales aux basses fréquences introduites par leur modèle tête et torse (*Head And Torso*, ou HAT). Les résultats indiquent que des IS secondaires existent en-dessous de 3 kHz, là où l'effet acoustique des pavillons est négligeable. Ces indices permettraient de localiser en élévation, avec une précision toutefois assez limitée, et ils seraient d'autant plus importants que la source est positionnée loin du plan médian. Cependant, Morimoto *et al.* [183] montrent que les HRTF aux basses fréquences ne comportent pas d'IS saillants pour la discrimination avant/arrière (ou frontalisation), en comparaison avec les IS disponibles aux hautes fréquences. Pour cerner la bande fréquentielle des IS d'intérêt, des tests de localisation ont été menées dans le plan médian en jouant sur la largeur fréquentielle des stimuli à localiser. Puisque les indices interauraux (ITD et ILD) sont quasiment nuls sur le plan médian, une dégradation des performances de localisation peut être directement imputée à l'absence d'IS utilisables dans la bande occupée par le stimulus. Plusieurs études montrent ainsi que les IS d'intérêt sont situés entre 4 kHz et 16 kHz, [85, 120, 131], c'est-à-dire qu'ils sont essentiellement induits par les pavillons. S'ils existent, les IS aux basses fréquences sont donc d'une importance mineure quand les stimuli contiennent de l'énergie aux hautes fréquences.

3.2.3 Aspects temporels

Hofman et Van Opstal [90] ont étudié l'influence de la durée des stimuli. Les auteurs montrent que pour une bonne localisation en élévation, le système auditif doit pouvoir évaluer le signal source sur une durée totale d'au moins 80 ms, et qu'il fonde son jugement sur une stratégie dite *multiple look*, c'est-à-dire qu'il évalue le spectre sur des observations successives à court terme, dont la durée a été évaluée à 5 ms. Cette durée de la fenêtre d'analyse est en accord avec les résultats de Viemester et Wakefield [256], et avec ceux de Jin [110]. Une analyse similaire est tirée de l'étude de Vliegen et Van Opstal [257], mais les auteurs observent par contre qu'un plateau est atteint dans les performances de localisation à partir d'une durée de stimulus de 30 ms.

3.2.4 Influence du niveau du stimulus

Hartmann et Rakerd [80] montrent que l'augmentation du niveau du stimulus peut dégrader les performances de localisation en élévation en champ libre : cette dégradation, appelée *negative level effect*, est observée sur des trains de clics très courts, mais pas sur des bruits blancs de même durée totale que les trains de clics.

Les auteurs peinent à expliquer cette différence. Alves-Pinto et Lopez-Poveda [4] ont eux étudié la capacité du système auditif à discriminer des stimuli de type bruit blanc, de stimuli comparables mais entachés de creux dans le spectre. Ce deuxième type de stimulus est, dans leur expérience, représentatif de l'effet possible du filtrage directionnel induit par le pavillon sur un stimulus source large bande. Les résultats indiquent que la discrimination de ces deux stimuli évolue de façon non monotone : elle se dégrade progressivement avec un niveau croissant, pour atteindre un seuil maximum à 70-80 dB, puis s'améliore pour des niveaux supérieurs. Les auteurs concluent qu'en termes de localisation sur la base des IS, donc notamment en élévation, on doit s'attendre à de meilleures performances à 60-70 dB qu'à 70-80 dB. Cette variation non monotone en fonction du niveau est corroborée par les résultats des tests de localisation menés par Vliegen et Van Opstal [257].

3.2.5 Influence de la largeur de bande du stimulus

On doit à Blauert [18] l'une des premières expériences montrant l'effet de largeur de bande sur la perception de la localisation. L'auteur a étudié la localisation d'une source réelle, en champ libre, dans le plan médian, émettant un stimulus de bande étroite. Les stimuli sont localisés à des positions complètement indépendantes de la position réelle de la source, mais dans des directions très corrélées avec la fréquence centrale des signaux. L'auteur reporte les tendances observées en termes de bandes directives : chaque fréquence est associée à une direction du plan médian. Voici leur évolution typique pour une fréquence croissante :

- les sons à 4 kHz sont localisés plutôt devant ;
- ils "migrent" vers l'arrière pour une fréquence croissante du stimulus ;
- les sons à 8 kHz sont reportés autour du zénith ;
- ils continuent à se décaler vers l'arrière ;
- à 12 kHz, les stimuli sont localisés complètement à l'arrière ;
- enfin autour de 16 kHz on observe un retour discontinu à l'avant.

Ces résultats sont moyennés d'après les observations sur plusieurs sujets, et ne permettent pas de juger des différences entre les individus. Butler [43] prouve que ces tendances sont générales, mais que les fréquences auxquelles ces phénomènes apparaissent sont différentes d'un individu à l'autre. Middlebrooks [167] étend ces résultats à toutes les directions de l'espace : l'auteur montre que la perception l'élévation est très liée à la fréquence centrale du stimulus, mais que la perception de l'azimut n'est pas affectée. Les confusions avant/arrière augmentent également quand le spectre est étroit. Jin montre que ces illusions s'estompent à mesure que la largeur de bande augmente [110].

3.2.6 Influence du profil spectral du stimulus

Les résultats présentés en 2.5 montrent que le système auditif tolère une diminution de la résolution spectrale des HRTF. Des expériences ont en outre été menées pour évaluer l'effet induit par la présence d'accidents prononcés dans le spectre des stimuli. L'idée était d'observer jusqu'à quel point le système auditif est en mesure de distinguer les accidents du spectre de la source des indices spectraux nécessaires à sa localisation. MacPherson [140] a utilisé des stimuli de spectre très irrégulier, dont l'amplitude variait jusqu'à $\pm$ 20 dB d'une bande fréquentielle à l'autre (1/3 d'octave). L'étude montre que la localisation en élévation reste assez précise, même si les reports de localisation sont plus dispersés que dans le cas d'un stimulus de spectre plat. D'autres expériences ont étudié l'effet de stimuli présentant une ondulation du spectre [145]. Une bonne perception de l'élévation est préservée pour des périodes d'oscillations très basses (< 0.5 période par octave) et très élevées (> 2 périodes par octave), mais on observe des dégradations dans les cas intermédiaires, autour d'une période par octave. Une autre expérience, menée par Qian et Eddins [208], indique que la limite basse se situe plutôt vers 0.1 période par octave. Ce résultat indique que lorsque les caractéristiques du spectre source ressemblent aux colorations utiles des HRTF, le système auditif n'est pas en mesure de séparer les deux pour former un percept fiable de l'élévation de la source. Une source sonore est donc localisée de façon idéale si elle offre de l'énergie sur toute la bande de fréquences où se situent les indices spectraux, et si les variations du spectre sont suffisamment neutres. C'est pourquoi la plupart des tests de localisation utilisent classiquement des stimuli tels que des bruits blancs.

3.2.7 Rôle des IS pour la localisation en azimut

Il apparaît que le spectre des HRTF contient potentiellement des indices de latéralisation autant que d'élévation (cf. Fig. 3.7, 3.10, 3.11). Une expérience intéressante a été menée par Van Wanrooij et Van Opstal [255] sur des sourds monauraux. Leurs résultats indiquent que certains de ces sujets utilisent les IS pour juger la latéralisation d'une source. En effet, leurs performances de localisation en azimut sont moindres quand leurs pavillons sont modifiés. Les sourds monauraux qui tirent partie des IS pour la latéralisation sont les mêmes que ceux qui les utilisent de façon efficace pour l'élévation. Jeppesen [108] a montré en synthèse binaurale, que l'introduction d'IS en conflit avec les indices interauraux, mène à une perception moins précise de l'azimut des sources virtuelles. Cependant, de nombreux résultats montrent que les IS jouent un rôle limité dans la localisation en azimut. L'emploi de stimuli de bande étroite [167, 182], ou des modifications sur les pavillons [92] affecte les IS tout en gardant les indices interauraux intacts : dans les deux cas, la perception de l'élévation est per-

turbée, tandis que celle de l'azimut reste précise. Middlebrooks [168] et MacPherson [144] avancent même que si les colorations spectrales montrent une évolution favorable à une localisation dans le plan horizontal, elles ont un impact mineur, voire inexistant, quand les indices interauraux sont disponibles. MacPherson interprète les performances des sourds monauraux, comme le résultat d'un apprentissage d'IS sur le long terme, palliant l'absence d'indices interauraux. Cette expérience ne reflète donc pas le fonctionnement normal du système auditif en écoute binaurale.

3.2.8 Traitement des IS : monaural ou binaural ?

Les IS sont généralement qualifiés de monauraux : cette dénomination reflète le fait que les colorations sont induites séparément pour chaque oreille, par opposition aux indices interauraux, qui n'existent que si les signaux aux deux oreilles sont disponibles. Cependant, la question suivante se pose : le traitement des indices monauraux par le système auditif est-t-il effectué indépendamment pour chaque oreille ou bien comme la différence des spectres gauche et droit ?

L'idée de l'existence d'un ILD spectral - appelons le ISD (*Interaural Spectral Difference*) - est séduisante. En effet, l'utilisation d'un tel indice aurait cet avantage par rapport à l'identification monaurale indépendante des IS, que le système auditif pourrait en extraire des caractéristiques directionnelles sans connaissance *a priori* sur le spectre source. Cependant, les différences perçues entre les oreilles sont quasiment nulles dans le plan médian, et pourtant les capacités de localisation en élévation y sont bonnes. Batteau [7] a donc proposé le principe de *binaural disparity* : la dissymétrie entre les deux pavillons pourrait créer des différences non négligeables entre spectres gauche et droit, et constituer un indice d'élévation dans le plan médian. Searle [229] a repris cette idée, et a proposé une étude démontrant l'existence, la perceptibilité, et l'utilisation effective de ces indices de dissimilitude. Nandy et Ben-Arie [188], Duda [62], et Janko *et al.* [107] ont développé des modèles informatiques de localisation permettant de tirer de l'ISD toutes les informations nécessaires à la localisation en azimut et en élévation. Les auteurs n'ont cependant pas apporté la preuve que le système auditif utilise réellement l'ISD pour localiser.

Les théories basées sur l'utilité de l'ISD ont très tôt été controversées. Hebrank [84] a relevé des failles importantes dans le protocole expérimental de Searle [229], remettant en cause ses conclusions. Middlebrooks [174], Musicant [187], et Carlile et Pralong [51] ont décrit les indices potentiels que porterait l'ISD. D'après eux, des différences interaurales existent dans le spectre, dans le plan médian, mais elles varient de façon non monotone, et leur niveau est faible. Jin *et al.* [111] ont évalué, sur la base de mesures perceptives, les informations directionnelles qu'apportaient d'une part les IS monauraux pris séparément, et d'autre part l'ISD. Les auteurs démontrent

que, mis à part une légère asymétrie avant/arrière, l'ISD véhicule peu d'informations sur l'élévation d'une source. Les indices monauraux, traités séparément, permettent eux de façon plus fiable de lever les ambiguïtés de position sur un cône de confusion. L'ISD ne permettrait donc pas de créer un lien systématique entre une caractéristique de l'ISD et une direction de l'espace, c'est-à-dire qu'il ne fournirait pas un *mapping* spatial [97] synonyme de qualité pour un indice de localisation. Une expérience menée par Jin [111] en synthèse binaurale, prouve que l'ISD seul se montre incapable de maintenir une perception correcte de l'élévation, face à des indices monauraux conflictuels. Tous ces résultats tendent à montrer que l'utilisation des IS est essentiellement monaurale, ou en tout cas que les différences spectrales interaurales ne suffisent pas à elles seules à porter une information spatiale fiable. Cependant, l'utilisation des deux oreilles apparaît tout de même nécessaire à une bonne localisation dans le plan vertical. Plusieurs études montrent que si l'oreille ipsilatérale est prédominante pour la détermination de l'élévation, l'oreille controlatérale y contribue aussi [92, 111, 141, 181]. Musicant [186], et Humanski et Butler [97] ont montré que lorsque les IS de l'oreille ipsilatérale sont inexistants, le système auditif pouvait tirer partie des indices de l'oreille controlatérale restés intacts. Il semble qu'à mesure que la source se décale latéralement par rapport au plan médian, la contribution de l'oreille ipsilatérale augmente, tandis que celle de l'oreille controlatérale diminue. Pour des azimuts supérieurs à environ 30° par rapport au plan médian (système de coordonnées polaire horizontal), l'oreille controlatérale ne contribue plus. Selon les résultats de MacPherson [141], les IS de chaque oreille sont pondérés au niveau central pour créer le percept de l'élévation, et les poids respectifs que prennent les indices gauche et droit sont fonctions de l'azimut perçu, donc de l'analyse des indices interauraux (ILD et ITD). Ce résultat va dans le sens des conclusions de Martin *et al.* [153], selon lesquelles des informations fiables sur l'azimut de la source sont nécessaires à une utilisation correcte des IS pour percevoir l'élévation.

3.2.9 Influence de connaissances *a priori* sur la source

On peut imaginer que le degré de familiarité de l'auditeur avec la source, et/ou les suppositions qu'il peut faire *a priori* sur son spectre, peuvent avoir un impact sur la localisation. C'est ce que MacPherson [140] a cherché à mettre en évidence dans le test décrit en 3.2.6, impliquant l'utilisation de stimuli à spectre irrégulier et de stimuli à spectre plus lisse. Deux conditions supplémentaires ont été introduites pour évaluer l'impact de la connaissance du stimulus à localiser. Celui-ci était précédé d'un autre stimulus : soit il s'agissait du même stimulus, soit il était différent. L'auteur montre que les résultats ne sont pas significativement différents selon chacune des conditions. Les études de Rakerd *et al.* [210] et de MacPherson et Middlebrooks

[145], utilisant différents types de stimuli, confirment que tant que le spectre source n'est pas trop accidenté, les sujets sont capables de localiser les sources même quand les stimuli sont variables d'un essai à l'autre.

3.3 Modèles d'utilisation des IS

L'analyse des résultats psychophysiques ainsi que l'observation des caractéristiques spectrales de HRTF ont donné naissance à plusieurs modèles reflètant l'utilisation des IS pour la localisation. Deux grandes familles se distinguent : les modèles basés sur une identification de caractéristiques locales du spectre, et ceux basés sur une reconnaissance de formes sur une large bande.

3.3.1 Modèles basés sur l'identification de caractéristiques locales du spectre

Creux spectraux

De nombreuses études se sont focalisées sur les creux spectraux étroits nommés C1 et C2 (cf. 3.1.3). Le creux C1 présente la particularité intéressante de se décaler fréquentiellement en fonction de l'élévation, et ce de manière monotone. Dans le plan médian, à l'avant, sa fréquence centrale passe de 6 kHz environ à -45 °, jusqu'à 10 kHz environ à 45 °. Au fur et à mesure que l'on se décale vers le plan vertical interaural, on observe un décalage de ces bornes d'environ 1 kHz à 2 kHz vers les hautes fréquences. Bien qu'il existe une variabilité dans ces valeurs, ces creux se retrouvent chez tous les individus, avec des profondeurs et des largeurs variables elles aussi. On représente figure 3.14 l'évolution de la fréquence centrale du creux C1 sur toute la sphère. Elle laisse entrevoir la possibilité d'une localisation précise dans l'hémisphère ipsilatéral, si le système auditif se base conjointement sur l'ITD pour estimer l'azimut et sur la fréquence de C1 pour l'élévation dans le plan sagittal. On note néanmoins une certaine ambiguïté du codage spatial de C1, liée à une relative symétrie avant/arrière (pas visible sur la figure).

Bloom [21] a étudié la localisation en champ libre d'un bruit large bande, filtré de manière à simuler le creux C1 seul, et émis par une source fixe dans le plan frontal. Il apparaît que les sujets perçoivent la source à une élévation sans rapport avec sa position réelle, et que l'élévation perçue est fonction de la fréquence centrale du creux. Ces phénomènes d'illusions sont nommés *hearing phantoms*. De plus, Bloom montre que les fréquences centrales des creux synthétiques créant une illusion d'élévation correspondent aux fréquences centrales des creux réellement observés dans le spectre des HRTF des sujets, pour les mêmes élévations. Des travaux similaires confirment ces conclusions [85, 264]. Plus récemment, Iida et Itoh [99] ont montré

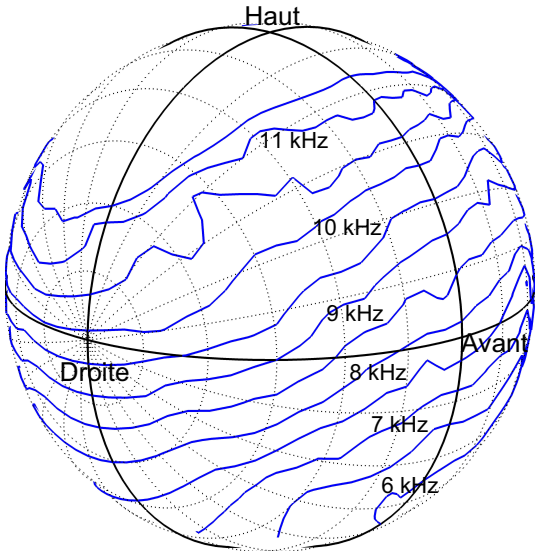


Figure 3.14 – Représentation sphérique de la fréquence centrale du creux C1 pour une oreille droite. Chaque ligne correspond à une fréquence constante. (Relevé effectué visuellement sur les HRTF du sujet n°5 de la base de HRTF privée d'Orange Labs, oreille droite.)

en synthèse binaurale, que l'utilisation d'une représentation synthétique des HRTF, ne conservant que les creux C1 et C2, préserve la perception de l'élévation dans le plan médian supérieur, avec cependant des dégradations pour certains individus. MacPherson [140] montre que pour obtenir des illusions d'élévation significatives, il faut utiliser des creux plus larges que les creux réellement observés dans les HRTF. De plus, ces illusions apparaissent de façon assez variable avec les individus. Des preuves neurophysiologiques soutiennent également l'idée que les creux spectraux sont importants pour la localisation en élévation. Poon et Brugge [204] ont étudié la sensibilité de fibres du nerf auditif d'un chat à la présence ou à l'absence d'un creux dans le spectre d'un bruit large bande. Leurs résultats indiquent que chacune des fibres nerveuses est capable, indépendamment, de signaler la présence d'un creux dans le spectre par leur taux de décharge. Moore *et al.* [180] ont étudié les capacités à détecter des creux à 8 kHz dans un spectre large, en fonction de la largeur de bande et de la profondeur du creux. Ils montrent que les creux observés dans les HRTF sont tout juste détectables. Par contre la détection de glissements dans la fréquence centrale de ces creux est tout à fait possible, suggérant qu'ils seraient des indices plus efficaces en localisation dynamique qu'en localisation statique.

***Covert Peaks Areas* (CPA)**

Des illusions similaires à celles décrites par Blauert (cf. 3.2.5) ont été relevées dans le plan horizontal pour des stimuli de bande étroite, mais en conditions monaurales,

c'est-à-dire avec une oreille bouchée. Butler et Flannery [42] reportent les résultats suivants pour une source fixe dans le plan horizontal (système de coordonnées polaire vertical) :

- Les stimuli autour de 4 kHz sont localisés à l'avant.
- On observe un décalage progressif de la source vers l'arrière pour une fréquence croissante.
- Les sons autour de 8 kHz sont localisés près de l'axe interaural, du côté de l'oreille intacte.
- On observe un retour discontinu vers l'avant à environ 9 kHz, et à nouveau décalage d'avant en arrière de 9 kHz à 12 kHz.

Pour analyser ces résultats, Butler [39] et Butler *et al.* [41] introduisent la notion de *covert peaks*, qui sont définis de manière rigoureusement identique aux "axes acoustiques" (cf. 3.1.3) introduits par Middlebrooks et Pettigrew [175] et Phillips *et al.* [201]. *Covert* signifie caché : contrairement aux creux décrits précédemment, ces caractéristiques des HRTF ne sont en effet détectables qu'en ayant connaissance de toutes les HRTF. Pour une fréquence donnée, le *Covert Peak Area*, ou CPA, est défini comme la portion de surface sur la sphère entourant le *covert peak*, et pour laquelle les HRTF présentent un module proche du module maximal (à 1 dB près). Butler *et al.* [41] étendent les résultats de Butler [43] et Butler et Flannery [42] en localisation monaurale à tout l'hémisphère ipsilatéral, et montrent que les reports de localisation de stimuli à bande étroite suivent fidèlement l'évolution spatiale des CPA, et que de plus, peu de reports sont effectués dans les zones de l'espace ne contenant pas de CPA. D'après Butler [39], Butler et Musicant [45], et Rogers et Butler [222], le système auditif détecterait et apprendrait l'évolution spatiale des CPA. Ayant intégré ainsi cette relation spatio-fréquentielle, le percept de l'élévation serait formé d'après l'identification des zones de fréquence les plus énergétiques dans le stimulus perçu. D'après Middlebrooks [167], il est cependant délicat de rapporter directement ces résultats, obtenus en bande étroite, au cas de la localisation des sons naturels, généralement de bande large, et pour lesquels le système auditif semble être le plus performant. Jin montre néanmoins que le modèle CPA prédit correctement la localisation de bruits blancs filtrés par des filtres passe-bande de diverses largeurs [110].

Combinaison de caractéristiques locales

On peut imaginer que c'est plutôt la combinaison de différentes caractéristiques locales du spectre - pics, creux, pentes - qui est associée à une direction particulière de l'espace, et que la tâche du système auditif consiste à identifier ces caractéristiques dans le spectre perçu. C'est le modèle qu'a évalué MacPherson [140]. Il s'avère assez

satisfaisant sur des stimuli à spectre lisse, mais très incorrect face à des accidents prononcés du spectre source. Les performances de localisation qui subsistent face à des spectres accidentés ne sont donc pas expliquées par ce modèle simple.

3.3.2 Modèles d'analyse large bande

Van Opstal et Van Esch [252] ont cherché à mettre en évidence les IS significatifs, par une approche bayésienne. Les auteurs collectent chez leurs sujets les élévations illusoires perçues pour de nombreux stimuli large bande, dont le spectre présente des oscillations aléatoires de période plus ou moins fine. Par un calcul de probabilités conditionnelles, on peut ainsi relier la perception de l'élévation dans le plan médian aux caractéristiques du spectre source, l'hypothèse étant que ce sont les oscillations du spectre source qui gouvernent la localisation. Les résultats montrent que ce sont les HRTF sur la totalité du spectre qui jouent un rôle dans le jugement de la localisation. Les auteurs concluent donc que les modèles basés uniquement sur des caractéristiques locales du spectre sont incorrects. Par ailleurs au niveau central, la dynamique des creux et des pics spectraux est largement réduite par le filtrage cochléaire, et un filtrage plus sévère encore des HRTF permet de conserver de bonnes performances de localisation en élévation en synthèse binaurale (cf. 2.5). C'est pourquoi il est plus probable que le système auditif analyse de profil spectral sur une partie plus large du spectre, comme le proposent les modèles suivants.

L'idée commune aux modèles basés sur une analyse large bande est que la formation du percept de localisation sur la base des IS serait le fruit d'une comparaison sur toute la bande entre le spectre perçu et des gabarits, ou *templates*, stockés "en mémoire", et associés à une direction de l'espace. Selon les modèles de localisation, ces *templates* peuvent être :

- les HRTF elles-mêmes, ou les DTF (*Directional Transfer Function*, c'est-à-dire les HRTF égalisées en champ diffus) traitées par un banc de filtres simulant le comportement de la cochlée,
- ou bien un ensemble de caractéristiques issues du spectre des HRTF.

Le système auditif opèrerait un processus de reconnaissance de formes (*template matching*), et les modèles suivants se distinguent par la manière dont est défini le calcul de similarité sous-jacent à ce processus.

Intercorrélation

Différents modèles utilisent comme mesure de similarité une opération d'intercorrélation. Middlebrooks [167], proposent comme *templates* les DTF complètes. L'élévation perçue serait celle de la DTF qui maximise la corrélation avec le spectre perçu. MacPherson [140] propose en plus d'intégrer l'ITD pour la perception de l'azimut,

et de considérer la somme pondérée des indices provenant de chaque oreille. Hofman et Van Opstal [90] reprennent la même idée en considérant les DTF avec une échelle logarithmique sur l'axe fréquentiel. Zakarauskas et Cynader [285] suggèrent que ce sont les positions fréquentielles des pentes du spectre qui sont importantes, et proposent donc d'utiliser comme *templates* les premières dérivées des DTF, à comparer avec le gradient du spectre perçu. MacPherson [140] a implémenté et évalué toutes ces modèles. Ils se montrent satisfaisants pour expliquer la localisation de stimuli à spectre large et plat, et reflètent également les illusions observées face à des stimuli de bande étroite. Cependant, face à des stimuli à spectre aléatoire et accidenté, leurs résultats sont mauvais, et ne permettent pas d'expliquer les capacités de localisation observées dans ces conditions.

Ecart-type

Langendijk et Bronkhorst [131] proposent une solution qui s'affranchit de différences globales de niveau afin de se focaliser sur les variations du spectre, porteuses de l'information spatiale. Selon leur modèle, la distance calculée entre le spectre perçu et les DTF traitées par un banc de filtres auditifs, choisies comme *templates*, est l'écart-type de leur différence, fréquence par fréquence. Ces *templates* correspondent aux informations spectrales disponibles au niveau central pour un stimulus de spectre large et plat : ce modèle, comme ceux présentés plus haut, traduit donc de façon implicite l'idée selon laquelle le système auditif ferait cette hypothèse *a priori* sur le spectre source. Ng [189] étend le modèle au cas extrême où le système auditif connaîtrait parfaitement le spectre de la source, et pourrait en tirer partie pour localiser, bien que les résultats psychophysiques disponibles contredisent cette hypothèse. Les *templates* sont alors les profils spectraux résultant du filtrage du spectre source par les différentes DTF. De plus, Ng inclut dans son modèle une étape probabiliste traduisant le processus de décision que constitue le report de localisation réalisé par un sujet. Pour cela, la distribution des mesures de similarités est normalisée, puis elle est utilisée selon la méthode de rejet pour générer des reports de localisation. Ng montre pour une série de stimuli de profils spectraux différents, que les performances de localisation des sujets sont en général meilleures que ce que prédit le modèle dans la condition "hypothèse d'un spectre source plat", mais moindres que ce qu'il prédit dans la condition "hypothèse d'un spectre source connu". Les deux hypothèses extrêmes considérées permettent donc d'encadrer les résultats psychophysiques, mais aucun des modèles n'est tout à fait satisfaisant. En particulier, ils ne traduisent pas bien les résultats psychophysiques observés pour des stimuli à spectre aléatoire et accidenté.

3.3.3 Extension du modèle CPA

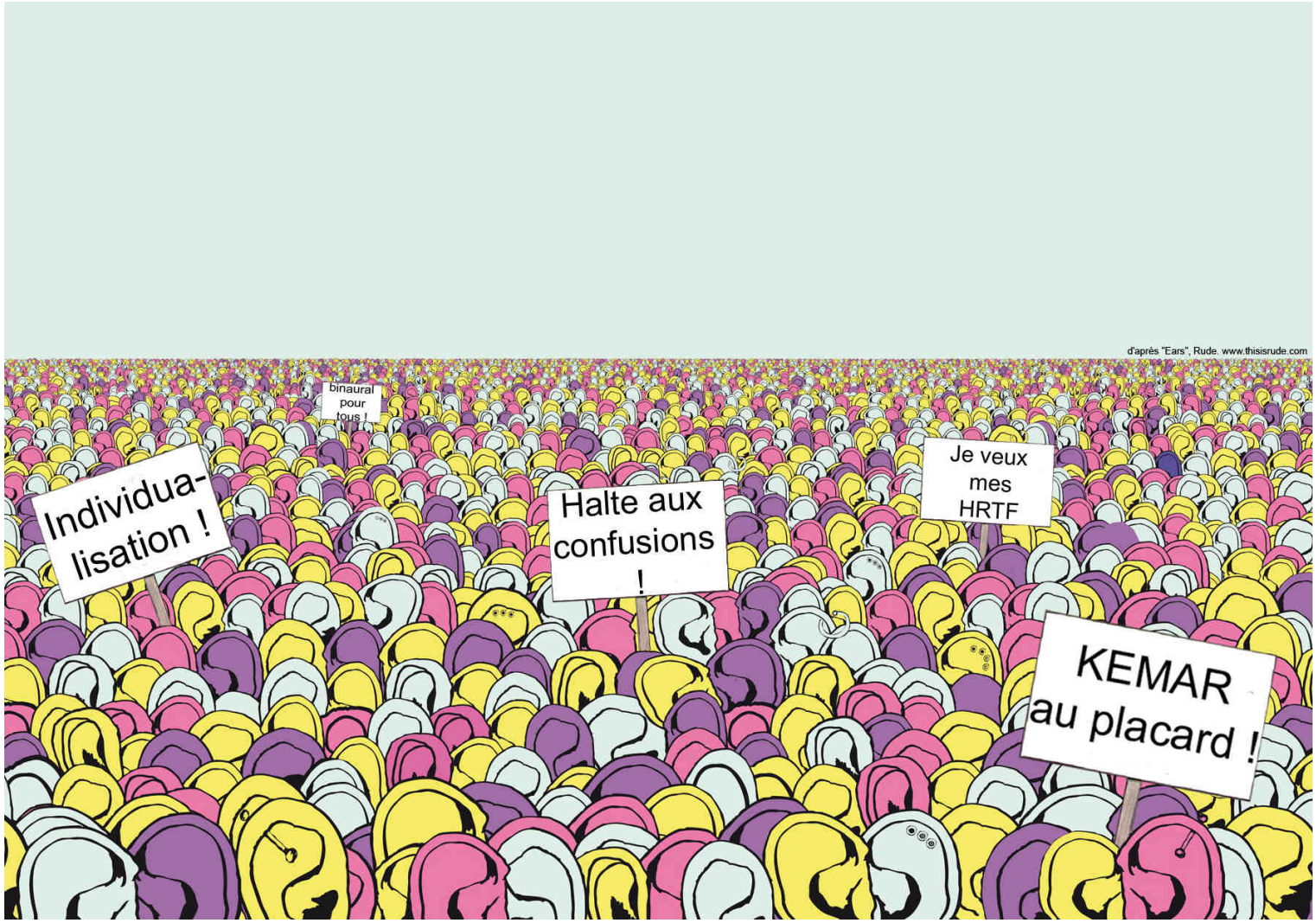
Jin [110] propose d'étendre à des stimuli de spectre quelconque le modèle basé sur l'identification des CPA par le système auditif. Tel quel, le modèle dit "CPA" ne permet pas de prédire les performances de localisation de stimuli à spectre large. En effet, on comprend en observant les trajectoires des axes acoustiques (cf. Fig. 3.12) que les CPA associés à de tels stimuli couvrent tout l'hémisphère ipsilatéral, et constituent donc un indice de localisation très ambigu. A l'origine, le modèle CPA est basé sur l'existence d'un apprentissage de la relation entre une bande de fréquence et une région de l'espace, ce qui suppose que le système auditif est capable de comparer l'énergie contenue dans une bande de fréquence avec celle du reste du spectre. Jin établit son modèle sur la capacité du système auditif à analyser le spectre perçu selon de nombreuses comparaisons entre les niveaux d'énergie contenus dans de multiples bandes de fréquence. Le modèle est donc appelé *Spectral Contrast Area*. Les bandes de fréquence considérées peuvent être de largeur diverses, et éventuellement se chevaucher. Jin définit une quantité appelée énergie normalisée, issue de la comparaison de l'énergie dans chaque bande avec l'énergie contenue dans ses deux plus proches voisines. Le modèle de Jin est basé sur la comparaison entre le jeu complet des énergies normalisées correspondant au spectre perçu et ceux correspondant aux DTF, stockés en mémoire, constituant les *templates*. Les résultats correspondant à chaque oreille sont combinés pour prédire la direction perçue. Les expériences menées par Jin montrent que cette analyse spectrale multi-échelles permet de bien prédire les résultats psychophysiques observés pour tous les types de stimuli, y compris ceux présentant un spectre très accidenté, ce qui faisait défaut aux autres modélisations.

Le fondement physiologique de ces traitements pourrait, selon Jin, se trouver dans la ceinture d'aires auditives entourant le cortex auditif primaire. D'après Rauschecker *et al.* [213], cette zone est composée, chez les primates, de neurones répondant préférentiellement à des sons d'une largeur de bande particulière. Les neurones sont organisés sur cette ceinture selon deux axes orthogonaux : l'un correspond à une largeur de bande préférée croissante, tandis que l'autre est associée à une fréquence centrale préférée croissante. Cette organisation serait favorable à l'analyse multi-échelles sur laquelle se fonde le modèle de Jin.

3.4 IS et stabilité perceptive d'un objet auditif

Carlile propose une réflexion intéressante sur l'interprétation des IS par le système auditif [47]. Quand un bruit blanc est filtré par des HRTF correspondant à diverses directions, et le résultat diffusé sur un haut-parleur pour une écoute en champ libre, on perçoit clairement le passage d'une HRTF à une autre, et on l'interprète comme

une différence de timbre entre les signaux. Par contre, lorsque le même type de filtrage est effectué en synthèse binaurale, ces mêmes différences objectives ne sont pas attribuées à un changement de timbre, mais interprétées comme résultant d'un changement de la position d'une même source sonore. Selon Carlile, ce serait le fruit de la capacité du système auditif à assurer une stabilité dans la perception d'un objet auditif. De façon analogue pour le système visuel, la couleur perçue d'un objet est liée à la lumière qu'il réfléchit, ce qui dépend largement de la composition spectrale de la source lumineuse. Malgré tout, une feuille d'arbre est toujours perçue de couleur verte, que ce soit à la lumière du jour au lever, ou au coucher du soleil, à midi, ou bien sous une lampe fluorescente. Le système visuel assure cette stabilité en analysant le ratio de l'énergie réfléchie à différentes longueurs d'onde. On ne sait pas encore s'il existe un mécanisme équivalent pour la localisation auditive, mais le modèle qui se rapproche le plus de ce principe est le *Spectral Contrast Area* proposé par Jin, qui apparaît comme le plus abouti parmi les modèles disponibles. Cette vision est néanmoins en partie réfutée par quelques expériences qui démontrent le système auditif est incapable à la fois de localiser et d'identifier des sources de spectres inconnus [139, 210]


d'après "Ears", Rude. www.thisisrude.com
binaural pour tous !
Individua-lisation !
Halte aux confusions !
Je veux mes HRTF
KEMAR au placard !

Chapitre 4

Individualisation des HRTF : nécessité et état de l'art

Le chapitre précédent a permis d'illustrer le rôle majeur des IS dans le processus de localisation auditive, et de comprendre l'impact de la morphologie d'un auditeur sur la création de ces colorations. Le processus de décodage des IS appris par le système auditif reste en partie obscur, mais on sait qu'il repose sur l'identification des colorations subies par le signal sonore entre la source et les tympans. Puisque les morphologies varient énormément d'un individu à l'autre, les IS connaissent des variations de même ampleur [160], et on s'attend donc à ce que l'utilisation de HRTF non-individuelles en synthèse binaurale ne conduise pas à la création d'un espace sonore virtuel suffisamment convaincant. Nous montrons dans ce chapitre qu'en effet, la synthèse binaurale non-individuelle est imparfaite et que les défauts de spatialisation observés ne sont pas inhérents à la technique elle-même, mais bien à l'utilisation de filtres binauraux inadaptés aux auditeurs. Cela implique donc que la génération, via la synthèse binaurale, d'espaces sonores virtuels de qualité nécessite l'adaptation des filtres binauraux à chaque auditeur : c'est le problème de l'individualisation des HRTF, et nous nous intéressons plus particulièrement à l'individualisation des IS. La mesure acoustique des HRTF étant ardue, différentes techniques alternatives d'individualisation ont été proposées. Nous les passons en revue en mettant en évidence leurs intérêts et leurs faiblesses.

4.1 Imperfections de la synthèse binaurale non individuelle

4.1.1 Méthodes et critères d'évaluation des VAS

L'évaluation de la synthèse binaurale est classiquement réalisée au moyen de tests psychoacoustiques. Adoptons donc le formalisme et la terminologie proposés par Blauert [19] pour décrire les différents éléments d'une expérience psychoacoustique. On distingue tout d'abord les évènements acoustiques (*sound events*, ou *acoustical events*) des évènements auditifs (*auditory events*). Les évènements acoustiques désignent les phénomènes physiques, attachés aux sources acoustiques dans l'espace physique, tandis que les évènements auditifs désignent leurs pendants perceptifs, dans l'espace auditif du sujet. Bien sûr, l'expérimentateur a accès uniquement à l'espace physique, s'il n'est pas lui-même sujet de l'expérience. Le sujet est soumis à un évènement acoustique, et doit fournir à l'expérimentateur une description des attributs perceptifs de l'évènement auditif généré. Le sujet peut donc être schématiquement représenté par un bloc de perception, dont découle l'évènement auditif, que seul le sujet connaît, et un bloc de description : le sujet traduit quantitativement les attributs perceptifs selon des règles préétablies. Dans notre problème, il s'agit de simuler, en synthèse binaurale sur casque, une source sonore réelle placée dans un

environnement anéchoïque. On vise donc à faire en sorte que les attributs perceptifs de l'évènement auditif soient identiques lors de l'écoute en champ libre d'une source réelle, et lors de l'écoute, en synthèse binaurale, de la source virtuelle correspondante, bien que les deux évènements acoustiques soient très différents. On représente figure 4.1 les expériences nécessaires à l'évaluation des VAS. Il faudrait idéalement comparer les attributs perceptifs des évènements auditifs h_{CL}, h_{Indiv} et $h_{NonIndiv}$, mais l'expérimentateur ne peut que comparer leurs descriptions respectives h'_{CL} , h'_{Indiv} et $h'_{NonIndiv}$, fournies par le sujet lui-même. Pour une source réelle à simuler, on doit évaluer et comparer les attributs perceptifs dans trois conditions :

- en écoute champ libre, la source étant réelle,
- en synthèse binaurale individuelle,
- en synthèse binaurale non-individuelle.

Les attributs qui nous intéressent sont :

- l'externalisation de la source, et/ou la distance égocentrique : la source est-elle perçue à l'intérieur de la tête, très proche de la tête, ou alors bien externalisée ?
- la direction perçue de la source.

Les autres attributs perceptifs liés à la spatialisation, tels que la présence, ou l'enveloppement, sont le plus souvent délaissés dans les expériences, car ils sont fortement liés à l'existence de réverbération dans la scène sonore, alors que l'évaluation est en général réalisée sur des VAS en condition anéchoïque. Le sujet doit traduire l'externalisation et la direction perçues, dans le cadre d'un test de localisation, dans un repère proposé par l'expérimentateur dans l'espace physique. Classiquement, la précision de la localisation s'exprime quantitativement via l'"erreur" angulaire commise entre la direction reportée et la mesure physique de l'angle. Une analyse plus fine des erreurs permet d'extraire des catégories pertinentes, notamment les confusions avant/arrière, et haut/bas [273]. Pour l'externalisation, le sujet peut verbaliser pour indiquer si la source est perçue en dehors de la tête, proche de la tête, ou bien nettement à l'extérieur de la tête. Il peut aussi, comme dans l'étude de Kim et Choi [119], reporter la distance perçue sur une échelle continue. Le test de localisation est largement utilisé pour l'évaluation des VAS, mais c'est un pis-aller, car il comporte des limites clairement identifiées [149]. En effet, en écoute binaurale, le sujet ne peut pas mettre au point de façon fiable une stratégie de report de localisation, car il n'a, dans son espace auditif, aucun référentiel calibré par rapport au référentiel physique. L'idéal serait que dans le même test le sujet puisse comparer la position perçue d'une source réelle en champ libre, et celle de la source virtuelle correspondante en synthèse binaurale. Malheureusement, l'utilisation d'un casque limite la possibilité d'une écoute simultanée en champ libre. Zahorik [**?**] a choisi d'utiliser un casque supra-aural qui, selon l'auteur, perturbe peu l'onde provenant des sources en champ libre, et plus récemment Romigh et Brungart [224] ont développé un dispositif

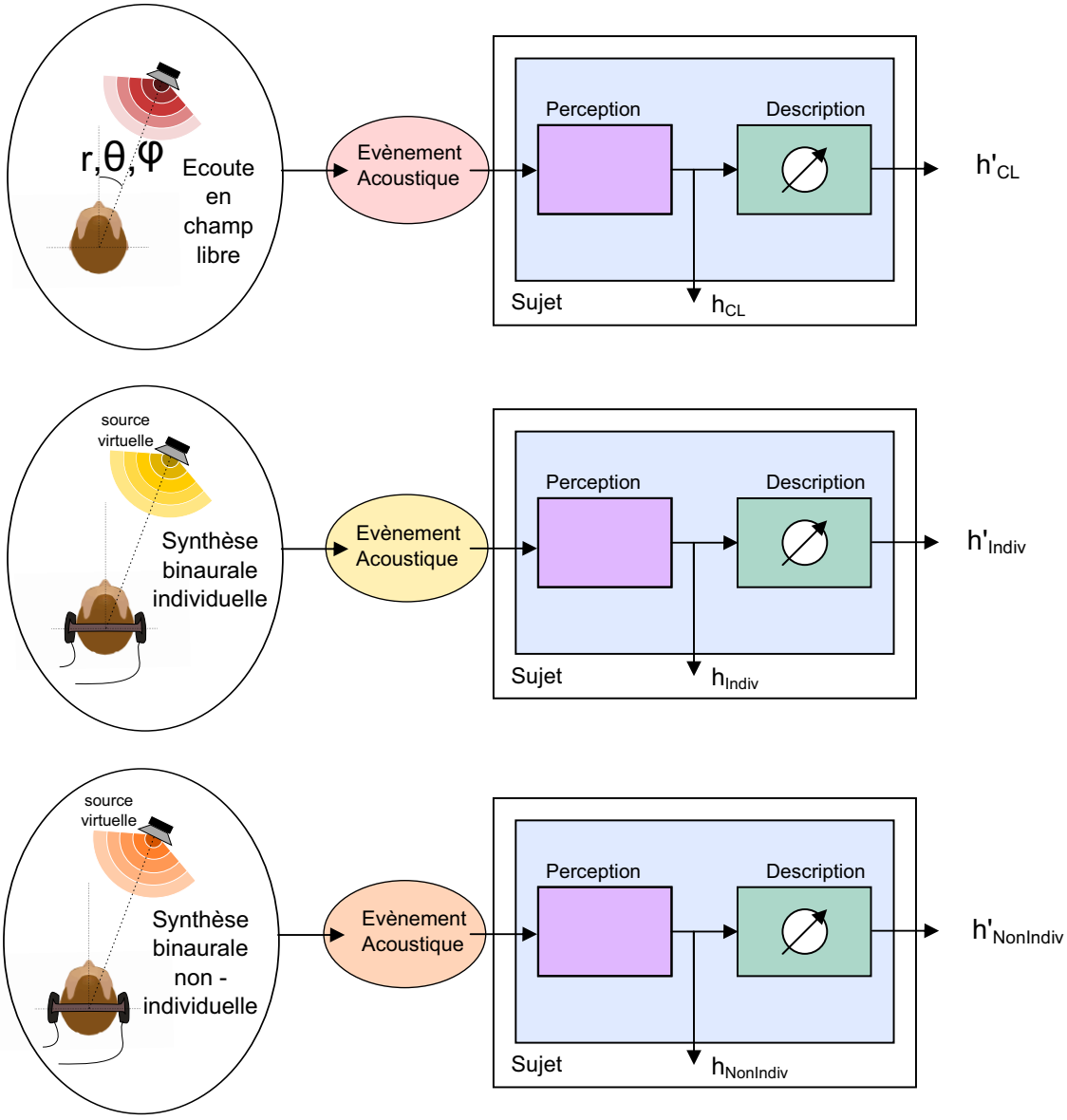


Figure 4.1 – Description des expériences psychoacoustiques selon le formalisme proposé par Blauert [19]

spécifique basé sur l'utilisation d'un casque ouvert. On peut aussi citer les travaux de Kulkarni et Colburn [127] : ils ont utilisé des écouteurs de type *tubephones*, qui laissent ouverts les conduits auditifs, affectant donc très peu l'écoute en champ libre. Ces transducteurs permettent donc une écoute simultanée d'environnements réels et virtuels, mais leur réponse médiocre aux basses fréquences reste un problème.

4.1.2 Evaluation en condition individuelle

Plusieurs études ont démontré la faisabilité pratique d'un VAS haute-fidélité en synthèse binaurale individuelle statique [50, 152, 272, 273]. Ces VAS ont été réalisés avec des HRTF mesurées sur chacun des auditeurs, et une calibration individuelle du casque. Dans ces conditions, les sources virtuelles sont perçues comme étant bien externalisées [81, 119, 272, 273]. En termes de localisation, les performances des sujets sont en général très légèrement dégradées par rapport à celles en champ libre. L'erreur moyenne de localisation est de 1 à 2 degrés supérieure, et c'est la localisation en élévation qui est en général la plus affectée [50, 273]. Les confusions avant-arrière, qui sont en moyenne d'environ 5% en champ libre, sont doublées pour les sources virtuelles [29, 50, 273]. Martin et McAnally [152] montrent cependant que l'illusion peut être parfaite : les performances de leurs sujets sont excellentes, et non significativement différentes des performances de localisation en champ libre. Romigh et Brungart ont aussi obtenu le même niveau de réalisme en synthèse binaurale individuelle dynamique [224].

Ces résultats psychophysiques sont corroborés par ceux d'études neurophysiologiques sur les collicules supérieurs (SC) et inférieurs (IC). Ces éléments du mésencéphale sont constitués de neurones se trouvant à la croisée des chemins neuronaux issus des olives supérieures médiane (MSO), et latérale (LSO), et des noyaux cochléaires dorsaux (DCN), connus pour être impliqués respectivement dans le traitement de l'ITD, l'ILD et des IS [254] (cf. Fig. 4.2). On observe que la dépendance spatiale des réponses de ces neurones (*Spatial Receptive Field*, ou SRF) n'est pas discriminable si l'on s'expose à une source en champ libre ou une source virtuelle en VAS individuel haute-fidélité (tests menés sur les IC du cobaye [11], et sur les SC [46] du furet). Les moyens du laboratoire permettent donc de créer des espaces sonores virtuels fidèles en synthèse binaurale individuelle : ils constituent le point de référence de cette technique de spatialisation.

4.1.3 Evaluation en condition non-individuelle

L'impact de l'utilisation de HRTF non-individuelles a été l'objet de plusieurs études [265, 266, 275]. On entend ici par synthèse binaurale en condition non-individuelle l'utilisation de HRTF non-individuelles, conjointement à une calibration

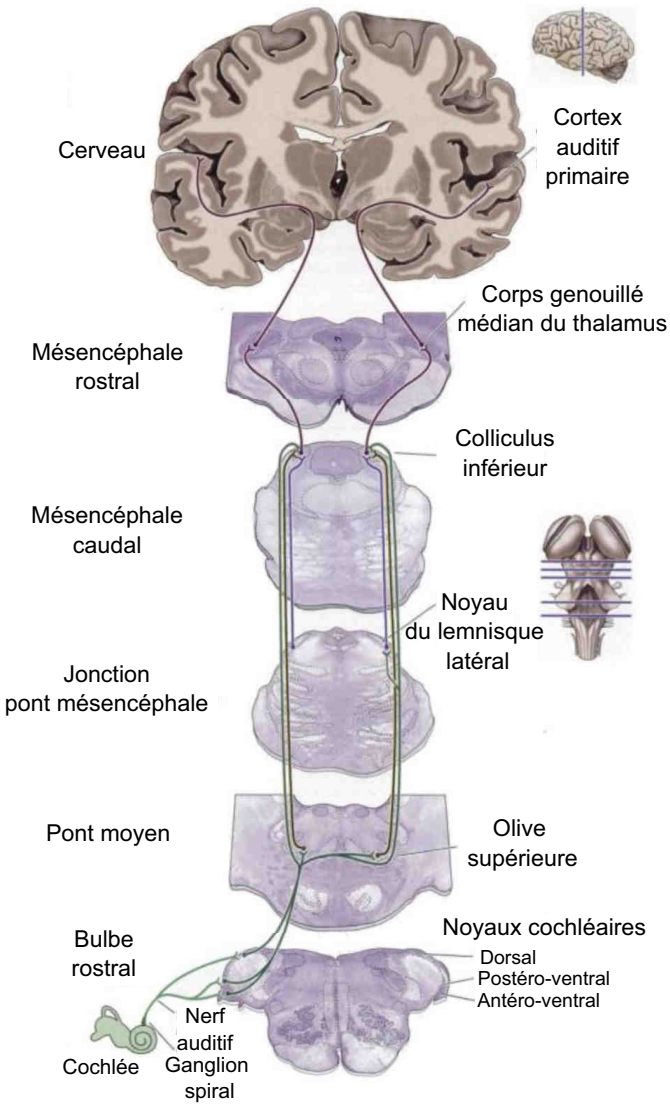


Figure 4.2 – Schéma des principales voies auditives (d'après [207])

individuelle du casque. Les éventuels problèmes liés à une calibration non idéale du casque ne sont pas inclus dans les résultats suivants.

En termes d'externalisation, Kim et Choi [119], et Völk *et al.* [258] ont montré l'apparition de dégradations liées à la non-individualisation. Les effets sont moindres pour les positions très latéralisées, la localisation intracrânienne apparaissant majoritairement pour des sources proches du plan médian. Les résultats de Völk *et al.* [258], obtenus pour des enregistrements binauraux, suggèrent que l'amélioration apportée par l'ajout de réverbération dans la scène sonore est bien plus significative que celle apportée par à l'utilisation de HRTF individuelles. Malgré tout, en présence de réverbération, l'apport des IS individuels semble bénéfique pour l'externalisation.

En termes de localisation, on observe une grande variabilité des résultats selon les sujets. Pour une minorité, il n'y a pas de dégradation des performances, mais pour la majorité des sujets, elle est effective. Dans ces cas, l'azimut des sources virtuelles reste généralement très bien estimé, alors que l'élévation est nettement moins bien perçue. Les confusions haut/bas et avant/arrière augmentent sévèrement, avec en général, des sources à l'avant perçues à l'arrière [266]. Les sujets rapportent également une perception très diffuse des sources, pourtant synthétisées comme ponctuelles [266]. Des études similaires ont été menées, non pas en synthèse binaurale, mais en utilisant des enregistrements binauraux effectués sur sujets humains ou sur têtes artificielles [155, 159, 178]. Elles révèlent que l'écoute d'enregistrements non individuels mène aussi à des artefacts de localisation : mauvaise perception de l'élévation et augmentation des confusions avant/arrière. De plus, il apparaît que la plupart des têtes artificielles, qui sont censées représenter des morphologies moyennes, donnent des performances médiocres : plus de 60% des enregistrements sur des sujets humains donnent de meilleurs résultats que tous les enregistrements sur têtes artificielles [178]. Bien que réalisées à partir d'enregistrements, et non de sons de synthèse, ces études viennent confirmer la pauvreté des indices de localisation non-individuels.

Ces imprécisions de la perception en synthèse binaurale non-individuelle sont probablement le fruit des changements de comportement observés au niveau neuronal entre ces deux conditions d'écoute. Au sein des IC, les SRF traduisent en condition individuelle une sensibilité marquée des neurones à des sources sonores situées dans une zone de l'espace réduite autour d'une "direction préférée". En condition non-individuelle, ces SRF se trouvent fondamentalement perturbés (résultats observés sur le cobaye [245]) : les SRF se décalent dans l'espace (en moyenne de 30°, et principalement d'avant en arrière), deviennent omnidirectionnels, ou encore se scindent autour de plusieurs "directions préférées" . Ces effets ont aussi été observés sur le furet, sur des neurones du cortex auditif primaire (A1), zone également impliquée dans la localisation auditive [184, 185]. On représente figure 4.3 les résultats de cette étude. Ces expériences illustrent bien le fait que les mécanismes neuronaux de la localisa-

tion sont calibrés par l'expérience aux indices de localisation individuels fournis par l'oreille externe [185]. Les artefacts de perception qui apparaissent en synthèse binaurale non-individuelle se situent précisément là où les IS interviennent : discrimination de l'avant et de l'arrière, et perception de l'élévation. Comme de nombreux travaux l'ont montré [160, 164, 174, 231, 272], il existe une grande variabilité des IS d'un individu à l'autre : elle touche la position fréquentielle, la largeur et l'amplitude des creux et des pics spectraux. Il semble ainsi naturel qu'en condition non-individuelle, le système auditif peine à extraire l'information spatiale d'IS qu'il n'a pas appris, et qui lui sont donc étrangers. Les artefacts que l'on observe seraient donc le fruit d'un décodage spatial erroné de la part du système auditif. L'utilisation de HRTF adaptées à chaque individu se révèle donc nécessaire pour générer un VAS fidèle en synthèse binaurale. Si l'on sépare le problème, avec d'un côté les indices temporels (ITD) et de l'autre, les IS, on comprend que c'est l'individualisation des IS qui est la plus critique, et c'est cet aspect qui est développé dans la revue de l'état de l'art suivante.

4.2 Individualisation des HRTF : Etat de l'art

Nous présentons ici, par familles, les différentes techniques d'individualisation proposées dans la littérature, en mettant en avant leurs atouts et leurs faiblesses.

4.2.1 Acquisition de HRTF par modélisation numérique

L'obtention de HRTF par modélisation numérique consiste à résoudre numériquement le problème acoustique de la propagation entre une source et des microphones placés à l'entrée des conduits auditifs. Cela nécessite l'acquisition de la morphologie de l'auditeur (buste et tête) en 3D par un scan laser, ou une image obtenue par IRM. Tous les paramètres du modèle - morphologie, positionnement des micros, impédances de surface - peuvent alors être contrôlés finement, et les effets de leurs variations directement évalués. On peut trouver une revue de ces techniques dans la thèse de Busson [38]. On distingue les méthodes suivantes :

- BEM et IBEM (*Boundary Element Method* et *Inverse Boundary Element Method*) Le principe de la BEM est de remplacer un problème différentiel aux limites du domaine d'étude par un problème intégral sur les bords du domaine. Le choix des impédances de surface traduira la prise en compte ou non des effets acoustiques des vêtements, de la peau et des cheveux. La BEM ne peut être utilisée que pour des surfaces closes, tandis que l'IBEM permet de travailler avec des surfaces ouvertes ou fermées, grâce à l'introduction de potentiels de couche.

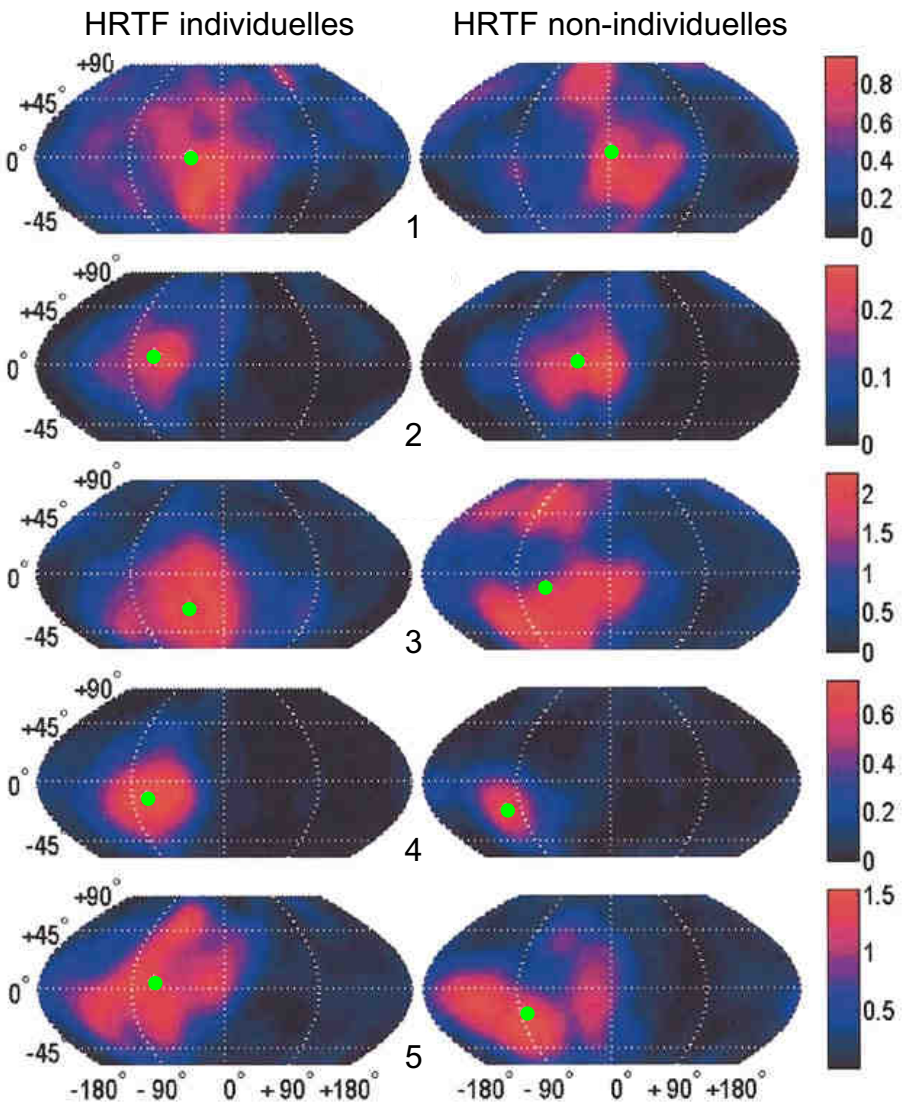


Figure 4.3 – Impact de l'utilisation de HRTF non-individuelles sur les SRF du cortex auditif primaire (A1) chez le furet (d'après [184]). On représente en couleurs la fréquence de décharge, évaluée sur 5 neurones en présentant des sources virtuelles dans 224 directions différentes (système de coordonnées polaire-vertical). Chaque ligne correspond à un neurone différent, chaque colonne à une condition : HRTF individuelles à gauche, non-individuelles à droite. Les points verts représentent la position des centres de gravité des SRF. La scission des zones spatiales de sensibilité maximale est manifeste en condition non-individuelle pour les neurones n°1, 3 et 5.

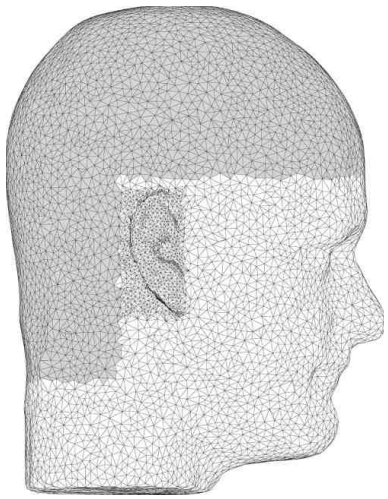


Figure 4.4 – Maillage de la surface d'une tête pour le calcul numérique des HRTF (d'après [117]). Pour un meilleur résultat en BEM, on utilise un maillage plus fin au niveau des pavillons, et on prend en compte l'impédance des cheveux

– FEM (*Finite Element Method*) Le volume d'étude est discrétisé en un grand nombre de petits éléments de volume. La pression acoustique est discrétisée sur tout le volume, et non pas seulement sur la surface. Cette technique permet de définir un couplage entre le fluide entourant la surface d'étude et l'objet immergé, c'est-à-dire entre l'air et le corps de l'auditeur pour notre application. La FEM est cependant difficilement applicable au calcul des HRTF, car on doit considérer un volume sphérique de 1 m au moins de rayon entourant le sujet, ce qui correspond à un maillage dont le coût de calcul est prohibitif.

La faisabilité du calcul précis de HRTF par ces méthodes numériques a déjà été largement démontrée, notamment par les travaux de Katz [116], et ceux de Kahana [114]. La limite fréquentielle de validité du modèle étant directement liée à la finesse du maillage, le coût de calcul peut devenir rapidement excessif. Pour limiter ce coût, on peut raffiner localement le maillage là où c'est nécessaire, au niveau des pavillons par exemple, et utiliser un maillage plus grossier ailleurs (cf. Fig. 4.4). On peut également envisager de ne pas effectuer le calcul entier pour chaque nouvel individu, mais de considérer les morphologies individuelles comme des petites variations d'un modèle pré-calculé [247]. Dans cette optique, la représentation paramétrique compacte de la morphologie des pavillons a motivé plusieurs études récentes [86, 87]. L'acquisition de scans 3D peut également être un frein à l'utilisation des méthodes numériques. C'est pourquoi une étude récente propose de générer un maillage individuel en réalisant des opérations de reconnaissance de formes et de *morphing* sur des maillages de têtes et d'oreilles issus d'une base de données [61]. Ainsi, seuls quelques

clichés photographiques de la morphologie du nouvel auditeur sont nécessaires pour obtenir un maillage 3D complet. Actuellement la modélisation numérique se révèle insatisfaisante pour l'acquisition individuelle de HRTF sur une large bande de fréquences. En effet, Kahana [114] a montré que le maillage des pavillons devait être très fin pour obtenir des résultats satisfaisants, ce qui entraîne des coûts de calcul très importants. La BEM se révèle en revanche intéressante aux basses fréquences, où cette méthode peut permettre de pallier les imperfections des systèmes de mesure, ou bien être utilisée en complément d'une autre méthode de modélisation pour le haut du spectre.

4.2.2 Reconstruction bayésienne de HRTF à partir de *hearing phantoms*

Hofman et Van Opstal [91] ont proposé une approche originale pour acquérir des HRTF individuelles. Il s'agit de faire écouter à un sujet des stimuli dont le spectre large bande présente de fortes variations, du même ordre que celles des HRTF. Ces stimuli sont délivrés par un seul et même haut-parleur positionné en face du sujet, mais ils engendrent des illusions de localisation, ou *hearing phantoms*, comme on a pu le montrer dans le chapitre précédent. La mesure consiste d'abord à acquérir les directions perçues de ces illusions (élévation dans le plan médian) via l'enregistrement des mouvements oculaires du sujet, qui a pour consigne de regarder dans la direction où il perçoit la source. L'opération est répétée de nombreuses fois, pour divers stimuli, de manière à pouvoir obtenir des résultats en termes de probabilités. La reconstruction est basée sur une approche bayésienne, c'est-à-dire via le calcul des probabilités conditionnelles liant les caractéristiques spectrales des stimuli et les directions des illusions correspondantes. Les auteurs montrent une bonne concordance entre les HRTF ainsi obtenues et les HRTF mesurées classiquement. De nombreux problèmes subsistent cependant dans cette méthode d'estimation des HRTF. D'abord le report de localisation n'est possible que dans la portion d'espace correspondant au champ visuel. De plus, elle nécessite de nombreuses répétitions, pour de nombreux stimuli, afin d'estimer au mieux les distributions de probabilité. On ne peut donc pas de considérer cette méthode comme une réelle technique d'individualisation. Néanmoins, ces expériences ouvrent peut-être la voie vers une nouvelle manière d'envisager l'individualisation : estimer d'après les illusions perceptives du sujet, les IS pertinents contenus dans ses HRTF.

4.2.3 HRTF non-individuelles issues d'une base de données

Différents tests de localisation menés dans les études sur la synthèse binaurale ont mis en évidence une grande disparité dans les performances de localisation, d'un

sujet à l'autre. Il ressort de ces travaux l'existence de "bons localisateurs", et une analyse fine de leurs HRTF révèle que leurs IS présentent une dépendance spatiale particulièrement prononcée [269]. Ce serait à l'opposé le manque d'IS spatialement discriminants qui expliquerait les performances médiocres des "mauvais localisateurs" [269, 271, 273]. Butler et Belendiuk [40] ont même avancé qu'il existe des sujets dont les HRTF présentent des indices tels que leur utilisation en synthèse binaurale non-individuelle procurerait une spatialisation meilleure que des HRTF individuelles. Wenzel *et al.* [265] réfutent cette idée : ils montrent que les mauvais localisateurs n'améliorent pas leurs performances en utilisant des HRTF de bons localisateurs. En revanche, Wenzel *et al.* affirment que les performances en synthèse binaurale non-individuelle restent peu dégradées par rapport à la synthèse binaurale individuelle, tant que les HRTF utilisées sont celles d'un bon localisateur. Bien que ces conclusions se basent au départ sur les résultats d'une seule expérience, non publiée, il est maintenant communément admis que l'utilisation des HRTF d'un bon localisateur est toujours un bon choix pour réduire les artefacts de la synthèse binaurale non-individuelle. Il nous semble cependant que cette tradition est fondée sur trop peu de résultats, et que l'argumentaire qui l'accompagne comporte des failles. En effet, si les "bons localisateurs" montrent de bonnes performances, c'est à la fois parce que leurs HRTF comportent des IS saillants, et spatialement discriminants, mais aussi et surtout parce que ces individus ont appris à bien les décoder, et à en extraire un jugement fiable de localisation. Cependant ces IS ne sont pas universels : ils peuvent être totalement différents de ceux d'un autre individu, qui, n'ayant pas appris à les décoder, ne pourra pas en extraire l'information pertinente. Le choix de HRTF non-individuelles constituant un bon compromis universel ne devrait donc pas se faire sur ces considérations *a priori*, mais, comme l'ont proposé Møller *et al.* [100, 156, 157], d'après une évaluation préliminaire de toutes les HRTF d'une base de données par un large échantillon d'individus. Le sujet élu, appelé sujet "typique", est celui dont les HRTF procurent au plus grand nombre une spatialisation satisfaisante. Choisir les HRTF d'un seul sujet comme compromis universel n'a rien d'une méthode d'individualisation, mais les performances acceptables ainsi obtenues montrent qu'on peut trouver une alternative à la mesure individuelle dans l'utilisation directe de HRTF non-individuelles. Si une base de données est constituée par la mesure de HRTF d'un grand nombre de sujets, dont les différentes morphologies sont représentatives d'une grande partie de la population, alors il est probable qu'un nouvel auditeur y trouve un ensemble de HRTF qui lui convienne. Les techniques que nous présentons ici utilisent ce principe, et se distinguent par la façon de sélectionner les HRTF dans la base de données, afin qu'elles soient au mieux adaptées à un sujet donné.

Seeber et Fastl [230] proposent que la sélection des HRTF soit réalisée d'après des tests psychoacoustiques successifs. Les HRTF sont d'abord jugées sur l'écoute d'une

source sonore (bruit blanc) évoluant sur une trajectoire horizontale. L'auditeur doit éliminer les jeux de HRTF qui sont les moins performants en termes de frontalisation, d'externalisation, et de largeur de la scène sonore, puis il doit classer les jeux de HRTF restants selon différents critères proposés par l'expérimentateur. L'opération prendrait une dizaine de minutes pour une douzaine de jeux de HRTF.

Iwaya [104] propose également de sélectionner les jeux de HRTF à partir de l'écoute de trajectoires dans le plan horizontal, mais avec un jugement par paires : les jeux de HRTF sont mis en comparaison deux à deux dans des matchs, organisés sous la forme d'un tournoi. Le sujet doit dans chaque match choisir les HRTF qui lui conviennent le mieux, et l'ensemble de HRTF finalement choisi est celui qui a remporté le plus de matchs. Selon les auteurs de ces deux études, les HRTF arrivant en tête selon ces méthodes donnent des résultats très peu inférieurs voire comparables à ceux obtenus avec des HRTF individuelles, en termes de taux de confusions avant/arrière. La portée de ces résultats est cependant limitée, car ces protocoles étant restreints au plan horizontal, ils laissent de côté la perception de l'élévation, qui est pourtant la plus critique en synthèse binaurale non-individuelle.

Pour rendre l'évaluation perceptive plus rapide encore, on peut envisager d'effectuer au préalable une classification des HRTF de la base de données, afin de ne présenter au sujet que les représentants d'un nombre réduit de classes. C'est cette possibilité qu'ont explorée Shimada *et al.* [238]. Après avoir mesuré des HRTF dans le plan horizontal pour un grand nombre de sujets, les auteurs décrivent ces données de manière compacte par une décomposition cepstrale de 16 coefficients. Une classification (*clustering*) est ensuite effectuée, basée sur la distance euclidienne entre les vecteurs de coefficients cepstraux. Ainsi, pour chaque direction, 8 clusters et leurs représentants sont calculés. Les sujets sont invités à sélectionner, tous jeux confondus, les HRTF-représentantes qui leur procurent une bonne spatialisation en termes de direction et d'externalisation. Une grande majorité des sujets (95%) parvient à faire ce choix aisément, ce qui démontre la pertinence du *clustering* comme mise en forme préalable d'une base de données. D'autres représentations compactes, et d'autres méthodes d'analyse de la similarité peuvent être envisagées avant l'étape de *clustering*. On peut citer les travaux de Wightman et Kistler [275], qui proposent d'abord une décomposition ACP (cf. Annexe I), comme celle proposée dans [121], puis une analyse de type *Multi Dimensional Scaling* (MDS) qui représente les données dans un espace à deux dimensions. L'étude montre que la dégradation des performances de localisation en synthèse binaurale non-individuelle est bien corrélée avec la distance objective dans cet espace. En d'autres termes, deux sujets sont d'autant plus "compatibles" en synthèse binaurale non-individuelle, que la distance entre leurs jeux de HRTF est faible, ce qui justifie le choix de cette méthode d'analyse pour la classification. Zotkin et al. [287, 288] proposent un autre mode de sélection, basé sur l'idée

suivante : si deux individus présentent des morphologies proches, alors les HRTF de l'un conviendront bien à l'autre. Les auteurs définissent 8 dimensions caractéristiques mesurables sur les pavillons grâce à une simple photographie. Une fois ces dimensions acquises sur chaque pavillon pour un nouvel auditeur, les HRTF sélectionnées sont celles du sujet de la base dont les pavillons présentent les dimensions les plus proches. Cette sélection est réalisée indépendamment pour chaque oreille. Selon cette méthode, les performances de localisation ne sont en moyenne améliorées que de 15% par rapport à des HRTF choisies au hasard dans la base de données, mais elles peuvent aussi être dégradées. La sélection de HRTF non-individuelles dans une base de données peut être une bonne solution, si l'on accepte d'emblée ses limites : les résultats de leur utilisation seront à coup sûr moins bons que ceux de la synthèse binaurale individuelle. Les méthodes de sélection par tests subjectifs, ou par préférence des sujets, semblent donner de bons résultats dans le plan horizontal, mais aucun résultat ne montre leur efficacité à assurer une perception satisfaisante de l'élévation. Par ailleurs la sélection de HRTF par rapprochement morphologique des sujets reste insatisfaisante : les performances obtenues avec les HRTF ainsi adaptées ne sont pas systématiquement meilleures qu'avec des HRTF choisies au hasard. On peut probablement imputer ces résultats à la mauvaise reproductibilité des données anthropométriques sur de simples images 2D, ou bien à la non pertinence du choix de ces dimensions comme représentation paramétrique des pavillons.

4.2.4 Transformation de HRTF non-individuelles

Frequency scaling

Une autre manière d'obtenir des HRTF personnalisées est de transformer de façon contrôlée les HRTF non-individuelles d'une base de données. C'est ce que proposent Middlebrooks [169, 170] et Middlebrooks *et al.*[173], dans la technique appelée *frequency scaling*, que nous traduirons par *scaling* fréquentiel. Cette méthode repose sur une idée simple. Les cavités du pavillon, peuvent être vues comme des résonateurs en parallèle, et on peut imaginer que si les dimensions des cavités sont modifiées d'un certain pourcentage, les caractéristiques des résonances seront modifiées d'un même facteur : cela se traduirait par une homothétie du profil spectral, et donc un décalage des creux et pics spectraux vers les basses fréquences si la taille augmente, et vers les hautes fréquences si la taille diminue. En échelle de fréquences logarithmique, cette homothétie devient une translation. Middlebrooks fait donc l'hypothèse que pour tout couple de sujets, il est possible de réduire ainsi les différences entre leur HRTF pour les deux oreilles et toutes les directions, par un seul et unique facteur d'homothétie, appelé facteur de *scaling*. Cela revient à considérer que la plus grande source de variabilité des HRTF d'un individu à l'autre est liée à des différences de

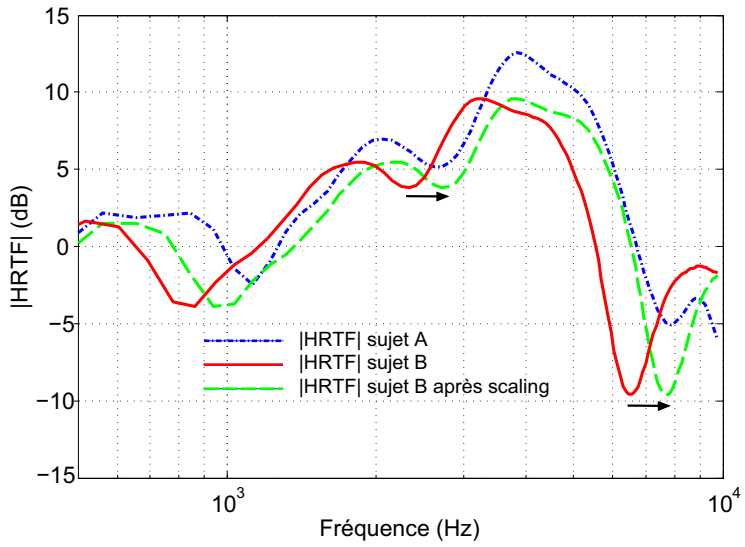


Figure 4.5 – Principe du *scaling* fréquentiel : une simple translation sur l'axe fréqentiel en échelle logarithmique permet d'aligner les profils spectraux, et ainsi de diminuer la différence globale entre les jeux de HRTF. Les deux HRTF correspondent à la même direction, mais appartiennent à deux sujets différents de la base privée de Orange Labs)

taille des pavillons. On illustre figure 4.5 le principe de la technique : la translation sur l'échelle des fréquences est effectuée de manière à assurer une superposition optimale des deux jeux de spectres, minimisant ainsi la distance inter-spectre. L'auteur montre que les différences inter-individuelles sont de cette manière globalement réduites de 15% pour plus de la moitié des sujets, et de plus de 50% pour 9% des sujets. Cependant il subsiste des différences irréductibles, car en réalité, les motifs spectraux des HRTF sont différents d'un individu à l'autre, et peuvent contenir plus ou moins de creux, plus ou moins de pics. Une simple homothétie du profil spectral ne peut traiter ces différences : en général, le *scaling* fréquentiel tend à faire coïncider les premiers creux spectraux. L'auteur valide par des tests de localisation la pertinence du *scaling* fréquentiel. On représente figure 4.6 les résultats d'une expérience menée sur 21 sujets. Les performances de localisation sont évaluées dans 3 cas : en condition individuelle, en condition non-individuelle, et après adaptation optimale des HRTF non-individuelles par *scaling* fréquentiel. L'erreur polaire désigne l'erreur quadratique moyenne de localisation en élévation dans le système de coordonnées polaire-horizontal (cf. Fig. 1), tandis que les erreurs de type *quadrant* correspondent aux erreurs de localisation en élévation de valeur supérieure à 90°. Exprimées en pourcentage, ces erreurs reflètent les confusions haut/bas importantes, et les confusions avant/arrière. Pour la majorité des sujets testés, une amélioration appréciable des performances est observée grâce au *scaling* fréquentiel, ce qui valide conjointe-

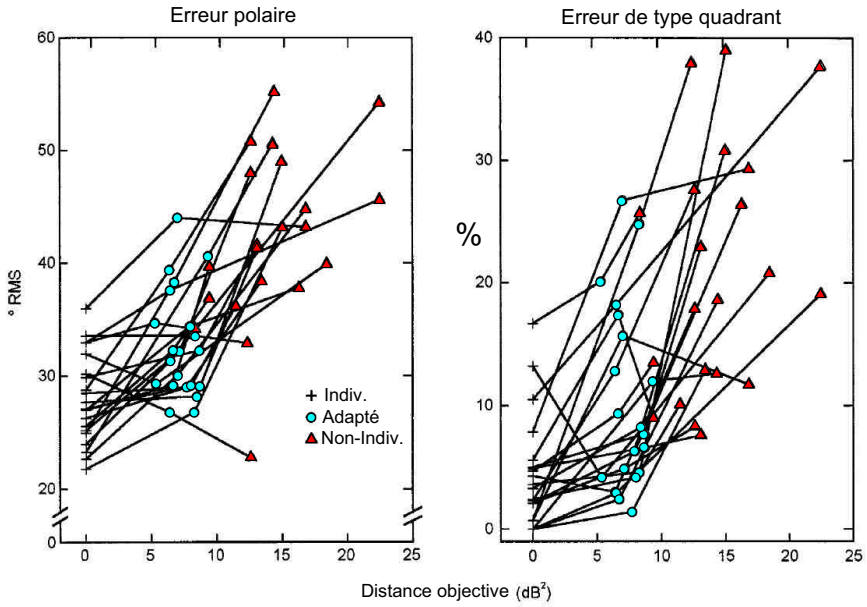


Figure 4.6 – Performances du *scaling* fréquentiel (d'après [170]). Pour 21 sujets, les erreurs de localisation sont représentées en fonction de la distance objective entre le jeu de HRTF utilisées et le jeu de HRTF individuelles de chacun des sujets (ISSD en dB^2, cf. chapitre 5). A gauche : erreur polaire (erreur quadratique moyenne en degrés). A droite : pourcentage d'erreurs de type *quadrant*. L'adaptation des HRTF par *scaling* fréquentiel réduit la distance objective, ce qui se traduit généralement par une amélioration de la précision de localisation en élévation, et une diminution des confusions haut/bas et avant/arrière.

ment le choix de ce mode d'adaptation des HRTF, et la distance objective choisie (ISSD, cf. 5.4).

Middlebrooks fait de cette technique une méthode d'individualisation dès lors qu'il propose deux moyens d'estimer le facteur de *scaling* optimal (ou facteur de *scaling* "signal"), mis en évidence précédemment : par une analyse des données anthropométriques, ou par un ajustement psychophysique.

Le facteur de *scaling* "signal" entre deux sujets, est bien corrélé avec le rapport des hauteurs de leurs pavillons et celui de la taille de leur tête [169]. Le facteur de *scaling* peut donc être simplement prédit d'après la comparaison entre de la morphologie d'un nouvel auditeur, et celle du sujet dont on connaît les HRTF. L'auteur montre que ce facteur de *scaling* "morphologique" approche bien le facteur de *scaling* "signal" à environ 5.8% près.

L'ajustement psychophysique de ce facteur peut être réalisé par l'auditeur lui-même, en réglant pour un jeu de HRTF donné, et pour quelques directions du plan médian, le facteur de *scaling* qui offre la spatialisation la plus fidèle. En une vingtaine de minutes, selon un protocole bien défini, les sujets parviennent à ajuster ce facteur "psychophysique", qui s'approche du facteur "signal" à environ 5.2% près. Le facteur de *scaling* "psychophysique" est celui qui se rapproche le mieux du facteur "signal", et qui permet d'atteindre les meilleures performances de localisation. Pour une réalisation pratique de cette méthode d'individualisation, on peut envisager le protocole suivant :

1. choix des HRTF d'un sujet de la base de données
2. mesure des dimensions du sujet et calcul du facteur de *scaling* "morphologique"
3. détermination d'une fourchette réduite de facteurs de *scaling* à tester, autour du facteur de *scaling* "morphologique"
4. évaluation du facteur de *scaling* "psychophysique"
5. transformation des HRTF selon le facteur de *scaling* "psychophysique"

Larcher [132] suggère que le choix de l'individu de la base de données dont les HRTF sont transformées doit être dicté par ses bonnes compétences en localisation, ou bien par la qualité de ses IS en termes de dépendance spatiale, les deux critères étant liés.

Bisection scaling

Martens [150] propose une méthode appelée *bisection scaling* : l'idée consiste à déterminer de quelle manière l'espace auditif d'un sujet est déformé par l'utilisation de HRTF non-individuelles, et de compenser cette déformation. Une procédure de calibration psychophysique est définie pour permettre au sujet de rendre compte de

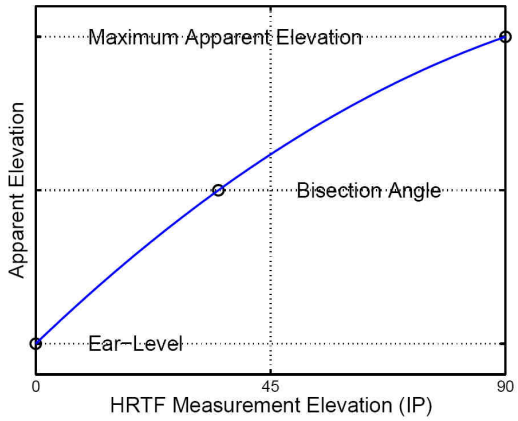


Figure 4.7 – *Look Up Table* obtenue pour un sujet par la technique de bisection scaling [150]. Les trois cercles représentent le lien entre les élévations (en coordonnées polaires horizontales) auxquelles ont été mesurées les HRTF non-individuelles (abscisses) et les élévations perçues par le sujet (ordonnées), sur un cône d'azimut constant. La courbe bleue est le résultat de l'interpolation entre ces trois points.

ces transformations. L'expérience est réalisée pour des directions appartenant à un cône, en rendant l'ITD artificiellement constant, afin que seules les différences spectrales subsistent d'une direction à l'autre. Les élévations décrites sont à comprendre dans le système de coordonnées polaire-horizontal (cf. Fig. 1). Le sujet détermine d'abord quelle HRTF correspond au niveau de ses oreilles, et ce indépendamment pour l'arrière et pour l'avant. La direction d'élévation 90° n'est pas évaluée : elle est dans la suite considérée comme une ancre externe. Le bisection scaling consiste à déterminer, d'après le jugement actif du sujet lui-même, quelles HRTF sont ressenties comme synthétisant les directions bissectrices entre les directions-ancres 0° et 90° à l'avant, et 180° et 90° à l'arrière. Le résultat de cette calibration est alors résumé par une table de correspondance (*Look Up Table* ou LUT) obtenue par interpolation entre ces trois points, et qui décrit le lien entre les élévations pour lesquelles les HRTF non-individuelles ont été mesurées, et les élévations ressenties par le sujet (cf. Fig. 4.7). L'intérêt de cette méthode réside dans la rapidité et la simplicité du protocole psychophysique. Le *bisection scaling* est une méthode encore inaboutie, aucune étude subjective ne l'ayant validée.

4.2.5 *Tuning* du spectre des HRTF

On peut envisager d'adapter des HRTF à un sujet en agissant directement sur le spectre des HRTF : amplifier ou atténuer le spectre d'amplitude sur telle ou telle bande de fréquence revient à modifier les IS, et peut donc permettre d'effectuer une adaptation individuelle. On parlera de *tuning* de HRTF.

Partant de l'idée que la qualité des IS réside dans leurs fortes variations spatiales, et dans l'existence de résonances et antirésonances marquées, plusieurs études ont exploré la possibilité d'amplifier, d'exagérer ces caractéristiques afin que les HRTF soient davantage porteuses de discrimination spatiale, sans chercher à approcher les HRTF du nouvel auditeur. Zhang *et al.*[286], Lee *et al.* [133], et Park et al.[193] proposent d'accentuer la saillance des IS en exagérant les différences spectrales observées entre les HRTF de directions symétriques avant/arrière. Gupta *et al.* [77] modélisent l'effet de pavillons particulièrement décollés, qui, selon les auteurs, procurent une meilleure discrimination avant/arrière, puis proposent d'appliquer artificiellement cet effet sur le spectre des HRTF. Les performances de localisation en termes de discrimination avant/arrière sont meilleures avec les HRTF modifiées selon ces quatre études. Silzle [241] choisit de laisser à un expert le soin de réaliser la phase de tuning, en lui donnant comme outils divers lissages et égalisations. Le but est de trouver un équilibre : améliorer la spatialisation, sans introduire de colorations excessives. Les résultats sont probants, validés par des tests subjectifs, mais il subsiste des faiblesses dans la synthèse de sources frontales, et l'étude se limite au plan horizontal. Ces cinq études démontrent la possibilité d'améliorer la localisation par des manipulations opérées sur le spectre des HRTF. Cependant ce ne sont pas vraiment des techniques d'individualisation, car les modifications apportées aux HRTF sont systématiques, et n'entrent pas en jeu dans une méthode d'adaptation à un sujet donné.

Tan et Gan [246] proposent une technique qui intègre le sujet dans un processus psychophysique de tuning des HRTF. L'auditeur est d'abord invité à choisir dans une base de données un jeu de HRTF qui lui conviennent bien. Cinq bandes de fréquences, couvrant tout le spectre, sont prédéfinies. Des combinaisons particulières d'une amplification ou d'une atténuation de l'énergie dans ces différentes bandes ont été identifiées par les auteurs comme procurant une perception spatiale de la source à l'avant, ou à l'arrière, et plus ou moins élevée. Le protocole est défini de telle manière qu'en ajustant le gain de ces combinaisons via un égaliseur (cf. Fig. 4.8), l'auditeur indique les réglages qui correspondent pour lui à une direction donnée. Cela a pour effet, en pratique, de modifier la position et l'amplitude des pics et des creux spectraux des HRTF non-individuelles. Les auteurs montrent qu'ainsi, les performances de localisation peuvent être améliorées, essentiellement en termes de perception frontale, et peu en termes d'élévation de la source.

Runkle *et al.* [226] proposent un nouveau protocole psychophysique pour adapter des HRTF à un sujet : les auteurs le nomment AST, pour *Active Sensory Tuning*. C'est une méthode d'optimisation adaptative qui tient compte au mieux des réponses du sujet, comme dans un test de la vision, où un ophtalmologiste affine le choix de verres correcteurs en fonction des réponses du patient. Il s'agit de mettre à l'épreuve diverses HRTF dans un test psychophysique, où d'après le ressenti du sujet ("c'est

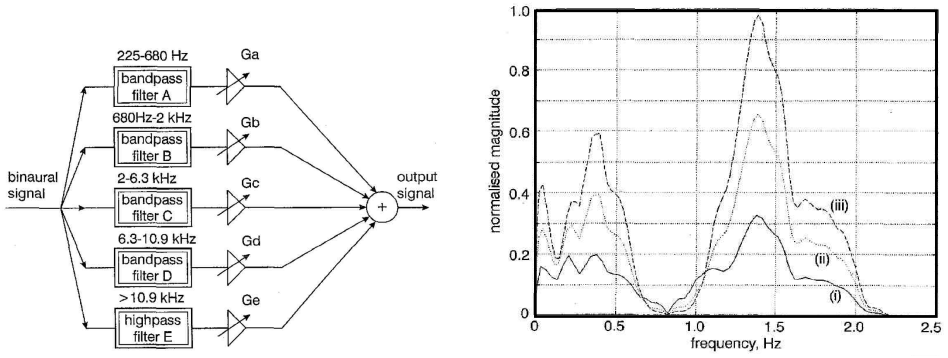


Figure 4.8 – A gauche : Egaliseur 5 bandes utilisé pour le tuning des HRTF. A droite : Effet sur une HRTF d'une combinaison d'atténuations/amplifications prédéfinie pour améliorer la frontalisation ((i) HRTF originale, (ii) atténuation/amplification de 6dB (iii) 12 dB). (d'après [246])

mieux", "c'est moins bien"), d'autres HRTF-candidates sont générées intelligemment, afin de converger le plus rapidement possible vers une solution personnalisée . Les auteurs proposent un algorithme génétique pour créer ces nouvelles HRTF-candidates. Cette technique nécessite pour être efficace de minimiser les degrés de liberté de modification des HRTF : l'alphabet avec lequel travaille l'algorithme génétique doit être limité en taille, et son choix motivé par des considérations perceptives. En pratique, les auteurs utilisent les paramètres d'un modèle pôle-zéro de HRTF [20]. Pour une direction cible, l'auditeur juge la qualité de la spatialisation pour diverses HRTF proposées : 4 HRTF doivent être élues parmi 8 à chaque génération. On représente figure 4.9 le diagramme-bloc du protocole. Les auteurs montrent qu'en général, la convergence vers une solution personnalisée s'effectue en une vingtaine de générations. Cette méthode paraît très puissante, et simple pour le sujet. L'algorithme génétique semble être une solution efficace pour proposer des candidats pertinents. Cependant, la direction cible dans le test de Runkle *et al.* est une source virtuelle, en synthèse binaurale, générée avec les HRTF d'une tête artificielle. Cette solution nous semble maladroite. En effet, cela revient à adapter les HRTF génériques avec comme référence la perception spatiale dans un espace auditif virtuel non-individuel, très probablement médiocre.

Les techniques d'individualisation par *tuning* du spectre des HRTF semblent être potentiellement efficaces. Les difficultés pratiques résident dans le choix et le contrôle des paramètres à ajuster, mais aussi et surtout dans le protocole du test psychophysique. Sur ce dernier point, l'AST proposé par Runkle *et al.* se distingue par son efficacité à tenir compte des réponses du sujet. Un problème critique qui subsiste est le contrôle des colorations. En effet, les opérations de *tuning* du spectre des HRTF peuvent améliorer la spatialisation, mais en même temps dégrader sérieusement le

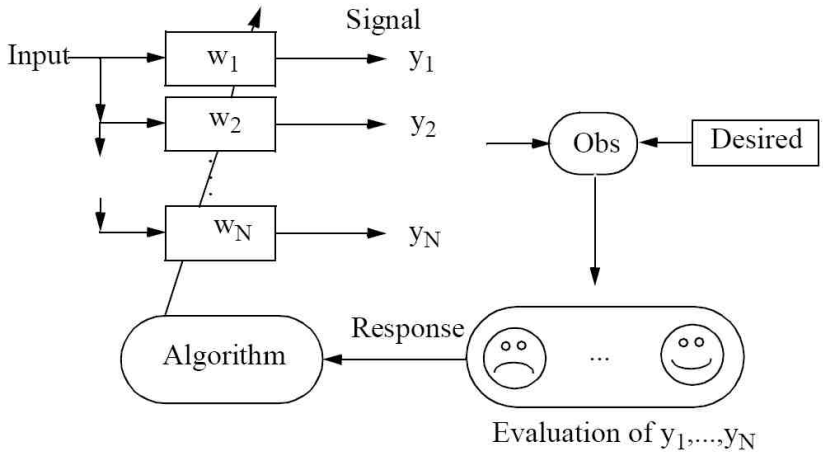


Figure 4.9 – Diagramme-bloc de la technique d'individualisation proposée par Runkle *et al.* [226]. On injecte en entrée une HRTF générique. L'algorithme génétique produit N jeux de paramètres w_i, pour composer les N stimuli candidats y_i. Le sujet évalue la correspondance perceptive de ces candidats avec une consigne (direction cible, ici *desired*), et en fonction de ce jugement, l'algorithme génétique propose une nouvelle génération de N candidats.

timbre du signal à spatialiser. Cette question de la neutralité des HRTF modifiées n'a été considérée que par Silzle [241], mais complètement éludée par les autres auteurs.

4.2.6 Modélisation des HRTF par apprentissage statistique

Une autre solution consiste à obtenir un modèle, grâce à une technique d'apprentissage statistique, qui relie les HRTF à des paramètres décrivant un individu. La première phase consiste à réaliser l'apprentissage sur une base de données contenant les HRTF mesurées dans toutes les directions de nombreux individus, ainsi que leurs paramètres individuels. Le modèle ainsi généré, il suffit de lui fournir en entrée les paramètres individuels nécessaires d'un nouvel auditeur, pour obtenir ses HRTF en sortie.

Rodriguez et Ramirez [218–220] imaginent un tel schéma d'individualisation, inspiré de celui proposé par Jin *et al.* [109], avec comme paramètres individuels des données anthropométriques, telles que celles utilisées par Zotkin *et al.* [287, 288]. Pour décrire de manière compacte les HRTF, une méthode d'extraction automatique des creux spectraux des HRTF est employée, celle proposée par Raykar *et al.* [214, 215], et les HRTF sont représentées par un modèle ARMA. Une analyse en composantes principales est réalisée sur ces données. Le processus d'apprentissage réside dans la régression linéaire multidimensionnelle qui est opérée entre les données anthropométriques d'intérêt, et les poids des décompositions ACP. Les paramètres

choisis dans l'étude se révèlent être d'une efficacité limitée pour prédire les HRTF, et les auteurs proposent pour améliorer leur technique d'utiliser des mesures telles que le volume des cavités du pavillon, ou bien les paramètres d'une description compacte du pavillon, comme la transformée de Fourier elliptique 3D [87]. Aucun test subjectif n'a été réalisé pour valider cette méthode d'individualisation. Une méthode similaire a été développée par Inoue *et al.* [102]. Les erreurs de reconstruction des HRTF sont importantes aux hautes fréquences, et du côté controlatéral. La validité subjective de la méthode n'a été montrée que dans le plan horizontal, dans un nombre très limité de directions, et seulement dans la bande de fréquence [0 kHz ; 8 kHz].

Une autre approche a été proposée par Busson [38]. L'auteur fait l'hypothèse que parmi toutes les HRTF, on peut en dégager un nombre limité, appelées HRTF représentatives, qui portent en elles suffisamment d'informations individuelles pour reconstruire les HRTF dans une direction quelconque. C'est un réseau de neurones qui permet de réaliser l'apprentissage statistique sur une base de données de la relation entre ces HRTF représentatives et les HRTF des directions voisines. Cette technique nécessite donc une mesure acoustique individuelle de ces HRTF représentatives. L'intérêt de la méthode est limité par le constat que les directions pour lesquelles les HRTF représentatives offrent une reconstruction optimale ne sont pas les mêmes d'un individu à l'autre. On peut alternativement choisir une solution sous-optimale, selon laquelle les HRTF représentatives sont réparties uniformément sur la sphère. Malheureusement, l'étude ne comporte ni évaluation subjective des résultats, ni comparaison objective avec les techniques classiques d'interpolation sur la sphère, qui se rapprochent le plus de ce dernier choix.

4.2.7 Mise à profit de la plasticité du système auditif

Le problème d'individualisation des HRTF peut être vu sous un angle différent : au lieu d'adapter des HRTF non-individuelles ou d'en créer sur mesure pour chaque auditeur, pourquoi ne pas exploiter la plasticité du système auditif, et sa capacité à "apprendre" des HRTF non-individuelles ? Hofman et al. [93] ont en effet démontré que si l'on appose des moulages sur les pavillons d'un sujet, modifiant ainsi ses IS, alors ses performances de localisation sont immédiatement dégradées, mais au bout de quelques semaines d'adaptation, le système auditif parvient à assimiler ces changements, et les performances retrouvent un niveau normal. De plus, il apparaît que le système auditif peut conserver simultanément les capacités à décoder les deux types d'IS, individuel et modifié. En effet, une fois les moulages retirés, les sujets retrouvent très rapidement leurs repères, et leurs performances de localisation. Cette expérience peut être transposée dans le domaine de la synthèse binaurale. En effet, apposer des moulages sur les pavillons, en écoute réelle, revient, en écoute virtuelle, à utiliser

des HRTF non-individuelles. Les dégradations initiales observées par Hofman *et al.* peuvent être comparées aux faiblesses de la synthèse binaurale non-individuelle, décrites précédemment. Blum *et al.* [24] ont donc exploré la possibilité d'accélérer, ou forcer l'apprentissage de HRTF non individuelles. La calibration du système auditif pour la localisation étant un processus multi-sensoriel, les auteurs ont imaginé un protocole, dans lequel l'auditeur peut explorer activement son espace auditif virtuel, en déplaçant simplement une balle, tenue dans la main, qui matérialise la source sonore (cf. Fig. 4.10), en synthèse binaurale dynamique. La direction de la balle est enregistrée en temps réel par rapport à la tête, et la source sonore qu'elle représente suit dans l'espace virtuel les mouvements que lui applique le sujet dans l'espace réel. La calibration est ainsi possible entre les deux espaces, et l'auditeur peut de cette manière apprendre rapidement à évoluer dans un espace auditif non-familier. Les auteurs choisissent de n'utiliser que la modalité sensori-motrice de l'apprentissage : les sujets ont donc les yeux bandés pendant l'apprentissage. Ils se créent ainsi une cartographie "auditivo-kinesthétique" de l'espace [23]. Les premiers résultats indiquent que cet apprentissage est effectif après une séance de 12 minutes seulement : une amélioration assez modeste, mais significative, est observée en termes de confusions avant/arrière, et de localisation en élévation. Pour évaluer l'intérêt de la technique, il conviendrait en plus de tester la durée de cet effet d'apprentissage, ce que n'ont pas proposé les auteurs dans ces travaux qui se voulaient exploratoires. Cette approche conforte l'idée selon laquelle la calibration visuelle n'est pas absolument nécessaire dans le processus d'apprentissage de localisation sur la base des IS, mais que d'autres modalités sensorielles peuvent intervenir (cf. Van Wanrooij [253], et Kacelnik *et al.* [113]).

Honda *et al.* [95] utilisent un jeu vidéo pour réaliser l'apprentissage de HRTF non-individuelles. Les HRTF sont préalablement sélectionnées pour convenir au mieux à l'auditeur, comme décrit par Iwaya [104]. Dans ce jeu, des sources sonores sont spatialisées en synthèse binaurale dynamique, et matérialisées par des abeilles qu'il s'agit de trouver, et d'écraser avec un maillet. Les modalités visuelle, motrice, et auditive participent donc conjointement à l'apprentissage. Les auteurs affirment qu'après un entraînement quotidien de 30 minutes pendant une semaine, les performances des sujets dans ce jeu atteignent celles obtenues en condition individuelle.

Zahorik *et al.* [284] choisissent eux une phase d'apprentissage basée sur un retour visuel. Le sujet écoute des stimuli en synthèse binaurale, tout en étant immergé dans un espace graphique virtuel, par le biais d'un casque immersif (*Head Mounted Display* ou HMD), disposant d'un système de *head-tracking*. La phase d'apprentissage ressemble à test de localisation : le sujet doit d'abord pointer le nez dans la direction où il perçoit chaque source virtuelle, présentée une seule fois. Après enregistrement de chaque réponse, un point lumineux s'allume, pour matérialiser la source sonore

Figure 4.10 – Photographie du dispositif d'apprentissage accéléré proposé par Blum *et al.* [24]. La balle tenue en main par le sujet matérialise la source sonore, diffusée virtuellement en synthèse binaurale dynamique

virtuelle, tandis que le stimulus est à nouveau diffusé sur casque. Le sujet doit alors pointer le nez vers ce point lumineux, correspondant à la réponse "juste", avant de passer au stimulus suivant. Les auteurs montrent que par ce protocole d'apprentissage, une diminution significative des confusions avant/arrière est possible. Par contre, aucune amélioration de la perception de l'élévation n'est observée, ce qui montre que les IS non-individuels n'ont pas été totalement assimilés par les sujets.

Chapitre 5

Adaptation morphologique de HRTF non-individuelles

On développe dans ce chapitre une solution d'individualisation des HRTF qui vise à adapter, pour un nouvel auditeur, les HRTF d'un autre individu issues d'une base de données. Le jeu de transformations des HRTF qui permet cette adaptation est prédit par un alignement morphologique entre les pavillons des deux sujets. La méthode constitue une extension du *scaling* fréquentiel de Middlebrooks (cf. 4.2.4), car une transformation supplémentaire - la rotation - est considérée.

5.1 Observations préliminaires et hypothèses de travail

Malgré les différences marquées qui existent entre les pavillons, d'un individu à l'autre, l'observation des HRTF sous forme d'une série de fonctions de directivité permet de dégager des comportements typiques. La trajectoire classique de l'axe acoustique pour une fréquence croissante, prenant la forme d'un α, révèle l'existence de modes de résonance similaires, quel que soit le pavillon considéré (cf. Fig. 3.12). On retrouve en particulier les deux comportements dipolaires dits vertical et horizontal.

On représente figures 5.1, 5.2, 5.3, et 5.4 les fonctions de directivité de trois sujets différents pour une fréquence croissante. Malgré l'existence de fortes ressemblances, d'un sujet à l'autre, entre les motifs adoptés par les lobes principaux, les profils spectraux des HRTF correspondantes peuvent apparaître très différents, car deux sources de variabilité viennent brouiller ces observations : des décalages sur l'axe fréquentiel, ainsi que des différences d'orientation spatiale.

On remarque tout d'abord entre les deux premiers sujets (figures 5.1 et 5.2) que des fonctions de directivité similaires correspondent à des fréquences différentes d'un individu à l'autre. De ce constat est née la méthode d'adaptation appelée *frequency scaling* (cf. 4.2.4) : les différences de taille entre les pavillons seraient à l'origine de ces décalages fréquentiels, et une simple homothétie le long de l'axe fréquentiel permettrait de réduire les différences entre les profils spectraux de deux individus. Middlebrooks [169] a montré que le facteur de l'homothétie optimale est bien corrélé au ratio des hauteurs d'oreille, ou des tailles de tête des sujets considérés. Ainsi, pour un individu présentant de grands pavillons, on observera sur le profil spectral de ses HRTF des creux et des pics disposés à des fréquences plus basses que pour un sujet ayant des pavillons plus petits. Larcher [132] s'est intéressée à l'alignement des profils spectraux aux hautes fréquences, et a montré une bonne corrélation du facteur de *scaling* optimal avec le ratio des hauteurs de conque. Middlebrooks montre par ailleurs que la réduction des différences inter-individuelles par *scaling* fréquentiel permet de diminuer en partie les artefacts liés à l'utilisation de HRTF non-individuelles : on observe notamment une amélioration de la perception de l'élé-

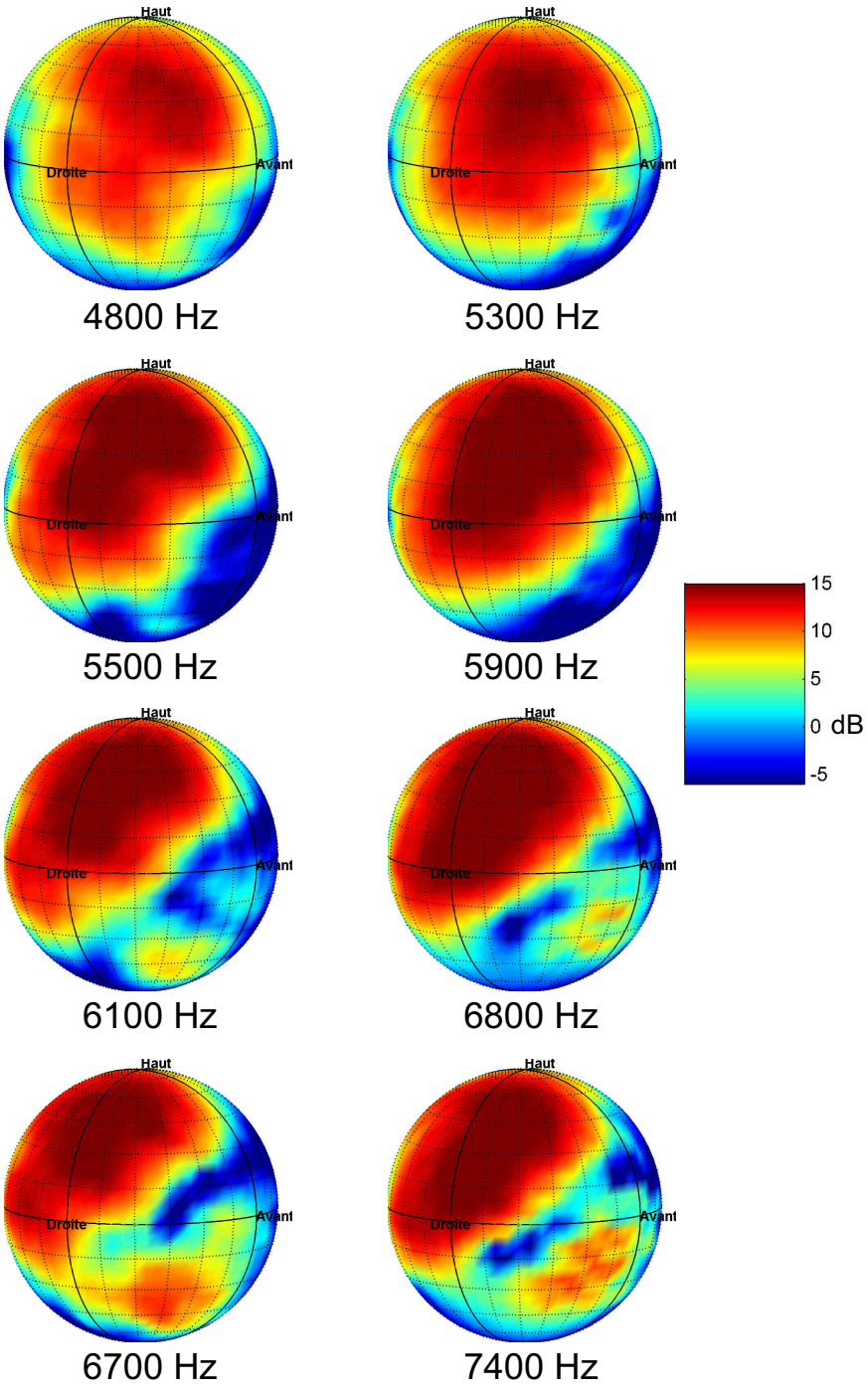


Figure 5.1 – Fonctions de directivité pour une fréquence croissante. A gauche : sujet n°1 de la base privée d'Orange Labs, oreille gauche. A droite : sujet n°27 de la base publique du CIPIC [57], oreille gauche. Des motifs similaires sont observés, mais avec un décalage fréquentiel. Les données sont symétrisées par rapport au plan médian (cf. *infra*).

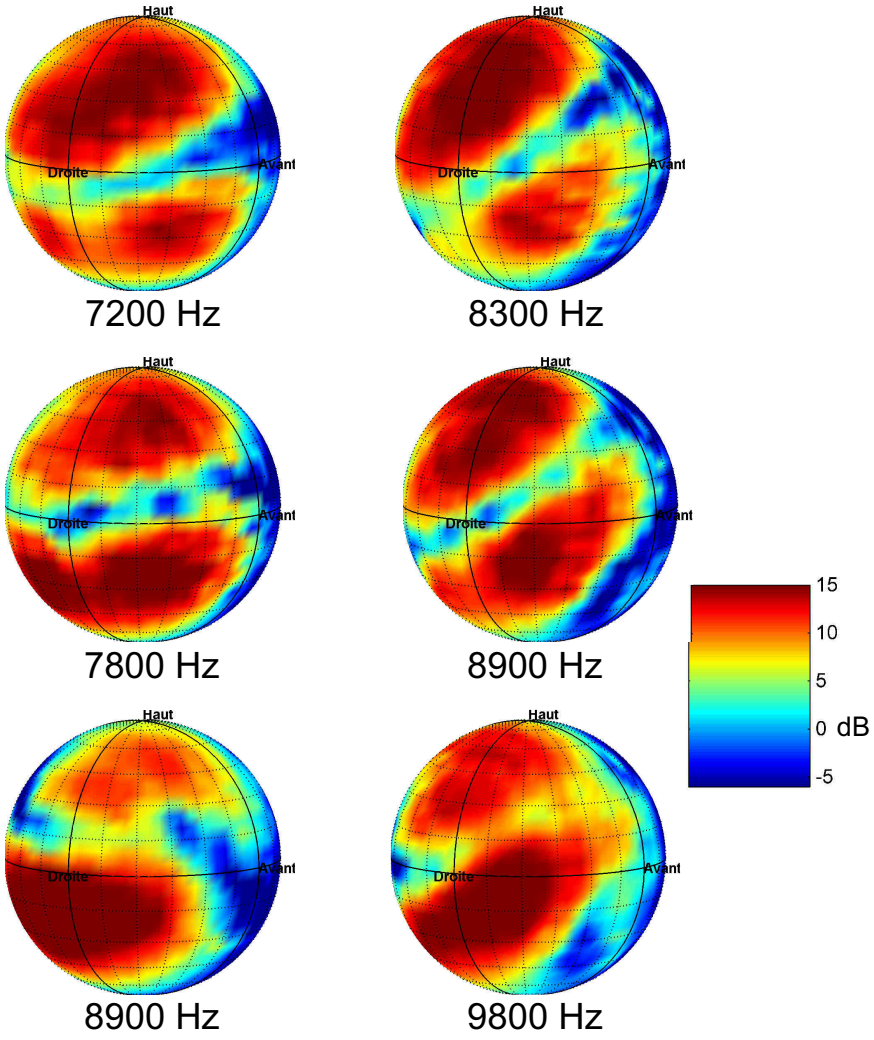


Figure 5.2 – Suite de la figure 5.1, pour des fréquences supérieures.

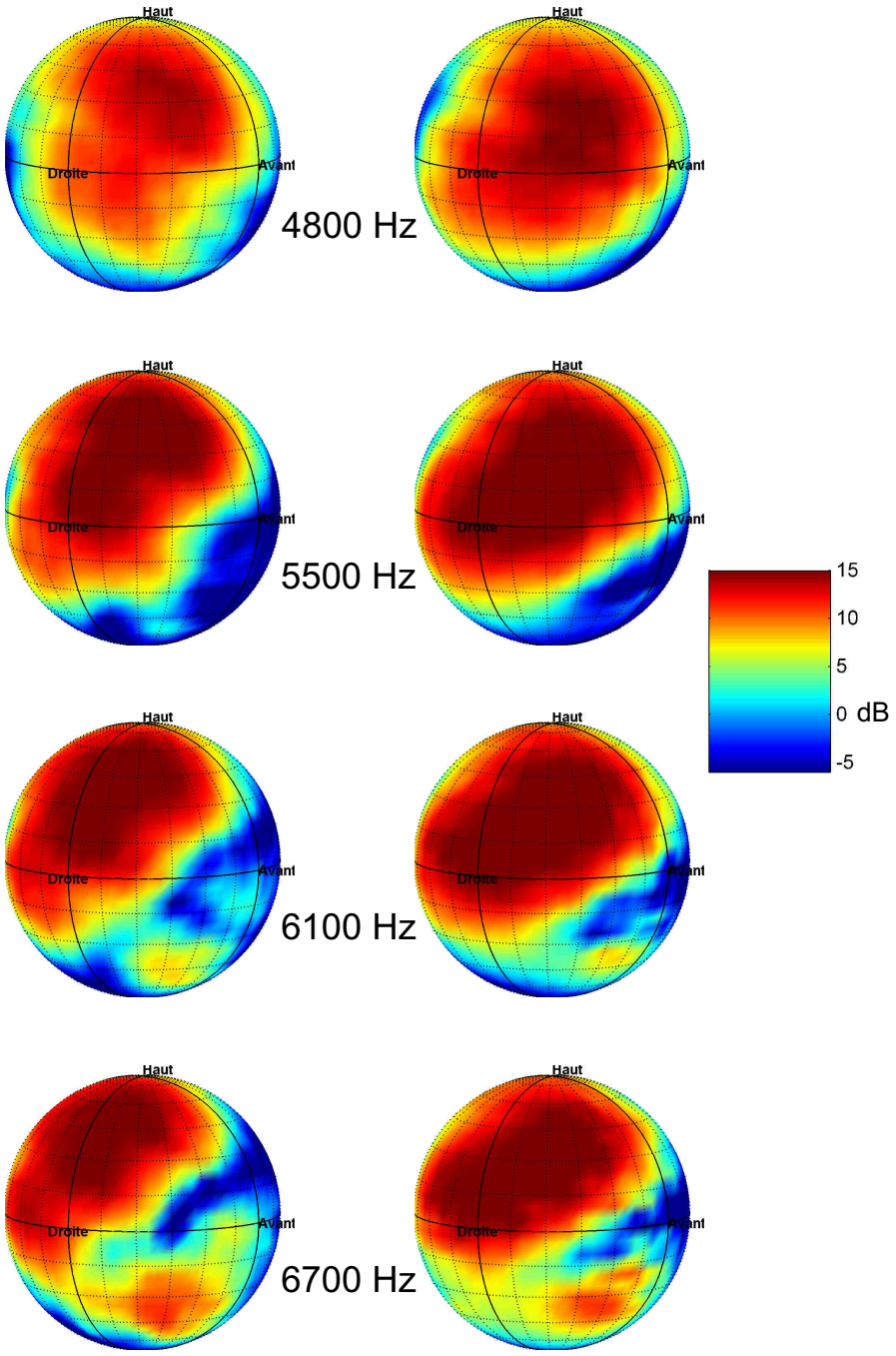


Figure 5.3 – Fonctions de directivité pour une fréquence croissante. A gauche : sujet n°1 de la base privée d'Orange Labs, oreille gauche. A droite : sujet *th* de la base publique de l'Université du Maryland [76], oreille gauche. Malgré des similarités dans la forme des lobes, des différences spatiales sont observables : l'orientation de la vallée séparant les deux lobes principaux est différente entre les individus. Les données sont symétrisées par rapport au plan médian (cf. *infra*).

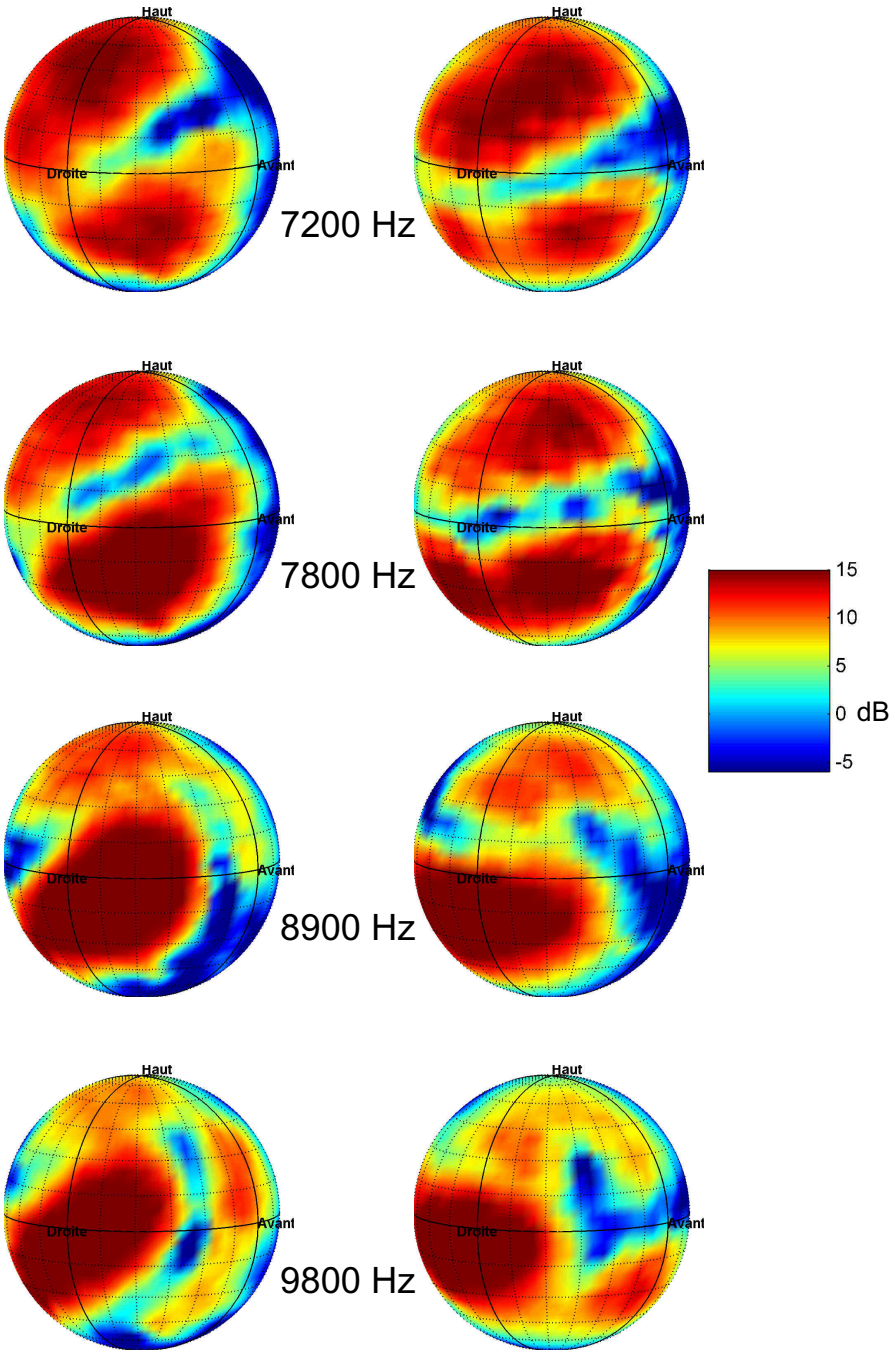


Figure 5.4 – Suite de la figure 5.3, pour des fréquences supérieures.

vation et une diminution des confusions avant/arrière. C'est en ce sens une méthode efficace d'individualisation des HRTF.

Les fonctions de directivité se distinguent également par des différences spatiales. On observe par exemple figures 5.3 et 5.4 une orientation différente, selon l'individu, de la vallée entre les lobes disposés verticalement. Elle est plus ou moins décalée, et plus ou moins inclinée par rapport au plan horizontal. Maki et Furukawa [148] se sont penchés sur le cas de la gerbille de Mongolie, et ont montré qu'une rotation du système de coordonnées peut contribuer à réduire ces différences spatiales inter-individuelles. Le principe de cette transformation est le suivant : après une rotation Γ du système de coordonnées, la HRTF H_Γ résultante dans la direction χ est égale à la HRTF H mesurée dans la direction $\Gamma^{-1}(\chi)$:

$$H_\Gamma(\chi) = H(\Gamma^{-1}(\chi)), \quad \forall \chi \in S^2 \tag{5.1}$$

En utilisant conjointement le *scaling* fréquentiel et cette opération globale de décalage par rotation, les auteurs montrent que l'on peut améliorer la méthode de Middlebrooks, c'est-à-dire réduire encore plus les différences objectives entre les jeux de HRTF de deux gerbilles. La technique de Maki et Furukawa devient une méthode d'individualisation dès lors que la rotation optimale du système de coordonnées est bien prédite par des quantités mesurables décrivant les différences d'orientation entre les pavillons des gerbilles.

Enfin, mis à part ces décalages spatiaux et fréquentiels, on remarque que les HRTF présentent des traits tout à fait individuels : la forme des lobes et leur amplitude diffèrent, les rendant parfois non alignables d'un individu à l'autre, ni par rotation, ni par *scaling* fréquentiel (à titre d'exemple, cf. Fig. 3.10 et 3.11). C'est ce que l'on peut appeler la part irréductible des différences entre les HRTF, celles liées à l'existence de caractéristiques très individuelles des pavillons.

Partant de ces observations et résultats, nous proposons une nouvelle méthode d'adaptation de HRTF non-individuelles, s'inspirant des travaux de Maki et Furukawa sur la gerbille de Mongolie. Son principe est décrit figure 5.5 : les HRTF issues d'une base de données sont choisies et transformées pour un nouvel auditeur, les paramètres de ces transformations étant calculés d'après le résultat d'une comparaison morphologique entre le nouvel auditeur et le propriétaire des HRTF. On choisit de se focaliser sur les indices spectraux de la localisation, au delà de 4 kHz. C'est pourquoi, comme l'ont suggéré Larcher [132] sur l'homme, et Maki et Furukawa [148] sur la gerbille de Mongolie, la comparaison morphologique doit être réalisée entre les pavillons des individus. La mise au point de cette technique d'individualisation pose plusieurs problèmes, auxquels nous apportons des solutions.

D'abord, il s'agit de **quantifier les différences morphologiques entre deux individus** (bloc *a* sur la figure 5.5). On ne s'intéresse qu'aux différences de taille

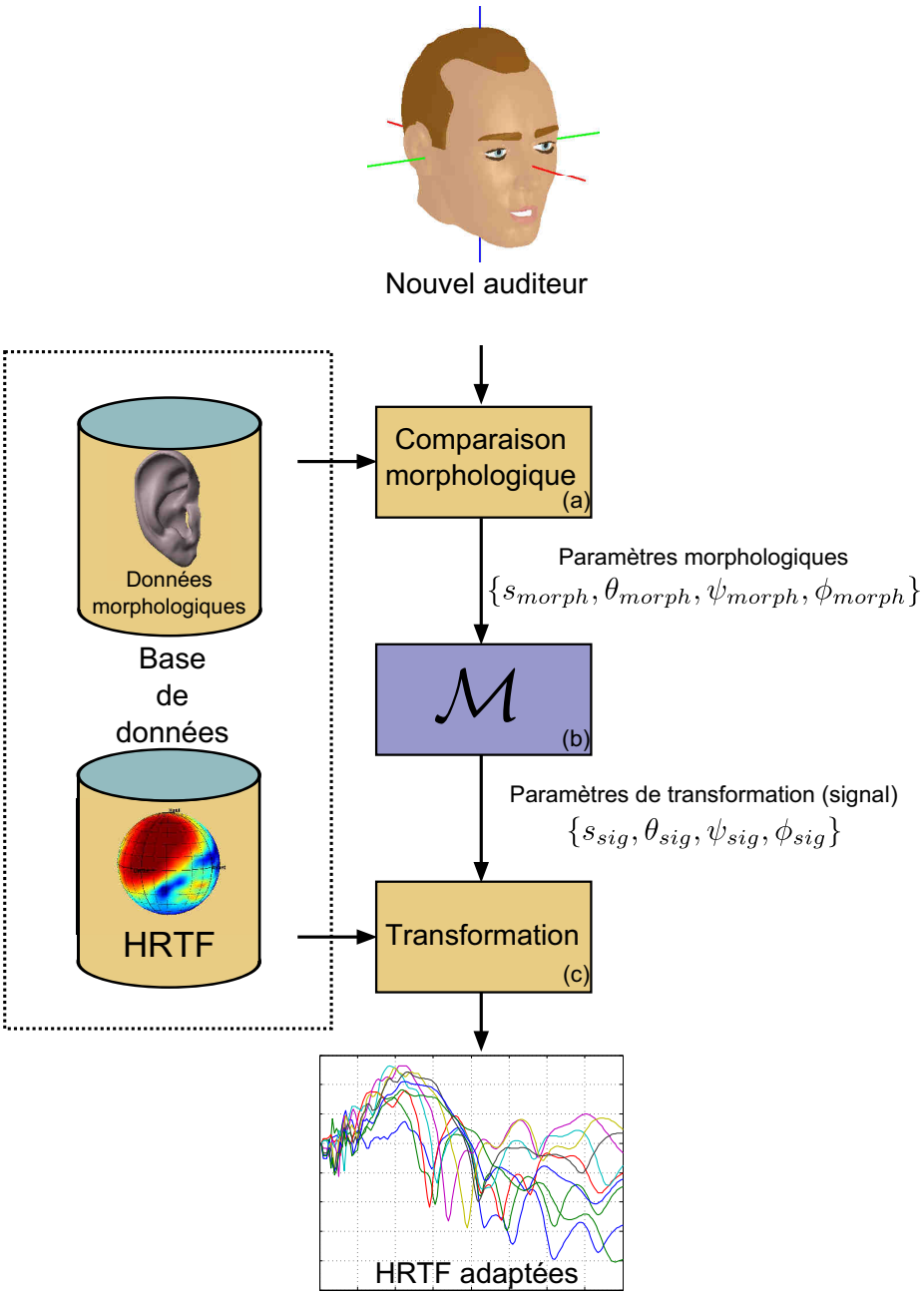


Figure 5.5 – Principe de la technique d'individualisation proposée. L'étape de transformation combine *scaling* fréquentiel et rotation du système de coordonnées. Les différents paramètres sont définis en 5.5.

et d'orientation observées entre deux pavillons donnés. Ainsi, il s'agit d'estimer quel facteur d'échelle et quelle rotation permettent d'aligner idéalement les surfaces en 3D de ces deux pavillons. Cette estimation est réalisée au moyen d'un algorithme approprié appelé ICP, ou *Iterative Closest Point* [17].

Par ailleurs, on doit se doter des **outils algorithmiques nécessaires à la transformation des HRTF** : *scaling* fréquentiel et rotation du système de coordonnées (bloc *c* sur la figure 5.5). Nos solutions s'inspirent largement de ceux proposés par Middlebrooks [169], puis Maki et Furukawa [148].

Enfin, au coeur de la technique d'individualisation proposée se situe l'élément permettant de déterminer les paramètres de transformation des HRTF à partir des résultats de la comparaison morphologique (bloc *b* sur la figure 5.5). Cet élément est mis au point grâce à l'analyse conjointe, pour un échantillon de sujets, des différences morphologiques entre deux individus, et des différences entre les HRTF de ces mêmes individus. Une étape préalable consiste donc à déterminer, indépendamment de la morphologie, les valeurs des paramètres de transformation qui minimisent la distance objective entre deux jeux de HRTF : ces paramètres sont dits "optimaux". Comme proposé par Middlebrooks [169], l'optimisation réalisée est la minimisation d'une distance objective globale, appelée ISSD (*Inter-Subject Spectral Difference*), entre les spectres d'amplitude de deux jeux de HRTF. Une fois ces paramètres optimaux déterminés, il suffit de les relier par régression aux paramètres résumant les différences morphologiques entre les sujets correspondants, afin d'établir la relation $\mathcal{M}$ qui permet ainsi de déterminer les paramètres de transformation des HRTF à partir du résultat de l'alignement morphologique (cf. Fig. 5.5).

Après une description des données disponibles pour l'expérience (en 5.2), on détaille respectivement en 5.3.1 et 5.4 les outils développés d'une part pour aligner les morphologies de deux pavillons, et d'autre part pour déterminer les paramètres de transformation optimaux entre deux jeux de HRTF. A partir des résultats obtenus, l'élément prédictif central de la méthode proposée est mis au point, puis évalué en 5.5.

5.2 Dispositif expérimental

Les expériences permettant de mettre au point et d'évaluer la technique proposée sont réalisées avec les données recueillies sur 6 sujets, employés d'Orange Labs.

5.2.1 Acquisition de la morphologie

L'algorithme de comparaison morphologique nécessite la connaissance en trois dimensions de la surface des pavillons d'oreille de chaque sujet. De plus on s'intéresse

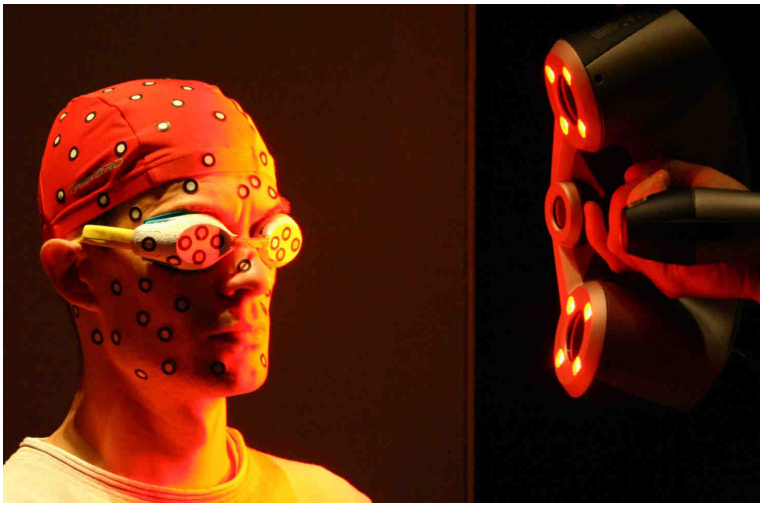


Figure 5.6 – Acquisition en 3D de la surface de la tête d'un sujet grâce à un scanner laser (*Creaform Handyscan 3D™*).

aux orientations spatiales des pavillons, c'est pourquoi il faut également acquérir la surface de la tête, car c'est par rapport à celle-ci que sont ajustés les axes du référentiel. Ces surfaces sont acquises au moyen d'un scanner laser (*Creaform Handyscan 3D*™) (cf. Fig. 5.6). Cet appareil présente l'avantage d'être manipulable à la main, ce qui permet d'acquérir la surface d'objets complexes. Les repères spatiaux utilisés pendant la phase d'acquisition sont matérialisés par de simples gommettes réfléchissantes disposées sur la tête du sujet (cf. Fig. 5.6 et 5.8). Le sujet peut donc se permettre de ne pas rester strictement immobile pendant le scan. Les cheveux diffusent le faisceau laser émis par le scanner, c'est pourquoi les sujets doivent porter un bonnet de bain. Pour protéger leurs yeux du faisceau laser, ils sont de plus équipés de lunettes opaques. Les méandres des pavillons sont cependant difficiles à acquérir *in vivo* : en effet, ces surfaces concaves sont mal détectées par le scanner. Il est alors nécessaire de réaliser un moulage des pavillons, dont la surface est ensuite scannée (cf. Fig. 5.7). L'opération de moulage est effectuée par un audioprothésiste. Les données issues des scans de la tête et des deux moulages de pavillons sont fusionnées dans un logiciel approprié. On obtient pour chaque sujet un maillage tel que celui représenté figure 5.10.

Le référentiel lié à chacune des têtes est déterminé comme suit. L'axe interaural est la droite qui passe par les sommets des *tragus* gauche et droit, nommés respectivement T_G et T_D. L'origine du référentiel est le centre de la tête, défini comme le milieu du segment $[T_G T_D]$. Le plan médian est le plan vertical perpendiculaire à l'axe interaural et passant par l'origine. Il reste alors une indétermination sur la définition du plan horizontal et du plan vertical interaural. On lève cette indétermination en

Figure 5.7 – Réalisation d'un moulage du pavillon de l'oreille. Les surfaces concaves, difficiles à atteindre *in vivo*, deviennent des surfaces convexes sur le moulage.

Figure 5.8 – Préparation avant l'acquisition 3D. Des gommettes réfléchissantes sont collées sur la surface à acquérir. Elles constituent des ancres spatiales nécessaires pour l'acquisition par le *Creaform Handyscan 3D™*.

Figure 5.9 – Obtention d'un point de référence sur le nez pour la détermination du repère de la tête.

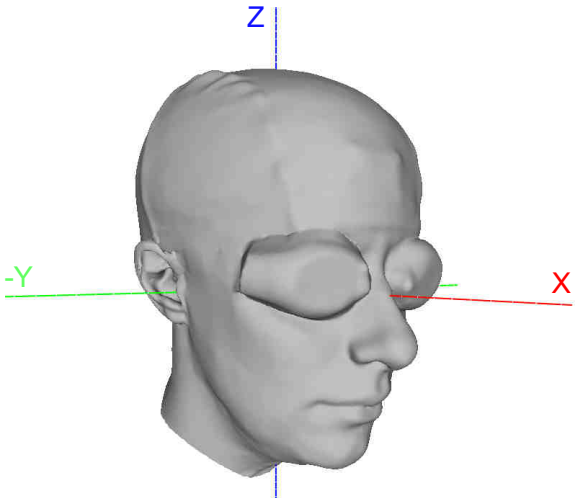


Figure 5.10 – Maillage 3D de la tête et des oreilles obtenu par scan laser.

prenant des photographies du profil droit de chacun des sujets, qui ont pour consigne de se tenir droits, le regard à l'horizon [1], afin de se rapprocher au mieux de la posture adoptée lors de la mesure des HRTF [2]. Sur les photographies, un point est alors repéré sur le nez de chaque sujet, à la même hauteur que le *tragus* (cf. Fig. 5.9). Ce même point est identifié sur le scan 3D correspondant, et l'axe médian est défini selon la droite qui relie l'origine à ce point. Le plan horizontal passe donc par les deux *tragus* et par le point repéré sur le nez. On connaît ainsi précisément la position et l'orientation de chacun des pavillons par rapport au référentiel de la tête, défini de façon à coïncider au mieux avec le référentiel de mesure des HRTF.

Dès lors, les surfaces des pavillons sont considérées indépendamment, et la surface de la tête n'est plus utilisée. Tous les pavillons d'oreilles gauches sont symétrisés par rapport au plan médian de telle sorte que les 12 pavillons disponibles soient tous comparables. Si l'on exclut la comparaison d'un pavillon gauche avec le pavillon droit d'un même sujet, on dénombre un total de 60 couples de pavillons comparables.

Pour l'expérience, on utilise 4 types de maillages issus des scans de pavillons. Ils correspondent à différentes découpes de la surface des pavillons, éliminant progressivement des éléments de leur anatomie. Le dernier niveau de découpe ne conserve que la surface de la conque (cf. Fig. 5.11). Ce paramètre d'étude est nécessaire, car l'importance relative des différents éléments du pavillon n'est pas claire d'après la littérature.

5.2.2 Préparation des HRTF

Les HRTF utilisées sont issues de la base privée d'Orange Labs. Chaque sujet considéré dispose de ses HRTF individuelles, mesurées et validées antérieurement à ces expériences [197]. Ces mesures ont été réalisées selon la technique du conduit bloqué, sur un échantillonnage fin de l'espace (965 directions, à partir du plan d'élévation -56.25°, système polaire vertical, cf. Fig. 1) [29]. Au début de la mesure, un ajustement a été effectué de façon à ce que la position de la tête du sujet par rapport au dispositif de haut-parleurs se rapproche au mieux de la position idéale [3]. Une fois cette position initiale enregistrée, la mesure des HRTF a été réalisée en mesurant en permanence la position et l'orientation de la tête du sujet. Ainsi, si le sujet ne restait pas parfaitement immobile, un réajustement du positionnement du haut-parleur de mesure permettait de compenser les décalages de sa tête par rapport à la position

1. Un repère visuel était disposé à cet effet sur le mur qui faisait face aux sujets.

2. La mesure des HRTF ayant été réalisée antérieurement à cette expérience, au cours des travaux de thèse de Pernaux [196], il existe une incertitude quant à la position réellement adoptée par les sujets lors de la mesure.

3. La position idéale de la tête est celle décrite figure 1 : l'axe interaural, l'axe médian et l'axe vertical sont confondus avec les axes du référentiel du laboratoire.

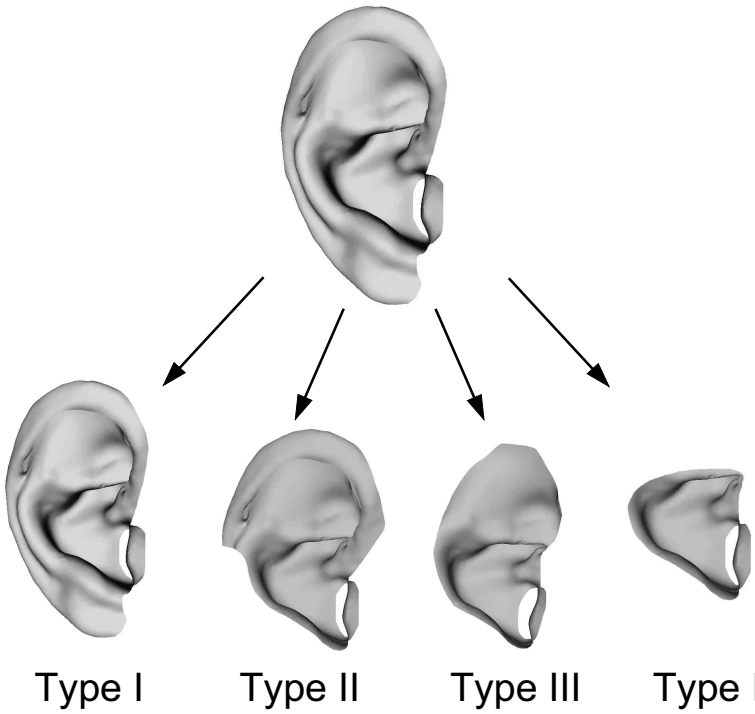


Figure 5.11 – On utilise 4 types de maillage, issus de découpes progressives du pavillon. La comparaison morphologique entre deux pavillons est réalisée sur des maillages de même type.

initiale.

Cependant, malgré tous les soins pris par l'expérimentateur, cette position initiale a pu être légèrement différente de la position idéale. Dans la mesure du possible, une telle déviation doit être corrigée pour s'assurer que le référentiel lié aux pavillons d'oreille et celui lié aux HRTF coïncident. On fait l'hypothèse que la tête du sujet était bien au centre du dispositif expérimental, mais qu'elle présentait éventuellement une déviation par rapport à son orientation idéale (cf. Fig. 5.12), constante au cours de la mesure. Une simple rotation du système de coordonnées doit alors suffire pour réaligner le jeu de HRTF sur le référentiel idéal. Comme proposé par Maki et Furukawa [148], on peut estimer cette rotation à partir du calcul de l'ITD. Dans l'hypothèse d'une orientation idéale de la tête, il est raisonnable de considérer l'ITD comme une fonction fortement antisymétrique par rapport au plan médian. C'est en effet ce que prédisent les modèles de têtes sphériques (cf. 1.1.1). Chercher la rotation qui corrige l'erreur d'orientation de la tête revient donc à chercher la rotation du système de coordonnées selon laquelle l'ITD est la plus antisymétrique par rapport au plan médian. Nous proposons pour cela une nouvelle méthode. L'ITD est d'abord estimée pour chaque direction en cherchant le maximum d'intercorrélation entre les enveloppes des HRIR gauches et droites [272]. L'ITD est alors une fonction

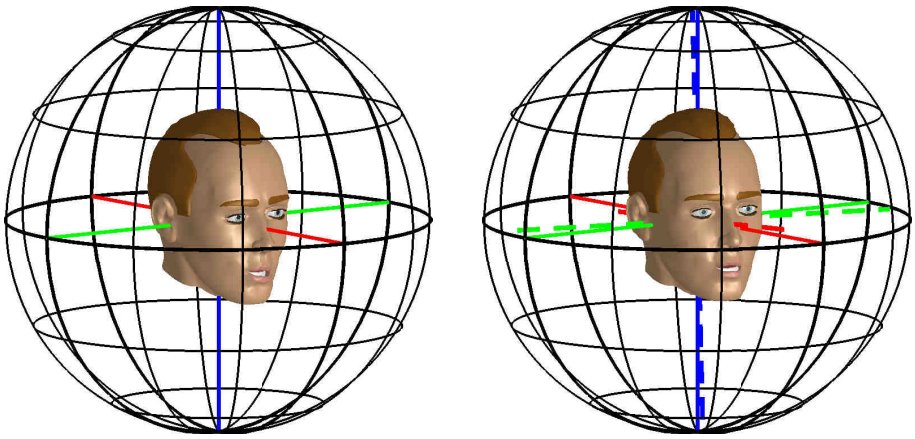


Figure 5.12 – Lors de la mesure des HRTF, l'orientation réelle de la tête du sujet (à droite) a pu dévier de l'orientation idéale (à gauche). On fait l'hypothèse que la tête était bien centrée à l'origine du dispositif de mesure.

définie sur la sphère, que l'on décompose sur une base d'harmoniques sphériques, selon la même procédure que pour les SFRS, décrite en 5.4. La décomposition en harmoniques sphériques complexes présente une particularité intéressante : la partie symétrique de la fonction décrite est portée par la partie réelle des coefficients de la décomposition, et donc dans notre problème, maximiser le caractère antisymétrique de la fonction d'ITD revient à minimiser leur énergie. L'optimisation est réalisée grâce à un algorithme de descente du gradient sur le groupe des rotations $SO(3)$, comme celui décrit en 5.4. L'analyse des erreurs d'orientation est ainsi réalisée pour les 6 sujets. Les rotations résumant ces erreurs sont analysées comme la composée de trois rotations selon les angles roll, pitch et yaw (cf. Fig. 5.14). Les valeurs obtenues pour ces trois angles tombent respectivement dans les intervalles $[-3.75°; 2.2°]$, $[-0.17°; 0.14°]$, et $[-1.86°; 1.43°]$, ce qui montre que les erreurs d'orientation commises pendant la mesure étaient très faibles. Néanmoins, on applique pour chacun des sujets la rotation du système de coordonnées nécessaire pour les corriger. On représente figure 5.13 l'évolution spatiale de l'ITD sur la sphère avant et après un tel recalage. On observe une amélioration de la symétrie, et un recentrage près du plan médian de la ligne de niveau iso-ITD à 0 μs. Des tests informels réalisés sur des HRTF issues d'autres bases de données révèlent que cette technique est très efficace, et permet de corriger les erreurs d'orientation importantes parfois observées. Après cette opération, les jeux de HRTF correspondant à des oreilles gauches sont symétrisés par rapport au plan médian : cette opération est équivalente à l'opération de symétrie effectuée sur les maillages 3D des pavillons. En procédant ainsi, tous les jeux de HRTF deviennent comparables, quelle que soit l'oreille considérée.

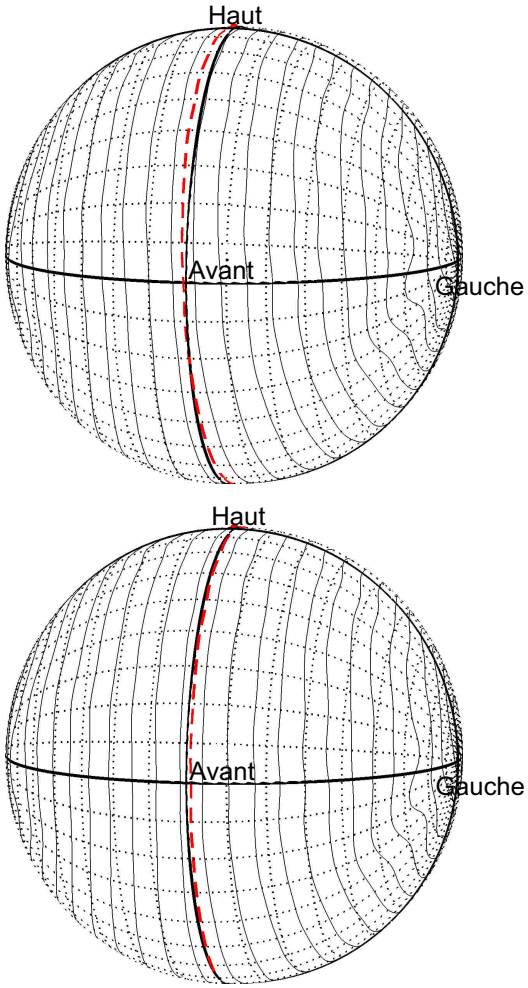


Figure 5.13 – Contours iso-ITD sur la sphère pour le sujet n°3 de la base privée de HRTF d'Orange Labs. En haut : avant la correction. En bas : après la correction. Un modèle de tête sphérique prédit l'existence de contours iso-ITD inscrits dans des plans sagittaux, avec en particulier la ligne iso-ITD à $0\,\mu s$ inscrite dans le plan médian. Après la correction, les contours iso-ITD s'approchent mieux de ces comportements.

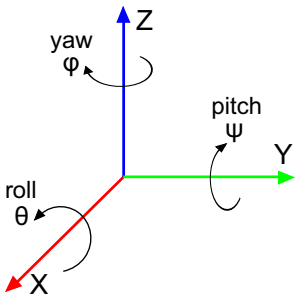


Figure 5.14 – Définition des 3 angles de rotation : roll, pitch et yaw.

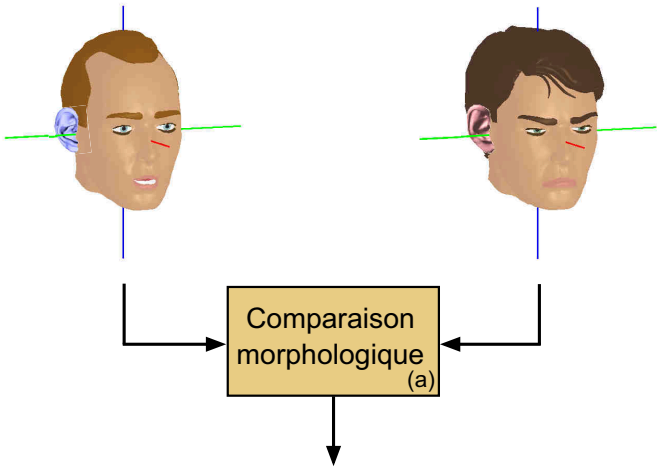


Figure 5.15 – A partir des scans 3D de la morphologie de deux sujets, le premier élément de la technique proposée extrait des paramètres quantitatifs décrivant le résultat de l'alignement entre leurs pavillons.

5.3 Alignement morphologique

5.3.1 Principe

On considère ici l'élément de la technique d'individualisation représenté figure 5.15, qui réalise une comparaison morphologique entre les pavillons de deux individus, dont les surfaces ont été préalablement acquises en 3D. On propose d'utiliser une technique d'alignement afin de résumer par des paramètres quantificatifs ces différences morphologiques. Idéalement, cette technique doit pouvoir se focaliser sur la forme intrinsèque des pavillons, qui est à l'origine des colorations spectrales d'intérêt. C'est pourquoi elle doit être capable, grâce à des degrés de liberté tels que la rotation et l'homothétie, de mettre en correspondance des surfaces des pavillons, malgré d'éventuelles différences de taille et d'orientation. L'ICP, ou *Iterative Closest Point* [17], a déjà montré son efficacité à aligner convenablement des surfaces de pavillon, dans une étude sur la variabilité des formes de la conque et du canal auditif [60]. Cet algorithme a pour objectif d'aligner deux surfaces définies par des nuages de points, en établissant de façon itérative une correspondance entre les deux jeux de points, puis en déterminant la transformation géométrique globale réduisant de façon optimale la distance totale entre les points ainsi appariés. Dans la forme classique de l'ICP, les transformations disponibles sont la composée d'une rotation et d'une translation. De telles transformations suffiraient pour compenser des différences d'orientation entre les pavillons, mais ne pourraient pas corriger des dif-

férences de taille. C'est pourquoi nous ajoutons l'homothétie comme degré de liberté supplémentaire, selon la formulation proposée par Umeyama *et al.*[251]. Les transformations disponibles pour l'alignement correspondent ainsi mathématiquement à des similitudes.

Soient deux ensembles de points décrivant les surfaces des pavillons à comparer, définis dans $\mathbb{R}^3$: $M = \{m_i\}_{i=1}^{N_m}$ est celui correspondant à un nouvel auditeur, et $P = \{p_i\}_{i=1}^{N_p}$ celui d'un sujet de la base de données, dont les HRTF sont donc connues. Aligner ces deux ensembles de points revient à déterminer la transformation par laquelle P se retrouve le plus proche de M. Ce problème revient mathématiquement à déterminer la rotation $\mathbf{R}$, la translation $\vec{t}$ et le facteur s, solutions du problème des moindres carrés suivant :

$$\min_{s,\mathbf{R},\vec{t}}(\sum_{i=1}^{N_p} \|(s\mathbf{R}\vec{p_i} + \vec{t}) - \vec{m}_j\|_2^2) \tag{5.2}$$

avec $\mathbf{R}^T\mathbf{R} = \mathbf{I}_3$, $\det(\mathbf{R}) = 1$, où $\mathbf{I}_3$ est la matrice identité. A chaque itération, l'algorithme comprend deux étapes :

- Etape 1 : on trouve les correspondances c entre les ensembles de points, après application de la transformation courante $(s_k, \mathbf{R}_k, \vec{t_k})$ sur l'un des pavillons :

$$c(i) = \underset{j\in(1,2,\ldots,N_m)}{\arg\min} (\|(s_k\mathbf{R}_k\vec{p_i} + \vec{t_k}) - \vec{m}_j\|_2^2) \tag{5.3}$$

- Etape 2 : on calcule la nouvelle transformation $(s^*, \mathbf{R}^*, \vec{t^*})$:

$$(s^*, \mathbf{R}^*, \vec{t^*}) = \underset{(s,\mathbf{R},\vec{t})}{\arg\min}(\sum_{i=1}^{N_p} \|s\mathbf{R}(s_k\mathbf{R}_k\vec{p_i} + \vec{t_k}) + \vec{t} - \vec{m}_{c(i)}\|_2^2) \tag{5.4}$$

  Puis on met à jour $s_{k+1}, \mathbf{R}_{k+1}, \vec{t}_{k+1}$

$$s_{k+1} = s^*s_k, \quad \mathbf{R}_{k+1} = \mathbf{R}^*\mathbf{R}_k, \quad \vec{t}_{k+1} = s^*\mathbf{R}^*\vec{t_k} + \vec{t^*} \tag{5.5}$$

L'expression de s^*, $\mathbf{R}^*$ et $\vec{t^*}$ est celle qui minimise la distance totale observée entre les points appariés selon c. Les calculs sont détaillés en Annexe C. Seuls les résultats utiles pour la mise en oeuvre de l'algorithme sont décrits ici.

Soient $\overrightarrow{p_i'} \triangleq s_k\,\mathbf{R}_k\,\vec{p}_i + \vec{t}_k$, $\vec{q}_i \triangleq \overrightarrow{p_i'} - \frac{1}{N_p}\sum_{j=1}^{N_p}\overrightarrow{p_j'}$ et $\vec{n}_i \triangleq \vec{m}_{c(i)} - \frac{1}{N_p}\sum_{j=1}^{N_p}\vec{m}_{c(j)}$. On calcule la matrice $\mathbf{H}$ (3×3) et sa décomposition en valeurs singulières (SVD) :

$$\mathbf{H} \triangleq \frac{1}{Np}\sum_{i=1}^{N_p}\vec{q}_i\vec{n}_i^T, \mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{5.6}$$

La mise à jour de la transformation est réalisée selon les expressions suivantes :

$$\mathbf{R}^* = \mathbf{U}\mathbf{\Xi}\mathbf{V}^T \tag{5.7}$$

$$s^* = \frac{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi})}{\frac{1}{N_p}\sum_{i=1}^{N_p} \parallel \vec{q}_i \parallel^2} \tag{5.8}$$

$$\vec{t}^* = \frac{1}{N_p}\sum_{i=1}^{N_p} \vec{m}_{c(i)} - \frac{1}{N_p}\sum_{i=1}^{N_p} s^*\mathbf{R}^*\vec{p}'_i \tag{5.9}$$

avec

$$\mathbf{\Xi} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{V})\det(\mathbf{U}) \end{pmatrix} \tag{5.10}$$

L'étape d'établissement de la correspondance consiste à trouver pour chaque point d'une surface, le point le plus proche appartenant à l'autre surface. La détermination exhaustive des distances par paires entre les deux ensembles de points est exclue, car elle entraînerait des coûts de calcul prohibitifs. C'est pourquoi on utilise une structure d'organisation des données adaptée, qui permet d'accéder plus rapidement au résultat. Il s'agit de la structure dite *kd-tree* pour *k-dimensional tree* [12] (cf. Annexe D). C'est un arbre binaire dont chaque feuille est à peu près à la même distance de la racine. Il est obtenu en découpant l'espace en une série de parallèlépipèdes, selon des plans parallèles aux axes du référentiel. La recherche dans un tel arbre du point le plus proche d'un point quelconque est un algorithme itératif simple et rapide. L'itération de l'ICP est arrêtée quand la distance entre les formes alignées atteint un seuil, ou bien n'évolue plus.

Finalement, les résultats utiles de cet alignement sont les valeurs finales de la rotation et du facteur d'homothétie : ils résument les différences d'orientation et de taille entre les deux pavillons [4].

5.3.2 Mise en œuvre

On représente figure 5.16 différentes itérations de l'algorithme de l'alignement entre deux pavillons (maillage de type I), matérialisés par des nuages de points. Le pavillon en bleu, qui subit les transformations, se rapproche progressivement du pavillon en rouge, qui lui reste intact. L'algorithme d'alignement se révèle assez efficace pour les surfaces de pavillons testées. On observe en effet une bonne mise en correspondance des cavités du pavillon. La conque semble jouer un rôle majeur dans cet alignement, tant sur l'orientation que sur le facteur de l'homothétie. Néanmoins,

4. La translation présente peu d'intérêt, car elle ne reflète que des différences entre la taille des têtes des sujets.

les résultats sont parfois différents selon le type de maillage. Pour analyser la qualité de l'alignement entre les différentes cavités, on définit une erreur dite locale, notée ε_{loc}, décrivant localement l'écart entre les surfaces des pavillons. On s'appuie pour cela sur le maillage points/arêtes/faces décrivant la surface des pavillons : pour chaque face d'un des pavillons, l'erreur locale est la distance moyenne entre chacun de ses sommets et la surface de l'autre pavillon.

On représente figures 5.17, 5.18, 5.19 et 5.20, les alignements obtenus pour les 15 premiers couples de pavillons considérées : il s'agit des pavillons gauches des 6 sujets de l'étude (symétrisés). Pour les couples n° 1, 2, 3, 4, 11, 12 et 15, l'erreur d'alignement ε_{loc} est assez similaire d'un type de maillage à l'autre, ce qui traduit le fait que les paramètres d'alignement sont proches. On observe des différences plus marquées entre les types de maillage pour les couples n° 5, 6, 8, 10, 13 et 14 : la présence ou non du rebord du pavillon influe nettement sur le résultat de l'alignement au niveau de la conque. C'est un effet inévitable lorsque les différences entre les formes de pavillons ne peuvent être réduites par une simple homothétie. Enfin, pour les couples n° 7 et 9, on observe pour les types I et III un échec de l'algorithme : les cavités des pavillons ne sont pas convenablement alignées. L'algorithme tombe dans un minimum local insatisfaisant, quelle que soit la position d'initialisation. Ce phénomène est observé uniquement sur les maillages de type I et III parmi les couples de pavillons non représentés (n°16 à 60). Darkner [60] a déjà pointé ce défaut potentiel de l'ICP, et a proposé une méthode d'alignement alternative, plus complexe à mettre en oeuvre, mais qui évite ces problèmes de minima locaux. Sa méthode inclut une étape préliminaire qui calcule la position d'initialisation optimale d'après des considérations probabilistes. La technique est totalement automatique et donc séduisante, mais il nous a semblé difficile d'y inclure le degré de liberté de l'homothétie, non considéré par l'auteur. Les problèmes d'alignement apparaissent toutefois assez rarement avec l'ICP dans notre expérience (4 couples sur 60), et ils ne touchent que deux types de maillage, c'est pourquoi l'algorithme est tout de même considéré comme valide, au moins dans cette étude exploratoire.

5.4 Transformations optimales des HRTF

5.4.1 Principe

On décrit ici les outils nécessaires à la détermination des paramètres optimaux de transformation (cf. Fig. 5.21). Comme dans l'étude de Maki et Furukawa, on considère le problème de la réduction des différences inter-individuelles entre deux jeux de HRTF, en utilisant conjointement comme transformations une rotation du système de coordonnées et une homothétie sur l'axe fréquentiel (*scaling* fréquentiel).

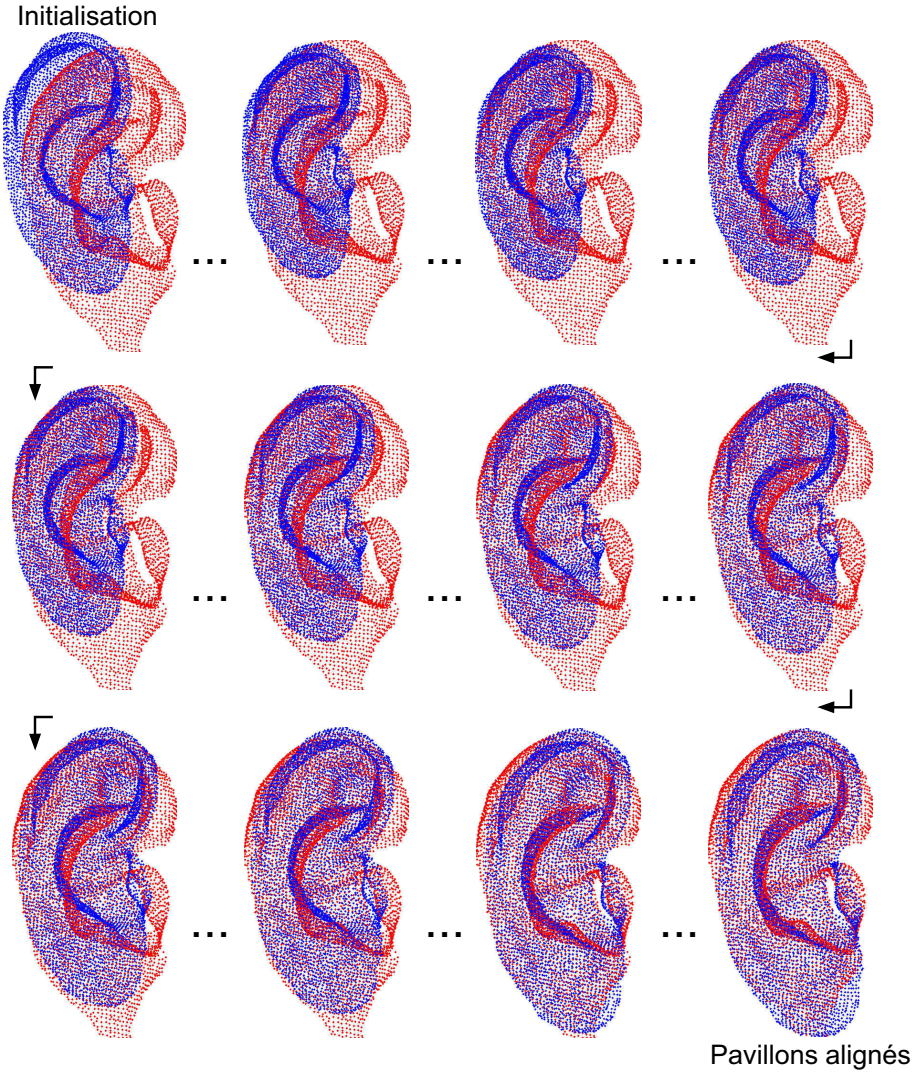


Figure 5.16 – Itérations de l'algorithme d'alignement de deux pavillons (maillage de type I). Le pavillon en bleu subit les transformations tandis que le pavillon en rouge reste intact. Le pavillon en bleu est ici artificiellement rétréci avant l'initialisation pour mieux illustrer l'évolution de l'algorithme.

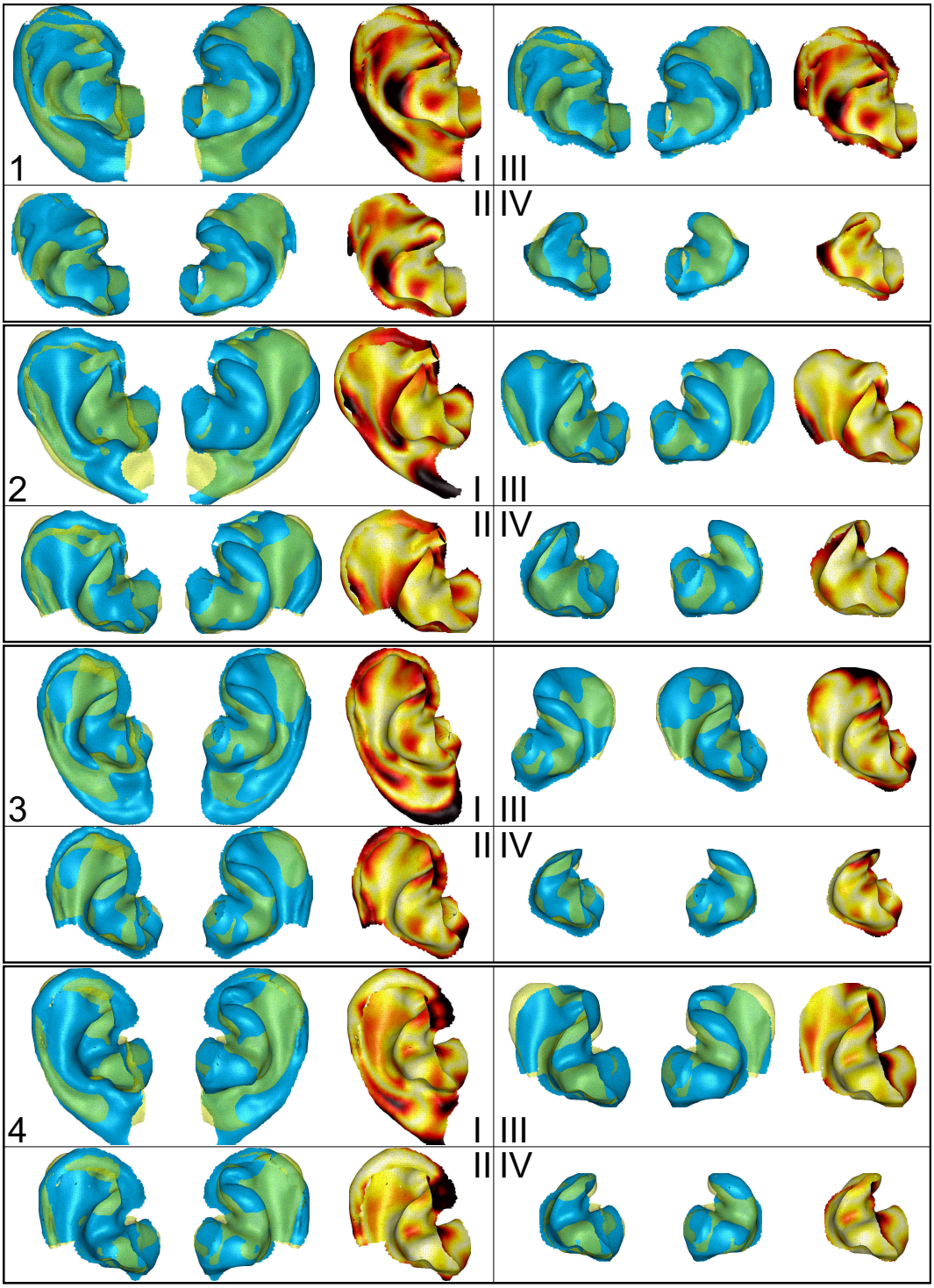


Figure 5.17 – Résultat de l'alignement morphologique, pour les couples de pavillons n°1 à 4 (oreilles gauches, symétrisées). Pour chaque type de maillage, on représente les deux surfaces entrelacées des pavillons (l'un en bleu opaque, l'autre en jaune translucide) vus de face, et de derrière, ainsi que l'erreur locale d'alignement ε_{loc} sur la surface du pavillon restant fixe, dont la valeur absolue est codée en couleur (faible sur les zones claires, élevée sur les zones sombres).

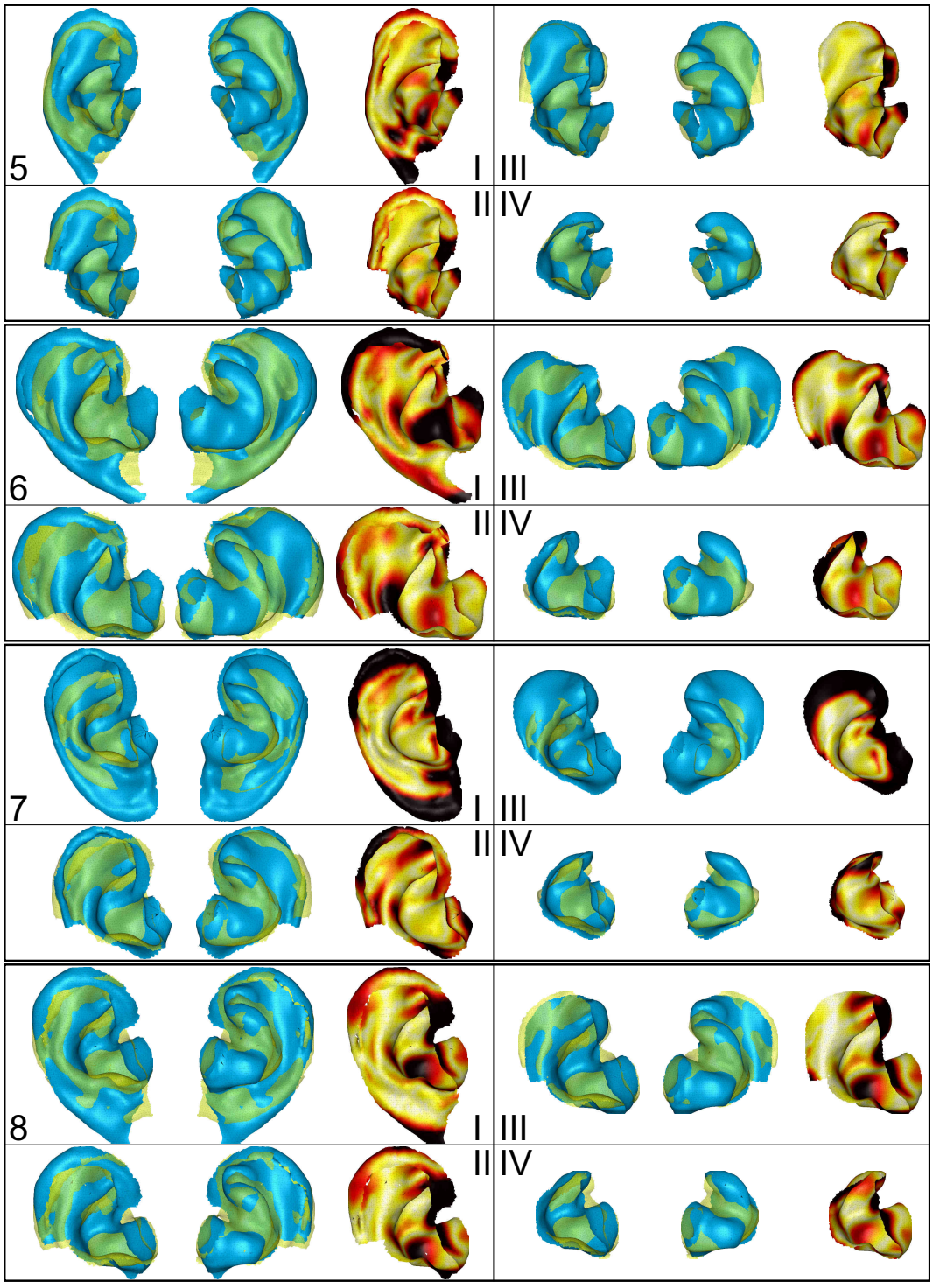


Figure 5.18 – Résultat de l'alignement morphologique, pour les couples de pavillons n°5 à 8 (oreilles gauches symétrisées, cf. Fig. 5.17 pour les détails).

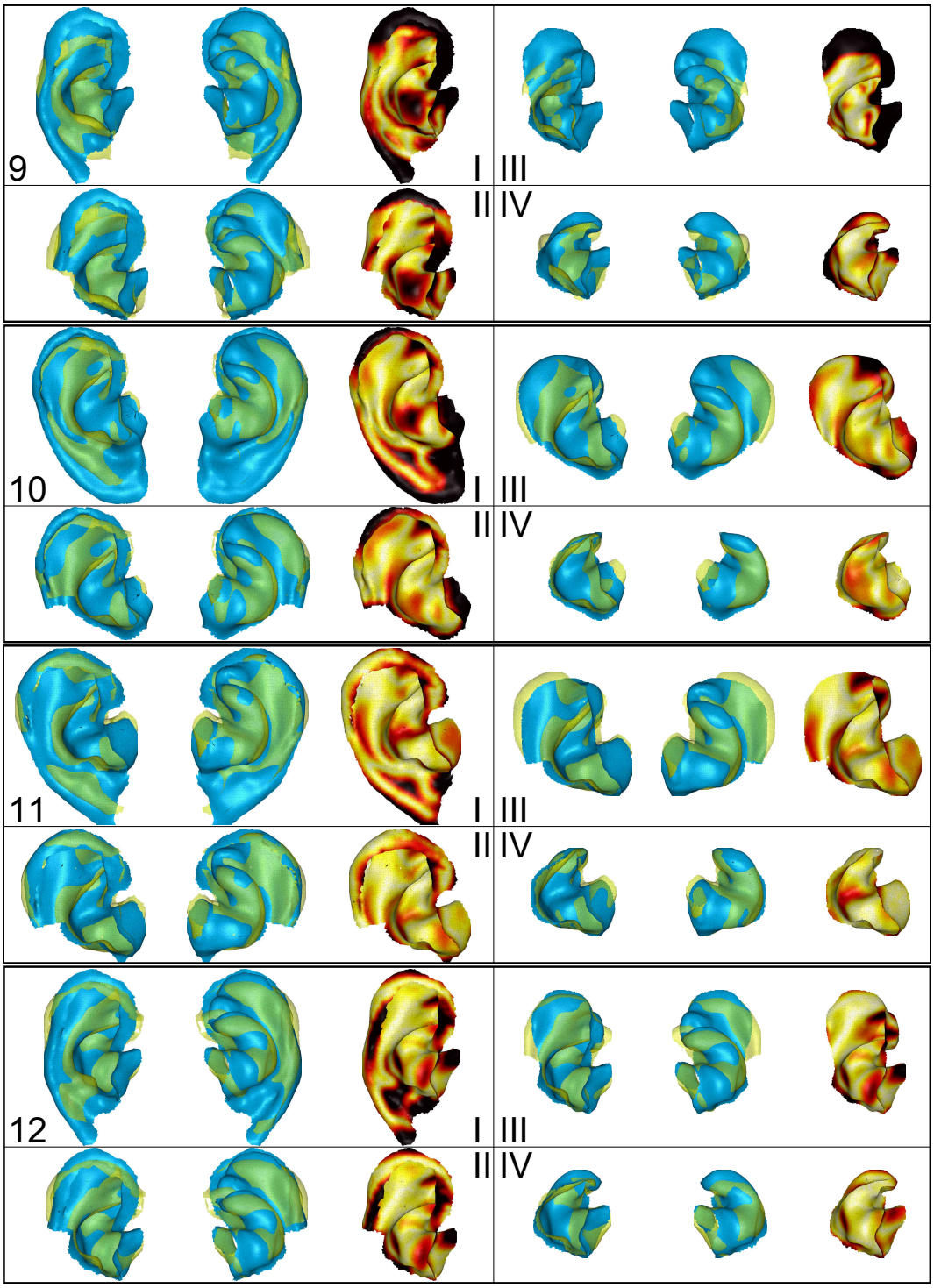


Figure 5.19 – Résultat de l'alignement morphologique, pour les couples de pavillons n°9 à 12 (oreilles gauches symétrisées, cf. Fig. 5.17 pour les détails).

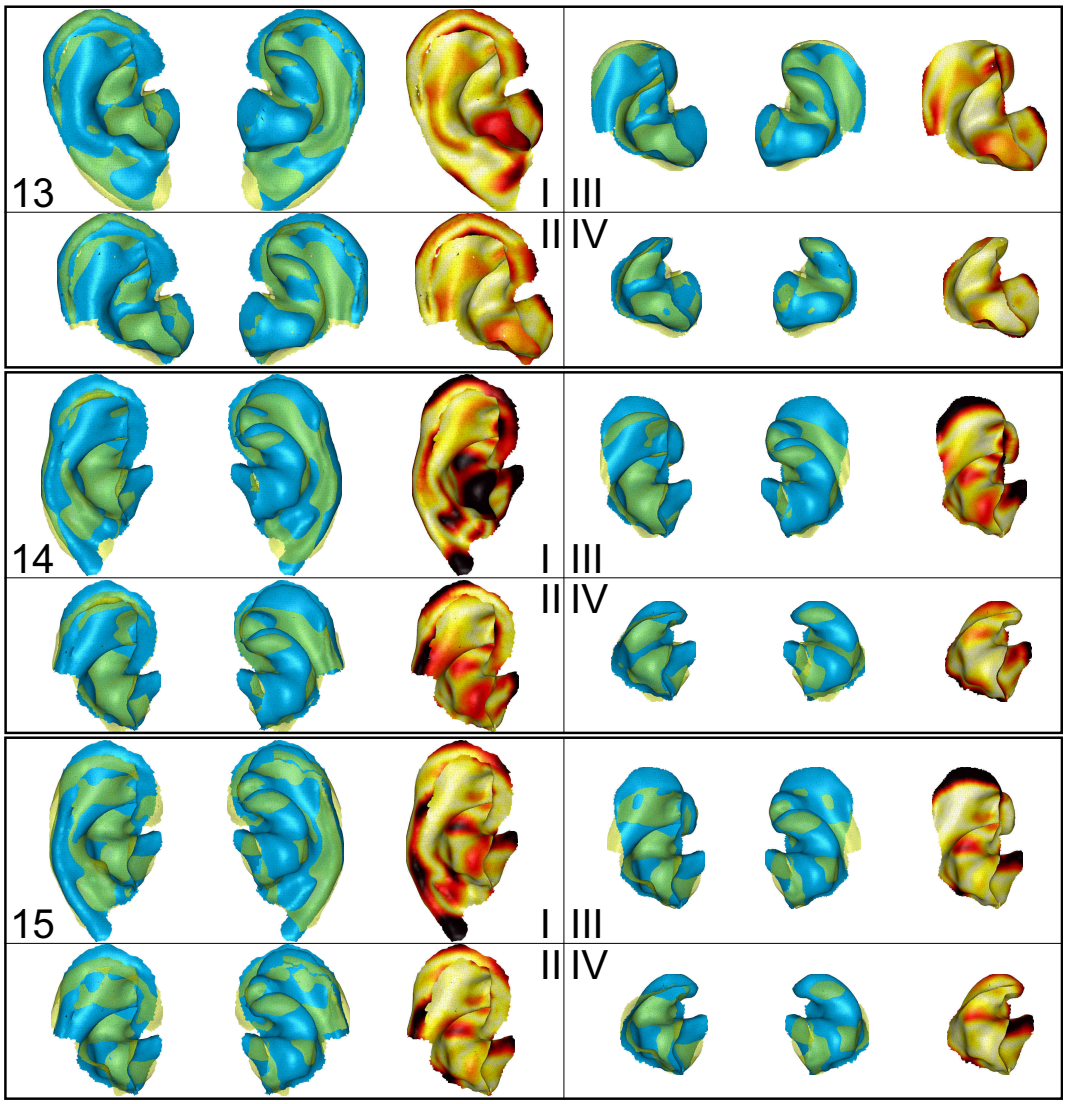


Figure 5.20 – Résultat de l'alignement morphologique, pour les couples de pavillons n°13 à 15 (oreilles gauches symétrisées, cf. Fig. 5.17 pour les détails).

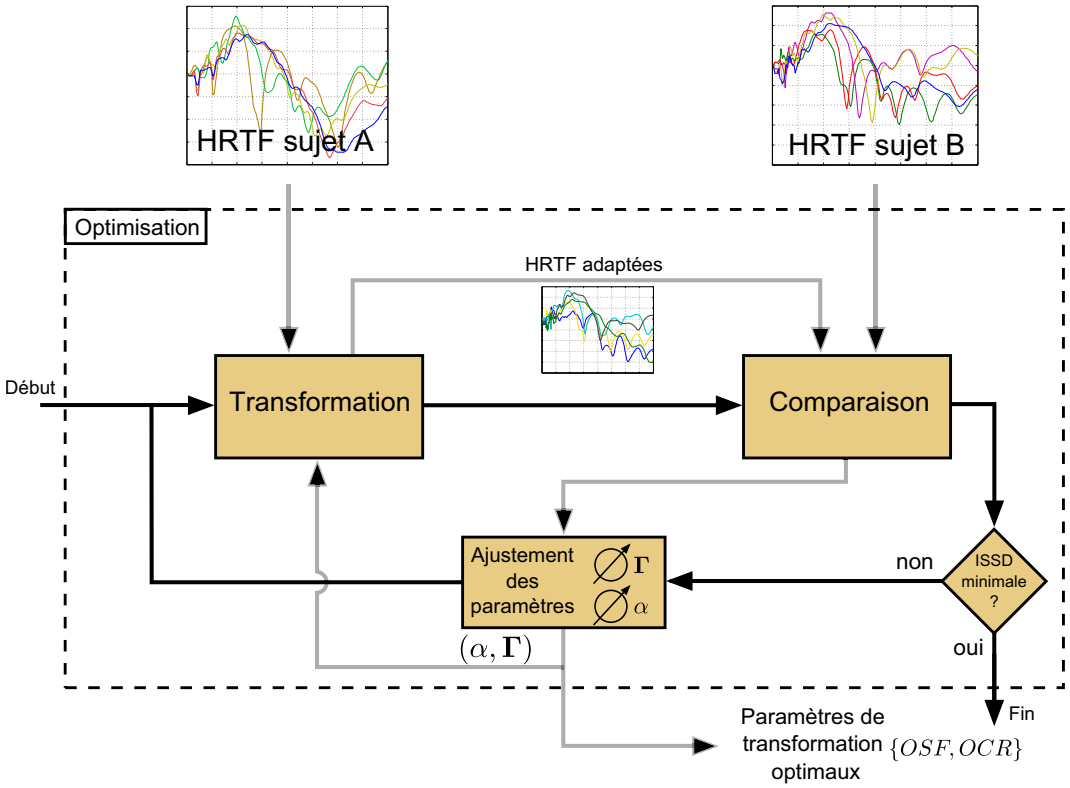


Figure 5.21 – Par minimisation de l'ISSD, on détermine les paramètres de transformation "optimaux", ceux permettant de rapprocher au mieux les HRTF du sujet A de celles du sujet B (les flèches grises représentent des flux d'informations, tandis que les flèches noires représentent l'enchaînement séquentiel des éléments de l'optimisation).

On utilise les dénominations proposées par Middlebrooks, et Maki et Furukawa [148, 169].

Directional Transfer Functions

Tout d'abord, en place des HRTF, on considère les *Directional Transfer Functions* (DTFs) [174], contenant uniquement la part des HRTF qui présente une dépendance directionnelle : les DTF sont les spectres d'amplitude en dB des HRTF égalisées par le champ diffus. De plus, on tient compte de la résolution fréquentielle limitée du système auditif en faisant passer les HRTF dans un banc de filtres gammatone, comme décrit en 2.5. Soient H les HRTF mesurées dans les directions $\{\chi_n\}_{n=1,N}$, et $\tilde{H}$ les HRTF ainsi filtrées. La DTF correspondant à une direction χ_k est définie, en fonction de la fréquence ν, selon la relation suivante :

$$DTF_{\chi_k}(\nu) = \frac{20.log_{10}(|\tilde{H}_{\chi_k}(\nu)|)}{\frac{1}{N}\sum_{n=1}^{N} 20.log_{10}(|\tilde{H}_{\chi_n}(\nu)|)} \tag{5.11}$$

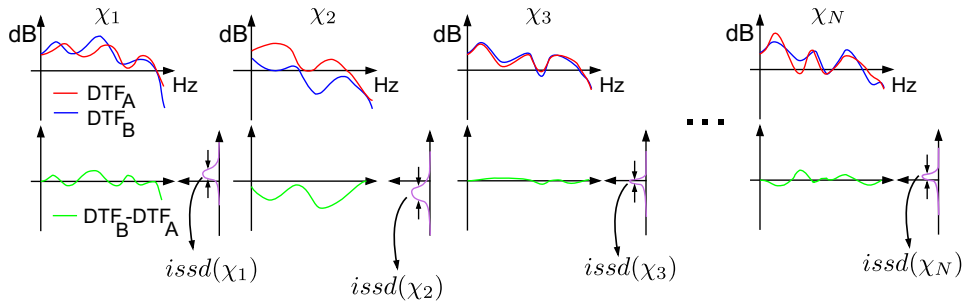
Inter-Subject Spectral Difference

On résume la différence globale entre les jeux de DTF de deux sujets [5] en calculant l'ISSD (*Inter-Subject Spectral Difference*), définie comme suit (cf. Fig. 5.22). Pour chaque direction, la DTF d'un sujet est soustraite, fréquence par fréquence, de la DTF de l'autre sujet, à la même direction. La variance de cette différence est calculée sur la bande de fréquence [4 kHz ; 13 kHz], sur laquelle se trouvent les IS d'intérêt. Pour une direction χ_n, cette variance exprimée en dB^2 est appelée ISSD directionnelle (notée $issd(\chi_n)$, cf. figure 5.22). L'ISSD globale est finalement définie comme la moyenne, pour toutes les directions disponibles, des ISSD directionnelles. Cette mesure de dissimilarité s'apparente à l'écart-type utilisé avec succès par Langendijk et Bronkhorst dans leur modèle de localisation auditive [131]. Middlebrooks a validé perceptivement le choix de l'ISSD [170] : l'adaptation de HRTF par une réduction de l'ISSD mène en synthèse binaurale à une diminution des confusions avant/arrière, et une amélioration de la perception en élévation (cf. 4.2.4). L'ISSD est donc une mesure objective corrélée à une distance perceptive.

Paramètres optimaux

Le facteur de l'homothétie optimale ou *Optimal Scale Factor* (OSF) est par définition le facteur de *scaling* fréquentiel, appliqué pour toutes les directions à un des jeux de DTF, qui minimise l'ISSD par rapport à un autre jeu de DTF [169]. La rotation optimale du sytème de coordonnées ou *Optimal Coordinate Rotation* (OCR)

5. L'ISSD est en fait calculée entre le jeu de DTF correspondant à une oreille d'un sujet et celui correspondant à une oreille d'un autre sujet.



$$ISSD_{A,B} = \frac{1}{N}\sum_{n=1}^{N} issd(\chi_n)$$

Figure 5.22 – Définition de l'ISSD entre les DTF de deux sujets A et B. Les ISSD directionnelles, obtenues pour chaque direction χ_n sont notées $issd(\chi_n)$.

est définie comme la rotation à appliquer au système de coordonnées d'un jeu de DTF pour minimiser de façon optimale l'ISSD par rapport à un autre jeu de DTF [148]. On cherche à optimiser simultanément les paramètres des deux transformations considérées - la rotation Γ et le facteur de *scaling* α - afin d'obtenir l'OCR et l'OSF.

Spatial Frequency Response Surfaces

En pratique, on considère sur tout l'espace les valeurs des DTF indépendamment pour chaque bin fréquentiel : on appelle *Spatial Frequency Response Surfaces* (SFRS) les fonctions de directivité résultantes [54], que l'on a représentées sur les figures 5.1, 5.2, 5.3, et 5.4. Pour une fréquence ν, et une direction χ_n sur la sphère, on obtient :

$$SFRS_\nu(\chi_n) = DTF_{\chi_n}(\nu) \tag{5.12}$$

Les SFRS sont décomposées sur une base d'harmoniques sphériques complexes, de façon à obtenir une représentation fonctionnelle compacte de ces données, nécessaire pour la recherche des paramètres optimaux [66]. On procède comme suit. Soit f une SFRS : c'est une fonction définie sur la sphère, et de carré intégrable ($f \in L^2(S^2)$). Supposons que cette fonction soit à bande limitée et de bande B, on peut alors décomposer f en une somme finie d'harmoniques sphériques $Y_l^m(\chi)$, où χ désigne une direction de la sphère, sous la forme :

$$f(\chi) = \sum_{l=0}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\chi)$$

En pratique, les coefficients $\hat{f}_l^m$ de la décomposition sont estimés d'après les valeurs prises par f sur N directions choisies de façon non régulière sur la sphère $\{\chi_n\}_{n=1,\dots,N}$

(les directions de mesure des HRTF) :

$$f(\chi_n) = \sum_{l=0}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\chi_n)$$

ce qui peut s'écrire sous forme matricielle :

$$\begin{aligned} \mathbf{f} &= \mathbf{Y}.\mathbf{C} \\ \mathbf{Y} &= \{Y_l^m(\chi_i)\}_{N\times B^2} \\ \mathbf{C} &= \{\hat{f}_l^m\}_{B^2\times 1} \\ \mathbf{f} &= \{f(\chi_i)\}_{N\times 1} \end{aligned}$$

Calculer la décomposition en harmoniques sphériques revient à déterminer le vecteur $\mathbf{C}$ qui minimise l'erreur entre les valeurs connues de $f(\chi_i)$ et celles estimées par $\mathbf{Y}.\mathbf{C}$:

$$\epsilon = (\mathbf{f} - \mathbf{Y}.\mathbf{C})^2 \tag{5.13}$$

ce qui est résolu par calcul du pseudo-inverse :

$$\mathbf{C} = (\mathbf{Y}^t.\mathbf{Y})^{-1}.\mathbf{Y}^t.\mathbf{f} \tag{5.14}$$

si $N \geq B^2$ (problème surdéterminé). Pour régulariser la décomposition, il est nécessaire de connaître les valeurs de la fonction de façon homogène sur toute la sphère. Classiquement les HRTF ne sont pas mesurées en dessous d'une élévation limite. Les données correspondant à quelques directions de cette calotte inférieure de la sphère sont donc obtenues au préalable par interpolation. Nous utilisons l'interpolation par spline de type plaque mince sur la sphère (*Spherical Thin Plate Spline* ou STPS) (cf. Annexe B).

Recherche de la rotation optimale du système de coordonnées

Pour un facteur de *scaling* α fixé[6], on utilise la descente du gradient pour faire évoluer la rotation Γ du système de coordonnées vers la valeur qui minimise l'ISSD entre les deux jeux de DTF. En appliquant cet algorithme pour une série de valeurs de α, on obtient l'OSF et l'OCR, qui sont respectivement les valeurs de α et de Γ pour lesquelles l'ISSD atteint un minimum global. L'algorithme de descente du gradient sur le groupe des rotations $SO(3)$ a été formalisé par Stein *et al.*[244] et Chirikjian *et al.* [56]. On cherche à minimiser une fonction $g_\alpha(\Gamma)$, qui dans notre

6. Contrairement à $SO(3)$, l'espace de recherche pour la valeur optimale de α est unidimensionnel, c'est pourquoi, en pratique, il est plus simple par exemple d'échantillonner par pas de 0.01 l'intervalle $[0.75\,;\,1.25]$, et de répéter la recherche de la rotation optimale autant de fois qu'il y a de valeurs de α considérées.

cas est la fonction continue décrivant l'ISSD pour une valeur fixée de α. A chaque itération, l'algorithme consiste à trouver la direction qui réduit au maximum la fonction considérée, en calculant son gradient. Dans le groupe $SO(3)$, ce sont les éléments de la base de l'algèbre de Lie de $SO(3)$ qui permettent de calculer des déplacements infinitésimaux à partir d'un point de $SO(3)$ [7] :

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} X_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} X_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Les dérivées directionnelles de g_α sont définies par :

$$X_i^R g_\alpha(\Gamma) = \frac{d}{dt}(g_\alpha(\Gamma \circ e^{tX_i}))|_{t=0} \quad i = 1, 2, 3 \tag{5.15}$$

Pour le calcul, on utilise les gradients approchés : les dérivées à droite $X_i^R g_\alpha(\Gamma), i = (1, 2, 3)$ sont remplacées par leurs approximations aux différences finies x_i :

$$x_i = \frac{g_\alpha(\Gamma \circ e^{tX_i}) - g_\alpha(\Gamma \circ e^{-tX_i})}{2t} \quad i = 1, 2, 3 \tag{5.16}$$

L'ensemble de ces dérivées directionnelles est un vecteur gradient ayant la direction de la plus grande pente, mais pointant dans le sens ascendant. On considère donc l'opposé de ce vecteur, et on met à jour l'élément courant Γ de $SO(3)$ comme suit :

$$\begin{aligned} \Gamma &\leftarrow \Gamma \circ exp\{-\epsilon \sum_{i=1}^{3} X_i [X_i^R f_\alpha(\Gamma)]\} \\ &\approx \Gamma \circ exp\left\{-\epsilon \begin{pmatrix} 0 & -x_1 & x_2 \\ x_1 & 0 & -x_3 \\ -x_2 & x_3 & 0 \end{pmatrix}\right\} \end{aligned}$$

où ϵ est un pas élémentaire. Cet algorithme mène au plus proche minimum local de la fonction g_α. C'est pourquoi, il convient de recommencer plusieurs fois avec différentes valeurs initiales de Γ, pour confirmer la validité de son résultat. L'intérêt de cet algorithme est le fait qu'il permet de converger précisément vers la solution sans calculer l'ISSD sur un échantillonnage fin et complet du groupe des rotations $SO(3)$.

Calcul de l'ISSD

Pour chaque couple de paramètres de transformation (α, Γ), le calcul de l'ISSD nécessite une stricte correspondance spatiale et fréquentielle entre les DTF transformées d'un sujet et les DTF intactes d'un autre sujet. On obtient les SFRS après

7. Un point de $SO(3)$ est une rotation.

rotation du système de coordonnées, aux directions sur lesquelles les HRTF ont été mesurées, en faisant subir à cet échantillonnage spatial une rotation Γ^{-1}, puis en évaluant les SFRS sur l'échantillonnage résultant, grâce à la décomposition en harmoniques sphériques. Le *scaling* fréquentiel de facteur α est lui réalisé en réorganisant les SFRS sur l'axe fréquentiel : une SFRS correspondant à l'origine à la fréquence ν est associée à la fréquence $\alpha.\nu$. On obtient finalement les valeurs des DTF sur les bins fréquentiels initiaux par interpolation spline du résultat sur l'axe fréquentiel. On peut alors calculer l'ISSD correspondant à la transformation de paramètres (α, Γ).

5.4.2 Mise en œuvre

On met en œuvre, grâce aux outils décrits en 5.4, le calcul des paramètres optimaux des transformations des HRTF. Pour évaluer l'apport du nouveau degré de liberté que constitue la rotation du système de coordonnées par rapport à l'état de l'art [169], on considère le problème de la réduction des différences inter-individuelles entre des jeux de DTF, pour trois types de transformation :

- *scaling* fréquentiel (ou *Scaling*)
  La seule transformation considérée est l'homothétie des DTF sur l'axe fréquentiel, comme proposé par Middlebrooks [169], c'est-à-dire qu'on fixe la rotation Γ à $\mathbf{I_3}$ (matrice identité 3×3). C'est donc pour des DTF mesurées aux mêmes directions que le rapprochement est réalisé entre les deux jeux de données. La recherche du facteur de *scaling* optimal est réalisée sur un échantillonnage fin de l'intervalle $[0.75\,;\,1.25]$. On note OSF_S le facteur de *scaling* optimal correspondant à cette configuration.
- rotation du système de coordonnées (ou *Rotation*)
  La seule transformation considérée est la rotation du système de coordonnées. Cela revient à fixer le facteur de *scaling* α à la valeur 1, et à optimiser la transformation en explorant seulement l'espace des rotations. La rotation optimale obtenue est notée OCR_R.
- *scaling* fréquentiel et rotation du système de coordonnées (ou *Scaling&Rotation*)
  On considère conjointement la rotation du système de coordonnées et le *scaling* fréquentiel. Pour chaque valeur de α fixée sur l'intervalle $[0.75\,;\,1.25]$, on recherche la rotation Γ offrant l'ISSD minimale. Les paramètres α et Γ permettant d'atteindre le minimum global sont respectivement notés OSF_{SR} et OCR_{SR}.

On représente figure 5.23 les valeurs de l'ISSD après adaptation pour chacune de ces expériences. en fonction de l'ISSD initiale, celle observée sans aucune adaptation, qui correspond au cas où $\alpha = 1$ et $\Gamma = \mathbf{I_3}$. On remarque que c'est la configuration *Scaling&Rotation* qui permet de réduire au mieux l'ISSD. A première vue, ce n'est

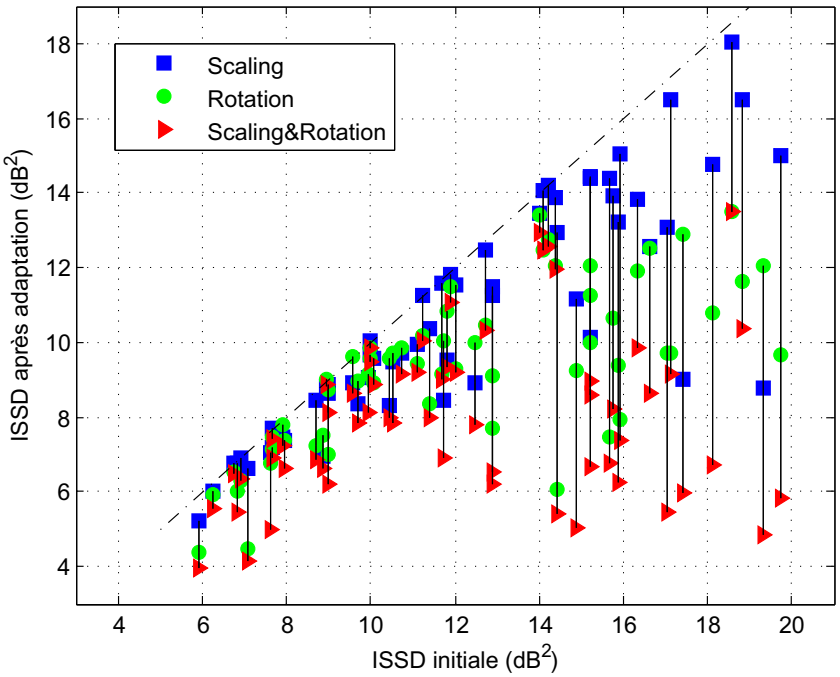


Figure 5.23 – ISSD optimale obtenue après adaptation selon chaque configuration en fonction de l'ISSD initiale, pour les 60 couples d'oreilles considérées (*Scaling&Rotation* (triangles rouges), *Scaling* (carrés bleus), *Rotation* (cercles verts)). Les résultats concernant un même couple d'oreilles sont reliés par une ligne noire.

pas une surprise, car c'est la configuration qui offre le plus de degrés de liberté. Néanmoins, l'ampleur de la réduction est un résultat intéressant : l'apport de la rotation est d'autant plus net pour les couples présentant des différences initiales élevées. Il est également à noter que la configuration *Rotation* offre généralement de meilleurs résultats que la configuration *Scaling*, ce qui prouve l'intérêt en soi de ce degré de liberté. On représente figure 5.24 un exemple de résultat de cette adaptation. Les caractéristiques locales des DTF sont effectivement mieux alignées dans la configuration *Scaling&Rotation*, ce qui illustre le résultat précédent.

5.5 Mise au point de la méthode d'individualisation et première évaluation

5.5.1 Sélection des paramètres morphologiques d'intérêt

L'intérêt de la méthode d'individualisation proposée réside dans la capacité des paramètres d'alignement morphologique à prédire les paramètres optimaux des transformations à appliquer à un jeu de HRTF pour les adapter à un nouvel auditeur (étape *b* figure 5.5). Afin d'établir le lien entre l'alignement morphologique des pa-

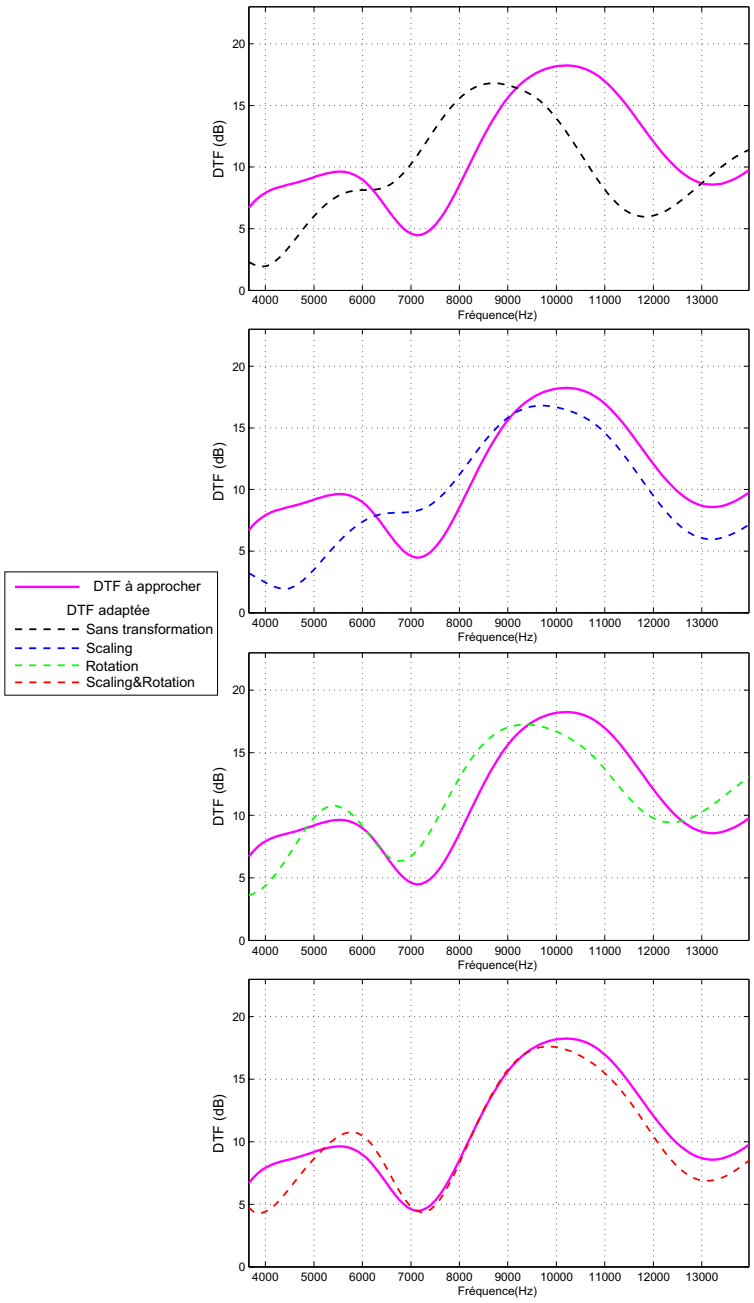


Figure 5.24 – Illustration de l'adaptation des DTF pour les différentes configurations considérées. De haut en bas : sans transformation, *Scaling*, *Rotation*, *Scaling&Rotation*. Les DTF de l'oreille droite du sujet n°1 de la base d'Orange Labs sont transformées pour se rapprocher des DTF de l'oreille droite du sujet n°5. La DTF en magenta, appartenant au sujet n°5 de la même base, est la cible à atteindre, et correspond à l'azimut 270°, et l'élévation -33.75° (système polaire vertical, cf. Fig. 1). Les autres DTF sont celles du sujet n°1 qui lui sont au mieux adaptées.

villons et les transformations nécessaires des HRTF, on étudie donc la corrélation entre leurs paramètres respectifs. Les opérations d'alignement morphologique (cf. 5.3) et de réduction de l'ISSD (cf. 5.4) mènent toutes deux à la définition d'un facteur d'homothétie (respectivement s et OSF), et à celle d'une rotation (respectivement $\mathbf{R}$ et OCR). Chaque rotation peut être décomposée en 3 rotations, autour des axes X, Y, and Z, selon les angles appelés respectivement roll θ, pitch ψ et yaw ϕ (cf. Fig. 5.14). Pour chaque couple de pavillons considérée, on combine ces paramètres en deux quadruplets $(\tilde{s}_{sig}, \tilde{\theta}_{sig}, \tilde{\psi}_{sig}, \tilde{\phi}_{sig})$, et $(s_{morph}, \theta_{morph}, \psi_{morph}, \phi_{morph})$: l'indice $morph$ correspond aux paramètres issus de la comparaison morphologique, l'indice sig correspondant aux paramètres des transformations signal à appliquer aux HRTF, et le symbole ˜ indique que ce sont les paramètres optimaux.

$$\begin{aligned} \tilde{s}_{sig} &= OSF_{SR} \\ s_{morph} &= s \\ \{\tilde{\theta}_{sig}, \tilde{\psi}_{sig}, \tilde{\phi}_{sig}\}_{XYZ} &= OCR_{SR} \\ \{\theta_{morph}, \psi_{morph}, \phi_{morph}\}_{XYZ} &= \mathbf{R} \end{aligned}$$

Une homothétie de facteur unitaire correspondant à une transformation nulle, on considère le logarithme des facteurs d'homothétie $\tilde{s}_{sig}$ et s_{morph}. On représente figures 5.25, 5.26, 5.27 et 5.28 les corrélations entre les paramètres signal et les paramètres morphologiques, pour les 4 types de maillage considérés. On observe généralement une dispersion importante en ce qui concerne les angles de rotation, mais quelques couples de paramètres se distinguent et révèlent une corrélation plus élevée, quel que soit le type de maillage considéré : il s'agit des couples $(\theta_{morph}, \tilde{\theta}_{sig})$, $(\theta_{morph}, \tilde{\psi}_{sig})$, et $(\varphi_{morph}, \tilde{\varphi}_{sig})$. La corrélation entre s_{morph} et $\tilde{s}_{sig}$ est par contre plus variable d'un type de maillage à l'autre. Il apparaît qu'on pourrait avantageusement combiner les résultats obtenus pour plusieurs types de maillage, car aucun ne se distingue nettement par des corrélation élevées pour tous les couples de paramètres à la fois. Cependant, il est préférable pour des raisons pratiques de n'avoir à considérer qu'un seul des 4 types de maillage. On choisit donc d'utiliser dans la suite de l'étude le maillage de type II, car c'est celui qui offre la meilleure corrélation entre les facteurs d'homothétie s_{morph} et $\tilde{s}_{sig}$ (0.79). De plus, l'ICP ne montre pas de problème d'initialisation avec ce type de maillage, tout au moins sur les cas testés dans cette expérience. Finalement on ne retient que les couples de paramètres cités précédemment : $(\theta_{morph}, \tilde{\theta}_{sig})$, $(\theta_{morph}, \tilde{\psi}_{sig})$, $(\varphi_{morph}, \tilde{\varphi}_{sig})$et $(s_{morph}, \tilde{s}_{sig})$, car ils offrent la meilleure corrélation, et impliquent chacun des paramètres signal qu'il s'agit de déterminer pour adapter les HRTF.

Ainsi, pour l'étape de prédiction morphologique des paramètres signal appropriés à appliquer à un jeu de HRTF, on propose de se baser sur une régression entre les paramètres des couples retenus : s_{morph} et $\tilde{s}_{sig}$, θ_{morph} et $\tilde{\theta}_{sig}$, θ_{morph} et $\tilde{\psi}_{sig}$, et φ_{morph}

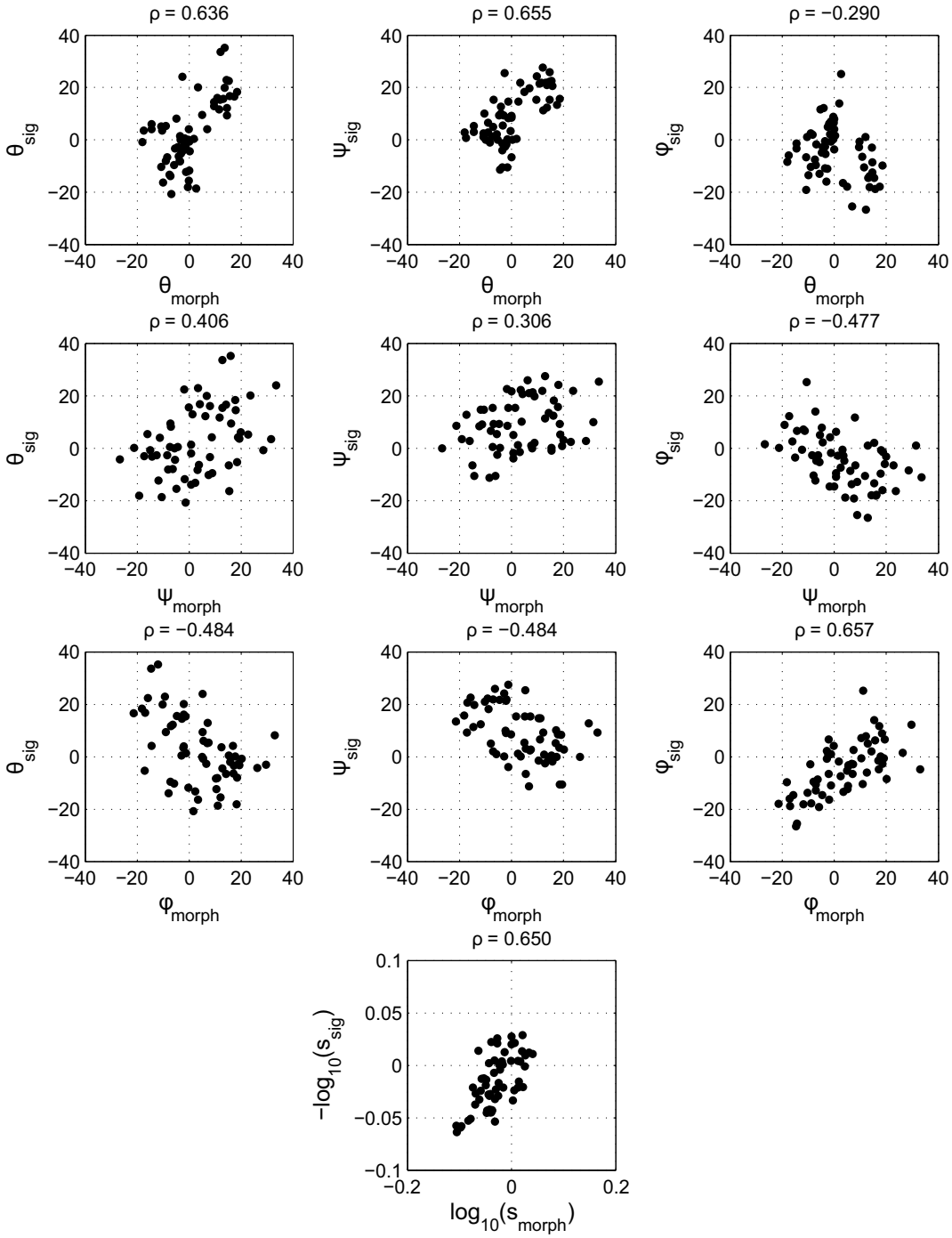


Figure 5.25 – Corrélations ρ observées entre les paramètres morphologiques et les paramètres signal, pour les maillages de type I.

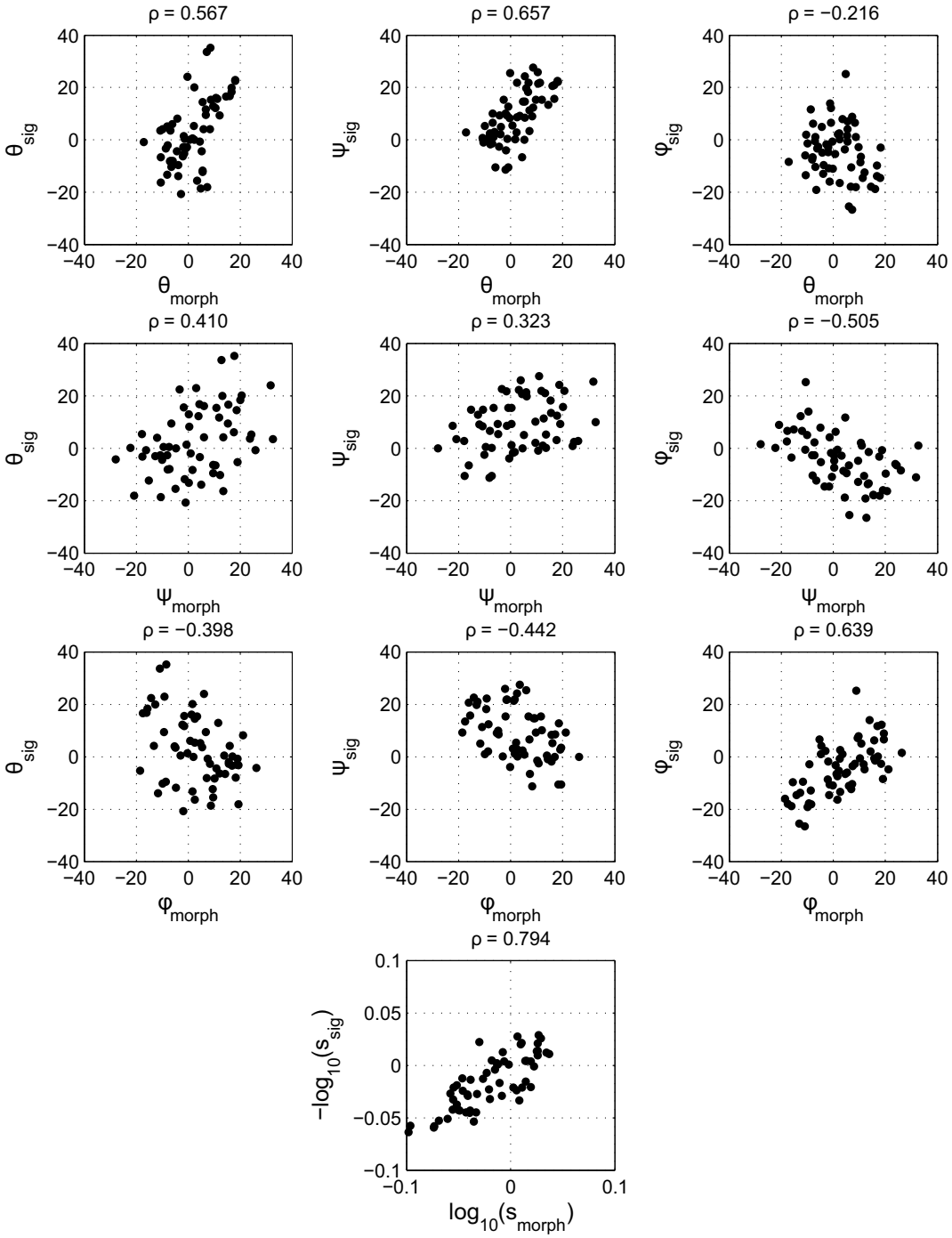


Figure 5.26 – Corrélations ρ observées entre les paramètres morphologiques et les paramètres signal, pour les maillages de type II.

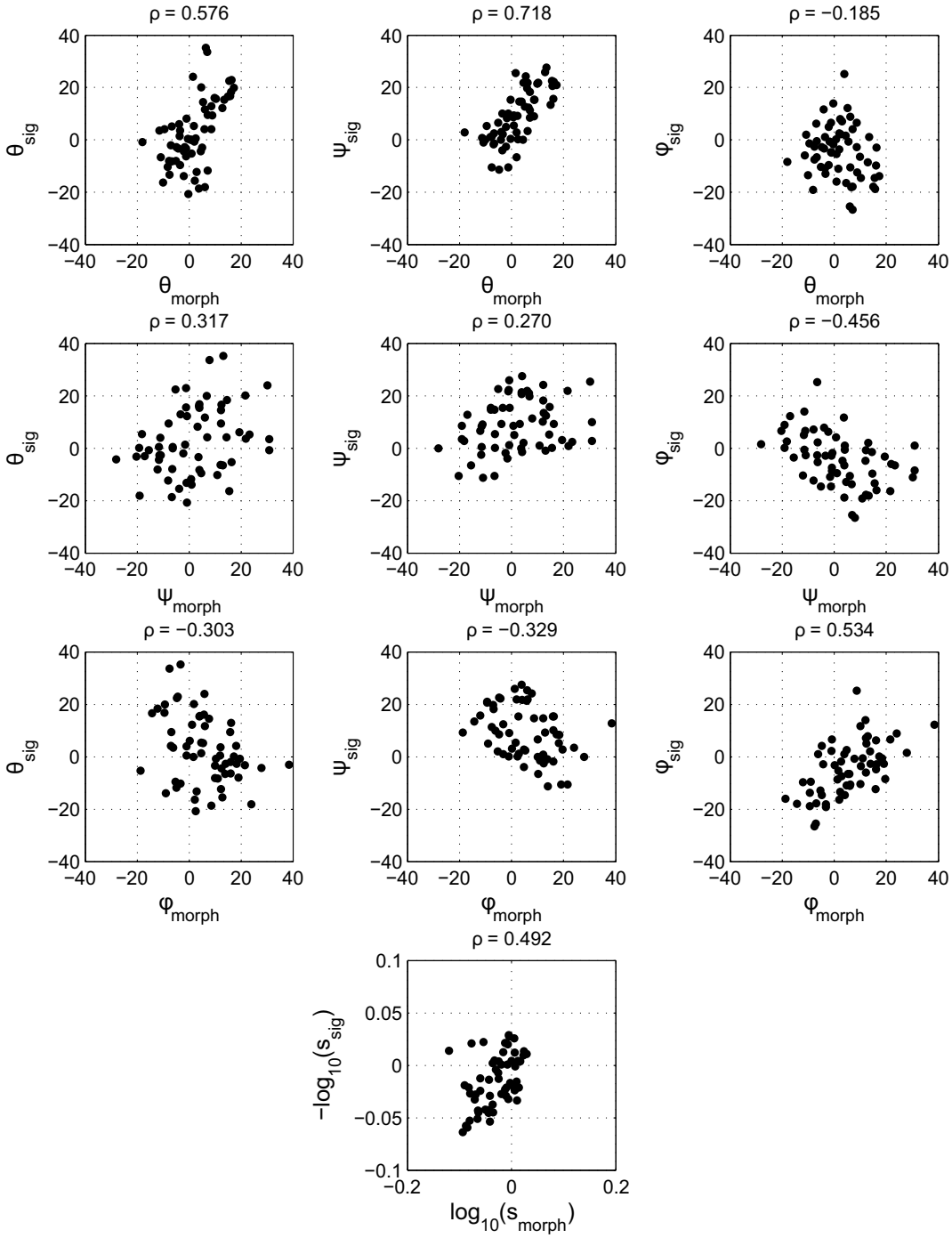


Figure 5.27 – Corrélations ρ observées entre les paramètres morphologiques et les paramètres signal, pour les maillages de type III.

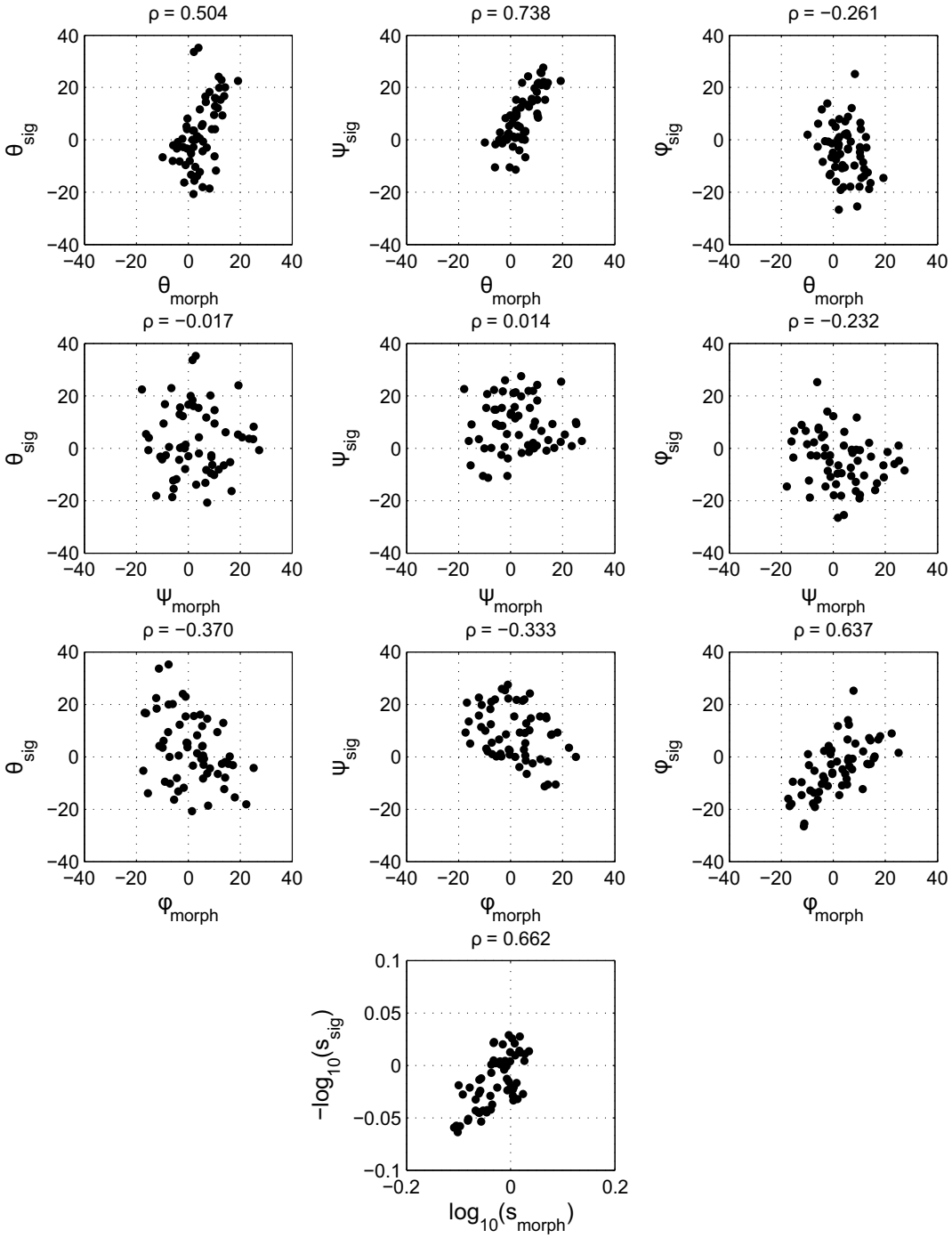


Figure 5.28 – Corrélations ρ observées entre les paramètres morphologiques et les paramètres signal, pour les maillages de type IV.

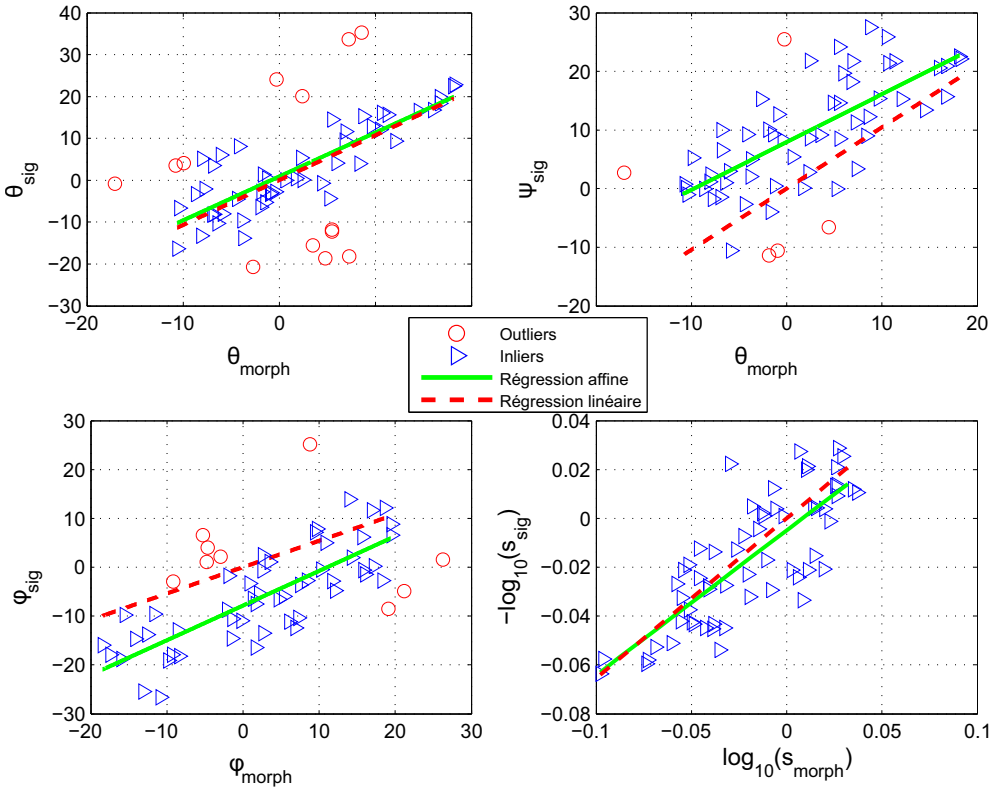


Figure 5.29 – Application de l'algorithme RANSAC sur les couples de paramètres retenus. Une classification est obtenue entre *inliers* (triangles bleus) et *outliers* (cercles rouges) sont détectés avant. Les droites représentent les résultats des régressions affines et linéaires sur les *inliers* (rsp. en vert et en pointillés rouges).

et $\tilde{\varphi}_{sig}$. Il apparaît cependant que les corrélations ne sont pas très élevées entre les angles. Ce phénomène est probablement inévitable au regard du nombre relativement faible de sujets considérés dans l'étude. La présence de points marginaux affecte la lisibilité des résultats, et risque également de rendre inefficace l'étape de prédiction morphologique. On choisit donc d'appliquer un algorithme dont le but est d'éliminer les points marginaux, et ainsi de "nettoyer" nos données. Cet algorithme est appelé RANSAC [69], pour *RANdom SAmple Consensus* : il est fondé sur l'hypothèse que le jeu de données est constitué d'un ensemble de points distribués selon un certain modèle (les *inliers*), et de points marginaux, qui ne correspondent pas au modèle (les *outliers*). Ces point marginaux peuvent notamment être issus d'un bruit dans l'estimation des valeurs physiques qui constituent leurs coordonnées, ou bien être le fruit d'erreurs. Dans notre cas, on peut supposer qu'il s'agit de couples d'oreilles intrinsèquement dissemblables, et dont les HRTF présentent des différences irréductibles par le jeu de transformations proposé. Etant donné un ensemble de *inliers*, même en quantité limitée et minoritaire, l'algorithme RANSAC permet d'estimer les paramètres d'un modèle approprié. On décrit l'algorithme en Annexe E. Comme aucune étude antérieure ne permet d'avoir du recul sur les données traitées dans cette expérience, on règle les paramètres de l'algorithme de façon empirique. Faisant l'hypothèse d'une relation affine entre les paramètres de chaque couple, on obtient la classification des données en *inliers* et *outliers*, et la régression représentées figure 5.29. On y trace également les droites matérialisant la relation affine obtenue par régression, et les droites obtenues dans l'hypothèse d'une relation purement linéaire (ordonnée à l'origine nulle).

On obtient sans surprise une meilleure adéquation avec les données dans l'hypothèse d'une relation affine, car les points retenus par le RANSAC ont été choisis selon cette hypothèse. Cette observation souligne néanmoins que le degré de liberté que constitue l'ordonnée à l'origine dans le cas affine a été avantageusement utilisé, et il semble bien dicté par la distribution des données. Il y a donc un biais dans les relations entre paramètres signal et paramètres morphologiques, qui est notamment très net pour les couples $(\varphi_{morph}, \tilde{\varphi}_{sig})$ et $(\psi_{morph}, \tilde{\theta}_{sig})$. On pourrait s'attendre à obtenir un biais plus faible. En effet, si d'emblée il existe une identité morphologique entre des pavillons (paramètres morphologiques nuls), cela doit correspondre à des jeux de HRTF très semblables, entre lesquels aucune adaptation n'est nécessaire (paramètres signal nuls).

Par ailleurs on observe une non-correspondance entre les angles : le pitch signal $\tilde{\psi}_{sig}$ est en effet mieux corrélé avec le roll morphologique θ_{morph}, qu'avec le pitch morphologique ψ_{morph}. Ce résultat reflète peut-être le fait que l'estimation du pitch morphologique est particulèrement entachée de bruit. Rappelons que cet angle correspond à une rotation autour de l'axe interaural. C'est le degré de liberté qui était

problématique lors de l'établissement du référentiel lié à la tête, et il a fallu passer par des photographies du profil des sujets pour déterminer la position du plan horizontal. Si la position de la tête était différente lors de la mesure des HRTF, alors il existe nécessairement un biais de rotation entre les deux référentiels, de surcroît de valeur différente pour chaque sujet. Ce biais peut intervenir de façon complexe dans les résultats observés, car ils ne traduisent que de façon relative les différences d'orientation des pavillons. Ainsi, l'impact d'un biais de rotation entre les référentiels signal et morphologique pour un seul sujet est susceptible d'affecter les résultats de 20 couples de pavillons (sur 60). On développe cette hypothèse en 5.5.4.

Cette non-correspondance entre les angles peut également être le reflet de phénomènes physiques bien réels, remettant en question nos *a priori* simplistes. Supposer l'existence d'une correspondance stricte entre les angles revient à considérer un problème acoustique dans lequel les pavillons flotteraient dans l'espace. En réalité l'influence de la tête ne peut pas être totalement négligée : bien qu'elle ne contribue pas, en tant qu'objet diffractant, de façon prépondérante aux colorations intéressantes du spectre, elle agit néanmoins comme un écran acoustique sur lequel sont disposés les pavillons. Le décalage des pavillons par rapport au centre de la tête, ainsi que leur inclinaison par rapport à la tête, ont un impact sur le comportement acoustique des pavillons eux-mêmes. L'influence de ces paramètres a été étudiée récemment grâce à des simulations numériques, dans les travaux de Plaskota et Dobrucki [202], et ceux de Iwaya et Suzuki [105] illustrés figures 5.30 et 5.31. Les premiers résultats suggèrent que l'angle de protrusion ζ affecte l'amplitude des caractéristiques spectrales des HRTF (creux et pics), et dans une moindre mesure les fréquences auxquelles elles apparaissent. L'impact de la position du pavillon est plus limité : on représente figure 5.31 les différences induites par des décalages de quelques centimètres sur le spectre d'amplitude des HRTF du plan horizontal à 2 kHz. Les auteurs indiquent que pour des fréquences supérieures, l'effet est encore moindre (communication privée, résultats non publiés). Les différences induites par ces phénomènes d'un individu à l'autre ne peuvent pas être réduites par le jeu de transformations que l'on s'autorise ici : rotations et homothéties. On cerne donc bien les limites de la technique proposée sur ce point.

5.5.2 Première évaluation

L'étape de régression décrite précédemment établit des relations analytiques entre les jeux de paramètres, que l'on résume par la relation $\mathcal{M}$ représentée figure 5.5. L'évaluation de la technique proposée consiste, pour chaque couple de pavillons considéré, à réaliser l'adaptation des HRTF en appliquant les transformations selon les paramètres signal $\{s_{sig}, \theta_{sig}, \psi_{sig}, \phi_{sig}\}$ prédits par $\mathcal{M}$ à partir des paramètres mor-

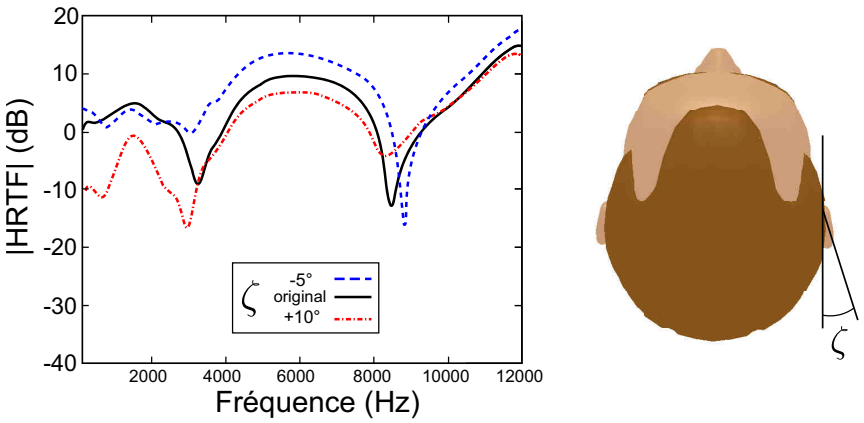


Figure 5.30 – Effet de l'angle de protrusion ζ du pavillon. HRTF dans la direction d'azimut 225° du plan horizontal (système polaire vertical). Résultats obtenus par simulation numérique d'après les travaux de Iwaya et Suzuki [105].

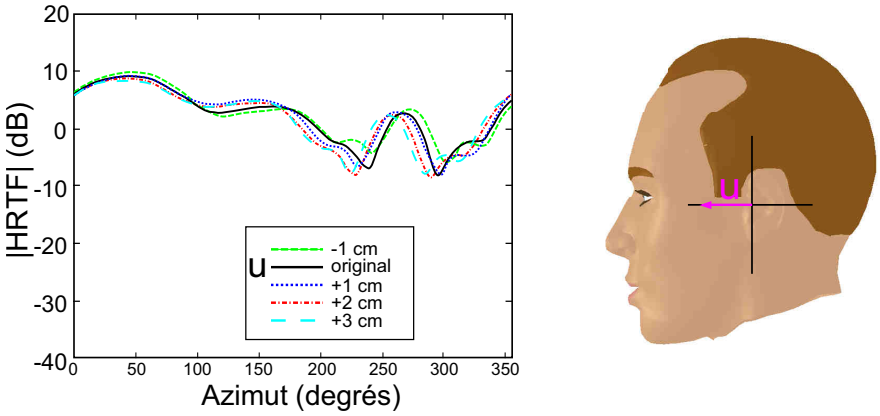


Figure 5.31 – Effet de la position du pavillon. Spectre d'amplitude des HRTF à 2 kHz dans le plan horizontal (système polaire vertical). Résultats obtenus par simulation numérique d'après les travaux de Iwaya et Suzuki [105].

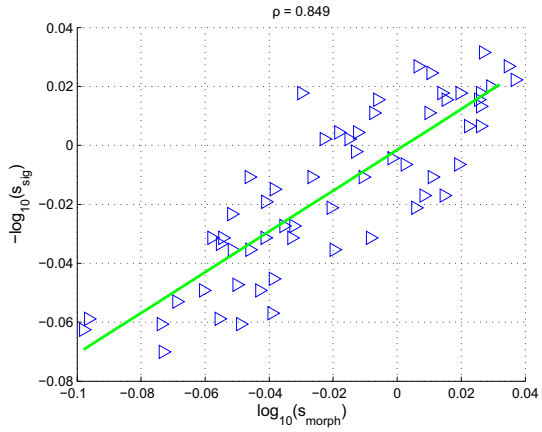


Figure 5.32 – Régression affine entre facteurs d'homothétie dans la configuration *Scaling*.

phologiques $\{s_{morph}, \theta_{morph}, \psi_{morph}, \phi_{morph}\}$ issus de l'alignement des maillages de type II :

$$\{s_{sig}, \theta_{sig}, \psi_{sig}, \phi_{sig}\} = \mathcal{M}\left(\{s_{morph}, \theta_{morph}, \psi_{morph}, \phi_{morph}\}\right) \tag{5.17}$$

L'ISSD atteinte après cette adaptation est évaluée, et elle peut être comparée à celle obtenue dans le cas limite d'une adaptation optimale, étudié en 5.4. Il convient de préciser que cette évaluation n'est pas une validation stricte du modèle proposé, car les données sur lesquelles elle est réalisée sont celles qui ont servi à sa mise au point. Cependant, il a été impossible d'impliquer dans cette expérience des sujets supplémentaires, pour lesquels à la fois la morphologie et les HRTF mesurées devaient être disponibles.

Parallèlement, il est nécessaire d'évaluer les performances du *scaling* fréquentiel simple, qui constitue l'état de l'art. On lie donc le facteur de *scaling* optimal OSF_S, calculé en 5.4 dans la configuration *Scaling*, au facteur d'homothétie s_{morph} obtenu pour le maillage de type II, sous leur forme logarithmique (cf. Fig. 5.32). Une régression affine entre ces données permet d'obtenir la relation analytique nécessaire pour l'adaptation des HRTF. L'ISSD issue de cette adaptation est également calculée.

On représente les résultats de ces expériences figure 5.33. La technique proposée se montre généralement meilleure que le *scaling* fréquentiel en termes de réduction de l'ISSD, en particulier pour les couples de pavillons présentant une ISSD initiale élevée. Les valeurs d'ISSD après adaptation s'approchent bien des valeurs optimales, ce qui révèle l'efficacité de la technique proposée. On observe néanmoins une augmentation de l'ISSD dans certains cas. Cela apparaît essentiellement pour des couples présentant une ISSD initiale faible. Le même phénomène a été rencontré par Middlebrooks [169], et Maki et Furukawa [148]. Middlebrooks a souligné qu'une faible

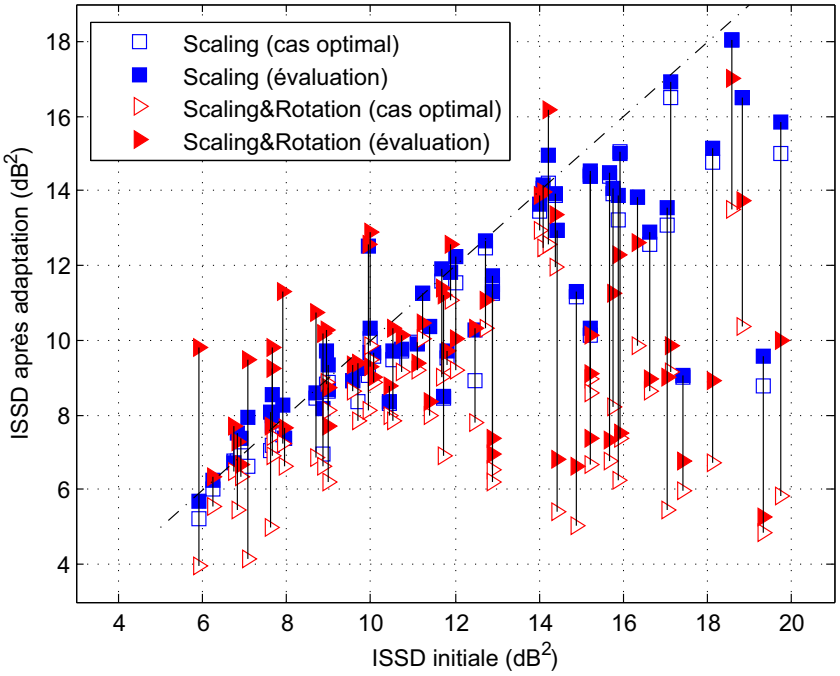


Figure 5.33 – ISSD obtenue après adaptation des HRTF en fonction de l'ISSD initiale, pour les 60 couples d'oreilles considérés. On représente en symboles pleins les résultats de l'évaluation des modèles d'individualisation, et en symboles creux les résultats optimaux (ceux représentés figure 5.23). Les résultats concernant un même couple d'oreilles sont reliés par une ligne noire.

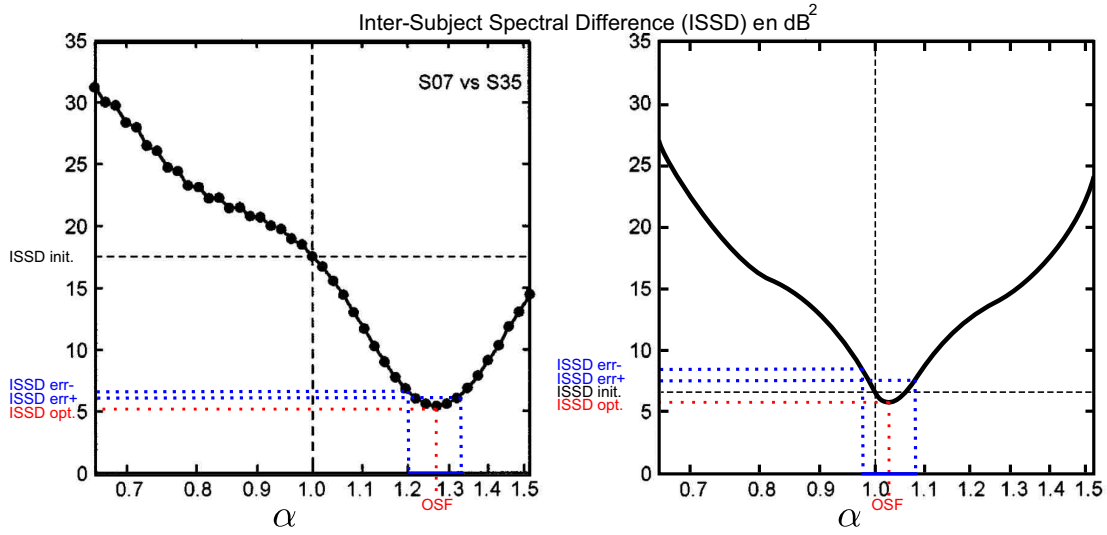


Figure 5.34 – Impact d'une erreur d'estimation du facteur de *scaling* optimal. A gauche : exemple issu des travaux de Middlebrooks [170]. On désigne par $\text{ISSD}_{\text{err+}}$ et $\text{ISSD}_{\text{err-}}$ les valeurs d'ISSD obtenues après adaptation dans les cas respectivement d'une surestimation et sous-estimation du facteur de *scaling*. Ces valeurs sont nettement plus faibles que celle de l'ISSD initiale, c'est pourquoi l'adaptation mène à une amélioration. A droite : exemple hypothétique de deux oreilles présentant une ISSD initiale faible, avec un facteur de *scaling* optimal proche de 1. L'impact d'une même erreur d'estimation est dans ce cas plus critique, car l'adaptation peut mener à une augmentation de l'ISSD.

erreur d'estimation du facteur de *scaling* peut rapidement mener à une solution sous-optimale. On illustre figure 5.34 un exemple d'une telle fonction, issue de [169]. L'impact d'une erreur d'estimation dépend des pentes de la fonction $ISSD(\alpha)$ autour de la valeur $\alpha = OSF$: il est d'autant plus important que ces pentes sont prononcées. Par ailleurs, une augmentation de l'ISSD peut apparaître d'autant plus facilement que l'OSF est proche de 1 : une telle situation traduit le fait qu'au mieux, une faible amélioration est à attendre de la technique d'adaptation. Les mêmes phénomènes apparaissent avec la technique proposée, mais de façon plus complexe, car l'ISSD dépend de 4 variables : le facteur de *scaling* et les trois angles de rotation du système de coordonnées. Une analyse des cas d'augmentation de l'ISSD révèle qu'il s'agit majoritairement de couples d'oreilles considérés comme marginaux par l'algorithme RANSAC, et qui ont donc été écartées lors de l'établissement de $\mathcal{M}$. Il semble donc naturel que la technique échoue pour ces cas particuliers. Bien que ce ne soit pas une validation, l'expérience montre qu'une simple régression affine, réalisée sur des données assez dispersées, permet de prédire une rotation du système de coordonnées qui va dans le sens d'une réduction des différences entre les jeux de HRTF, et ainsi d'une adaptation adéquate.

5.5.3 Choix des HRTF de la base de données

Il apparaît que la technique proposée n'offre pas systématiquement une adaptation efficace : on observe en effet des cas critiques pour lesquels l'ISSD après transformation optimale n'est que peu réduite par rapport à l'ISSD initiale (cf. Fig. 5.23). Il serait donc intéressant de se doter d'un critère, basé sur la morphologie, et capable de prédire la qualité finale de l'adaptation. Un tel critère permettrait de choisir dans la base de données le jeu de HRTF le plus prometteur. Dans cette optique, on définit, à partir de l'erreur locale d'alignement ε_{loc} décrite en 5.3.2, l'erreur "globale" d'alignement, notée ε_{glob} : les erreurs ε_{loc} locales sont pondérées par l'aire des faces correspondantes et additionnées, et le résultat divisé par l'aire totale. *A priori*, la valeur de ε_{glob} devrait être un bon indicateur du succès de l'adaptation, car si elle traduit la part irréductible des différences morphologiques entre deux pavillons après leur mise en correspondance, elle devrait également être liée à la part irréductible des différences entre les jeux de HRTF après une adaptation optimale, soit l'ISSD calculée en 5.4. On représente donc figure 5.35 cette ISSD en fonction de l'erreur globale d'alignement, obtenue entre les maillages de type II. La corrélation obtenue est faible (0.25), et même si on remarque généralement un accroissement de l'ISSD avec l'erreur globale d'alignement, cette dernière n'est pas un prédicteur fiable des performances finales. Un meilleur critère de prédiction reste donc à déterminer pour pouvoir mettre en œuvre la technique proposée. On peut relativiser ce manque en soulignant que, contrairement au cas du *scaling* fréquentiel, les différences sont ici bien réduites même dans les cas où l'ISSD initiale est élevée. On peut donc envisager d'appliquer la technique d'adaptation de façon aveugle pour quelques sujets d'une base de données, puis de proposer au nouvel auditeur de sélectionner perceptivement le jeu de HRTF adaptées le plus efficace pour la spatialisation en synthèse binaurale, selon une des méthodes décrites en 4.2.3.

5.5.4 Réflexions sur l'impact de décalages angulaires entre les référentiels signal et morphologique

Une étude informelle a été réalisée afin d'évaluer l'impact d'un décalage angulaire entre les plans horizontaux des deux référentiels d'étude : celui des scans de têtes et celui lié à la mesure des HRTF. L'idée de l'expérience est de déterminer les décalages angulaires qui peuvent avoir été commis entre ces plans horizontaux, par une procédure d'optimisation. On simplifie le problème en ne considérant que des rotations autour de l'axe interaural d'angle Δ_ψ. Faisant l'hypothèse simpliste que dans l'idéal, on devrait obtenir une corrélation maximale entre les angles $(\theta_{morph}, \tilde{\theta}_{sig})$, $(\psi_{morph}, \tilde{\psi}_{sig})$, et $(\varphi_{morph}, \tilde{\varphi}_{sig})$, on recherche pour chaque sujet, par la simulation, les décalages $\{\Delta_{\psi i}\}_{i=1...6}$, qui, une fois compensés, maximisent ces corrélations. Il

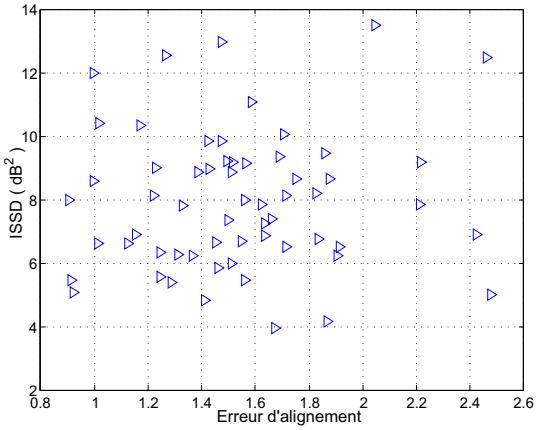


Figure 5.35 – ISSD obtenue après adaptation optimale en fonction de l'erreur globale d'alignement des pavillons ε_{glob}. Bien qu'une légère tendance montre que l'ISSD croît avec l'erreur d'alignement, la corrélation est assez faible (0.25).

manque une contrainte au problème ainsi posé : on rajoute donc à ce cas idéal à atteindre le fait que la régression linéaire, donc sans biais, entre ces angles mène à une droite de pente unitaire [8]. La simulation des décalages hypothétiques est réalisée en multipliant pour chaque couple de pavillons, la matrice de rotation d'alignement (OSR_{SR}) à gauche et à droite par les matrices de rotation Δ_{ψ_i} représentant le décalage pour chacun des sujets. L'optimisation est donc réalisée conjointement sur les 6 rotations d'angles Δ_{ψ_i}, et on obtient des angles qui tombent tous dans l'intervalle $[2°; 11°]$ (en valeur absolue). La corrélation entre ψ_{morph} et $\tilde{\psi}_{sig}$ est améliorée (0.65), tandis que les corrélations restent inchangées entre θ_{morph} et $\tilde{\theta}_{sig}$, et entre φ_{morph} et $\tilde{\varphi}_{sig}$. Cette expérience ne prouve pas que ce sont ces valeurs de décalage angulaire qui ont été commises entre les référentiels. Cependant ces valeurs sont plausibles, ce qui suggère qu'une non-coïncidence des référentiels, même de faible ampleur, peut contribuer à la dispersion des résultats, ainsi qu'à la non-correspondance entre certains angles. Un soin particulier doit donc être apporté à cette étape : idéalement, la mesure des HRTF et l'acquisition de la morphologie doivent être réalisés conjointement pour constituer la base de données.

5.6 Discussion

La technique d'adaptation proposée se montre efficace et dépasse nettement les performances de l'état de l'art, et ce malgré les hypothèses simples qui la sous-

8. Imposer une pente de la régression proche de 1 revient à ne sélectionner que les solutions $\{\Delta_{\psi_i}\}$ qui assurent que les angles morphologiques soient du même ordre de grandeur que les angles signal.

tendent. On observe surtout un progrès dans le cas des couples de pavillons présentant des différences initiales élevées entre les HRTF. Néanmoins, les biais, et la non-correspondance entre certains angles signal et morphologiques ont une origine qui reste incertaine. Elle peut résider dans l'existence de phénomènes complexes liés à la présence de la tête, et non considérés dans le modèle. Ces résultats révèlent peut-être également des problèmes de biais entre les deux référentiels de l'étude : celui lié au scan de la tête, et celui attaché aux HRTF mesurées. Toutes les précautions ont été prises pour les faire coïncider, mais il peut subsister des erreurs, car la position réelle adoptée par les sujets lors des mesures de HRTF n'est ici pas parfaitement connue.

D'un point de vue méthodologique, on peut s'interroger sur la pertinence de la régression linéaire entre les angles roll, pitch et yaw représentant les rotations considérées. On pourrait s'attendre à de meilleurs résultats si l'on évitait de passer par une telle décomposition, c'est-à-dire si l'on conservait les rotations sous leur forme matricielle. Malheureusement nous avons cherché sans succès des outils mathématiques permettant une régression entre deux matrices de rotation, c'est-à-dire une régression qui soit menée de $SO(3)$ dans $SO(3)$.

La technique proposée s'est focalisée sur l'adaptation du spectre des HRTF aux hautes fréquences, délaissant les autres indices de localisation. Middlebrooks a proposé de transformer l'ITD non-individuelle, en la multipliant par un gain déterminé par le ratio entre les dimensions de la tête des sujets. Puisque la technique nécessite la mesure de la morphologie du nouvel auditeur, on peut plutôt envisager d'utiliser un modèle d'ITD tenant compte à la fois de la taille de la tête, mais aussi du positionnement de ses oreilles, comme le modèle développé par Busson [38]. Le spectre des HRTF aux basses fréquences pourrait être déterminé par un modèle simple de tête sphérique comme celui proposé par Duda et Martens [63], dont on pourrait adapter les paramètres à la morphologie de l'auditeur.

L'acquisition en 3D de la tête et des oreilles du nouvel auditeur peut sembler complexe à mettre en œuvre. Pour remplacer le scan laser, on peut envisager d'utiliser la méthode proposée par Dellepiane *et. al* [61] : il s'agit d'obtenir la morphologie en 3D d'un sujet, à partir seulement de 5 photographies de sa tête sous différents angles de vue. Des opérations de *morphing* permettent d'ajuster à chaque nouveau sujet des modèles 3D de têtes et de pavillons d'oreille issus d'une base de données (cf. Fig. 5.36).

Une validation objective stricte de la technique proposée reste à réaliser. Il faudrait pour cela disposer des morphologies et des HRTF d'autres sujets, non considérés dans l'ensemble d'entraînement. Une validation perceptive serait également intéressante. L'ISSD a été choisie comme distance objective entre des jeux de HRTF car Middlebrooks a prouvé qu'une amélioration de la spatialisation accompagne la réduc-

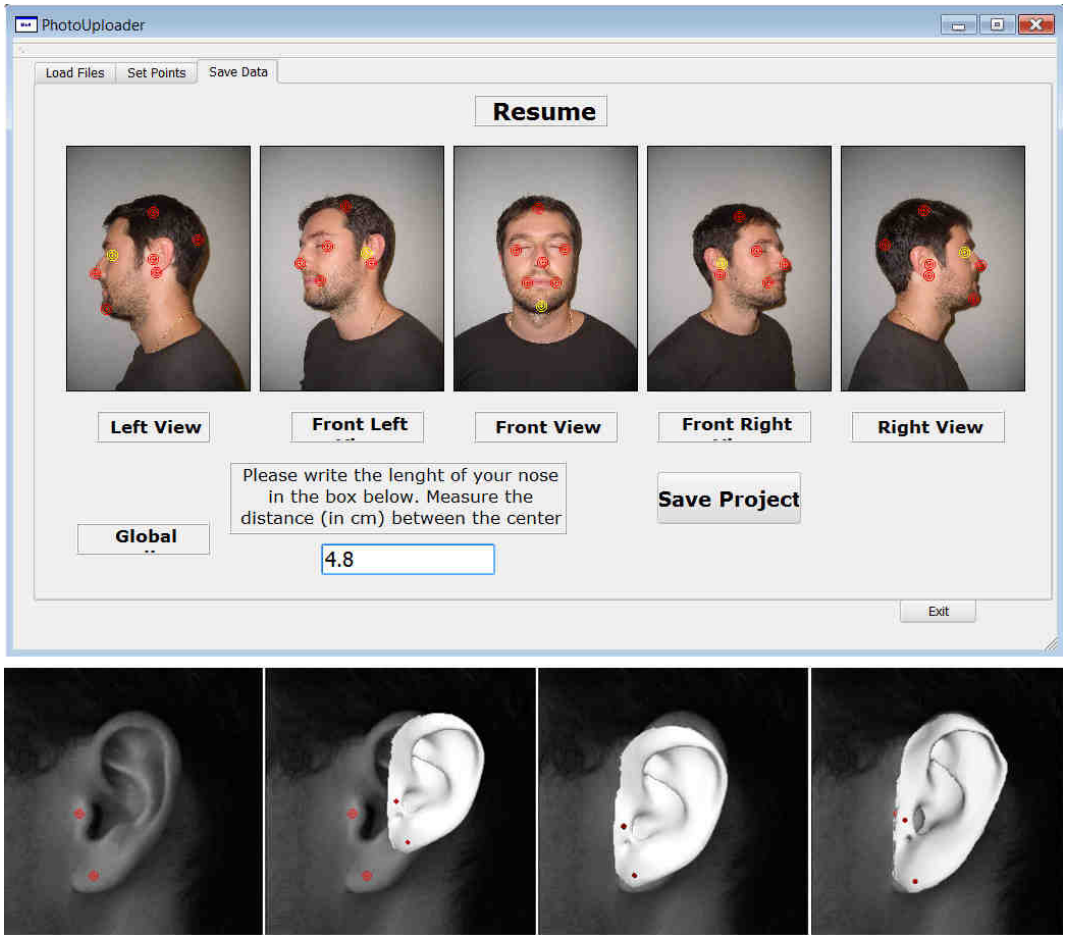
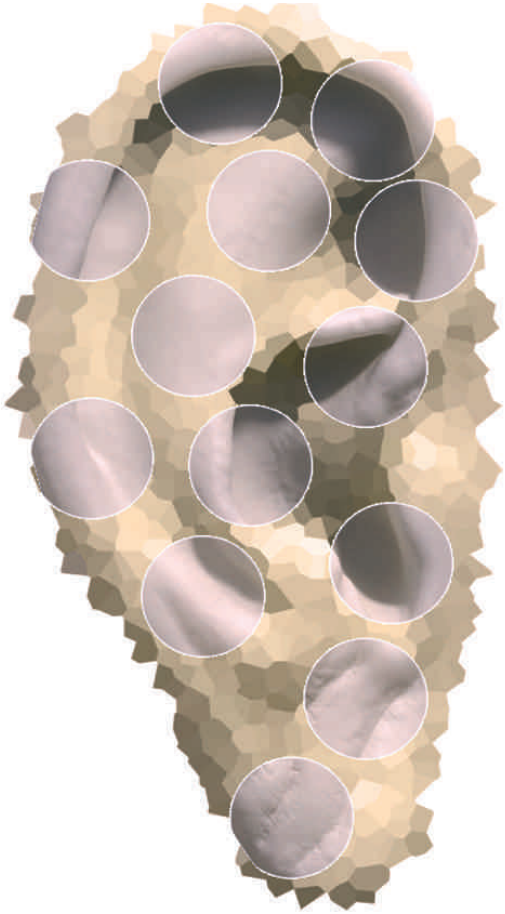


Figure 5.36 – Obtention de la morphologie en 3D d'un nouvel auditeur à partir de photographies (d'après [61]). En haut : interface d'acquisition des photographies sous différents angles de vue. En bas : ajustement par *morphing* d'un modèle 3D d'oreille sur l'image 2D de l'oreille du nouvel auditeur.

tion de cet indice de dissimilarité. Cependant, si l'on sait que la technique proposée dépasse nettement les performances du *scaling* fréquentiel selon ce critère, on ne mesure pas le gain associé en termes perceptifs, et il faudrait donc l'évaluer. Enfin, dans la lignée des travaux de Middlebrooks, on pourrait envisager d'obtenir selon un protocole psychophysique simple les paramètres des transformations à appliquer aux HRTF. Middlebrooks a montré pour le facteur de *scaling* que c'était possible, et facile pour l'auditeur. Dans la technique proposée, le problème est plus complexe, car il y a 4 paramètres à régler par oreille. Même si l'on simplifie le protocole en considèrant conjointement la transformation des HRTF des oreilles gauche et droite, l'espace dans lequel le sujet doit chercher les paramètres optimaux reste trop grand. Une étude préalable des morphologies des pavillons issus d'une large base de données pourrait permettre de dégager des relations classiques entre les angles roll, pitch et yaw, et ainsi réduire la dimension de cet espace de recherche.



Chapitre 6

Reconstruction individuelle de HRTF à partir de mesures allégées

Ce sont les inconvénients de la mesure acoustique des HRTF qui ont motivé la recherche de techniques alternatives pour obtenir des filtres binauraux adaptés à chaque auditeur. Pour acquérir convenablement les HRTF, il semble absolument nécessaire de se situer dans un environnement anéchoïque, et d'assurer un contrôle précis de la position égocentrique des haut-parleurs. Ces aspects techniques entraînent un coût important pour la mise en place d'un dispositif de mesure, mais c'est essentiellement la durée d'une session de mesure qui pose problème pour l'inconfort qu'elle impose au sujet. Des solutions de traitement du signal ont été proposées par Majdack *et al.* [147] afin de réduire au maximum cette durée, pendant laquelle le sujet doit rester quasiment immobile. Il s'agit d'exciter simultanément avec des sinusoïdes de fréquence glissante des haut-parleurs positionnés selon différentes directions. Un entrelacement des signaux temporels permet de s'affranchir judicieusement des éventuelles non-linéarités de la chaîne de mesure, et d'assurer l'identification du haut-parleur lors de l'opération d'inversion des signaux mesurés aux oreilles. Les auteurs affirment qu'un tel dispositif permet de diviser par quatre le temps de mesure par rapport aux solutions classiques. Pour aller encore plus loin, on peut envisager d'acquérir les HRTF pour un nombre limité de directions, puis de générer les données dans les positions intermédiaire. Les bases de données utilisées à des fins scientifiques sont généralement obtenues selon un échantillonnage spatial extrêmement fin de la sphère entourant l'auditeur (environ 1000 directions), mais une telle résolution n'est pas absolument nécessaire, car de nombreuses techniques d'interpolation sont satisfaisantes à partir d'un nombre plus limité de mesures. Malgré tout, quelle que soit la technique, on conçoit aisément que ses performances se dégradent à mesure que les informations sur lesquelles elle s'appuie se raréfient. On fait ici l'hypothèse que dans le cas d'un nombre réduit de directions mesurées, les faiblesses des techniques classiques d'interpolation résident en partie dans le fait qu'elles agissent de façon aveugle. Il paraît en effet préférable d'intégrer un maximum d'informations *a priori* dans un problème où les données traitées sont en quantité critique. On adopte donc une stratégie tenant compte de la nature des données à reconstruire : on propose une nouvelle technique de reconstruction fondée sur un processus de reconnaissance de formes, et qui utilise avantageusement des informations issues de l'analyse d'une base de données, constituée de HRTF mesurées finement sur de nombreux sujets. La mise au point de la technique est réalisée en tirant parti des observations décrites au chapitre 3 sur l'évolution spatio-fréquentielle du spectre d'amplitude des HRTF.

On présente d'abord une brève revue de l'état de l'art des techniques d'interpolation, avant de détailler la technique de reconstruction proposée. Une évaluation objective et une évaluation subjective sont finalement décrites : elles démontrent

l'apport de notre méthode par rapport aux techniques existantes.

6.1 Interpolation

Le problème de l'interpolation des HRTF a fait l'objet de nombreuses études, et diverses techniques ont été mises au point. Les plus pertinentes sont détaillées ici. L'intérêt recherché dans ces techniques est bien sûr leur capacité à reconstruire fidèlement les HRTF à partir d'un nombre minimal de mesures. Ce nombre limite dépend naturellement de la technique utilisée.

Minaar *et al.* [179] ont cherché à reconstruire les HRTF dans le domaine temporel, par interpolation linéaire des composantes à phase minimale (cf. 2.4). Les HRTF utilisées dans leur étude ont été mesurées sur une tête artificielle, selon un échantillonnage très fin de la sphère (2° en élévation et en azimut, dans le système polaire-vertical, soit 11975 paires de HRTF). Le nombre minimal de mesures nécessaires est déterminé en évaluant la capacité de sujets à discriminer, en synthèse binaurale, une source spatialisée grâce aux HRTF originales d'une source spatialisée grâce aux HRTF reconstruites (selon une procédure *Three Alternatives Forced Choice* ou 3-AFC). Les auteurs concluent que 1100 mesures sont nécessaires pour assurer une reconstruction parfaitement transparente. Ce résultat est cependant très discutable, car on peut identifier des biais dans leur expérience. En effet, les HRTF d'une tête artificielle ne permettent pas, en général, d'assurer une spatialisation optimale, et on peut ainsi douter de la qualité des sources virtuelles générées lors des tests perceptifs (externalisation insuffisante des sources et/ou élévation mal perçue). Les différences spectrales entre les HRTF originales et les HRTF reconstruites ont probablement été perçues en partie comme des différences de timbre. Le protocole psychoacoustique n'est donc pas adapté au problème, car trop discriminant. Il aurait été plus judicieux d'évaluer l'impact des dégradations du spectre des HRTF sur la perception spatiale des sources, au moyen de tests de localisation par exemple. Le nombre de 1100 directions de mesure n'a donc rien à voir avec le résultat recherché, mais doit plutôt être considéré comme une limite haute, c'est-à-dire qu'un échantillonnage spatial plus fin est parfaitement inutile.

Utilisant une technique d'interpolation similaire, Langendijk et Bronkhorst [130] proposent une évaluation subjective qui permet de déterminer si la discrimination entre les HRTF originales et reconstruites est le fruit d'une interprétation des différences spectrales comme des différences de timbre, ou bien en plus comme des différences spatiales entre les sources qu'elles génèrent. Il apparaît que ce sont d'abord les différences de timbre qui interviennent : une résolution de 6° en azimut et en élévation permet une reconstruction parfaitement transparente, tandis qu'autour de 10° à 15°, des différences de timbre sont perçues, la spatialisation restant fidèle. Au-delà

de 20°, la spatialisation est elle aussi affectée.

Hartung *et al.* [82] ont proposé une comparaison objective et subjective des performances de plusieurs techniques d'interpolation spécifiquement adaptées à des données recueillies sur la sphère, et agissant dans le domaine fréquentiel ou dans le domaine temporel. Les auteurs montrent que c'est l'interpolation par spline de type plaque mince sur la sphère (*Spherical Thin Plate Spline* ou STPS) (cf. Annexe B), appliquée séparément sur le spectre de phase, et le spectre d'amplitude, qui offre les meilleurs résultats à partir d'un échantillonnage régulier de pas 15° en azimut et 10° en élévation (système polaire-vertical). L'étude n'a cependant pas déterminé le nombre minimal de mesures nécessaires en entrée de cette technique. Cette limite a été explicitement recherchée par deux études. Carlile *et al.* [48] ont proposé de reconstruire le spectre d'amplitude par interpolation STPS des poids de la décomposition ACP du spectre d'amplitude des HRTF mesurées, comme proposé par Chen *et al.* [53] (cf. Annexe I). Grâce à une décomposition en filtre à phase minimal et retard pur, les indices temporels sont traités séparément : le retard interaural est obtenu par interpolation STPS des retards évalués dans les directions mesurées. Des tests de localisation montrent que des dégradations significatives de la spatialisation apparaissent quand l'interpolation est réalisée avec moins de 150 mesures.

Martin et McAnally [151] ont développé une technique s'appliquant à des HRTF mesurées sur un échantillonnage régulier, selon laquelle le spectre d'amplitude et le spectre de phase sont interpolés séparément. Pour une direction quelconque de l'espace, la reconstruction des HRTF implique la somme pondérée des HRTF des directions de mesure les plus proches. Des tests perceptifs montrent la validité de la technique jusqu'à une résolution de 20° en élévation et en azimut (système polaire-horizontal), ce qui représente environ 120 à 130 directions de mesure, si l'on exclut la calotte sphérique inférieure à l'élévation -20°.

6.2 Technique proposée

6.2.1 Description

Le principe général de la technique proposée est détaillé figure 6.1. La première étape de la technique proposée consiste en l'analyse d'une base de données de HRTF, mesurées selon un échantillonnage fin, et sur de nombreux sujets. L'objectif de cette étape est de dégager des comportements typiques, et de fournir des informations utiles pour la reconstruction des HRTF d'un sujet quelconque. On propose de se focaliser sur la structure spatiale du spectre d'amplitude des HRTF, en considérant les données sous forme de SFRS. Comme le chapitre 5 a permis de l'illustrer, de fortes similarités existent d'un sujet à l'autre, mais elles peuvent être masquées par

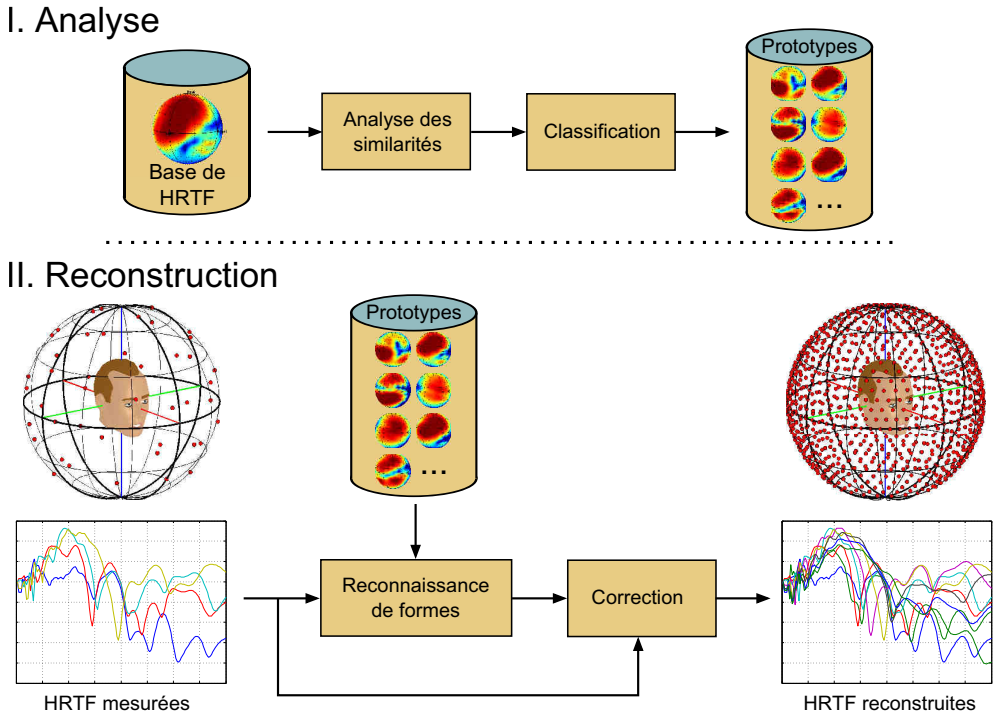


Figure 6.1 – Schéma de principe de la technique proposée.

des décalages sur l'axe fréquentiel, ainsi et que par des décalages spatiaux. Une analyse des similarités entre les SFRS doit donc utiliser une mesure appropriée, s'affranchissant de ces phénomènes : la mesure de similarité proposée est basée sur le calcul de l'intercorrélation normalisée sur la sphère. S'appuyant sur cette mesure, on réalise une classification des SFRS, selon une technique de classification spectrale normalisée. On définit, à partir de cette classification, un nombre réduit de SRFS prototypiques, résumant les principales formes de SFRS observées dans la base de données. Ces prototypes constituent les informations *a priori* utilisées lors de l'étape de reconstruction.

La seconde étape, dite de recontruction, est réalisée pour chaque nouvel auditeur, dont on a mesuré les HRTF dans un nombre réduit de directions. Un processus de reconnaissance de formes est utilisé pour remplacer les données mesurées par le jeu de SFRS prototypiques les plus semblables, au sens de la mesure de similarité utilisée lors de l'étape d'analyse. Finalement, l'erreur commise dans cette opération de substitution, et qui est connue dans les directions de mesure, est compensée globalement lors d'une phase dite de correction, basée sur une interpolation STPS.

6.2.2 Analyse

Analyse des similarités

Introduisons l'intercorrélation normalisée entre deux fonctions définies sur la sphère : soient h et $h' \in L^2(S^2)$, l'intercorrélation normalisée $C_R(h,h')$ entre h et h' est définie pour une rotation R par :

$$C_R(h,h') = \frac{\int_{S^2} \breve{h}(\chi)\overline{\Lambda_R(\breve{h'})(\chi)\mathrm{d}\chi}}{\sqrt{\int_{S^2}|\breve{h}(\chi)|^2\mathrm{d}\chi \int_{S^2}|\breve{h'}(\chi)|^2\mathrm{d}\chi}} \tag{6.1}$$

où $R \in SO(3)$,

$$\Lambda : \quad L^2(S^2) \rightarrow L^2(S^2) \tag{6.2}$$

$$\Lambda_R(h')(\chi) = h'(R^{-1}(\chi)) \tag{6.3}$$

et $\breve{h}$ est le résultat du centrage de h autour de sa moyenne spatiale :

$$\breve{h}(\chi) = h(\chi) - \frac{1}{4\pi}\int_{S^2} h(\chi)\mathrm{d}\chi \tag{6.4}$$

Cet outil mathématique est bien connu dans les problèmes d'appariemment rotationnel (*rotational matching*) [56, 242]. L'intercorrélation peut être calculée pour une paire quelconque de SFRS. Notons que par définition les SFRS sont de moyenne spatiale nulle (cf. 5.4.1), donc $h = \breve{h}$. Un passage dans le domaine des harmoniques sphériques permet d'obtenir les valeurs de l'intercorrélation rapidement pour une série de valeurs de $R \in SO(3)$ [123] (cf. Annexe F). Retenons que l'échantillonnage de $SO(3)$ sur lequel l'intercorrélation est évaluée est d'autant plus fin que la bande B de la décomposition en harmoniques sphériques est large.

L'intercorrélation normalisée sur la sphère apparaît comme un outil intéressant pour analyser la similarité entre deux SFRS, car on y trouve les deux degrés de liberté nécessaires à l'analyse. D'abord elle permet de détecter des similarités malgré l'existence d'un décalage rotationnel entre les SFRS, de la même manière que l'intercorrélation entre deux signaux temporels permet une analyse qui s'affranchit d'un éventuel retard. De plus, si l'on compare des SFRS correspondant à des fréquences différentes, on s'affranchit également des décalages fréquentiels observés d'un individu à l'autre. On définit donc la mesure $\Upsilon(h,h')$ de similarité entre deux SFRS h et h' comme la valeur maximale prise par l'intercorrélation normalisée $C_R(h,h')$ quand la rotation R parcourt le groupe $SO(3)$ [1] :

$$\Upsilon(h,h') = \max_{R \in SO(3)} C_R(h,h') \tag{6.5}$$

On illustre figure 6.2 le principe du calcul de la similarité Υ entre deux SFRS.

1. On approchera $\Upsilon(h,h')$ par la valeur maximale prise par $C_R(h,h')$ sur l'échantillonnage discret de $SO(3)$ correspondant à une bande B donnée (cf. Annexe F).

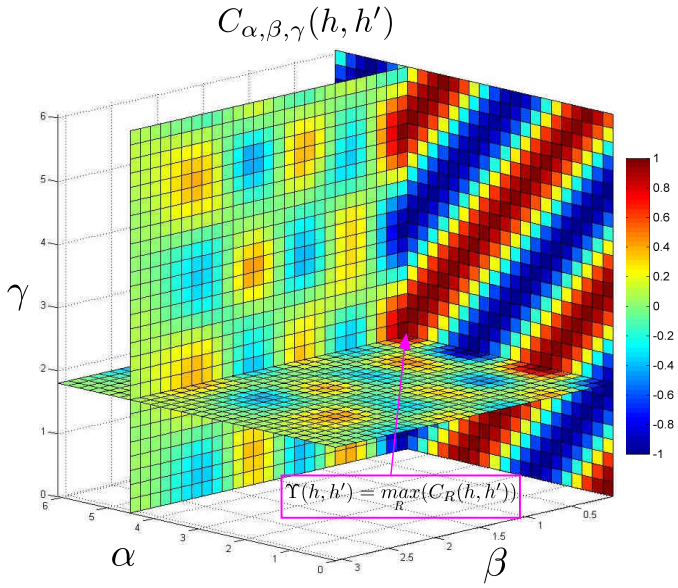


Figure 6.2 – Définition la similarité entre deux SFRS h et h'. Les angles α, β et γ sont les angles d'Euler : toute rotation de $SO(3)$ peut être vue comme la composée de 3 rotations, et donc représentée par un triplet (α,β,γ).

On représente figures 6.3, 6.4, 6.6, 6.8, et 6.9 les valeurs des similarités entre les SFRS de plusieurs paires de sujets, sur la bande de fréquence [4 kHz - 13 kHz].

En particulier, on illustre figure 6.3 la similarité entre les SFRS d'une même oreille (oreille gauche du sujet n°1 de la base d'Orange Labs). Sur la diagonale, Υ est donc l'autocorrélation normalisée, qui atteint naturellement la valeur 1. Cet exemple montre que le choix de cette mesure de similarité est pertinent, dans le sens où l'on observe un constraste intéressant entre les SFRS correspondant à des fréquences différentes.

On représente figure 6.4 les similarités entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs, et les SFRS de l'oreille de gauche du sujet n°27 de la base du CIPIC [57]. En complément, on représente figure 6.5 l'angle Θ_Υ de la rotation R_Υ qui permet pour une paire de SFRS (h,h') d'atteindre le maximum de l'intercorrélation [2] :

$$R_\Upsilon(h,h') = \underset{R\in SO(3)}{\arg\max}\, C_R(h,h'). \tag{6.6}$$

On avait noté en 5.1 une forte similarité entre les SFRS de ces deux oreilles, mais avec une homothétie nécessaire sur l'axe fréquentiel. Cette observation est traduite sur la figure 6.4 par une sorte de "rift" rectiligne, semblable à celui que l'on peut voir figure 6.3, mais décalé par rapport à la diagonale, et d'amplitude moindre. Par ailleurs, on

2. Toute rotation R du groupe $SO(3)$ peut être réduite comme une rotation unique autour d'un axe. L'angle Θ_Υ est précisément défini comme l'angle de cette rotation.

observe figure 6.4 que le maximum de la similarité Υ le long de ce rift correspond à une rotation R_Υ d'angle minimal Θ_Υ, presque nul. Cela traduit le fait qu'aucune rotation du système de coordonnées n'est nécessaire pour réduire les différences entre ces deux jeux de données, comme l'observation préliminaire des SFRS permettait de le suggérer. Ainsi le facteur de *scaling* optimal peut être approché par la pente de la droite passant au plus près de ce rift.

On représente figures 6.6 et 6.7 les résultats entre l'oreille gauche du sujet n°1 de la base de Orange Labs, et l'oreille gauche du sujet *th* de la base de l'Université du Maryland [76, 162]. Comme l'observation des SFRS permettait de le suggérer, un faible décalage fréquentiel est observé entre les HRTF des deux oreilles, ce qui se traduit par un rift proche de la diagonale. L'angle Θ_Υ de la rotation correspondante est par contre non nul : la similarité entre les SFRS est donc maximale moyennant une rotation du système de coordonnées d'un des jeux de données.

Ces deux derniers exemples correspondent à des paires d'oreilles pour lesquelles les différences entre les HRTF peuvent être efficacement réduites de façon globale par la technique décrite chapitre 6. Il existe également des cas critiques, pour lesquels les HRTF présentent des dissimilarités plus complexes, probablement liées à des morphologies de pavillon intrinsèquement différentes. Le profil des similarités entre les SFRS ne présente alors pas de rift tel que celui observé précédemment, mais des taches éparses (cf. Fig. 6.8 et 6.9).

Notons enfin l'existence de couples de SFRS présentant une similarité élevée, mais au prix d'une rotation R_Υ d'angle important (parfois proche de 180°). De tels angles de rotation ne peuvent trouver leur origine dans des différences physiques d'orientation entre les pavillons des individus considérés. C'est pourquoi il convient lors de l'analyse de la base de données d'écarter ces cas particuliers.

D'après ces observations, on propose d'analyser la base de données comme suit :

1. Pour chaque oreille de chaque sujet, N_s SFRS régulièrement espacées sur la bande de fréquences [4 kHz - 13 kHz] sont considérées. Les SFRS des oreilles gauches sont symétrisées par rapport au plan médian de façon à rendre les SFRS des oreilles gauches et droites toutes comparables.

2. Pour chaque paire d'oreilles, les SFRS sont comparées uniquement aux SFRS de fréquences voisines : pour une SFRS h_ν correspondant à la fréquence ν, on calcule donc sa similarité avec les SFRS $h'_{\nu'}$ de l'autre oreille, correspondant à la fréquence ν', telle que $\nu' \in [\nu - \nu_0(\nu)/2; \nu + \nu_0(\nu)/2]$. On fait croître la largeur de cet intervalle avec la fréquence pour pouvoir capturer les éventuels effets de décalages fréquentiels précédemment décrits. La valeur $\nu_0(\nu)$ est réglée de sorte que, dans le cas où deux jeux de SFRS sont tels que leurs différences peuvent être annulées par *scaling* fréquentiel, l'analyse des similarités peut

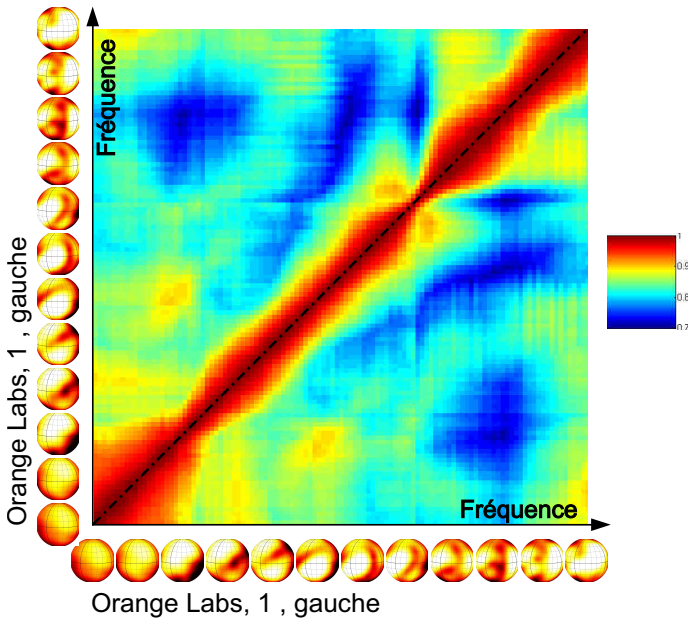


Figure 6.3 – Similarité Υ calculée entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs.

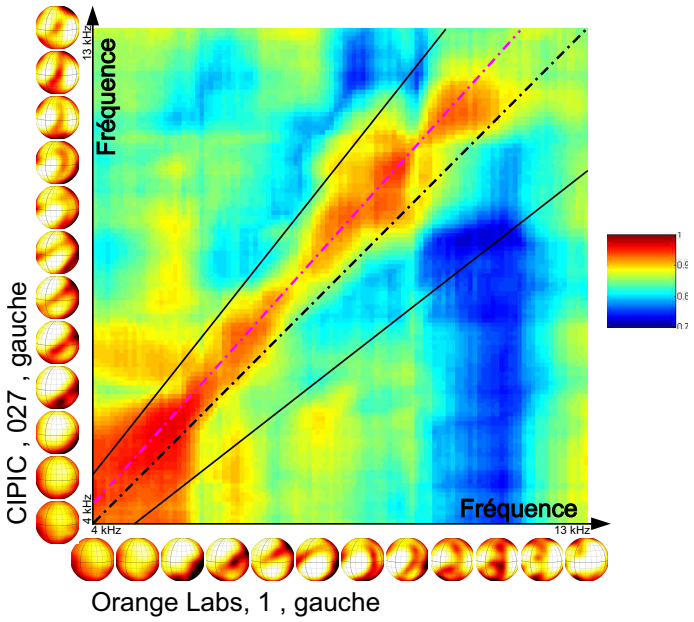


Figure 6.4 – Similarité Υ calculée entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs et celles de l'oreille gauche du sujet n°27 de la base du CIPIC [57]. La ligne en pointillés magenta représente la position du "rift" de similarité maximale observé quand il existe une relation homothétique simple entre les jeux de HRTF, et qui correspond le mieux aux données comparées. Les lignes noires en trait plein représentent les limites de l'intervalle fréquentiel de l'analyse réalisée en pratique.

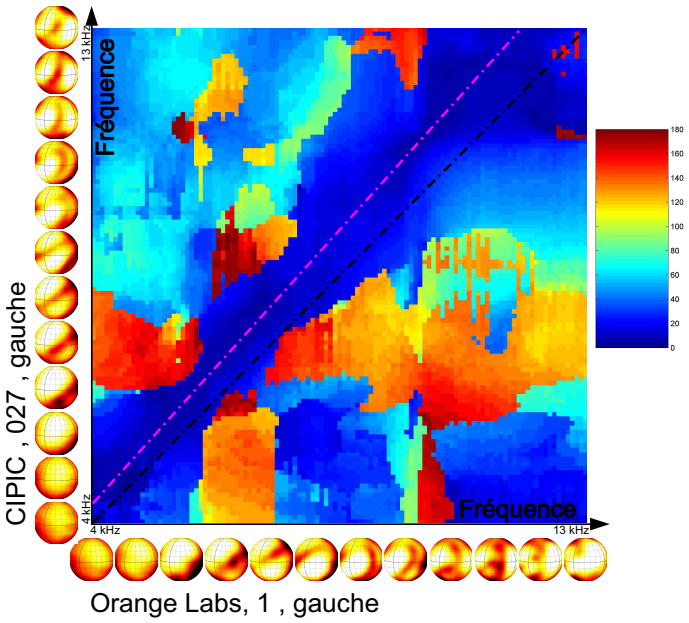


Figure 6.5 – Angle Θ_Υ de la rotation R_Υ, relative à la similarité calculée entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs et celles de l'oreille gauche du sujet n°27 de la base du CIPIC [57] (cf. 6.4 pour les détails).

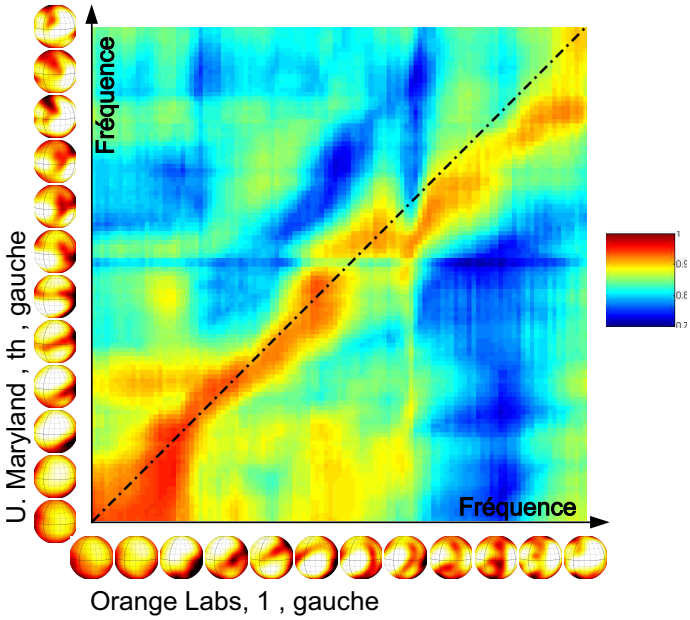


Figure 6.6 – Similarité Υ calculée entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs et celles de l'oreille gauche du sujet *th* de la base de l'Université du Maryland [76, 162].

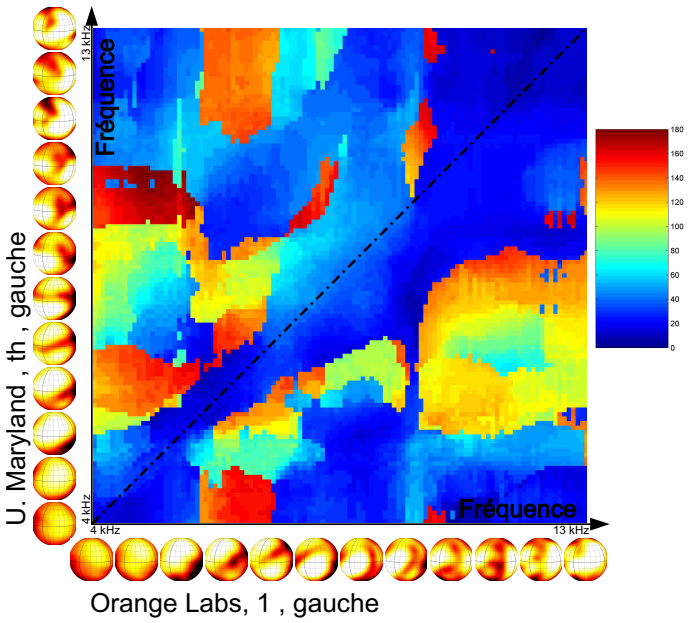


Figure 6.7 – Angle Θ_Υ de la rotation R_Υ, relative à la similarité calculée entre les SFRS de l'oreille gauche du sujet n°1 de la base d'Orange Labs et celles de l'oreille gauche du sujet *th* de la base de l'Université du Maryland [76, 162].

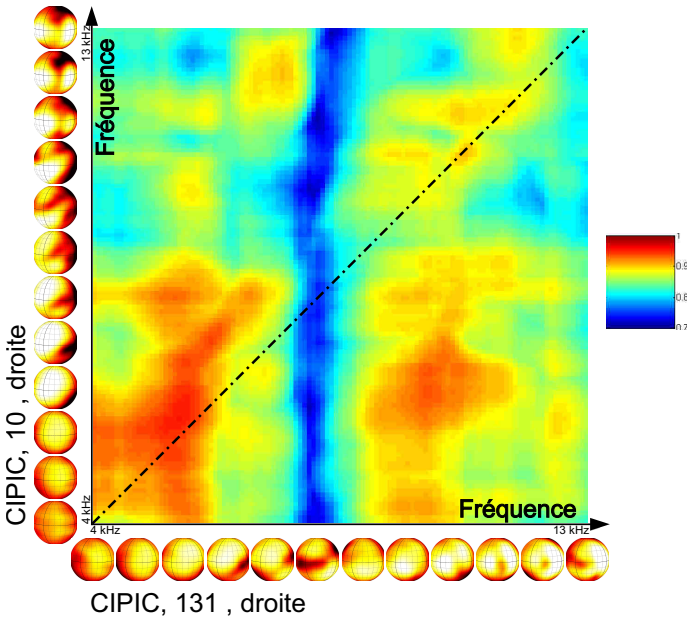


Figure 6.8 – Similarité Υ calculée entre les SFRS de l'oreille droite du sujet n°131 et celles de l'oreille droite du sujet n°10 de la base du CIPIC [161].

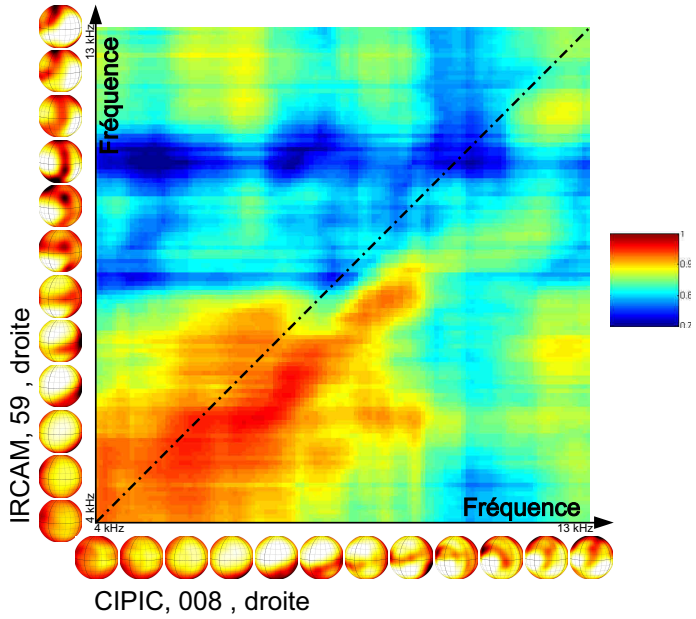


Figure 6.9 – Similarité Υ calculée entre les SFRS de l'oreille droite du sujet n°8 et celles de l'oreille droite du sujet n°59 de la base de l'IRCAM [103].

capturer le sommet du rift caractéristique correspondant à des facteurs de *scaling* optimaux situés dans l'intervalle [0.75 - 1.25] (cf. Fig. 6.4).

3. On ne conserve que les similarités entre SFRS atteintes moyennant une rotation R_Υ d'angle inférieur à 40°, de façon à écarter les appariements dont l'origine physique est peu plausible.

Classification

L'analyse des similarités précédemment décrite est un préalable nécessaire à la classification de l'ensemble des données. Puisque les relations de similarité ne sont pas connues de façon exhaustive entre chaque SFRS et toutes les autres, on choisit naturellement de représenter le résultat de l'analyse sous la forme d'un graphe (cf. Fig. 6.10). Chaque nœud de ce graphe correspond à une SFRS, tandis que chaque arête reliant deux nœuds matérialise l'existence du calcul de similarité entre ces SFRS. En outre, chaque arête est pondérée par la valeur de la similarité entre ses sommets. Une technique efficace de classification pour des données ainsi organisées est la classification spectrale normalisée ou *normalized spectral graph clustering* [260]. L'objectif de cette technique est de trouver une partition du graphe telle que les arêtes entre des groupes différents ont une faible pondération (faible similarité inter-*cluster*), tandis que les arêtes entre les nœuds d'un même groupe ont un poids élevé (grande similarité intra-*cluster*). L'algorithme est le suivant :

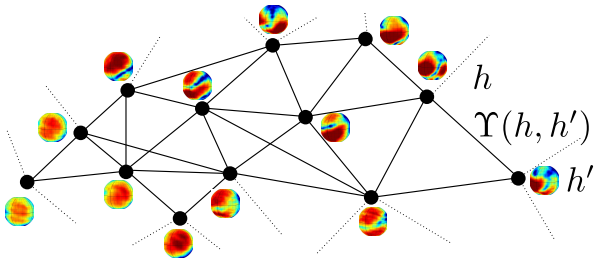


Figure 6.10 – Organisation des données analysées sous forme d'un graphe.

1. Une matrice d'adjacence symétrique $\mathbf{W}$ de dimension $n \times n$ est constituée, où n est le nombre total de nœuds du graphe. Sur chaque ligne de $\mathbf{W}$, correspondant à un nœud du graphe, seuls k éléments sont non nuls : ils correspondent aux plus proches voisins de ce nœud, au sens de la similarité Υ. Chaque élément w_{ij} de $\mathbf{W}$, s'il est non nul, est égal à la valeur de la similarité Υ entre les SFRS d'indices i et j.
2. Le Laplacien normalisé $\mathcal{L}$ du graphe est calculé :

$$\mathcal{L} = \mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}$$

   où $d_i = \sum_{j=1}^{n} w_{ij}$ sont les éléments de la matrice diagonale $\mathbf{D}$ appelée matrice de degré, et $\mathbf{I}_n$ est la matrice identité.
3. On calcule les κ premiers vecteurs propres $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\kappa$ de $\mathcal{L}$.
4. Soit $\mathbf{U} \in \mathbb{R}^{n \times \kappa}$ la matrice dont les colonnes sont des vecteurs propres : $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\kappa]$.
5. Pour $i = 1, \ldots, n$, soit $\mathbf{y}_i \in \mathbb{R}^\kappa$ le vecteur correspondant à la i-ème ligne de $\mathbf{U}$.
6. L'algorithme de classification *k-means* [146] est appliqué à l'ensemble des points $\{y_i\}_{i=1,\ldots,n}$ de l'espace $\mathbb{R}^\kappa$, pour former les *clusters* $\mathbf{C}_1, \ldots, \mathbf{C}_\kappa$. L'algorithme *k-means* est détaillé en Annexe G.
7. Finalement les κ *clusters* de SFRS $\{\mathbf{A}\}_{i=1,\ldots,\kappa}$ qui nous intéressent sont tels que $\mathbf{A}_i = \{h_j | y_j \in \mathbf{C}_i\}$: chaque *cluster* $\mathbf{A}_i$ est composé de toutes les SFRS h_j dont le vecteur correspondant $\mathbf{y}_j$ appartient à $\mathbf{C}_i$.

On obtient schématiquement un découpage du graphe tel que celui représenté figure 6.11. Une approche physique de cette technique de classification a été proposée par Shi et Malik [237] : le graphe représente un système complexe de masses et de ressorts connectés les uns aux autres, où chaque nœud est associée à une masse, et chaque arête à un ressort. Les pondérations des arêtes sont les raideurs des ressorts, tandis que la masse de chaque nœud est la somme de tous les poids des arêtes qui lui

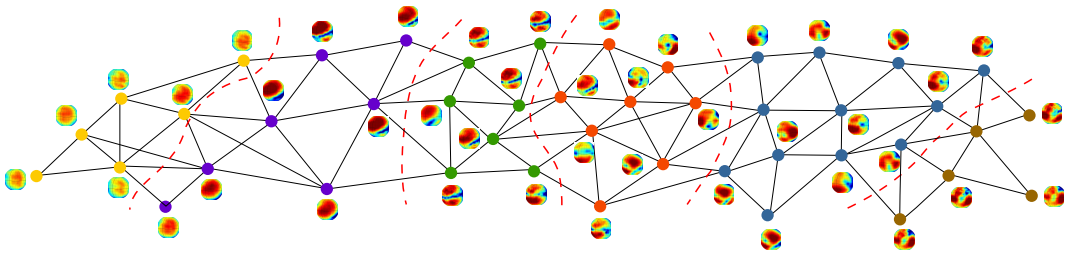


Figure 6.11 – La classification consiste à découper le graphe en *clusters*, de telle manière que la somme des poids des arêtes coupées soit minimal.

sont connectés. Si l'on excite ce système mécanique, les groupes de masses interconnectées de façon assez rigide par des ressorts de raideur élevée oscillent ensemble. A mesure que la fréquence de l'oscillation forcée augmente, les ressorts de faible raideur connectés à de tels groupes s'étendent plus fortement, et les *clusters* se dégagent alors naturellement. La détermination du partitionnement d'un tel graphe passe donc par l'analyse modale de ce système mécanique : il apparaît que les modes fondamentaux sont précisément décrits par les vecteurs propres du Laplacien $\mathcal{L}$, calculés dans l'algorithme de classification [237].

Détermination des prototypes

La classification permet de constituer des *clusters*, partitionnant l'ensemble des données en regroupant les SFRS semblables. On propose de définir chacune des κ SFRS prototypiques comme la moyenne des SFRS de chaque *cluster*. Cependant, si les SFRS d'un même *cluster* sont particulièrement semblables, c'est moyennant une rotation d'alignement. Ainsi, dans chaque *cluster*, une étape préalable d'alignement de toutes les SFRS sur l'une d'entre elles est nécessaire. Cette SFRS particulière, ou représentante, est choisie comme étant celle qui maximise la similarité moyenne avec les autres éléments du *cluster*. Chaque rotation d'alignement est en première approximation donnée par la rotation R_Υ, celle qui permet d'atteindre le maximum de l'intercorrélation. Néanmoins, la précision de cet alignement dépend du pas angulaire de l'échantillonnage discret du groupe $SO(3)$. On peut approcher plus précisément l'alignement optimal en explorant continument le groupe $SO(3)$ selon un algorithme de descente du gradient, tel que celui décrit en 5.4.1. Il convient alors d'initialiser l'algorithme avec la rotation R_Υ.

Une fois les SFRS alignées dans chaque *cluster*, la moyenne peut être calculée. Finalement, les κ SFRS prototypiques sont stockées en mémoire, en vue d'être utilisées pour l'étape de reconstruction, et ordonnées selon la moyenne des fréquences auxquelles les SFRS du *cluster* correspondaient originellement.

6.2.3 Reconstruction

L'étape de reconstruction comporte deux phases :

1. une phase de reconnaissance de formes, qui identifie pour chaque SFRS mesurée le prototype le plus proche ;
2. une phase de correction, qui permet d'annuler l'erreur commise dans les directions de mesure.

Reconnaissance de formes

Supposons que les HRTF d'un nouvel individu ont été mesurées sur un échantillonnage grossier de la sphère. Les SFRS $\hbar$ correspondantes sont considérées, et celles de l'oreille gauche sont symétrisées. Pour chacune de ces SFRS, une comparaison est effectuée avec les SFRS prototypiques : c'est la SFRS prototypique h_{prot} offrant la similarité $\Upsilon(\hbar, h_{prot})$ maximale qui est retenue et associée au bin fréquentiel de $\hbar$. De façon à réduire les calculs, la comparaison n'est effectuée qu'avec les SFRS prototypiques h_{prot} correspondant à un *cluster* de fréquence moyenne proche du bin fréquentiel de $\hbar$, selon le même principe que lors de l'analyse. Par ailleurs, un seuillage angulaire, tel que celui utilisé dans l'étape d'analyse, permet d'écarter les associations incorrectes, car correspondant à une rotation R_Υ d'angle trop important. Comme dans la phase de création des prototypes, un alignement rotationnel de la SFRS prototypique choisie est réalisé pour l'adapter à l'orientation de la SFRS individuelle mesurée. On nomme $\mathring{h}_{prot}$ le résultat de l'alignement de h_{prot}.

Rappelons que le calcul de la similarité entre SFRS mesurées et SFRS prototypiques est précédé d'une décomposition de ces données sur une base d'harmoniques sphériques. Le degré maximal B qui peut être atteint dans cette décomposition est intimement lié au nombre de directions pour lesquelles les HRTF sont disponibles (cf. 5.4.1). Ce degré maximal représente en quelque sorte l'analogue, dans le cas d'un spectre spatial, de la fréquence de Nyquist pour un signal temporel. Le degré B décroît donc à mesure que l'échantillonnage spatial de mesure devient plus grossier, et le repliement spatial est de plus en plus probable d'intervenir dans la décomposition. C'est pourquoi, comme c'est le cas pour toutes les méthodes de reconstruction du type interpolation, on s'attend à ce que la technique proposée soit de moins en moins performante, ce qui est naturel dès lors que l'information individuelle disponible se raréfie.

Correction

Si l'on remplace simplement chaque SFRS mesurée par son prototype le plus proche aligné par rotation $\mathring{h}_{prot}$, les données ainsi reconstruites sont potentiellement

entachées d'erreurs dans les N_{mes} directions de mesure $\{\chi_i\}_{i=1,\dots,N_{mes}}$. L'objet de l'étape de correction est précisément de rendre nulles ces erreurs $h_{err}(\chi_i)$, que l'on peut directement calculer comme la différence, dans les directions de mesure, entre les données d'entrée $\hbar(\chi_i)$ et les SFRS prototypiques alignées $\mathring{h}_{prot}(\chi_i)$. On fixe un échantillonnage cible $\{\chi_i^*\}_{i=1,\dots,N_{cible}}$ sur lequel on veut obtenir *in fine* les HRTF. On obtient pour chaque bin fréquentiel les SFRS reconstruites h_{rec} dans ces directions en retranchant h_{err} de $\mathring{h}_{prot}$ (les valeurs $h_{err}(\chi_i^*)$ étant obtenues par interpolation STPS, et les valeurs $\mathring{h}_{prot}(\chi_i^*)$ d'après la décomposition en harmoniques sphériques de $\mathring{h}_{prot}$) :

$$h_{rec}(\chi_i^*) = \mathring{h}_{prot}(\chi_i^*) - h_{err}(\chi_i^*), \quad i = 1, \dots, N_{cible}$$

Enfin les SFRS de l'oreille gauche sont à nouveau symétrisées, pour revenir au système de référence originel.

La littérature nous indique que malgré l'existence d'une contribution conjointe des IS des oreilles gauche et droite dans la perception de l'élévation d'une source, les IS de l'oreille controlatérale contribuent de moins en moins à mesure que la source se déplace vers le côté ipsilatéral (cf. 3.2.8). Cette contribution devient nulle dès que la valeur absolue de l'azimut θ d'une source est supérieure à 30° (système de coordonnées dit *polaire horizontal*, cf. Fig. 1) [181]. Par ailleurs les IS de l'oreille controlatérale dans cette zone de l'espace sont particulièrement accidentés. Cela est dû conjointement à l'existence d'effets physiques complexes apparaissant dans l'ombre de la tête, mais également au rapport signal à bruit plus réduit dans cette zone de l'espace lors de la mesure. Toutes ces remarques nous poussent à écarter du modèle les HRTF dans ces directions. Ainsi, pour toutes les directions de mesure appartenant à cette calotte controlatérale, les valeurs des SFRS $\mathring{h}_{prot}$ sont remplacées par les données mesurées $\hbar$ avant le calcul de h_{rec}.

6.3 Evaluation objective

La technique proposée repose sur la préexistence d'une grande base de données de HRTF. Nous constituons celle-ci d'après la fusion de 4 bases de données différentes : celle du CIPIC [161], la base *Listen* de l'IRCAM [103], celle de l'Université du Maryland [162], et enfin celle d'Orange Labs. Au total, on dispose donc des HRTF de 109 sujets : les HRTF des deux oreilles de 101 sujets servent à constituer la base de données sur laquelle on effectue l'analyse, tandis que les données des 8 sujets restants sont mises de côté, pour l'évaluation objective. Douze échantillonnages spatiaux de taille décroissante sont considérés, comptant de 130 à 20 directions environ. Les performances de la technique proposée sont comparées, selon différents critères, à celles de l'interpolation STPS proposée par Hartung [82], ainsi qu'à celles de la

méthode hybride proposée par Carlile *et al.* [48], basée sur l'interpolation STPS des poids d'une décomposition ACP.

6.3.1 Constitution de la base de données

Les HRTF des 4 bases de données originales ont été mesurées avec des fréquences d'échantillonnage temporel diverses. Une première étape consiste donc à obtenir ces données sur un échantillonnage fréquentiel commun, par une interpolation spline sur l'axe fréquentiel appliqué indépendamment pour chaque direction. De plus, les échantillonnages spatiaux différent également, c'est pourquoi une interpolation STPS est effectuée de façon à disposer de toutes les HRTF aux mêmes directions. On choisit l'échantillonnage cible de 1602 directions, obtenu selon la méthode de projection octahédrique proposée par Bronkhorst [29]. Les HRTF de la calotte inférieure, que la mesure n'a pas permis de déterminer, sont obtenues lors de cette phase d'interpolation. Bien que le résultat n'ait aucune valeur physique, ces données sont nécessaires pour calculer convenablement la décomposition en harmoniques sphériques. Cette décomposition est obtenue jusqu'à l'ordre 30. De façon à rendre toutes les SFRS comparables, les SFRS correspondant à des oreilles gauches subissent une symétrie par rapport au plan médian. Enfin, pour chaque oreille des 101 sujets, seules $N_s = 37$ SFRS régulièrement espacées sur la bande de fréquences [4 kHz - 13 kHz] sont retenues.

6.3.2 Echantillonnage spatial

De façon à évaluer l'impact de la raréfaction des données en entrée de la technique proposée, 12 sous-ensembles de taille décroissante sont constitués à partir des données mesurées pour chaque oreille des 8 sujets retenus pour évaluer le modèle (sujets n°1 et 5 de la base d'Orange Labs, sujets *dz*, *eg*, et *nm* de la base de l'Université du Maryland, et sujets n°12, 40 et 124 de la base du CIPIC). Pour conserver une distribution spatiale homogène, les directions retenues dans les échantillonnages spatiaux originaux sont choisies comme étant les plus proches des directions d'un échantillonnage obtenu comme solution du problème de *sphere covering* [79]. On obtient ainsi 12 échantillonnages différents, contenant respectivement environ 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, et 130 directions [3]. Pour réaliser les comparaisons

3. Ce sont les nombres de directions des échantillonnages "consignes" obtenus par *sphere covering* sur toute la sphère, mais comme les bases de données originales ne comportent pas de HRTF dans certaines zones de l'espace, les 12 échantillonnages utilisés contiennent en pratique un peu moins de données. Pour les données issues de la base d'Orange Labs, ces 12 échantillonnages sont constitués respectivement de 19, 27, 36, 45, 56, 65, 73, 82, 91, 101, 109, et 121 directions. Pour les données issues de la base du CIPIC : 19, 27, 36, 45, 56, 65, 73, 82, 90, 101, 109, et 119 directions. Pour les données issues de la base de l'Université du Maryland : 19, 27, 36, 45, 55, 65, 72, 82, 90, 99, 108, et

avec les prototypes, une décomposition en harmoniques sphériques est nécessaire. Il faut alors générer par interpolation STPS les HRTF dans quelques directions de la calotte inférieure. La décomposition est calculée jusqu'au degré maximal que permet le nombre de directions disponibles. Enfin, les SFRS des oreilles gauches sont symétrisées par rapport au plan médian.

6.3.3 Réglages

Le réglage de quelques paramètres est nécessaire dans la technique proposée. Ils sont choisis de façon empirique d'après les résultats de tests préalables non détaillés dans ce document. On fixe à $k = 40$ le nombre de plus proches voisins retenus pour composer la matrice d'adjacence $\mathbf{W}$. Le nombre de *clusters* recherchés dans la phase d'analyse est fixé à $\kappa = 300$, ce qui entraîne la création d'autant de SFRS prototypiques. Les données correspondant aux bandes de fréquences non considérées par la technique proposée - en-dessous de 4 kHz et au-dessus de 13 kHz - sont obtenues par simple interpolation STPS.

La technique proposée par Carlile est évaluée parallèlement. La base de vecteurs propres nécessaire dans cette technique est générée d'après l'ACP de HRTF mesurées sur les oreilles de 6 sujets, non inclus dans l'ensemble des 8 sujets choisis pour l'évaluation (cf. Annexe I). Seuls les 8 premiers vecteurs propres et les poids associés sont retenus. L'analyse des valeurs propres montre qu'ils englobent 95% de la variance totale. Rappelons que la reconstruction des données pour une direction quelconque de l'espace est obtenue par interpolation STPS des poids associés à la décomposition ACP.

6.3.4 Prototypes

On représente en Annexe H les 300 SFRS prototypiques générées par l'analyse de la base de données, classées par ordre croissant de la fréquence moyenne des *clusters* dont elles sont issues. Rappelons que ces prototypes ont été formés comme la moyenne de SFRS détectées comme semblables indépendamment de leur fréquence et de leur orientation spatiale. Grossièrement, on peut donc considérer que chaque prototype représente la SFRS associée à une certaine forme intrinsèque de pavillon, et pour un échelon donné de leur évolution fréquentielle. On observe que les prototypes obtenus constituent un éventail de formes variées, et présentent une gamme de changements subtils dans la forme des lobes de directivité. Cet alphabet semble donc bien adapté à l'utilisation qui en est faite dans la phase de reconstruction.

117 directions.

6.3.5 Critères d'évaluation

Quatre critères objectifs sont définis pour quantifier et comparer les performances de reconstruction de chacune des techniques étudiées.

Racine carrée de l'erreur quadratique moyenne (*Root Mean Square Error* ou RMSE)

L'erreur RMSE, que l'on notera ϵ_{rmse}, exprimée en dB, est évaluée à partir de la différence fréquence par fréquence entre les spectres d'amplitude des HRTF originales H_0 et reconstruites H_{rec}, soit sur N bins fréquentiels $\{\nu_i\}$ de la bande d'intérêt ([4kHz - 13kHz]). Le résultat est moyenné pour les M directions d'évaluation $\{\chi_j\}$: il s'agit de l'ensemble des directions disponibles dans la base de données originale, hormis celles choisies comme données d'entrée des techniques de reconstruction. Le calcul est décrit par la relation suivante :

$$\epsilon_{rmse} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[H_0(\nu_i,\chi_j) - H_{rec}(\nu_i,\chi_j)\right]^2}$$

ISSD

L'ISSD décrite en 5.4.1 est utilisée comme critère objectif décrivant la qualité de reconstruction des HRTF, sur la bande fréquentielle [4 kHz - 13 kHz]. Rappelons que Middlebrooks a montré le lien entre cette erreur et une dégradation sensible de la spatialisation en synthèse binaurale [170].

Erreur maximale évaluée par tiers d'octave du côté ipsilatéral

Langendijk *et al.* [130] proposent une mesure d'erreur tenant compte grossièrement de la résolution fréquentielle limitée du système auditif. D'abord, le spectre d'amplitude entre 200 Hz et 16 kHz est traité par un banc de filtres 1/3 d'octave : il en résulte $\tilde{H}_0$ et $\tilde{H}_{rec}$. La différence maximale ϵ_{max} exprimée en dB est calculée parmi les bandes de fréquence considérées :

$$\epsilon_{max}(\chi_j) = \max_i \left[\tilde{H}_0(\nu_i,\chi_j) - \tilde{H}_{rec}(\nu_i,\chi_j)\right]$$

L'évaluation est réalisée sur toutes les directions de mesure ipsilatérales de la base de données originale, hormis celles choisies comme données d'entrée.

CPA

Puisque la technique proposée se focalise sur la structure spatiale des données, on introduit une séries de critères permettant de quantifier la fidélité de la reconstruction des motifs spatiaux des SFRS. Le modèle nommé *Spectral Contrast Area*,

proposé par Jin (cf. 3.3.3) semble être le plus apte à simuler les capacités du système auditif à exploiter les IS. On propose donc d'évaluer la précision avec laquelle sont reconstruites les *Covert Peak Area* (CPA). Les CPA sont ici définies comme les portions de la sphère sur lesquelles le spectre d'amplitude des HRTF est à 1.5 dB près égal au maximum spatial observé. On s'intéresse à la position et à l'étendue des CPA pour une série de bins fréquentiel (28 bins sur la bande [4 kHz - 13kHz]). Deux quantités objectives sont introduites : l'angle θ_{CPA} entre les centroïdes des CPA des HRTF originales et reconstruites, ainsi que le pourcentage $\cap_{CPA}$ des CPA d'origine recouvert par les CPA des données reconstruites.

6.3.6 Résultats

On représente respectivement figures 6.12 et 6.13 l'erreur RMSE et l'ISSD en fonction du nombre de directions de mesure. L'évolution de ces deux critères est très similaire : pour chacun d'eux, c'est la technique que nous proposons qui permet d'atteindre l'erreur la plus faible. L'amélioration par rapport aux autres techniques est d'autant plus importante que l'échantillonnage spatial de mesure est grossier. Si l'on prend comme référence les valeurs de l'erreur RMSE et de l'ISSD commises par la technique de Carlile *et al.* (PCA8 sur les figures) pour 130 directions mesurées, on remarque que ces mêmes valeurs sont atteintes par la technique proposée pour moins de 60 directions. Soulignons également que l'interpolation STPS montre de bonnes performances, proches de la technique proposée pour des échantillonnages de plus de 80 à 90 directions, mais divergentes pour des échantillonnages plus grossiers. Néanmoins les performances de cette interpolation simple dépassent nettement celles de la technique basée sur une décomposition ACP. Langendijk *et al.* [130] ont montré à propos de l'erreur nommée ϵ_{max}, que la valeur 2.5 dB constitue la limite approximative au-delà de laquelle des dégradations interviennent en termes de spatialisation. C'est pourquoi on représente figure 6.14, tous sujets et toutes directions confondus, le pourcentage des valeurs de ϵ_{max} dépassant la limite de 2.5 dB. Selon ce critère, la technique de Carlile *et al.* montre des performances nettement plus mauvaises que les autres techniques : quelle que soit la finesse de l'échantillonnage, on observe plus de 75% des valeurs au-delà de cette limite. En comparaison, pour la technique que nous proposons ce pourcentage reste inférieur à 30% pour des échantillonnages de mesure comportant plus de 60 directions. Pour des échantillonnages plus grossiers, la reconstruction est de 10 points meilleure à celle obtenue par interpolation STPS.

On illustre figure 6.15 pour un sujet et une fréquence donnés les performances de reconstruction des CPA pour les trois méthodes évaluées. On représente figure 6.16, tous sujets et bins fréquentiels confondus, la moyenne de l'angle θ_{CPA}. Sur

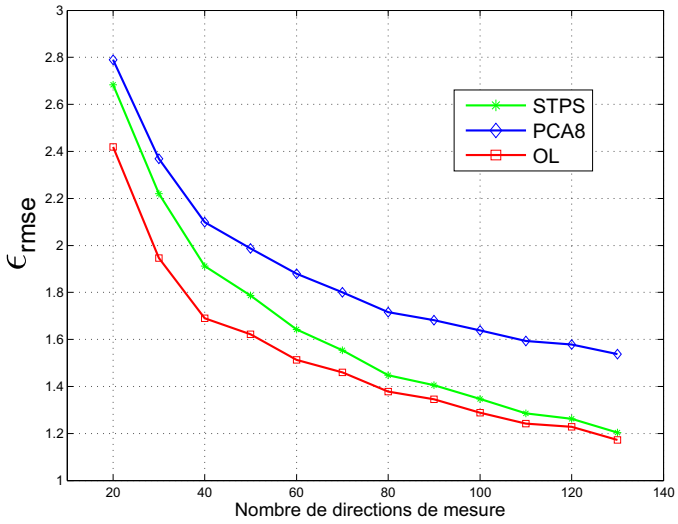


Figure 6.12 – Erreur RMSE ϵ_{rmse} en fonction du nombre de directions de mesure. On désigne par STPS l'interpolation simple, PCA8 la technique de Carlile *et al.*, et enfin OL (pour Orange Labs) la nouvelle technique proposée. Les résultats représentés sont issus du moyennage des résultats des deux oreilles des 8 sujets évalués.

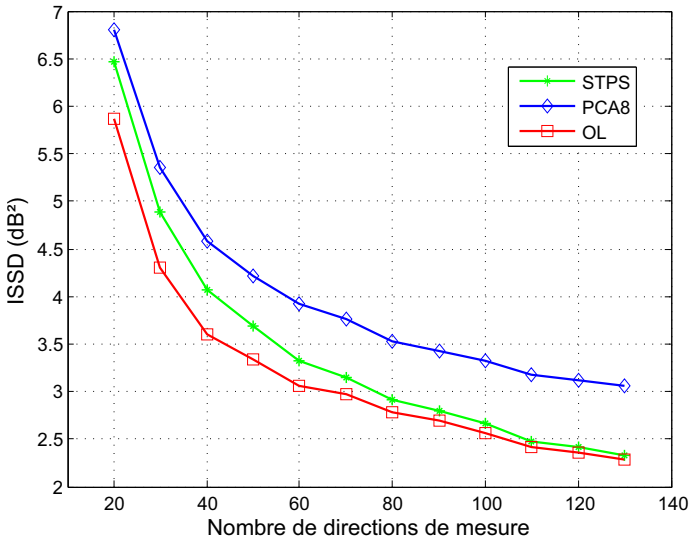


Figure 6.13 – ISSD en fonction du nombre de directions de mesure (cf. Fig. 6.12 pour les détails). Les résultats représentés sont issus du moyennage des résultats des deux oreilles des 8 sujets évalués.

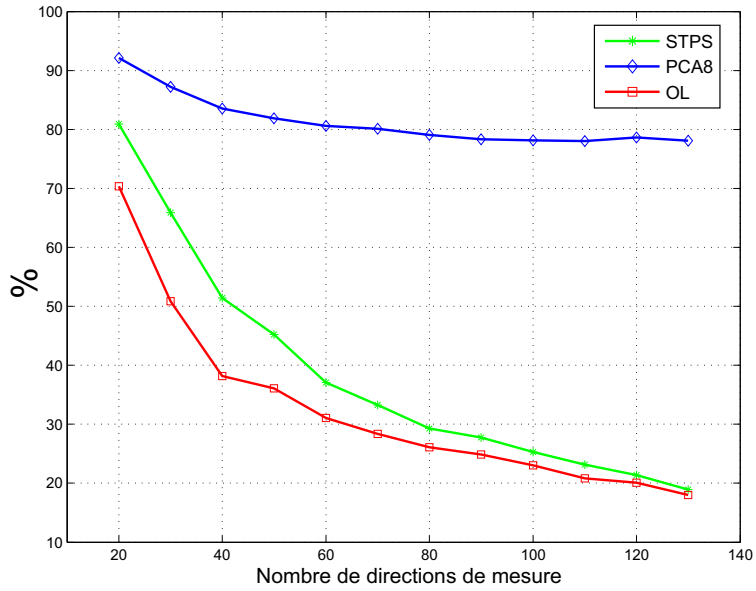


Figure 6.14 – Erreur maximale évaluée par tiers d'octave du côté ipsilatéral ϵ_{max} en fonction du nombre de directions de mesure : pourcentage de valeurs d'erreur supérieures à 2.5 dB, tous sujets et directions ipsilatérales confondue (cf. Fig. 6.12 pour les détails).

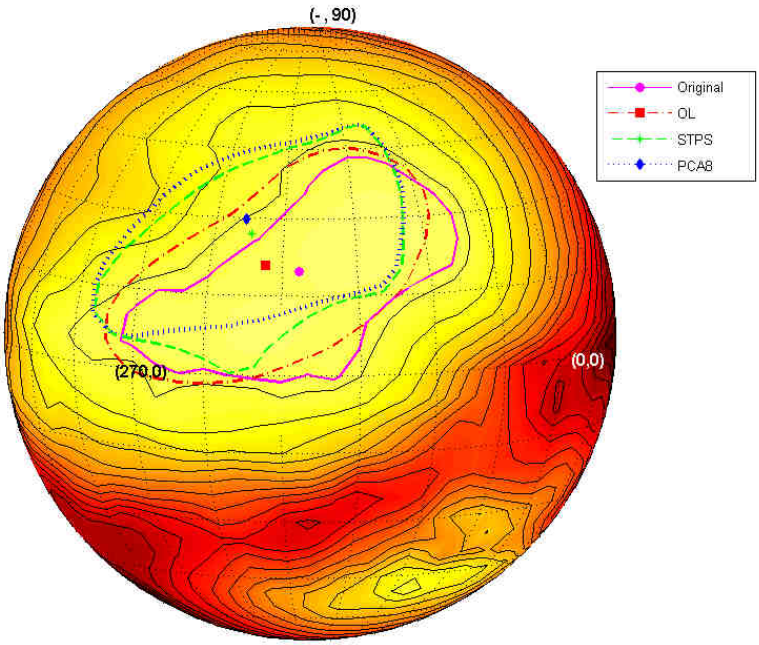


Figure 6.15 – Illustration de la reconstruction des données en termes de CPA pour un échantillonnage de 40 directions de mesure (SFRS de l'oreille droite du sujet n°1 de la base de HRTF d'Orange Labs à 7 kHz, système de coordonnées dit *polaire vertical*).

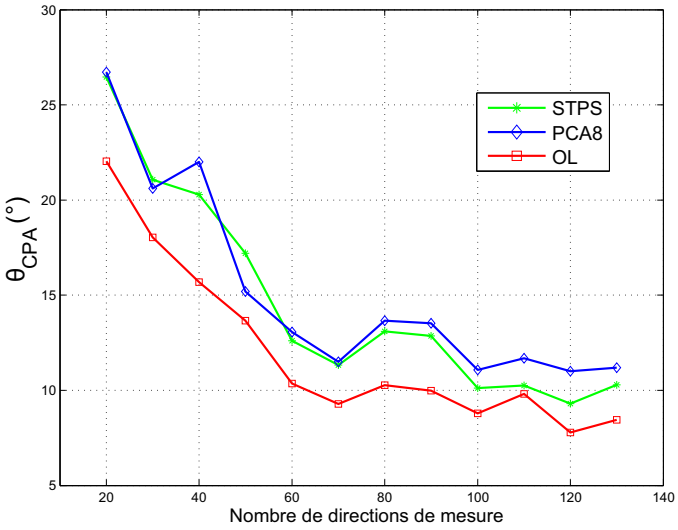


Figure 6.16 – Angle θ_{CPA} moyen entre les directions des centroïdes des CPA originales et reconstruites, en fonction du nombre de directions de mesure (cf. Fig. 6.12 pour les détails). La moyenne est calculée tous sujets et fréquences confondus.

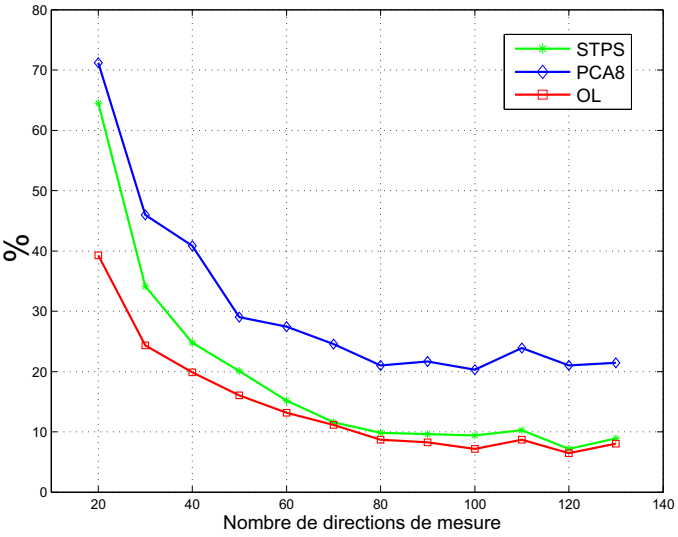


Figure 6.17 – Pourcentage des valeurs de l'angle θ_{CPA} dépassant 10° en fonction du nombre de directions mesurées (cf. Fig. 6.12 pour les détails). Le pourcentage est calculé tous sujets et fréquences confondus.

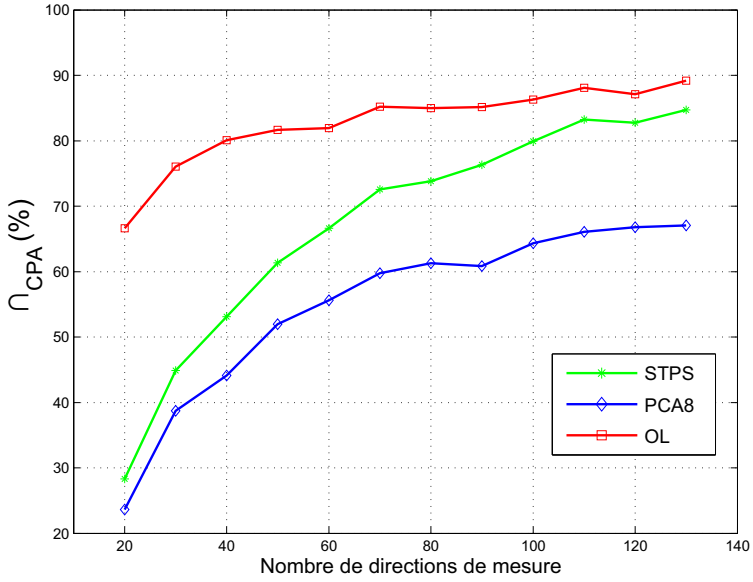


Figure 6.18 – Moyenne de la surface d'intersection $\cap_{CPA}$ exprimée en pourcentage de l'aire des CPA originales, en fonction du nombre de directions mesurées (cf. Fig. 6.12 pour les détails). La moyenne est calculée tous sujets et fréquences confondus.

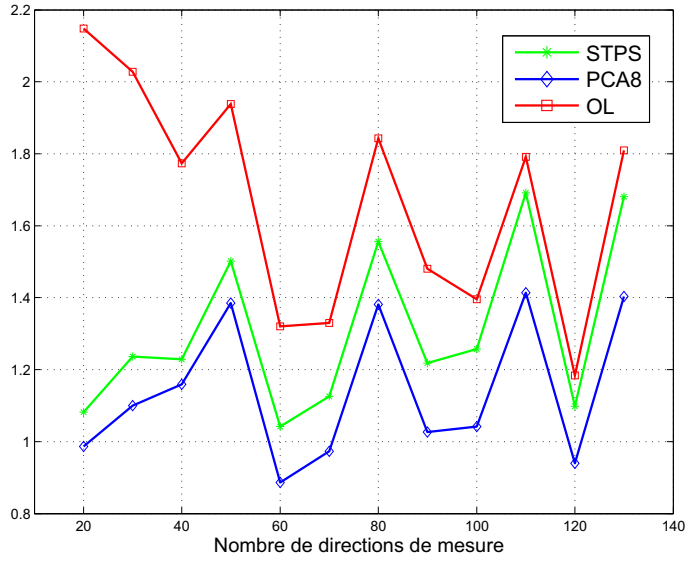


Figure 6.19 – Moyenne des ratios entre l'aire des CPA originales et celle des CPA reconstruites, en fonction du nombre de directions mesurées (cf. Fig. 6.12 pour les détails). La moyenne est calculée tous sujets et fréquences confondus.

ce point, la technique que nous proposons se révèle nettement meilleure que les deux autres méthodes évaluées : une amélioration de l'ordre de 2° à 5° est observée. Ces bons résultats sont confirmés par la distribution des valeurs de θ_{CPA} avant moyennage : on représente figure 6.17, tous sujets et bins fréquentiels confondus, le pourcentage des angles θ_{CPA} prenant une valeur supérieure à 10°, en fonction du nombre de directions de mesure. Cette valeur de 10° est choisie comme valeur représentative du MAA (*Minimum Audible Angle*) pour la discrimination de l'azimut et de l'élévation [176, 190]. La technique que nous proposons montre une amélioration de 10 points par rapport à la méthode de Carlile *et al.*. Au mieux, cette dernière atteint la valeur de 20% pour un échantillonnage de 130 directions de mesure, valeur atteinte par la technique proposée pour un échantillonnage de 40 directions. On représente figure 6.18 le pourcentage $\cap_{CPA}$ de la surface des CPA originales recouvert par celles des données reconstruites, et figure 6.19 le ratio entre leur aires. Dans le cas de la technique proposée, l'intersection représente entre 80 et 90% de l'aire des CPA, même pour un échantillonnage de 40 directions mesurées. Pour la méthode de Carlile *et al.*, ce pourcentage décroît de 65% à 45% quand le nombre de directions de mesure décroît de 130 à 40. L'interpolation STPS offre des performances proches de la technique proposée pour des échantillonnages fins, et plus proches de celles de la méthode de Carlile *et al.* pour des échantillonnages grossiers. Ce résultat doit être contrebalancé par l'observation de la figure 6.19, qui montre que la technique proposée a tendance à produire des CPA d'aire 1.8 à 2 fois supérieure à celle des CPA originales, tandis que ce facteur reste autour de 1 à 1.4 pour les deux autres méthodes. Cela peut expliquer en partie pourquoi le recouvrement spatial est plus important.

6.4 Evaluation subjective

Les résultats de l'analyse objective montrent clairement l'apport de la technique proposée par rapport à l'état de l'art. Néanmoins une évaluation subjective est nécessaire pour déterminer le nombre minimal de mesures requises en entrée de la méthode, afin d'assurer une spatialisation convenable en synthèse binaurale.

6.4.1 Protocole expérimental

Classiquement, l'évaluation subjective d'un jeu de HRTF est réalisée grâce à un test de localisation en synthèse binaurale statique. C'est en particulier la méthode adoptée par Wightman et Kistler [273] et Carlile *et al.* [48, 49]. Dans ce cas, les résultats sont analysés en termes de taux de confusion avant/arrière, ainsi qu'en termes de précision globale de localisation, qui peut être résumée par une quantité statis-

tique nommée *Spherical Coefficient Correlation* ou SCC [48]. Ce type de protocole constitue désormais la référence pour l'évaluation d'un VAS, mais il comporte un inconvénient important : la synthèse binaurale statique est peu écologique, car tous les indices dynamiques sont éliminés. Ces indices se révèlent pourtant très utiles en situation d'écoute naturelle, et ils participent à la perception extra-crânienne des sources et à la discrimination avant/arrière. Le choix de la synthèse binaurale statique est donc malvenu, dans la mesure où l'apparition d'artefacts de perception peut être imputée au protocole lui-même, et non aux HRTF qu'il s'agit d'évaluer, comme l'a fait remarquer Carlile [47]. Nous proposons donc d'utiliser la synthèse binaurale dynamique. Traditionnellement, cette technique de diffusion est écartée à cause de la complexité de sa mise en œuvre, mais aussi parce que les indices dynamiques peuvent être considérés comme gênants, au sens où il sont susceptibles de troubler l'analyse des résultats. En particulier, si le sujet a la liberté de bouger la tête pendant l'écoute des stimuli, il peut facilement résoudre les confusions avant/arrière, et il devient donc impossible d'évaluer la capacité des HRTF elles-mêmes à fournir les informations nécessaires à cette discrimination spatiale. En revanche, la qualité des IS en termes de codage de l'élévation peut tout à fait être évaluée. En effet, il a été montré que les indices dynamiques se révèlent incapables de fournir des informations suffisantes sur l'élévation d'une source pour contrebalancer des indices spectraux incorrects [68].

Nous proposons dans ce cadre un nouveau protocole pour évaluer la qualité de la spatialisation offerte par un jeu donné de HRTF, qui s'apparente à ceux proposés par Chen [52] et Yairi *et al.* [280] dans des optiques différentes. Si les stimuli sont diffusés suffisamment longuement, et en synthèse binaurale dynamique, on sait qu'un auditeur peut atteindre une précision optimale en amenant la source en position frontale, qui est la zone de l'espace selon laquelle la précision est maximale. On pourrait donc choisir de ne diffuser les stimuli que brièvement, mais on préfère favoriser une écoute prolongée, de façon à ce que les sujets perçoivent les scènes sonores dans leur globalité. On ne s'intéresse donc pas à la précision de localisation qu'un sujet peut atteindre dans un temps limité, mais plutôt à la rapidité avec laquelle il peut localiser une source avec une précision donnée. On fait l'hypothèse que la mesure du temps de réponse constitue une évaluation de la qualité de la spatialisation.

Plusieurs méthodes ont été proposées pour permettre au sujet de reporter la direction perçue d'une source sonore virtuelle [198]. Parmi ces méthodes, la plus écologique est celle qui consiste pour le sujet à orienter le visage dans la direction de la source [49]. En effet, selon Poincaré [203], *"localiser un objet, cela veut dire simplement se représenter les mouvements qu'il faudrait faire pour l'atteindre ; [...] il ne s'agit pas de se représenter les mouvements eux-mêmes dans l'espace, mais uniquement de se représenter les sensations musculaires qui accompagnent ces mouvements [...]. "*. Il semble donc naturel de donner au sujet la possibilité de traduire son per-

cept de localisation par la mise en œuvre de ces mouvements, dont on peut attendre qu'ils le mèneront face à la source sonore d'autant plus rapidement que le percept spatial initial aura été net et en rapport avec sa direction. Les stimuli à localiser sont diffusés continuellement jusqu'à la validation, qui intervient automatiquement quand la précision voulue est atteinte de façon stable. Le test fait intervenir différents jeux de HRTF : les HRTF individuelles des sujets, issues directement de la mesure, et les HRTF reconstruites à partir de la technique proposée. Les dégradations éventuelles de la spatialisation engendrées par un jeu donné de HRTF sont à chercher dans les différences observées en termes de temps de réponse par rapport au cas où les HRTF individuelles sont utilisées, considéré ici comme la condition de référence [4]. Le protocole proposé offre aux sujets la possibilité d'explorer activement les scènes sonores, et ainsi éventuellement de se familiariser avec un jeu de filtres binauraux. De plus, la validation du succès de la tâche de localisation est automatique, et il peut donc apparaître un apprentissage par rétroaction (ou *feedback*) du lien existant entre les indices de localisation fournis par les HRTF testées et les directions de l'espace correspondantes. Plusieurs expériences ont en effet souligné l'existence d'une certaine plasticité du système auditif face à des filtres binauraux non-individuels (cf. 4.2.7). Cet effet d'adaptation est en général considéré comme un biais à éviter si l'on cherche à évaluer la qualité brute des HRTF. Cependant, si cet effet bénéfique intervient dans un usage pratique de la synthèse binaurale, il est parfaitement illogique de l'occulter lors d'une évaluation psychophysique. Nous faisons le choix d'assumer les conséquences potentielles de l'apprentissage par rétroaction, et il faut donc garder à l'esprit que l'évaluation perceptive porte ici sur la qualité de HRTF avec lesquelles les sujets auront eu l'occasion de se familiariser brièvement.

Le déroulement de l'expérience est décrit figure 6.20. Une série de positions de test sont choisies à distance fixe du sujet, et tout autour de lui. Pour chacune d'elles une première source à atteindre, dite frontale, est générée et positionnée dans la direction (0°,0°) du repère absolu, c'est-à-dire directement devant le sujet quand il adopte la position de repos. A chaque instant, le système mesure la position de la tête du sujet au moyen du dispositif de *head-tracking* nécessaire à la diffusion de la synthèse binaurale dynamique. Il suffit que le sujet pointe l'axe médian de façon stable dans une direction incluse dans un cône d'angle limité, centré sur la direction de la source, pour que le système valide automatiquement le succès de la tâche de localisation. La première source s'arrête alors, et une seconde source, dite source de

4. Idéalement, si l'on suit les principes énoncés en 4.1.1, on devrait d'abord valider la qualité du VAS créé avec ces HRTF individuelles, grâce à une évaluation des performances des sujets en écoute naturelle en champ libre, qui doit toujours être considérée comme la condition de référence. Il faudrait pour cela utiliser un système de haut-parleurs distribués sur une sphère. Nous ne disposions pas d'un tel système pour cette expérience.

test, est ensuite générée et placée dans la direction de test. L'objectif du sujet est à nouveau d'orienter la tête dans sa direction, mais cette fois-ci le plus rapidement possible. Le dispositif de test enregistre la trajectoire suivie par le sujet pendant que la source de test est active, c'est-à-dire jusqu'à sa validation. Cette séquence source frontale - source de test est répétée autant de fois qu'il y a de directions à tester. L'intérêt de la source frontale, générée avant chaque source de test, est de forcer le sujet à orienter sa tête dans une direction initiale quasiment identique, ce qui est nécessaire pour l'analyse des trajectoires et des temps de réponse. Par ailleurs, on sait que le fait de pointer le visage dans une direction avec l'axe médian n'est pas naturel, car en situation réelle on a tendance à combiner les mouvements des yeux avec ceux de la tête pour s'orienter dans une direction donnée, en particulier si elle est très haute ou très basse. Ce problème a déjà été soulevé dans plusieurs études utilisant la même technique de report de localisation [49]. Chaque présentation de la source frontale, est donc l'occasion pour le sujet de s'améliorer dans la tâche de pointage, car il peut se remémorer le port de tête qui lui permet de faire coïncider son axe médian avec la direction (0°,0°) du repère absolu.

6.4.2 Mise en œuvre

Dispositif de test

L'interface de test est réalisée sous *Virtools*®, un logiciel de création de mondes 3D en temps réel. Un système de *head-tracking* magnétique, le *Polhemus Fastrak*®, envoie à l'interface les informations d'orientation et de position de la tête de l'auditeur. La synthèse binaurale dynamique est mise en œuvre grâce à ces informations, qui sont exploitées par des briques logicielles de spatialisation développées antérieurement [197]. Pendant le test, la représentation en 3D d'une tête est affichée à l'écran, et mime les mouvements du sujet en temps réel (cf. Fig. 6.21).

Le système de *head-tracking* est composé de 4 éléments :

- la partie électronique, qui communique avec l'ordinateur via un port série ;
- le moteur, positionné horizontalement, qui crée le champ magnétique dans lequel évoluent les objets dont on veut mesurer la position ;
- un pointeur, muni d'un bouton, et dont l'extrémité permet de mesurer précisément un point de l'espace ;
- un capteur, positionné sur la tête, et chargé de suivre ses mouvements.

Pendant le test, le sujet est installé sur une chaise tournante. Il a donc la liberté de combiner des mouvements de la tête et des mouvements du corps pour faire face aux sources virtuelles. On dispose un casque réglable sur sa tête, sur lequel est fixé le capteur de position. Le capteur étant ainsi parfaitement solidaire de la tête, il mesure bien l'évolution de sa position et de son orientation.

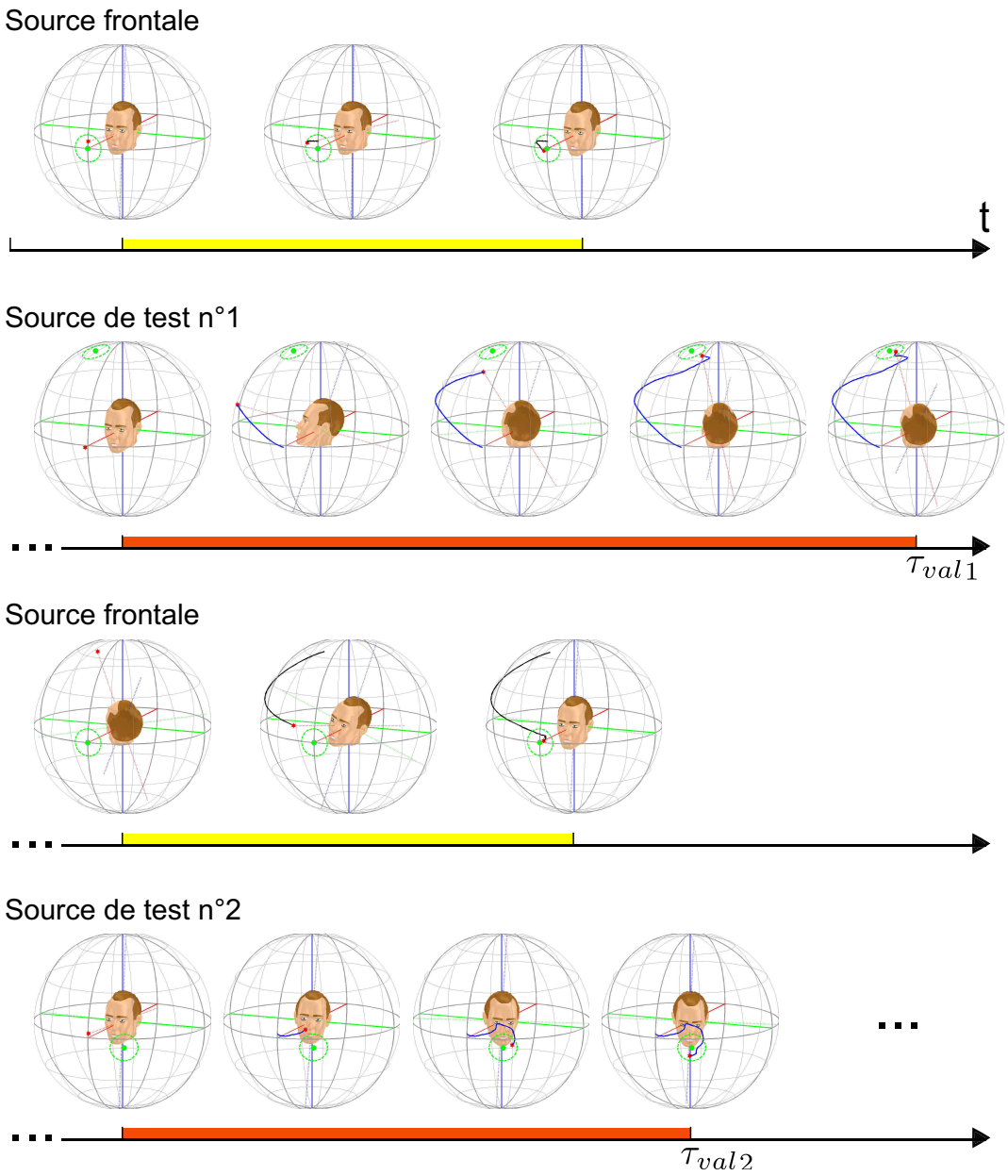


Figure 6.20 – Description séquentielle du test de localisation. Pour chaque source de test on enregistre la trajectoire adoptée par l'extrémité de l'axe médian du sujet (en bleu), ainsi que le temps τ_{val} nécessaire pour obtenir la validation. Les points verts matérialisent la direction de chacune des sources, et les cercles en pointillés qui les entourent représentent les bases des cônes de validation. Les sujets doivent pointer l'axe médian de façon stable dans ces cônes, pendant une durée donnée, pour que le système valide leur succès. Les zones colorées sur l'axe temporel indiquent les périodes d'activité des sources sonores.

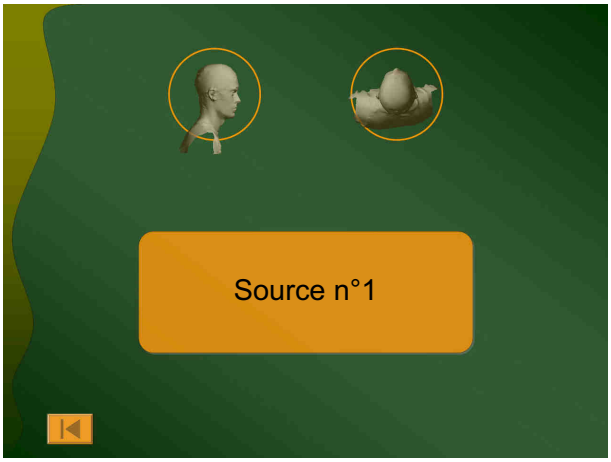


Figure 6.21 – Interface de test réalisée sous *Virtools*®. La représentation en 3D d'une tête mime les mouvements du sujet en temps réel.

Une étape de calibration est nécessaire pour chaque sujet, au début de chaque session de test. Il s'agit de définir les axes du référentiel associé à la tête du sujet. Pour cela le sujet est invité à se maintenir immobile, le regard à l'horizon, et perpendiculaire à l'écran qui lui fait face. L'expérimentateur mesure alors à l'aide du pointeur la position de chacun de ses *tragus*. Le système déduit la position du centre de la tête ainsi que l'axe interaural. L'axe médian est obtenu sans mesure supplémentaire : il est défini comme l'axe passant par le centre de la tête, perpendiculaire à l'axe interaural, et inscrit dans un plan horizontal, c'est-à-dire parallèle à celui sur lequel est posé le moteur magnétique. La période d'échantillonnage pour l'enregistrement des positions des capteurs est de 30 ms, ce qui constitue aussi la précision avec laquelle sont connus les temps de réponse.

La synthèse binaurale est diffusée sur un casque circum-auriculaire ouvert, de type *Sennheiser HD600*®, dont les HPTF ont été mesurées pour chacun des sujets du test (cf. *infra*). La carte son utilisée est de type *TerraTec Phase 26 USB*™, et interfacée selon un protocole ASIO. Les briques logicielles de synthèse binaurale dynamique sont paramétrées au début de chaque série de positions testées, avec un jeu de HRTF stockées sous formes de filtres FIR de 256 échantillons. Le système est spécialement adapté à la base privée d'Orange Labs, ce qui impose l'utilisation d'un jeu de 965 paires de filtres, correspondant aux directions de l'échantillonnage spatial de cette base de données. Le pas d'échantillonnage spatial étant assez réduit (5.625°), aucun dispositif d'interpolation n'est mis en œuvre. Afin d'assurer l'illusion qu'une source reste fixe dans le répère absolu quels que soient les mouvements de l'auditeur, la paire de HRTF choisie en temps réel est celle correspondant à la direction la plus proche, dans l'échantillonnage spatial disponible, de la direction égocentrique de la

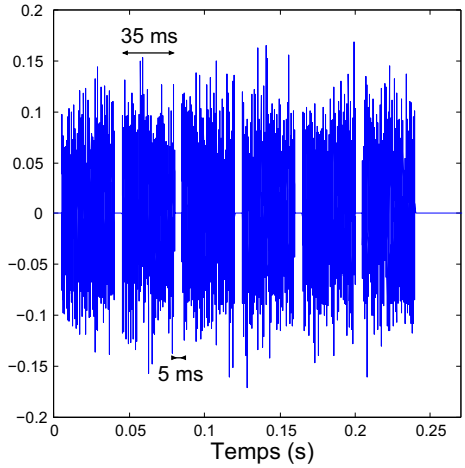


Figure 6.22 – Stimulus utilisé pour générer les sources virtuelles.

source, c'est-à-dire dans le référentiel lié à la tête de l'auditeur. La commutation entre deux filtres est réalisée selon la technique de fondu-enchaîné (ou *cross-fade*), sur une durée de 30 ms. La latence globale du dispositif de synthèse binaurale dynamique a été évaluée à 70 ms au maximum (25 ms pour le système de *head-tracking*, et 45 ms pour l'interface logicielle et la partie audio). Une telle latence est donc imperceptible d'après les travaux de Brungart *et al.* [35]. De plus, la résolution angulaire avancée par le constructeur du *head-tracker* est très fine : de l'ordre de 0.15°. Le stimulus de chaque source est commun, et créé grâce à un train de bruits blancs d'une durée totale de 270 ms, joué en boucle jusqu'à validation. Le stimulus est décrit figure 6.22 : il s'agit d'une série de 6 salves de bruits blancs gaussiens de 35 ms, espacées de 5 ms de silence, chaque série étant suivie de 30 ms de silence. Chaque salve débute et se termine par une rampe cosinusoïdale de 1 ms (en $\cos^2(t)$), afin de prévenir l'apparition de clics. Le niveau sonore de diffusion des signaux binauraux est préalablement ajusté sur la carte son, et une fois pour toutes, par l'expérimentateur. Pour cela, une source frontale est générée avec le stimulus décrit précédemment, et les HRTF individuelles des sujets participant au test. Le niveau mesuré sur l'oreille gauche d'une tête artificielle (modèle *Brüel & Kjær HATS, type 4128-C*) est compris entre 66 dB et 70 dB selon les HRTF utilisées.

En pratique, deux paramètres doivent être ajustés pour permettre aux sujets d'obtenir une validation automatique quand la direction de la source est correctement pointée : la largeur du cône de validation et la durée pendant laquelle le sujet doit y rester. D'après les résultats de tests préliminaires, on fixe à 9° le rayon du cône de validation, et à 750 ms la durée minimale pendant laquelle le sujet doit y maintenir, sans en sortir, son axe médian. Ces valeurs ont été ajustées de façon à ce qu'un

balayage rapide de l'espace ne mène pas trivialement à la validation. Une validation par chance n'est pas exclue, mais ne peut donc être obtenue qu'en effectuant des mouvements lents de la tête, ce qui s'accompagne nécessairement d'un temps de réponse assez long. Si le sujet ne parvient pas à s'orienter dans le direction de la source en moins de 30 s, la source est interrompue, et le système passe à la position de test suivante. Deux signaux sonores distincts sont utilisés pour signifier au sujet soit la validation, soit le dépassement du temps imparti.

Génération des filtres

De façon à contrôler finement les stimuli générés aux tympans des auditeurs, une calibration du casque est nécessaire. On sait que cette opération peut être périlleuse, car elle est susceptible d'entraîner des artefacts plus graves que ceux que l'on cherche à éviter (cf. 2.8). Ce sont en général des problèmes de répétabilité des mesures des HPTF qui sont incriminés. On représente figure 6.23 le spectre d'amplitude des HPTF de l'oreille gauche de chacun des 5 sujets du test, mesurées au cours des travaux de thèse de Pernaux [197] pour 10 positionnements du casque *Sennheiser HD600*, selon la méthode conduit bloqué. Pour obtenir chacun des 10 positionnements, le sujet a été invité à placer lui-même le casque de façon confortable sur ses oreilles. La répétabilité est variable selon les sujets : en particulier, l'amplitude des résonances et antirésonances connaît parfois de fortes variations d'une mesure à l'autre. Pour le sujet n°6, leur position fréquentielle montre également des variations. Les moyennes des spectres mesurés sont beaucoup moins accidentées, et semblent donc constituer une évaluation raisonnable de la réponse individuelle du casque dans l'optique d'une inversion. Ainsi, pour chaque oreille et chaque sujet, la moyenne du spectre d'amplitude des 10 HPTF mesurées est calculée, oreille par oreille, et retranchée de celui des HRTF à tester.

Pour former des filtres complets, la phase doit être également reconstituée. Carlile *et al.* ont proposé, dans une décomposition des HRTF en filtre à phase minimale et retard pur, d'obtenir les retards interauraux par interpolation STPS des retards estimés dans les directions de mesure. Une technique alternative d'implémentation a été proposée par Kulkarni *et al.* [128] : les HRTF utilisées sont à phase linéaire, et pour chaque direction, on contrôle le retard interaural τ_{gd} en jouant sur la différence entre les pentes des spectres de phase gauche et droit. Ce retard interaural est estimé d'après des HRTF mesurées, par la technique du maximum d'intercorrélation [5]. C'est ce type d'implémentation que nous retenons. On pourrait alors simplement obtenir la valeur de τ_{gd} dans une direction quelconque par interpolation des valeurs estimées

5. Le retard estimé $\hat{\tau}_{gd}$ est simplement l'argument maximum de la fonction d'intercorrélation entre les HRIR gauche et droite

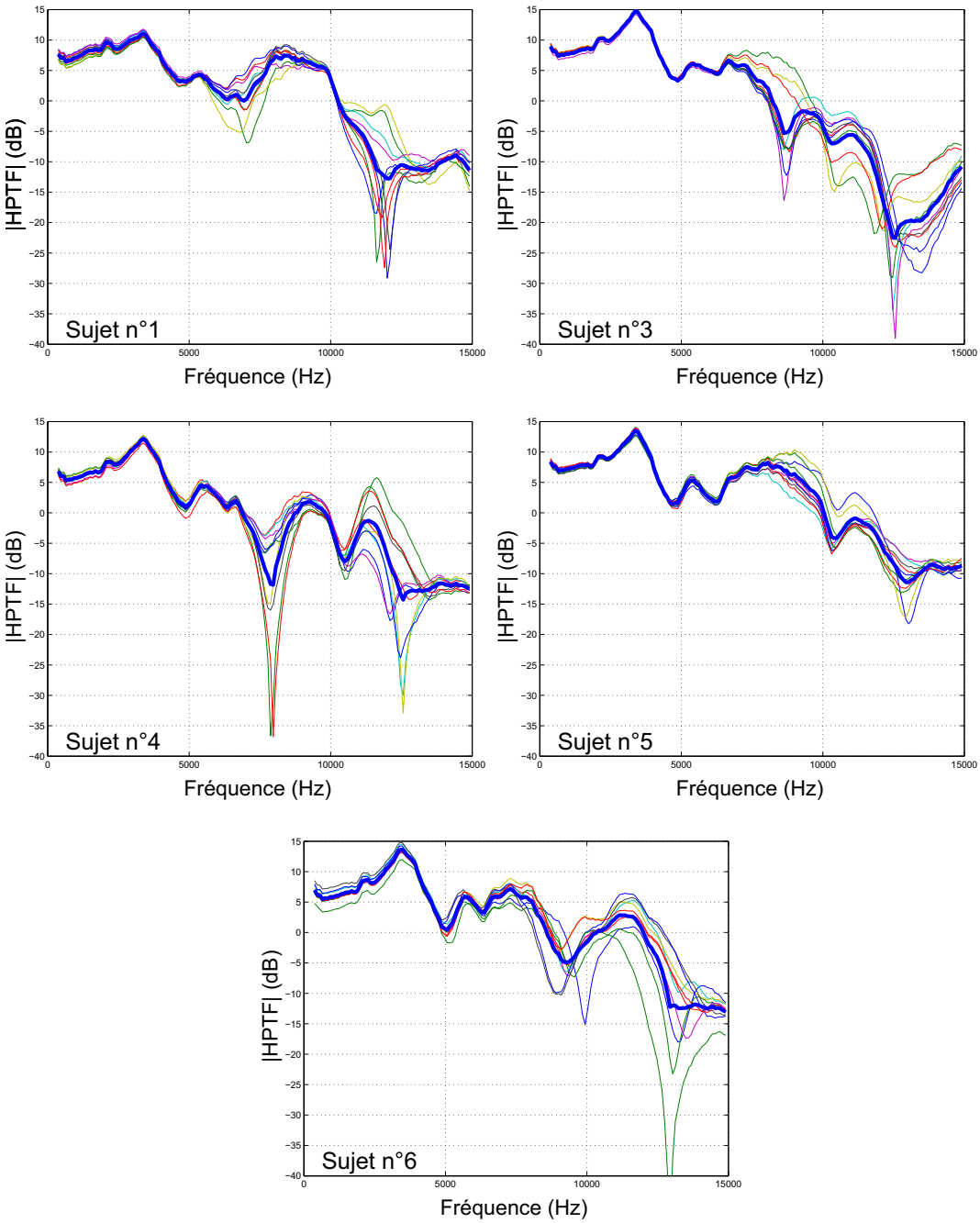


Figure 6.23 – Spectre d'amplitude des HPTF des oreilles gauches des sujets participant au test. Les courbes en trait fin correspondent aux 10 positionnements différents du casque, tandis que les courbes en trait épais représentent la moyenne de ces 10 mesures.

dans les directions mesurées. Néanmoins, des résultats de tests non présentés dans ce document montrent que l'erreur de reconstruction est importante pour les échantillonnages de mesure très grossiers. On propose donc, selon la même philosophie que pour le spectre d'amplitude, de s'appuyer sur un modèle, constituant en quelque sorte une information *a priori* sur la forme spatiale de l'ITD. On utilise le modèle d'ITD proposé par Larcher, qui est l'extension à tout l'espace du modèle de Woodworth (cf. 1.1.1). Pour une direction de l'espace, d'azimut θ et d'élévation ϕ dans le système de coordonnées dit *polaire vertical*, l'ITD est donnée par la relation :

$$ITD_{Larcher} = \frac{a}{c}(\arcsin(\cos\phi.\sin\theta) + \cos\phi.\sin\theta)$$

où a est le rayon de la sphère modélisant la tête, et c la célérité du son dans l'air. En pratique, on utilise un rayon unique : $a = 8.7$ cm. Une fonction d'erreur τ_{gderr} est calculée, dans les N_{mes} directions de mesure $\{\chi_i\}_{i=1,\ldots,N_{mes}}$, comme la différence entre l'ITD fournie par ce modèle, et le retard interaural τ_{gdmes} estimé :

$$\tau_{gderr}(\chi_i) = ITD_{Larcher}(\chi_i) - \tau_{gdmes}(\chi_i), \quad i = 1 \ldots N_{mes}$$

τ_{err} est obtenu pour une direction quelconque de l'espace par interpolation STPS, et le retard τ_{gd} implémenté est finalement donné par la relation :

$$\tau_{gd}(\chi) = ITD_{Larcher}(\chi) - \tau_{gderr}(\chi), \quad \forall \chi \in S^2$$

Les filtres FIR sont obtenus simplement par transformée de Fourier inverse des HRTF ainsi constituées. Aucun effet de salle n'est ajouté, ainsi les sources virtuelles sont diffusées en conditions anéchoïques. Notons, que préalablement à toutes ces opérations, on réalise un recalage des données par rotation du système de coordonnées, de façon à corriger d'éventuels problèmes de positionnement du sujet lors de la mesure des HRTF. La technique de recalage, basée sur une analyse de l'antisymétrie spatiale de l'ITD, est décrite en 5.2.2.

Conditions de test et organisation

L'évaluation subjective est menée sur 5 sujets de la base d'Orange Labs, ne présentant pas de problèmes d'audition : ce sont les sujets n°1, 3, 4, 5, et 6. Il s'agit de trois femmes et deux hommes, âgés de 35 à 45 ans environ. Ces sujets sont différents de ceux retenus pour l'évaluation objective, c'est pourquoi la technique de reconstruction proposée est à nouveau mise en œuvre. La base de données est donc constituée des HRTF des deux oreilles de 104 sujets, les données des 5 sujets testés étant exclues. Les autres paramètres sont les mêmes que ceux décrits en 6.3.

L'évaluation des capacités de reconstruction est réalisée pour 6 échantillonnages de mesure différents, correspondant respectivement à 19, 27, 45, 65, 82, 121 directions (cf. Fig. 6.24), choisies parmi les 965 directions de l'échantillonnage original (cf.

Fig. 6.25), comme décrit en 6.3.2. Ces conditions sont respectivement appelées R19, R27, R45, R65, R82, R121. Par ailleurs, les HRTF individuelles issues directement de la base d'Orange Labs constituent une condition supplémentaire d'évaluation, appelée I. Enfin, pour 3 des 5 sujets considérés (sujets n°1, 4 et 5), 3 jeux de HRTF supplémentaires sont utilisés comme conditions de contrôle : il s'agit de HRTF non-individuelles, issues de la base d'Orange Labs, donc connues sur le même échantillonnage spatial que les autres jeux de HRTF. Ces 3 jeux de HRTF sont choisis parmi les 7 disponibles en fonction des valeurs d'ISSD qu'ils présentent par rapport au jeu de HRTF individuelles. On retient au sens de cette distance objective le jeu le plus semblable (condition NI1), le jeu le plus différent (condition NI3), et un troisième jeu, d'ISSD moyenne [6] (condition NI2).

Une série de tests correspond à un ensemble de 35 directions de test. Ces directions appartiennent à l'échantillonnage de mesure de la base d'Orange Labs, mais ne font partie d'aucun des 6 échantillonnages de mesure considérés en entrée de la méthode de reconstruction. Elles sont réparties de façon uniforme tout autour du sujet, entre les plans d'élévation -30° et 60° [7] (cf. Fig. 6.26). Ces élévations limites sont fixées pour des raisons pratiques : il est en effet assez inconfortable pour les sujets de pointer le visage dans des directions très basses ou très élevées.

De façon à évaluer un éventuel effet d'apprentissage au cours du temps, et pour établir l'analyse statistique sur une quantité de données suffisante, l'expérience est répétée 5 fois pour chaque couple direction/jeu de HRTF : on nomme *essai* chacune de ces répétitions. Le test consiste donc au total à localiser (6+1) jeux de HRTF $\times$ 35 directions $\times$5 essais = 1225 sources de test pour 2 sujets (sujets n°3 et 6), et (6+1+3) jeux de HRTF $\times$ 35 directions $\times$ 5 essais = 1750 sources de test pour les 3 autres (sujets n°1, 4 et 5). Rappelons que ces sources sont présentées de façon entrelacée avec autant de sources frontales. Le test est organisé en séries correspondant chacune à un essai : lors d'une série de tests, un jeu unique de HRTF est testé, et les 35 directions de test sont séquentiellement présentées au sujet. Une session de test est composée de 4 à 5 séries de test et dure de 25 à 40 minutes. La première série de chaque session est une série d'apprentissage, non présentée comme telle aux sujets, et utilisant les HRTF individuelles. L'ordre de présentation des directions dans une série, est rendu aléatoire, de même que l'ordre de présentation des jeux de HRTF dans une même session [8]. Le test a duré en tout de 4h30 à 6h pour chaque sujet, et s'est déroulé

6. L'ISSD est en fait évaluée séparément entre les données des oreilles gauches et celles des oreilles droites, et c'est la moyenne de ces deux valeurs qui est considérée pour choisir les 3 jeux de HRTF non-individuelles.

7. Plus précisément, parmi les 965 directions directions possibles, on retient les 35 directions les plus éloignées de toutes les directions des 6 échantillonnages de mesure considérés, et situées entre les plans d'élévation -30° et 60° (système de coordonnées dit *polaire vertical*).

8. Néanmoins les séries correspondant aux 3 jeux de HRTF non-individuelles ont été rassemblées

sur une période de 2 mois, avec un espacement temporel irrégulier entre les sessions, planifiées selon les disponibilités des sujets.

Erreurs objectives des HRTF utilisées

L'évaluation objective présentée en 6.3 a permis de cerner des tendances générales sur l'évolution de l'erreur de reconstruction commise par la technique proposée. Il peut être utile pour étayer cette analyse subjective de connaître l'erreur objective entre les HRTF individuelles des sujets et les différents jeux de HRTF.

On représente figure 6.27 la valeur moyenne de l'ISSD pour la bande de fréquentielle [4 kHz - 13 kHz] observée sur les oreilles gauches et droites pour chacun des sujets, et chaque condition de test. L'évolution de l'ISSD en fonction du nombre de mesures individuelles est conforme aux tendances générales observées en 6.3. Les valeurs d'ISSD des HRTF non-individuelles (conditions NI1, NI2 et NI3) sont supérieures au double de l'ISSD correspondant aux HRTF reconstruites (conditions R19 à R121).

La génération complète des filtres binauraux a nécessité la réintroduction d'indices temporels sous la forme de l'ITD décrite en 6.4.2. Cette opération est également sujette à erreur, et il convient de l'évaluer. Afin de donner un sens perceptif aux erreurs d'ITD, on s'appuie sur la connaissance du seuil de discrimination de l'ITD (*Just Noticeable Difference* ou JND). La JND de l'ITD est minimale pour des sources proches du plan médian, et maximale pour des sources proches de l'axe interaural. On s'appuie sur les résultats de l'étude [122] pour obtenir la JND associée à une valeur d'ITD donnée. Pour chaque jeu de HRTF, on s'intéresse au pourcentage des directions pour lesquelles l'erreur commise sur l'ITD dépasse la limite de perceptibilité. On représente les résultats figure 6.28 pour chacun des sujets et les différentes conditions de test. On observe une évolution parfois non monotone en fonction du nombre de HRTF mesurées, ce qui suggère que le choix des directions de mesures a un léger impact sur la reconstruction des indices temporels. Là encore, pour les conditions non-individuelles, le pourcentage des directions pour lesquelles l'erreur d'ITD est perceptible est supérieur à la plus mauvaise valeur observée pour les HRTF reconstruites selon la méthode proposée.

6.4.3 Résultats : HRTF reconstruites et HRTF individuelles

Observations préliminaires

On représente figures 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35, 6.36, et 6.37 les trajectoires et la vitesse angulaire de l'extrémité de l'axe médian des sujets n°3, 5 et

dans les toutes dernières sessions, de façon à ce que le test soit comparable entre les 5 sujets pour les conditions R19 à R121, et I.

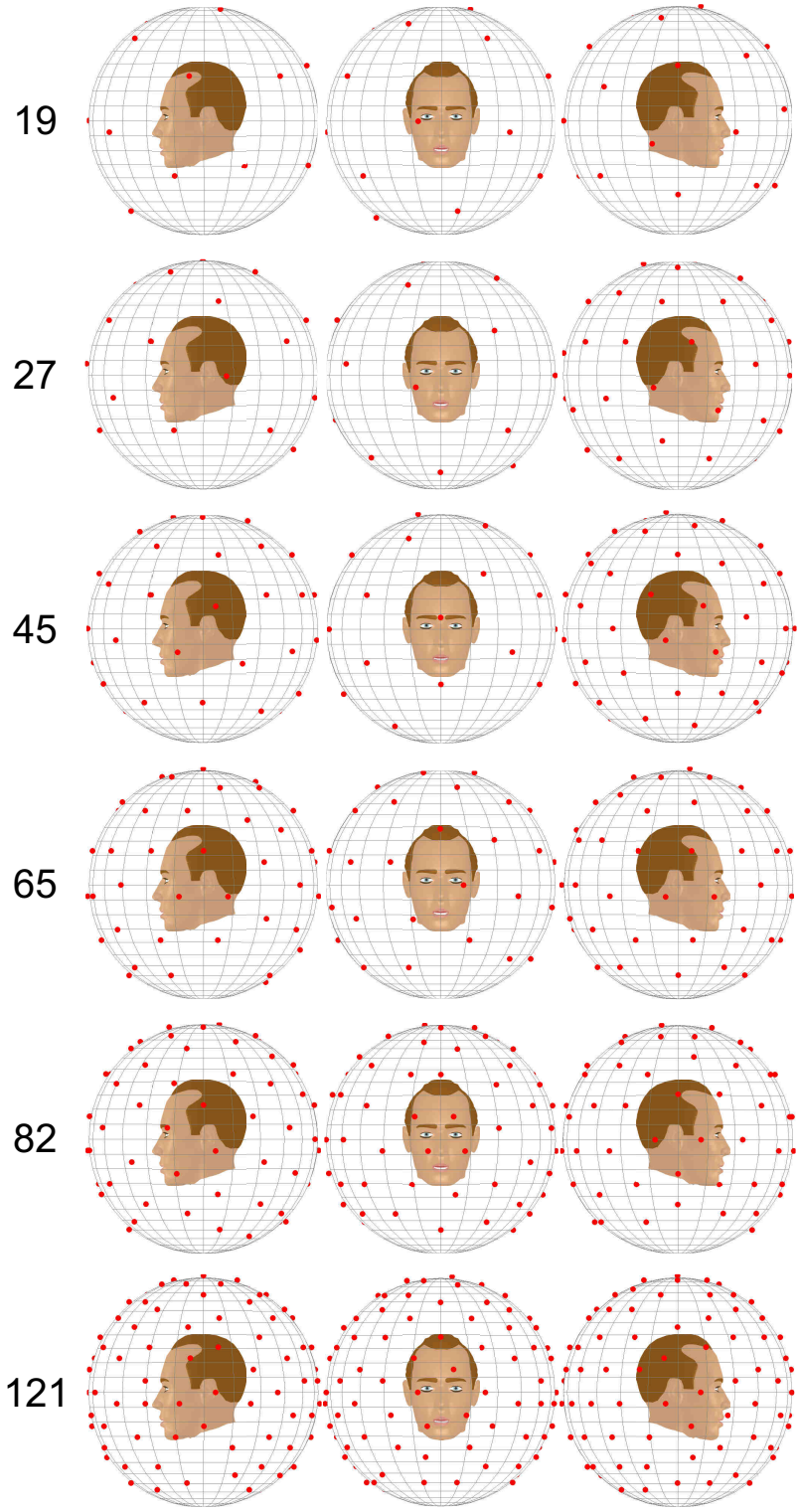


Figure 6.24 – Echantillonnages spatiaux utilisés en entrée de la technique de reconstruction, et évalués dans le test perceptif. Chaque ligne correspond à un échantillonnage, et le nombre total de directions de mesure est indiqué à gauche.

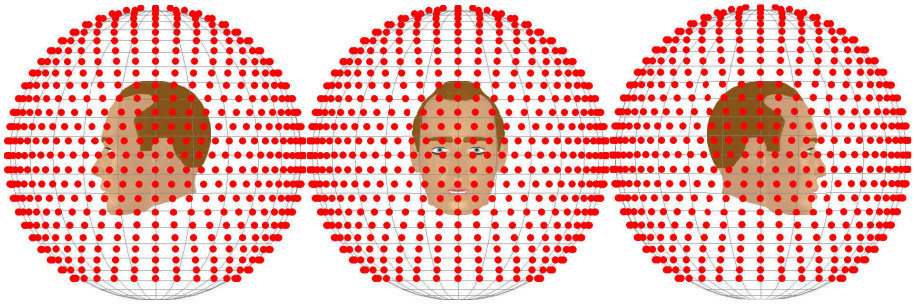


Figure 6.25 – Echantillonnage spatial original de la base de HRTF d'Orange Labs (965 directions).

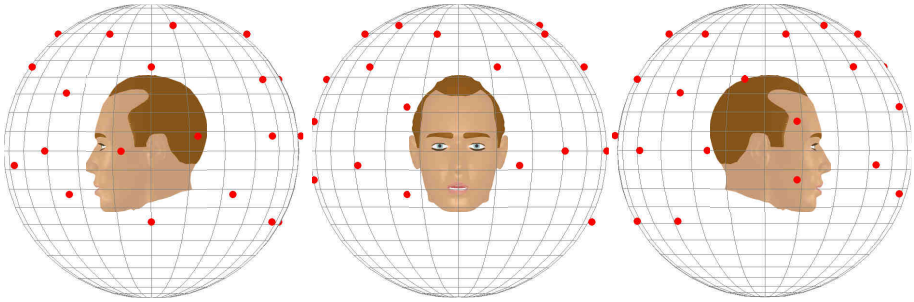


Figure 6.26 – Directions des sources de test : 35 au total sont choisies entre les plans d'élévation -30° et 60°.

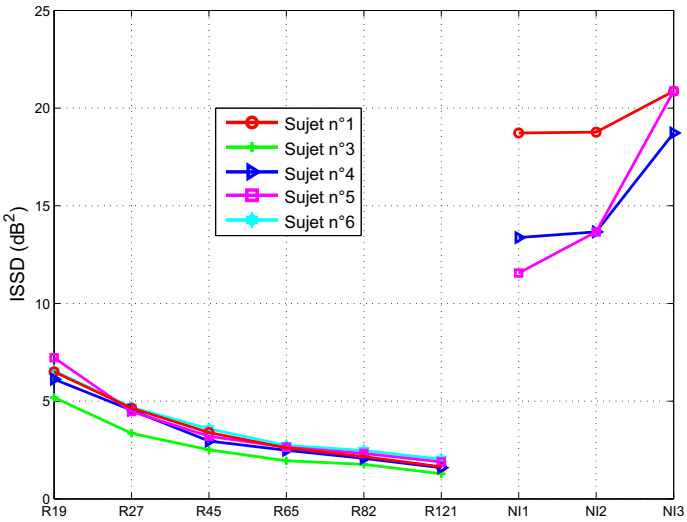


Figure 6.27 – ISSD entre les HRTF individuelles des sujets retenus pour l'évaluation et les différents jeux de HRTF utilisés (ISSD moyenne sur les oreilles gauches et droites).

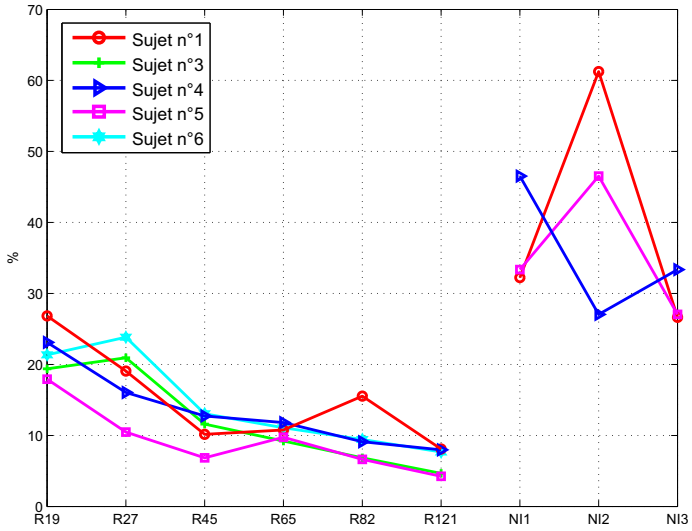


Figure 6.28 – Pourcentage des directions pour lesquelles l'erreur d'ITD dépasse la JND.

6, correspondant à 3 des 35 directions testées, et différents essais (n°1, 3 et 5), pour les 7 jeux de HRTF testés (conditions R19, R27, R45, R65, R82, R121, et I).

Dans la phase initiale du mouvement, après un temps de réaction très court, un pic de vitesse prononcé apparaît, menant rapidement le sujet dans une zone voisine de la direction de la source. On observe rarement un profil de vitesse à plusieurs pics, ou bien relativement plat, synonymes d'une hésitation, ou d'une analyse dynamique prolongée de la scène sonore. Cela suggère que la décision se fonde essentiellement sur le percept spatial formé quand le sujet est dans sa position initiale. Les trajectoires adoptées jusqu'au voisinage de la direction de la source sont très similaires d'une condition à l'autre. Leur orientation est généralement oblique pour des sources particulièrement hautes ou basses, ce qui suggère que tant l'azimut que l'élévation sont correctement perçus dès le début. Il semble que la part la plus variable du temps de réponse s'écoule quand le sujet est orienté dans une zone de l'espace assez proche de la direction à atteindre, pour laquelle la source est donc perçue comme quasiment frontale.

A première vue, il existe une variabilité des performances en termes de temps de réponse, d'un sujet à l'autre. L'observation des profils de vitesse suggère que ces différences sont partiellement explicables par des comportements individuels différents. Tous les sujets avaient la consigne de trouver le plus rapidement possible les sources virtuelles, mais il leur était également demandé d'atteindre une certaine constance dans la volonté d'accomplir cette tâche, tout au long du test. Chaque sujet a ainsi trouvé son propre rythme, comme le montrent les distributions des vitesses angulaires maximales atteintes, représentées figure 6.38. Cette observation suggère que

l'analyse des résultats ne peut être menée tous sujets confondus, mais seulement individuellement. Dans l'étude de Chen [52], qui se limitait à des sources dans le plan horizontal, de fortes différences inter-individuelles ont aussi été observées. Malheureusement l'auteur n'a utilisé qu'un seul et unique jeu de HRTF pour tous les sujets du test, c'est pourquoi rien ne permet de conclure si ces différences sont dues à des capacités intrinsèquement différentes entre les sujets, ou bien si c'est le reflet d'une qualité variable de la spatialisation d'un sujet à l'autre, ce qui est fréquent en conditions non-individuelles.

Comme attendu, on décèle un effet d'apprentissage. On voit par exemple figures 6.31 et 6.33 que les trajectoires adoptées mènent plus directement à la direction de la source, et en un temps plus court pour l'essai n°5 que pour l'essai n°1. C'est aussi ce que révèle le nombre de dépassements du temps imparti : très peu sont observés (entre 0 et 14, selon le sujet, sur 1225 sources), et sauf dans deux cas, ils sont apparus lors du premier essai (cf. Tableau 6.1).

Il est difficile de conclure, par l'analyse des trajectoires, sur l'existence de confusions avant/arrière, car les sujets peuvent les résoudre grâce aux indices dynamiques générés par des petits mouvements de tête. Il semble en tout cas que la discrimination est rapidement obtenue : pour chaque sujet, la vitesse maximale atteinte est en effet plus élevée pour des directions de sources situées dans l'hémisphère arrière, ce qui dénote la volonté franche d'atteindre rapidement une direction perçue d'emblée comme lointaine (cf. Fig. 6.38).

Au cours du test, les sujets étaient invités à traduire librement par écrit leurs impressions en termes de spatialisation, pour chacune des séries de 35 directions de test, sans savoir de quelle condition il s'agissait. Les sujets ont été familiarisés avec la synthèse binaurale et ont déjà participé à des tests de localisation. Ils connaissaient donc bien les problèmes de confusions avant/arrière et d'externalisation dont souffre parfois cette technique. Aucun de ces défauts n'a été signalé pour les conditions I et R19 à R121. En revanche, plusieurs sujets ont émis un jugement sur l'étendue spatiale des sources. Les conditions R19 et R27 sont les seules qui aient été critiquées en ces termes : les sources virtuelles ont souvent été qualifiées de diffuses, larges, ou de position imprécise dans l'espace.

Analyse

Temps de réponse normalisé

Bien que les sujets aient retrouvé leur position de repos avant que chaque source de test soit diffusée, la direction enregistrée à l'instant $t = 0$, auquel le stimulus a débuté, est plus ou moins éloignée de la direction (0°,0°). Afin que tout soit comparable, on choisit donc d'utiliser comme origine temporelle l'instant τ_{init} auquel l'axe

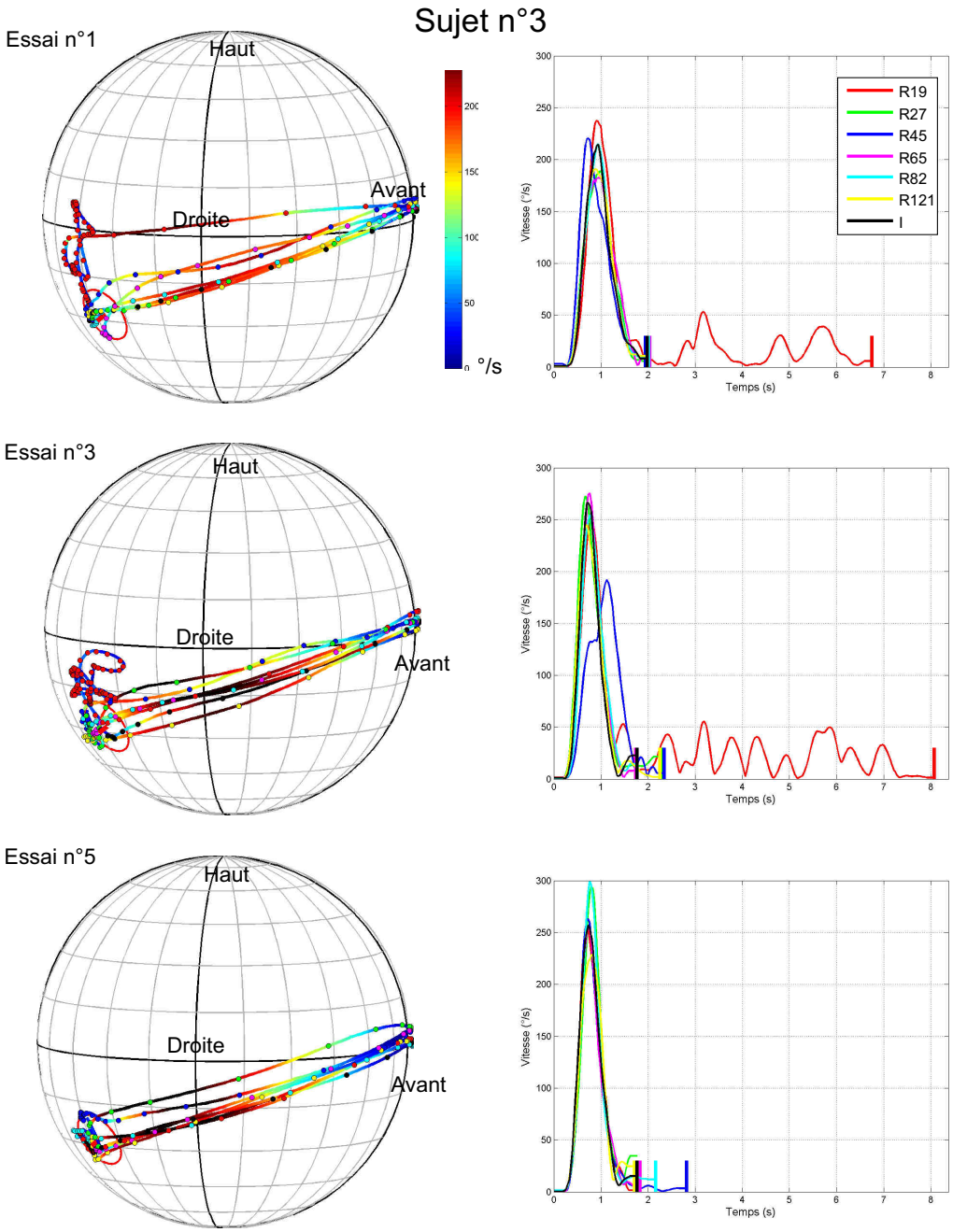


Figure 6.29 – A gauche : trajectoires adoptées par l'axe médian du sujet n°3 pour la direction n°1 (azimut 229°, élévation -28.15°, système polaire vertical), pour les essais 1, 3 et 5, et pour les conditions de test R19, R27, R45, R65, R82, R121 et I. La vitesse angulaire est codée en couleur le long de ces trajectoires. La direction de la source est matérialisée par un trait rouge, et le cône de la validation qui l'entoure par un cercle rouge. A droite : vitesse angulaire correspondante. Les traits verticaux à la fin de chaque courbe marquent l'instant de la validation.

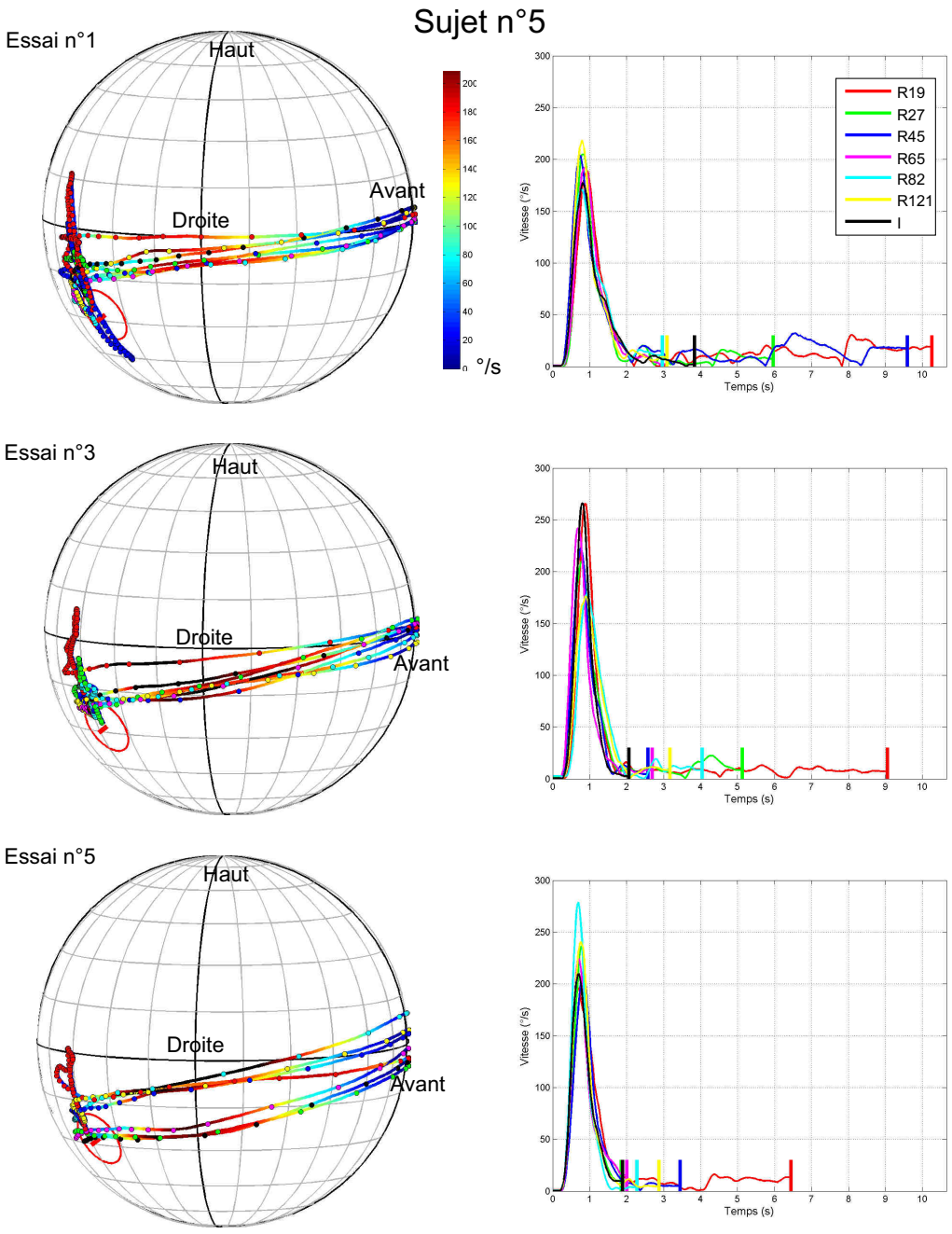


Figure 6.30 – Comportement du sujet n°5 pour la direction n°1. Voir figure 6.29 pour les détails.

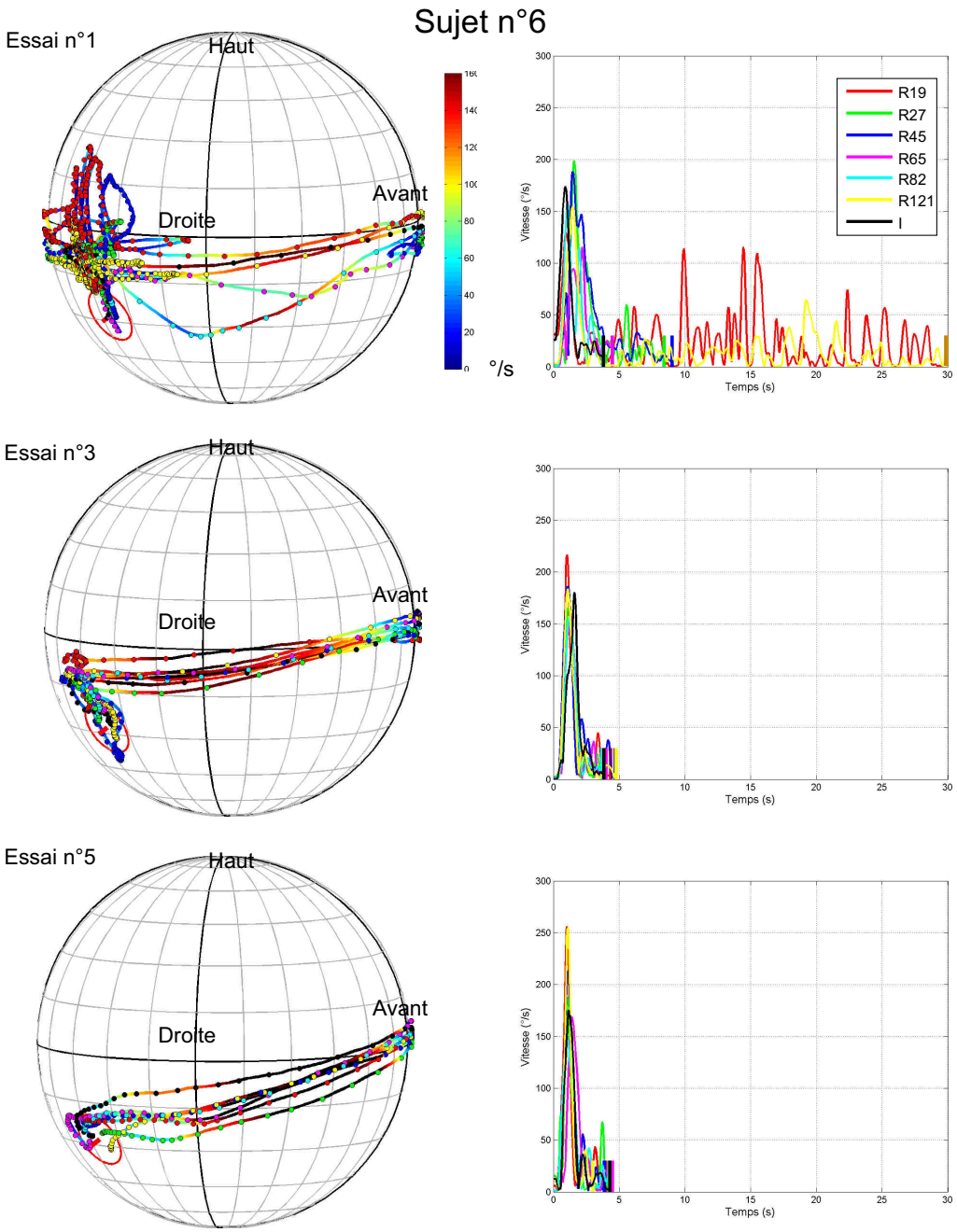


Figure 6.31 – Comportement du sujet n°6 pour la direction n°1. Voir figure 6.29 pour les détails.

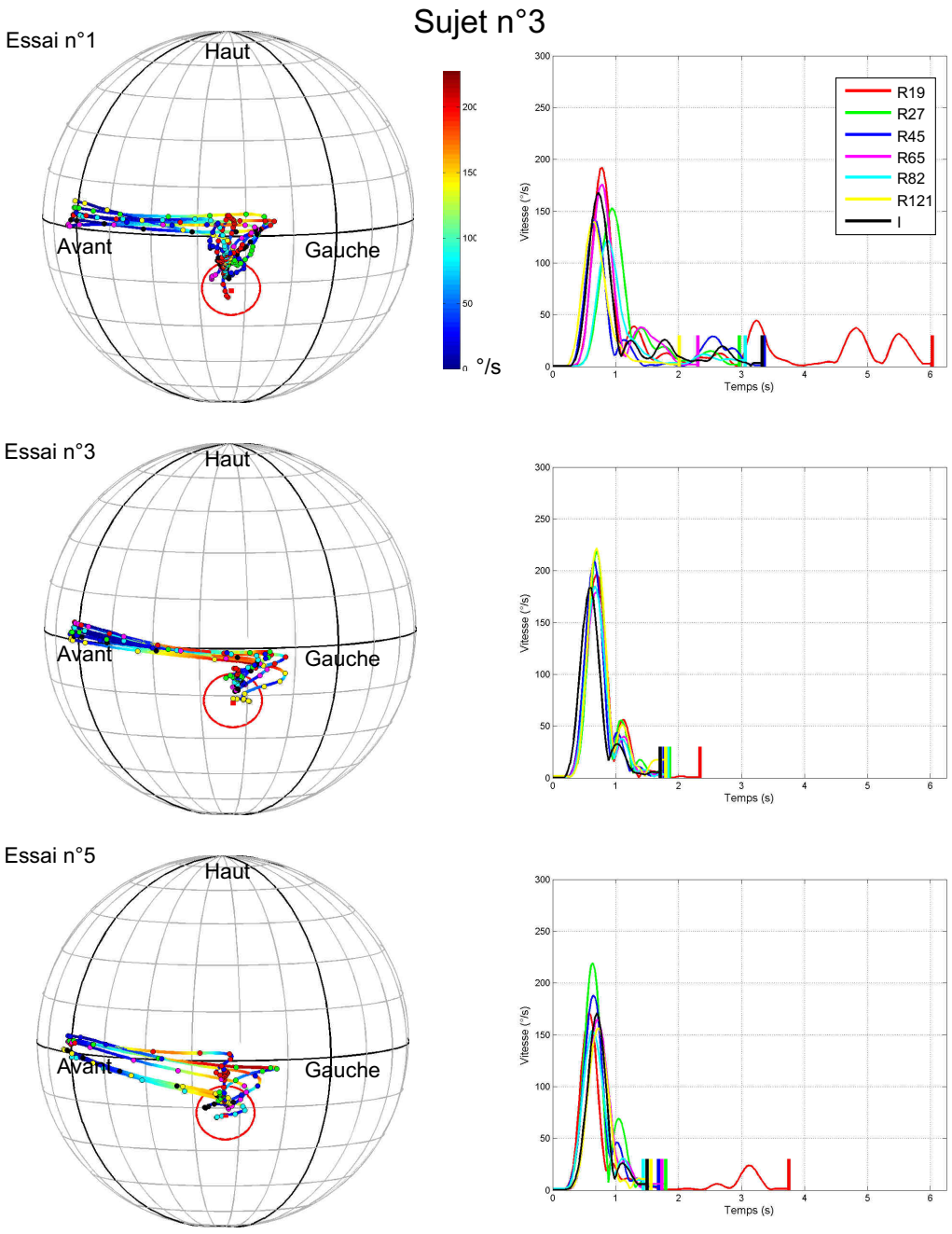


Figure 6.32 – Comportement du sujet n°3 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.29 pour les détails.

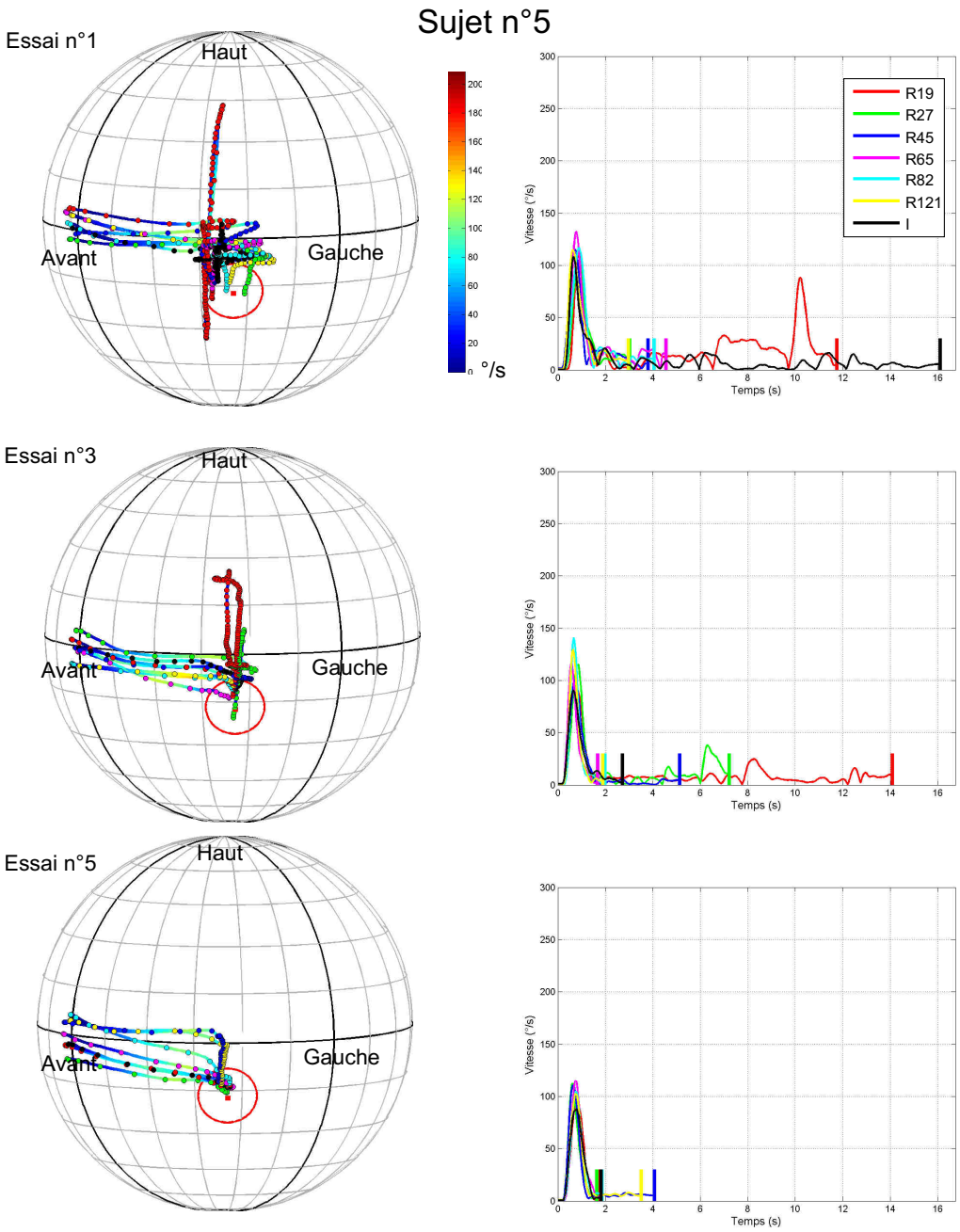


Figure 6.33 – Comportement du sujet n°5 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.29 pour les détails.

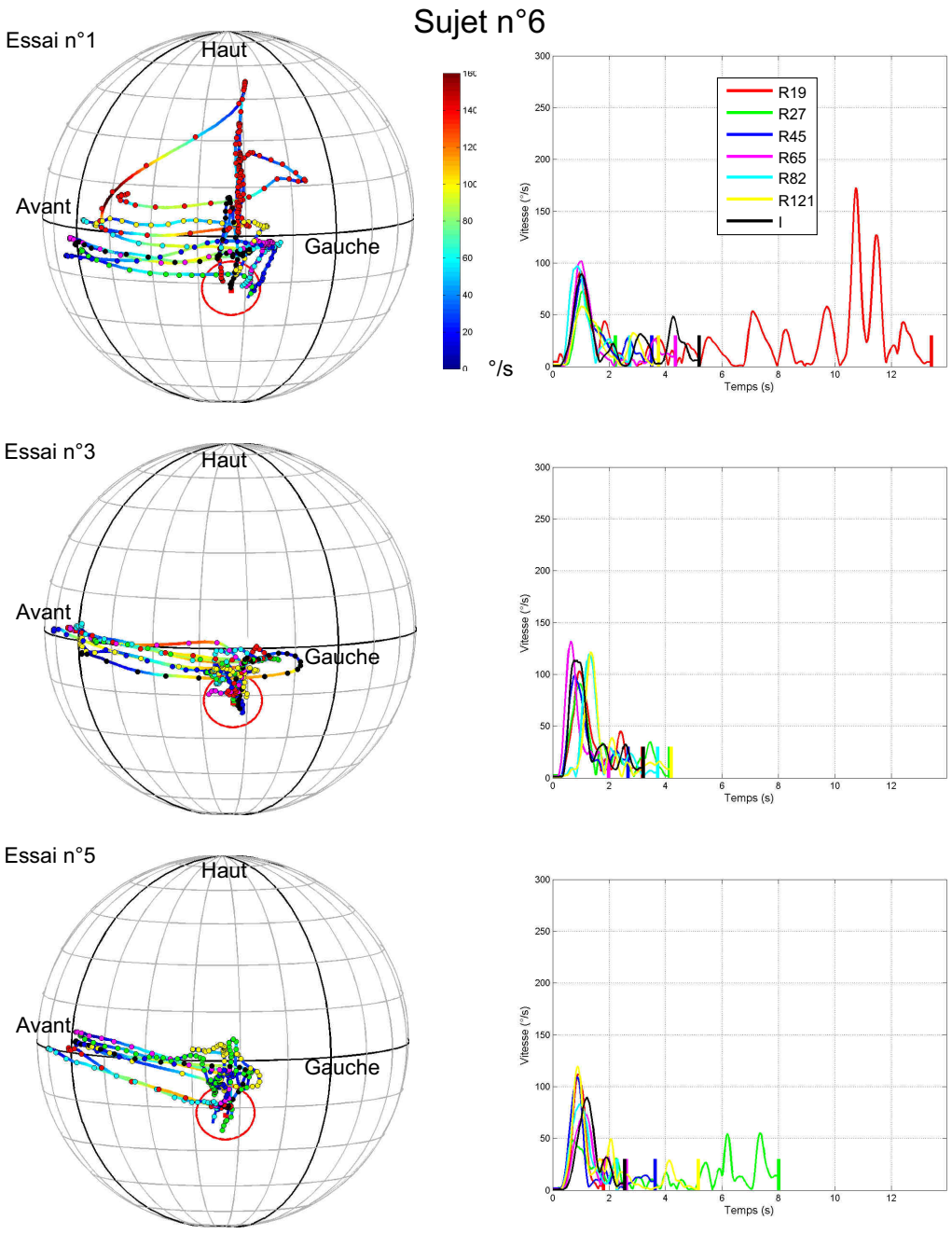


Figure 6.34 – Comportement du sujet n°6 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.29 pour les détails.

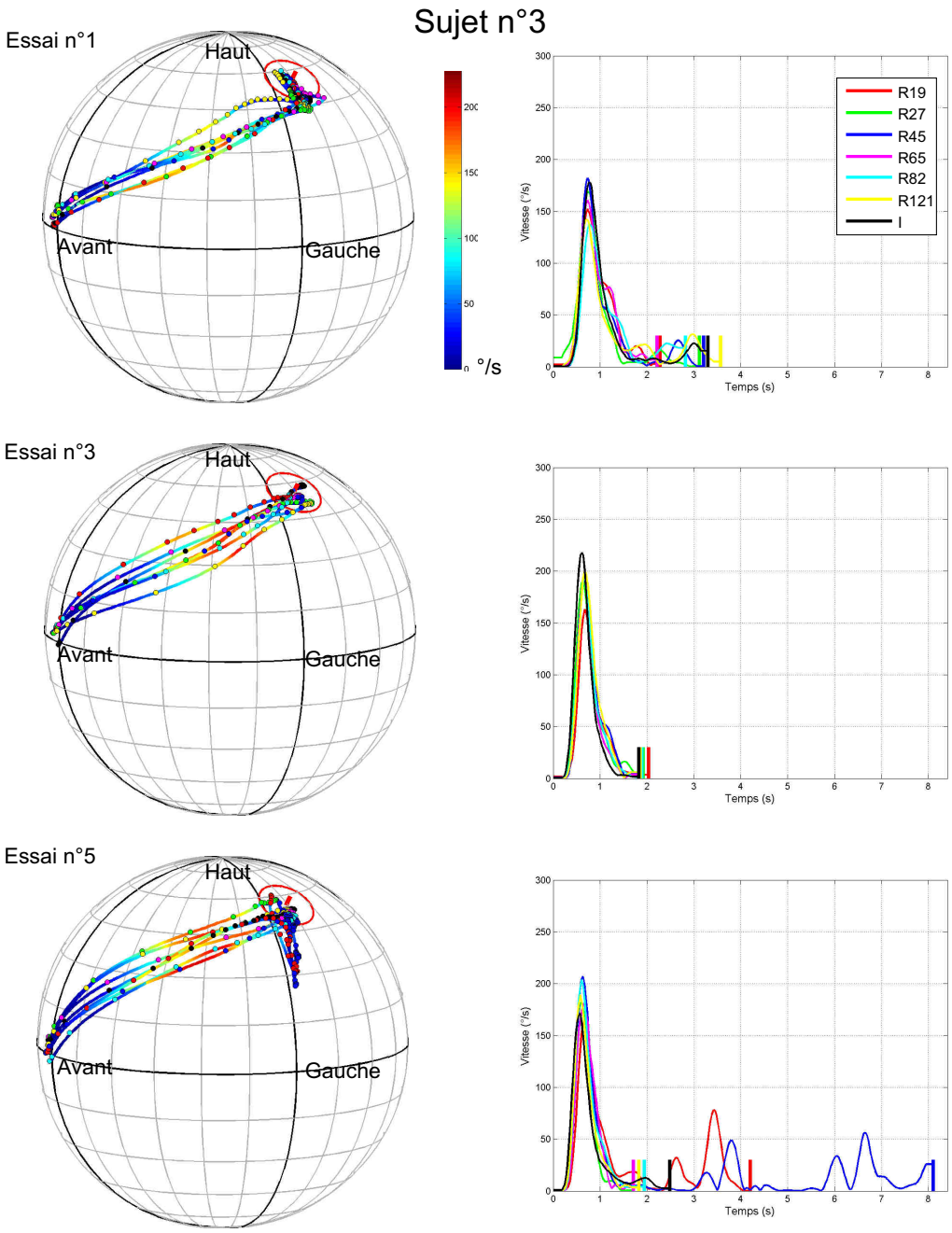


Figure 6.35 – Comportement du sujet n°3 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.29 pour les détails.

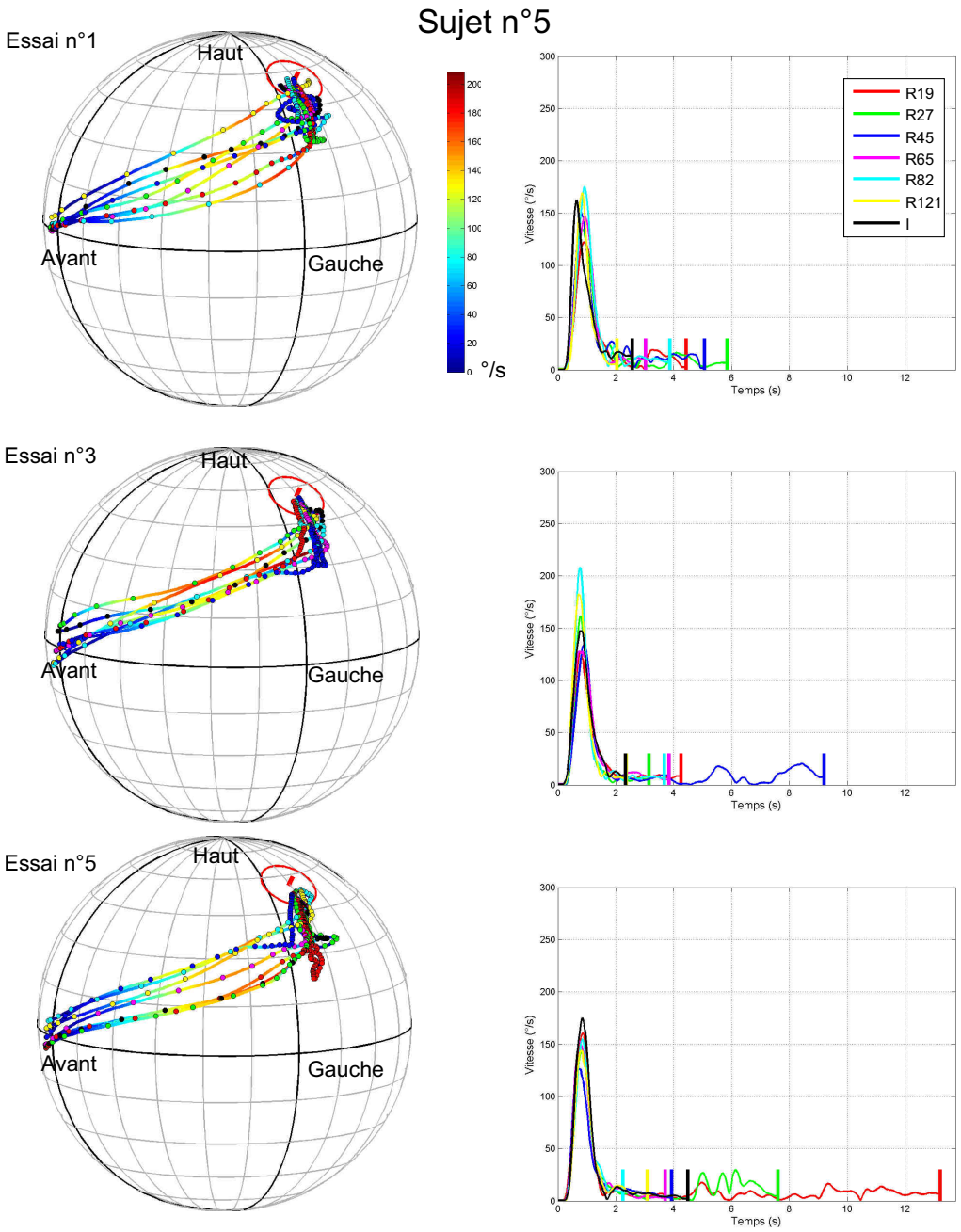


Figure 6.36 – Comportement du sujet n°5 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.29 pour les détails.

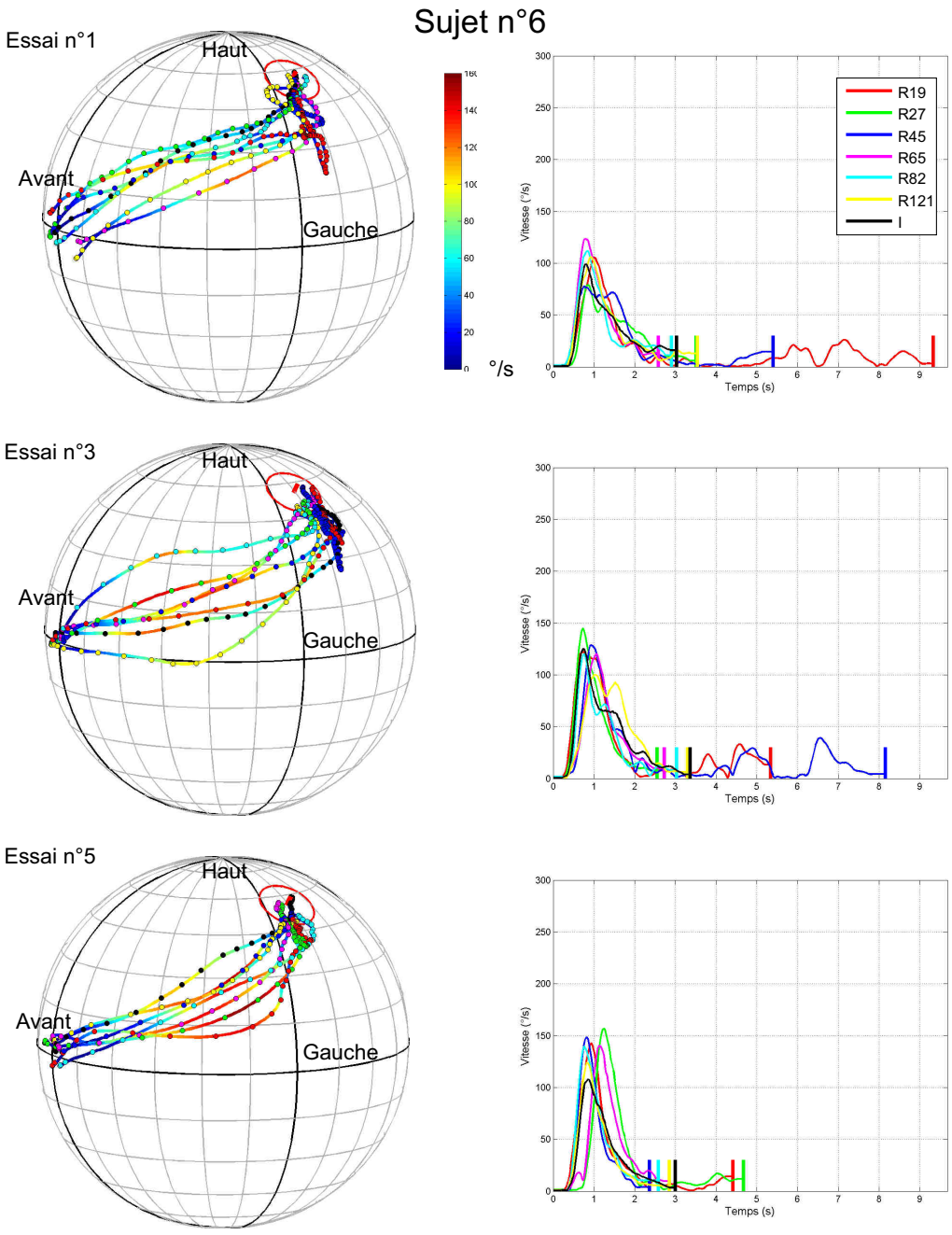


Figure 6.37 – Comportement du sujet n°6 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.29 pour les détails.

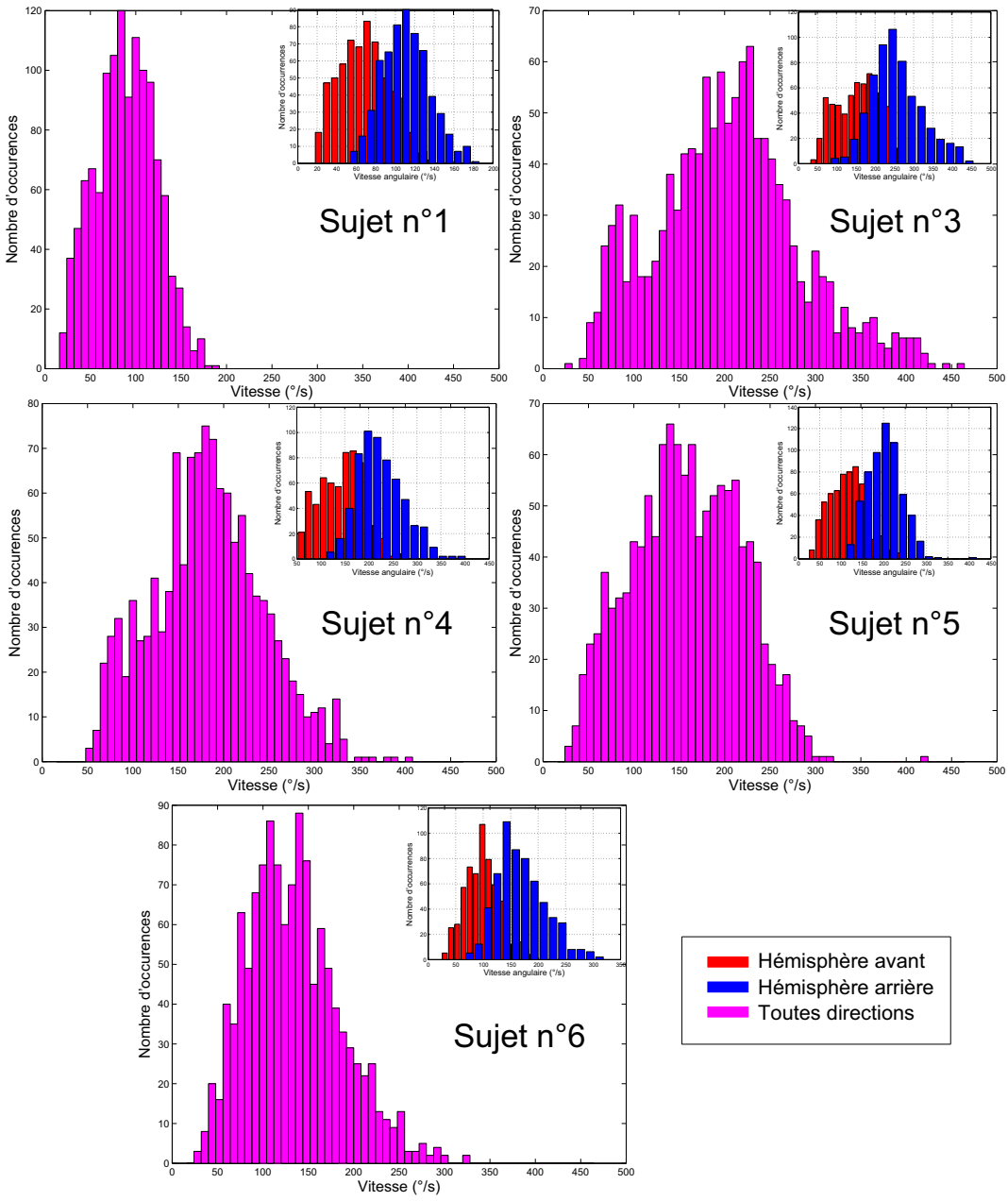


Figure 6.38 – Distributions des vitesses angulaires maximales observées le long des trajectoires, pour chaque sujet, tous essais et conditions confondus (R19 à R121 et I) (en magenta). Un histogramme supplémentaire permet de distinguer les valeurs observées en fonction de l'hémisphère dans lequel se situait la source à atteindre (avant en rouge, arrière en bleu).

Sujet n°	Essai n°	Condition						
		R19	R27	R45	R65	R82	R121	I
1	1	0	1	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0
5	1	0	0	1	0	0	0	0
	2	1	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	1	0	0	0	0	0	0
	5	0	0	0	0	0	0	0
6	1	10	0	0	0	0	4	0
	2	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0

Table 6.1 – Nombre de dépassements du temps imparti (30 secondes) en fonction du sujet, de l'essai et de la condition de test.

médian sort du cône de 9° de rayon, centré sur la direction (0°,0°) (cf. Fig. 6.40). Par ailleurs, on considère la durée τ_{rep} séparant la sortie du cône initial, de l'entrée dans le cône entourant la direction de la source [9], soit :

$$\tau_{rep} = \tau_{val} - \tau_{init} - 0.750$$

où τ_{val} est le temps de validation, et les 750 ms correspondent au laps de temps qui s'écoule entre l'entrée dans le cône entourant la source et la validation.

De façon à évaluer la qualité des jeux de HRTF dans tout l'espace, on a introduit 35 directions de test. En synthèse binaurale dynamique, les conclusions ne peuvent être tirées que de façon globale, car pour une direction de source donnée, le sujet peut en faire varier la direction égocentrique par ses propres mouvement, et ainsi tirer parti des indices de localisation fournis par un ensemble de paires de HRTF associées à des directions différentes. On cherche donc à analyser conjointement les résultats correspondant aux 35 directions de test. Cependant, on conçoit aisément que le temps nécessaire pour atteindre une source est influencé par l'écart angulaire entre la position de repos et la direction de cette source. La localisation de chacune des 35 directions de test correspond donc à une tâche distincte, c'est pourquoi une analyse mêlée des temps de réponse nécessite une normalisation préalable. A première vue, un bon candidat pour cette normalisation pourrait être l'écart angulaire entre la direction (0°,0°) et chaque direction de test. Cependant, cet écart ne tient pas compte des capacités de mouvement limitées des sujets, notamment en termes de rotation autour de l'axe interaural. Il semble donc plus judicieux d'intégrer dans la normalisation les capacités optimales observées pour chaque sujet, face à chacune des directions de test $\{\chi_i\}_{i=1,\dots,35}$. On considère la longueur d_{rep} de la portion de trajectoire parcourue pendant la durée τ_{rep} (cf. Fig. 6.40). Tous essais et toutes conditions confondus, on retient la plus courte : $\{d_i = min(d_{rep}(\chi_i))\}_{i=1,\dots,35}$. Finalement, on fonde l'analyse sur le temps de réponse normalisé $\tilde{\tau}_{rep}$, défini pour la direction χ_i, par la relation : $\tilde{\tau}_{rep} = \tau_{rep}/d_i$. On représente figure 6.39 pour chaque sujet, et les 35 directions, la corrélation entre le temps τ_{rep} minimal observé, toutes conditions confondues, et la distance minimale observée d_i. La corrélation est naturellement forte, et la relation quasi-linéaire entre les deux démontre l'existence d'un comportement assez constant des sujets face aux différentes directions de test. Notons néanmoins que pour les sujets n°1 et 4, un comportement plus lent est détecté pour les directions de test particulièrement élevées. Cette observation est confirmée par les commentaires des sujets eux-mêmes, qui ont avoué éprouver des difficultés à pointer le visage dans ces directions. Mis à part ces cas particuliers, c'est essentiellement la distance à parcourir qui détermine le temps minimal nécessaire pour

9. On considère l'instant de la dernière entrée dans le cône, celle qui précéde la validation.

accomplir la tâche de localisation, et cela justifie bien l'idée d'une normalisation des résultats concernant chaque direction par cette quantité.

Ajustement d'une distribution ex-gaussienne

Classiquement, les valeurs des temps de réponse dans une expérience psychophysique présentent une distribution asymétrique resserrée sur la gauche, c'est-à-dire avec une valeur médiane inférieure à la moyenne. C'est précisément le cas dans cette expérience, comme on l'illustre figure 6.41 pour $\tilde{\tau}_{rep}$. Dans certaines études [52, 280], cette particularité est négligée, et c'est la moyenne ou la médiane qui sont utilisées comme descripteurs statistiques des données récoltées. D'autres stratégies visent à corriger cette asymétrie, soit en éliminant les valeurs observées au-delà d'un certain seuil, soit en transformant les données par une fonction analytique qui les rapproche d'une distribution normale ($f(x) = log(x)$ ou $1/x$ [26, 27]). Cela revient à considérer que les données observées sont l'expression d'un processus gaussien, et que les temps de réponse élevés ne sont que des observations marginales, et ne méritent donc pas d'être considérés dans l'analyse. Cependant, il est plus probable que la distribution des données soit fondamentalement asymétrique. Il a en effet été montré dans d'autres expériences psychophysiques que l'amplitude de cette asymétrie pouvait contenir des informations sur les effets des manipulations expérimentales [83, 88, 166, 212]. Toute transformation des données est donc susceptible d'affecter l'analyse, en éliminant des caractéristiques pertinentes. Bien qu'elle présente une certaine robustesse, l'analyse des variances (ANOVA) repose sur l'hypothèse que les données suivent une distribution normale, ce qui disqualifie cette méthode pour l'analyse des temps de réponse.

Hohle [94] et Ratcliff [211] ont proposé de modéliser les temps de réponse comme la somme d'une variable aléatoire normale (ou gaussienne) et d'une variable aléatoire exponentielle, dont la distribution résultante est nommée ex-gaussienne [36]. On représente figure 6.42 la forme d'une telle distribution, qui est caractérisée par trois paramètres : μ et σ, qui sont respectivement la moyenne et l'écart-type de la composante normale, et τ la moyenne de la composante exponentielle. Les moments centraux m_1 (moyenne), m_2 (variance) et m_3 (coefficient de dissymétrie, ou *skewness*) sont définis à partir de ces paramètres selon les relations : $m_1 = \mu + \tau$, $m_2 = \sigma^2 + \tau^2$ et $m_3 = 2\tau^3$ (cf. Annexe L). On propose donc d'analyser les données récoltées en cherchant les paramètres μ, σ et τ selon lesquels les distributions des temps de réponse sont idéalement approchées par une ex-gaussienne $g(\mu, \sigma, \tau)$. L'ajustement est efficacement réalisé grâce à une estimation par maximisation de la vraisemblance (*Maximum Likelihood Estimation* ou MLE), méthode qui a l'avantage d'être stable et robuste (cf. Annexe L). Considérons le temps de réponse normalisé $\tilde{\tau}_{rep}$. On dispose pour chaque sujet et chaque couple condition-essai d'un échantillon de 35 valeurs de temps de réponse, unique observation issue de la distribution

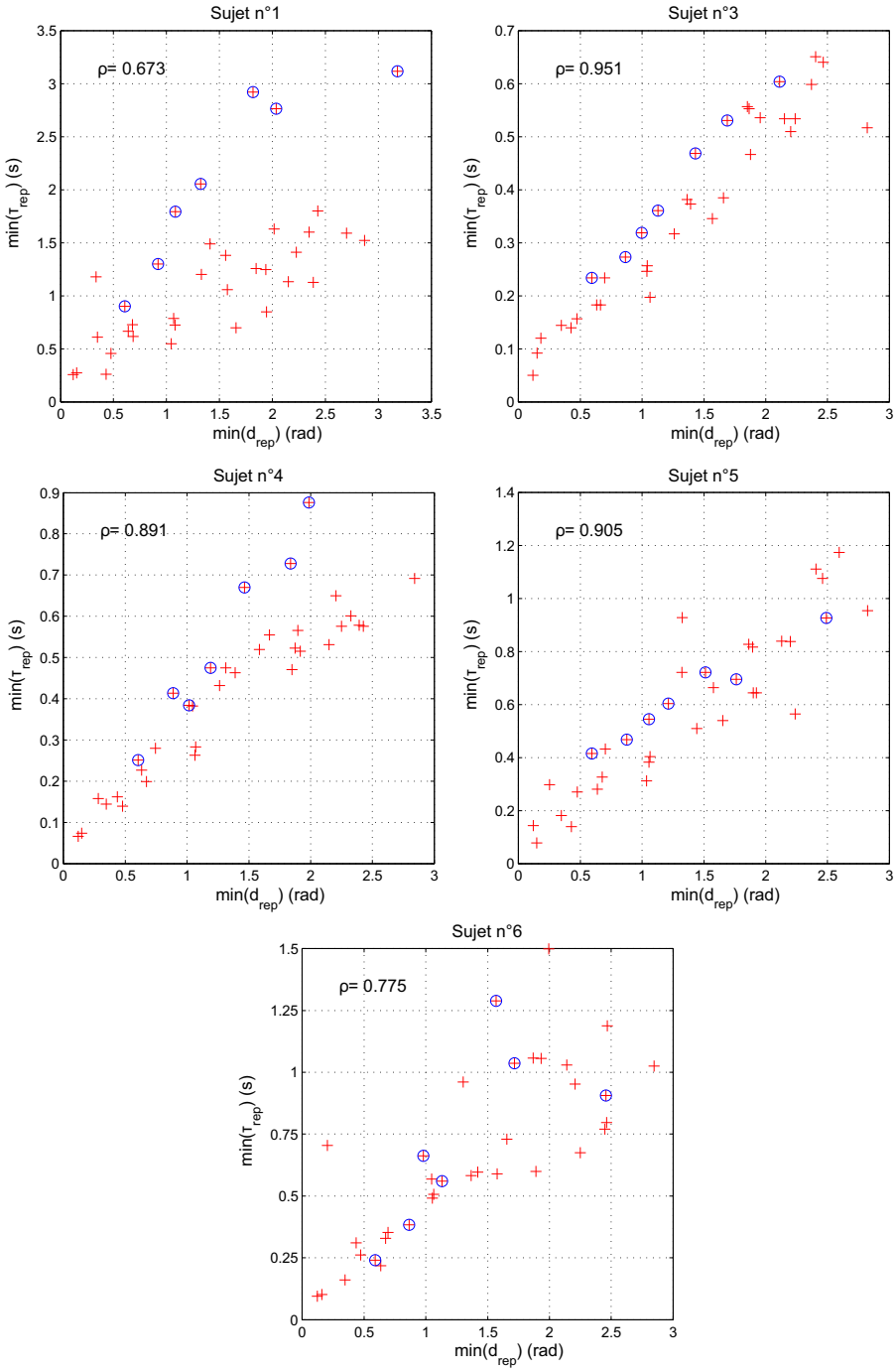


Figure 6.39 – Corrélation ρ entre la distance minimale parcourue pour atteindre une source $min(d_{rep})$ et le temps de réponse minimal $min(\tau_{rep})$, tous essais et conditions confondus (hormis NI1, NI2 et NI3). Les cercles bleus correspondent aux directions de test dont l'élévation est supérieure à 50°.

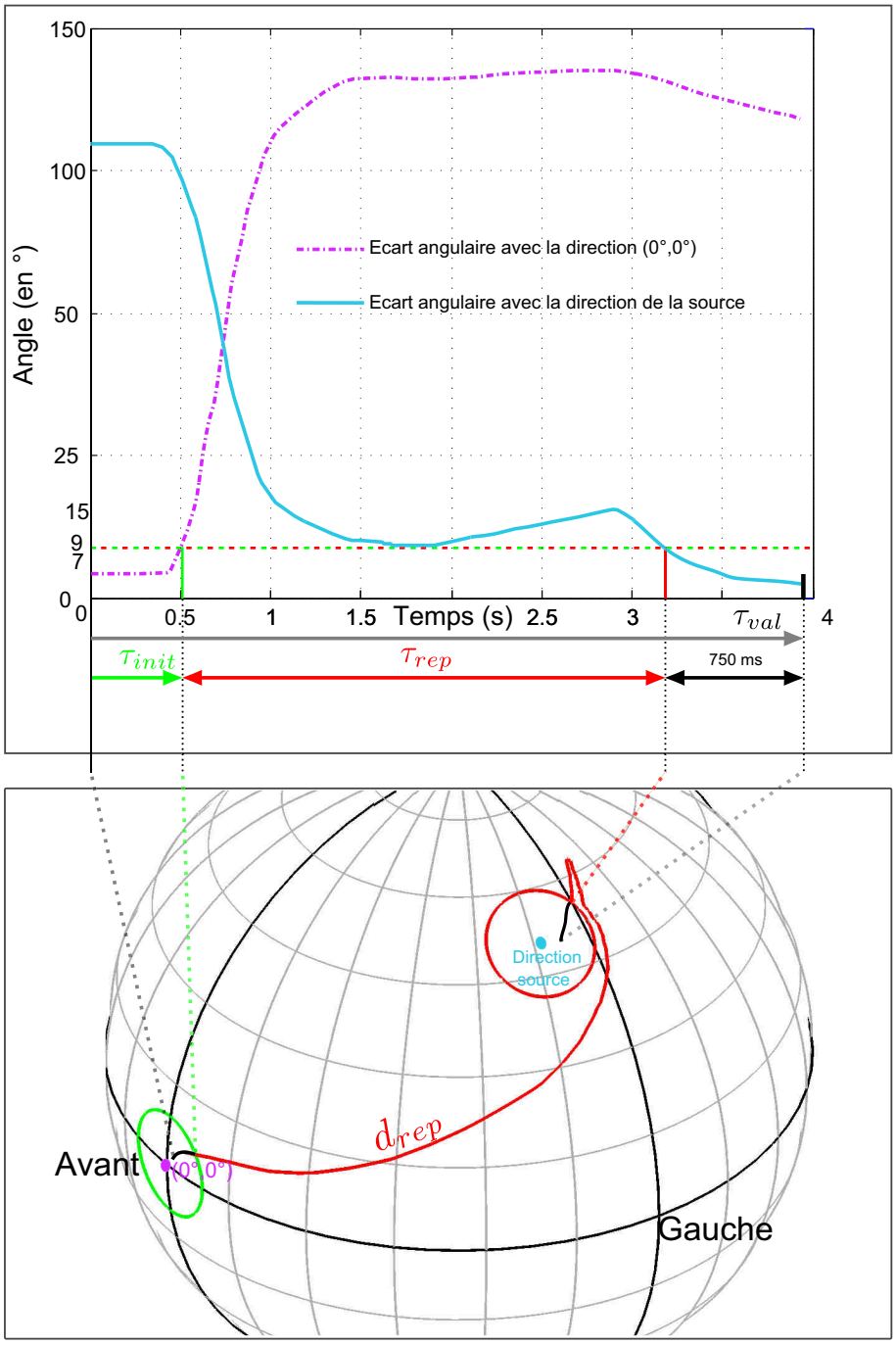


Figure 6.40 – Quantités utilisées pour la définition du temps de réponse normalisé.

sous-jacente qui nous intéresse. De façon à évaluer la qualité de l'estimateur que constitue le MLE en termes de dispersion, il nous faudrait néanmoins davantage d'échantillons. On se sort de cette situation en utilisant une technique d'inférence statistique appelée *bootstrapping*[10][65]. On considère que l'échantillon observé décrit la distribution empirique des données, et qu'on peut l'exploiter pour générer des pseudo-données. C'est donc une technique de rééchantillonnage : N_b échantillons de même taille que l'échantillon observé sont constitués par tirage aléatoire avec remise, à partir des valeurs issues de l'échantillon original. L'estimation des paramètres μ, σ et τ est donc réalisée N_b fois, c'est-à-dire sur chacun de ces pseudo-échantillons. On obtient alors des distributions de ces paramètres qui permettent de dégager des intervalles de confiance. En pratique, on commence par éliminer toutes les valeurs des temps de réponse normalisés correspondant à un dépassement du temps imparti, et on génère par *bootstrapping* $N_b = 1000$ pseudo-échantillons à partir de chaque ensemble de temps de réponse normalisés correspondant à un triplet sujet-condition-essai. On représente figures 6.43, 6.44, et 6.45 les paramètres d'ajustement relatifs aux distributions de $\tilde{\tau}_{rep}$.

L'observation de ces paramètres permet d'appréhender la forme générale des distributions des temps de réponse. Le paramètre μ connaît des variations peu lisibles en fonction des conditions du test, et les intervalles de confiance à 95% sont larges. Le paramètre σ lui aussi est connu avec une faible précision, mais reste extrêmement faible. La composante normale de la distribution ex-gaussienne est donc en général très resserrée autour de sa moyenne. Le paramètre τ est d'un ordre de grandeur supérieur aux deux autres paramètres, ce qui lui donne une importance prépondérante dans chacun des moments centraux. La distribution globale prend quasiment la forme d'une distribution exponentielle décalée sur l'axe temporel de la valeur μ. Comme c'était le cas dans d'autres études psychophysiques basées sur l'analyse de temps de réponse [96, 228], c'est ici le paramètre τ qui apparaît comme le plus apte à révéler lisiblement des différences perçues entre les conditions de test. On décèle une tendance à l'augmentation de τ quand le nombre de mesures diminue en entrée de la technique de reconstruction des HRTF. En moyenne, il existe donc un accroissement du nombre d'occurrences des temps de réponse normalisés particulièrement longs à mesure que l'erreur objective de reconstruction augmente.

Par ailleurs, on observe un apprentissage, qui se manifeste par une amélioration des performances au cours de l'expérience, c'est-à-dire une diminution de τ. On représente figure 6.46, pour chaque essai, la moyenne sur les 5 sujets des τ moyens[11], en fonction des conditions du test. Une diminution nette de τ apparaît entre le premier et le troisième essai, puis une relative stabilité est observée lors des essais suivants.

10. *To pull oneself up by one's bootstrap* : se tirer d'un mauvais pas.
11. Le paramètre τ moyen est le résultat de la moyenne sur les N_b estimations de ce paramètre.

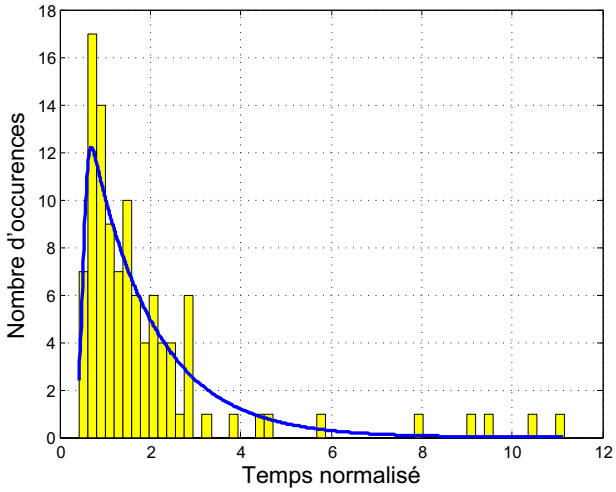


Figure 6.41 – Distribution des temps de réponse normalisés pour les essais 3, 4 et 5 mêlés, dans la condition I. La distribution asymétrique observée est typique des temps de réponse dans les expériences psychophysiques. On représente en superposition de l'histogramme la distribution ex-gaussienne la mieux adaptée.

On considère donc dans la suite de l'analyse que les données récoltées lors des trois derniers essais sont celles qui traduisent les performances optimales des sujets. Pour chaque sujet et chaque condition, les temps $\tilde{\tau}_{rep}$ observés pour les essais n°3, 4 et 5 sont rassemblés, et l'ajustement statistique décrit précédemment est à nouveau réalisé, mais cette fois sur des échantillons de taille triple. On obtient les résultats représentés figures 6.47. La tendance du paramètre τ se confirme. Le comportement du sujet n°3 semble cependant assez constant d'une condition à l'autre, et ce avec des valeurs de τ bien inférieures à celles des autres sujets, qui révèlent donc des performances supérieures.

L'évaluation de la technique de reconstruction proposée consiste à déterminer pour chaque sujet si les comportements observés avec les HRTF reconstruites (conditions R19 à R121) sont différents de celui observé en conditions individuelles (condition I), que l'on considère comme la référence à atteindre. On représente donc figure 6.48 une comparaison pour chacune des conditions R19 à R121 entre les distributions ex-gaussiennes ajustées sur les distributions des temps de réponse normalisés, en considérant conjointement les données des essais n° 3, 4 et 5. On observe des variations d'un sujet à l'autre : les conditions qui s'écartent nettement de la référence ne sont pas systématiquement les mêmes.

Afin de conclure sur la significativité des différences de comportement, on rassemble les temps de réponse normalisés récoltés pour les essais n° 3, 4 et 5, et on compare dans leur globalité les distributions statistiques. Pour chacune des conditions R19 à R121, on considère que le comportement du sujet s'écarte de façon significative

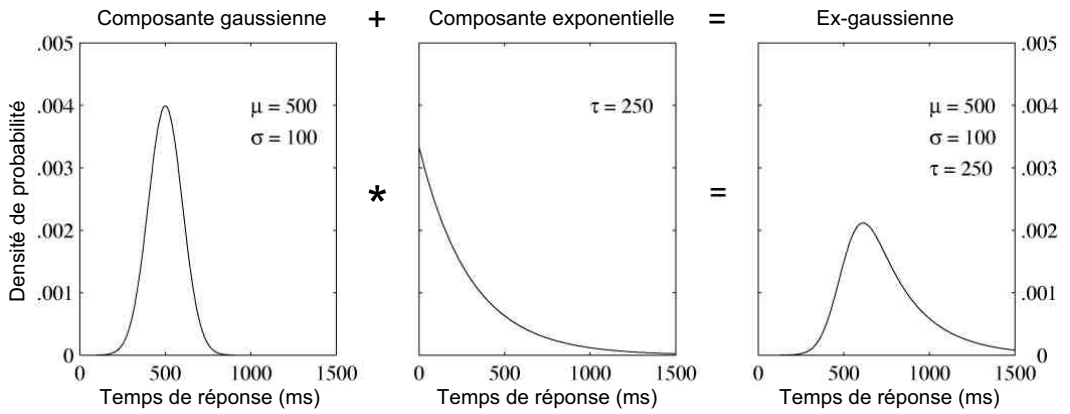


Figure 6.42 – Définition et densité de probabilité d'une variable aléatoire ex-gaussienne (d'après [129])).

de celui observé dans la condition individuelle I si les deux distributions correspondantes sont elles-mêmes significativement différentes. On cherche ces différences au moyen d'une série de tests d'hypothèse, réalisés par la méthode de permutation [65]. Ce faisant, aucune hypothèse n'est posée sur la forme des distributions traitées. De façon générale, nommons R la condition correspondant à des HRTF reconstruites (conditions R19 à R121). Il s'agit de tester l'hypothèse d'identité :

$$H_0 : F_R \equiv F_I \tag{6.7}$$

où F_R et F_I sont les distributions dont sont issus les échantillons observés, respectivement selon la condition R et selon la condition I. On note X et Y les variables aléatoires correspondantes, dont on ne dispose que d'un échantillon pour chacune (rsp. $\mathcal{E}_R$ et $\mathcal{E}_I$) :

$$X \sim F_R, \quad \mathcal{E}_R = \{x_1, \ldots, x_m\} \tag{6.8}$$

$$Y \sim F_I, \quad \mathcal{E}_I = \{y_1, \ldots, y_n\} \tag{6.9}$$

On choisit comme statistique de test S la différence entre les moyennes des échantillons [65] :

$$S = \overline{\mathcal{E}_R} - \overline{\mathcal{E}_I} \tag{6.10}$$

On note s_{obs} la valeur observée de cette statistique, c'est-à-dire celle correspondant aux échantillons disponibles. On se place ensuite sous l'hypothèse H_0 : les distributions de X et Y sont alors les mêmes, et on peut considérer l'échantillon combiné $\mathcal{E} = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$, et estimer F_R et F_I par la distribution empirique F_{n+m} de $\mathcal{E}$. On génère par permutation $r = 100000$ pseudo-échantillons

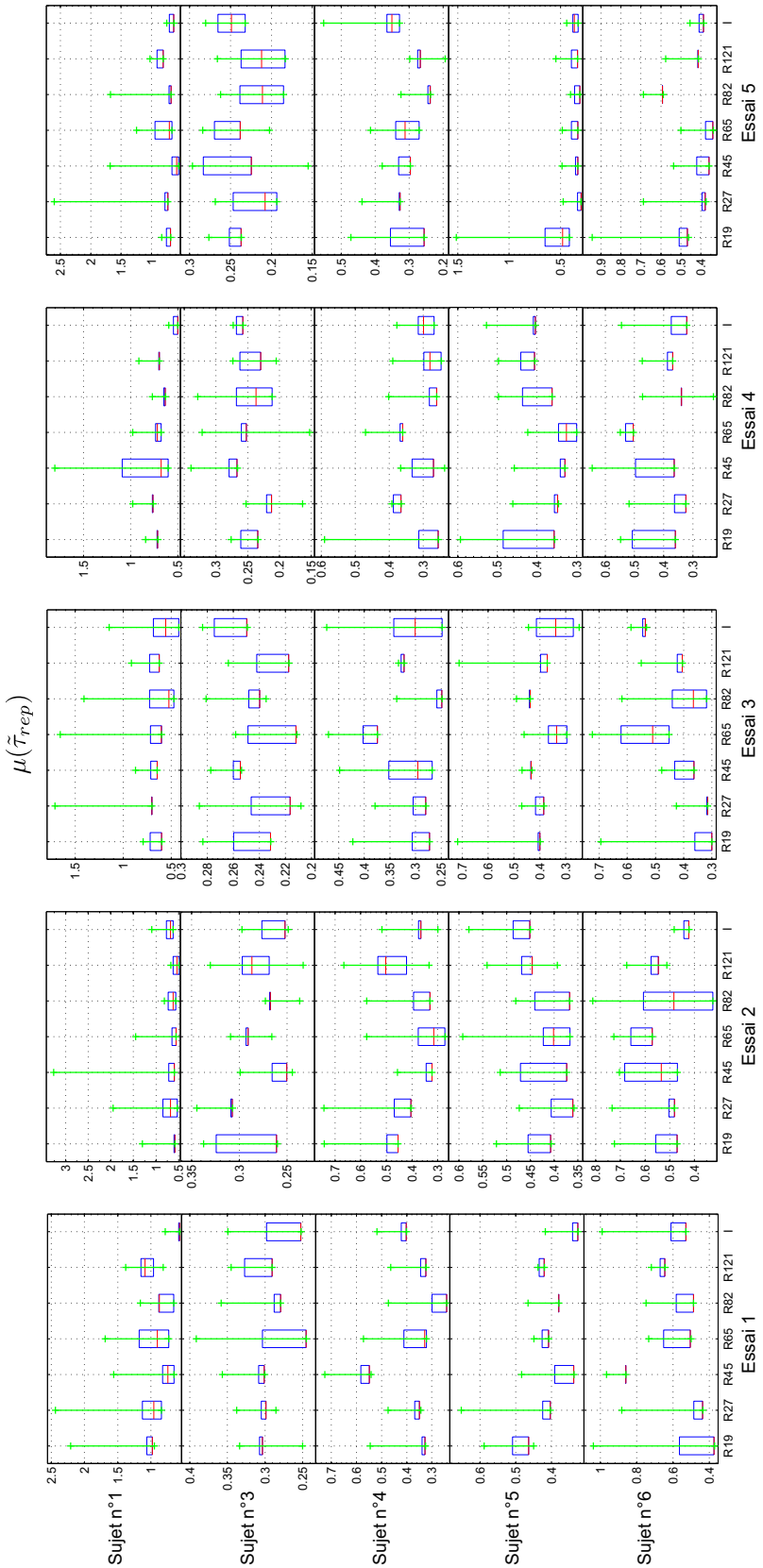


Figure 6.43 – Paramètre μ de la distribution ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les différents sujets, chaque essai et chaque condition (l'échelle en ordonnée diffère d'un sujet à l'autre).

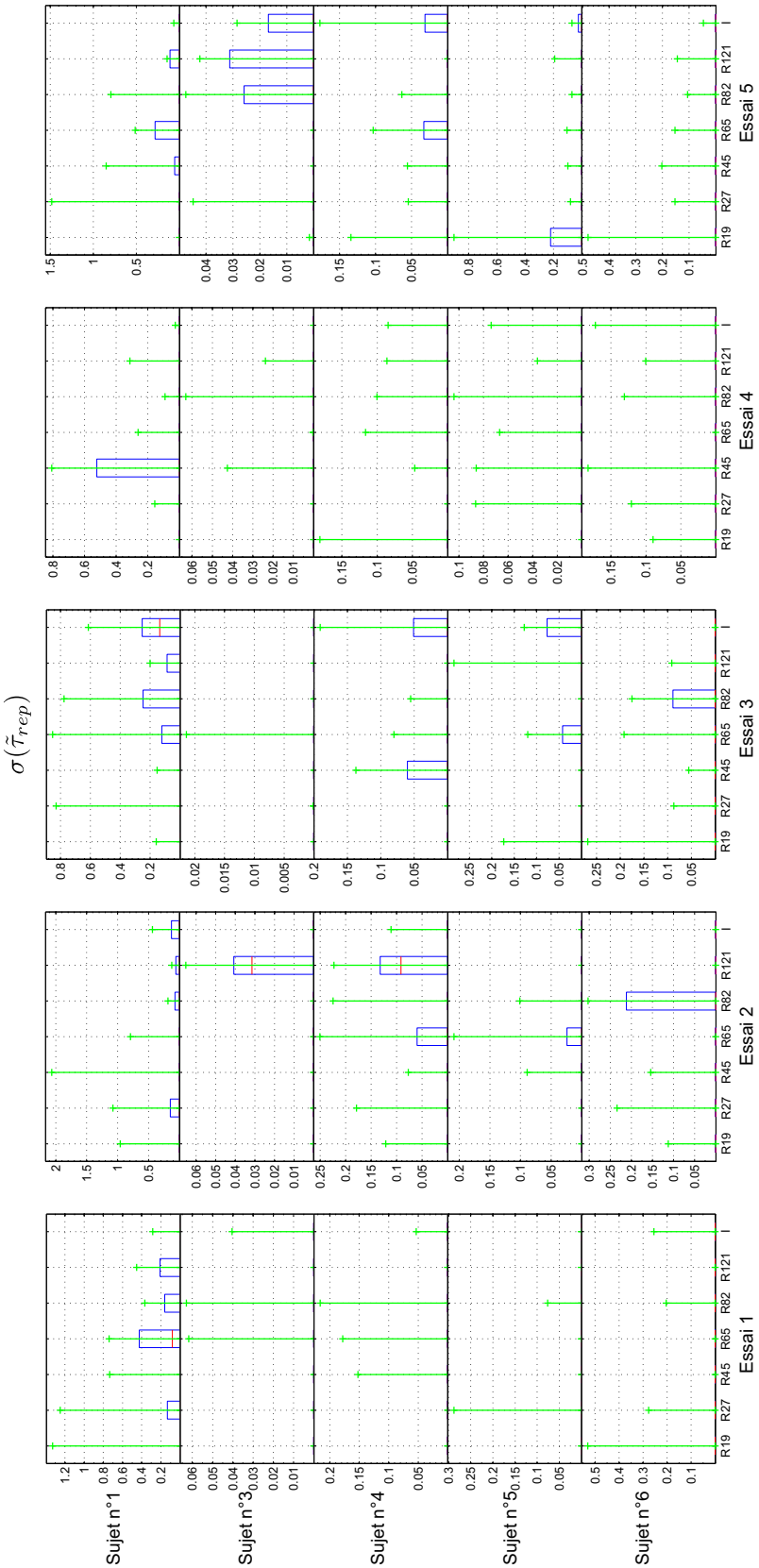


Figure 6.44 – Paramètre σ de la distribution ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les différents sujets, chaque essai et chaque condition (l'échelle en ordonnée dépend du sujet et de l'essai).

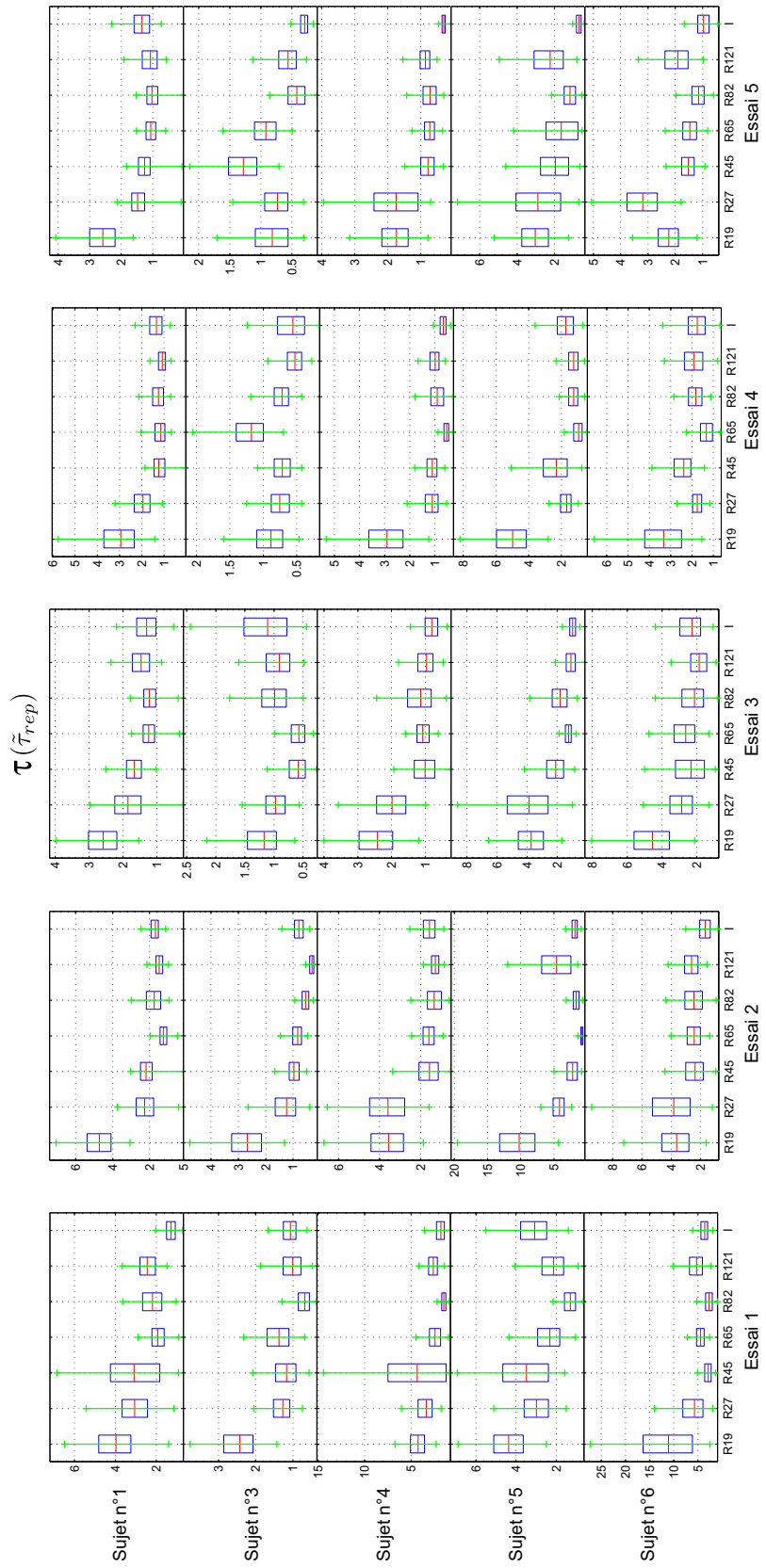


Figure 6.45 – Paramètre τ de la distribution ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les différents sujets, chaque essai et chaque condition (l'échelle en ordonnée dépend du sujet et de l'essai).

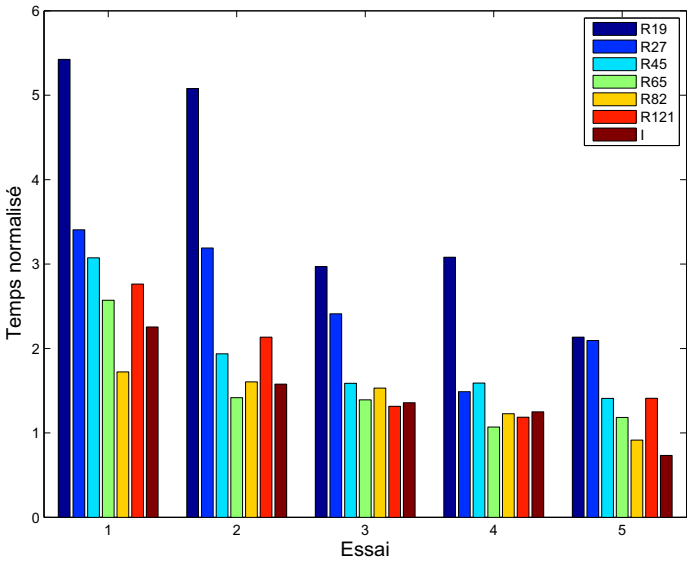


Figure 6.46 – Illustration du phénomène d'apprentissage : pour chaque essai et chaque condition les paramètres τ moyens sont moyennées sur les 5 sujets.

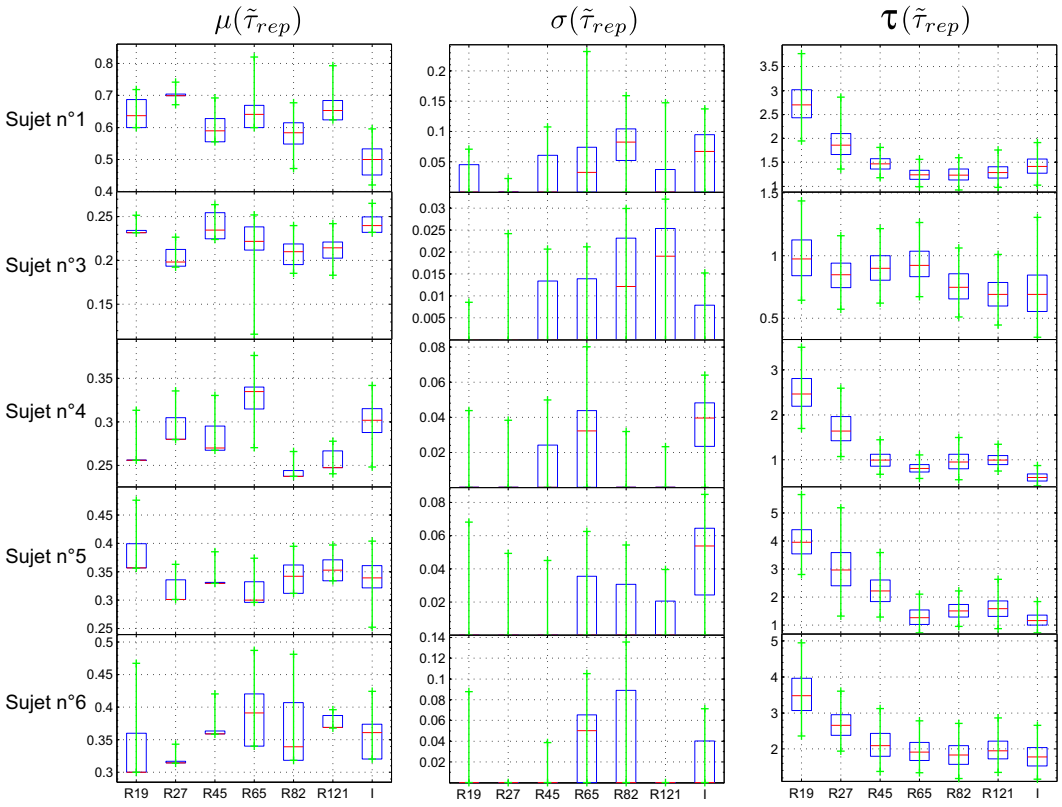


Figure 6.47 – Paramètres μ, σ et τ de la fonction ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les différents sujets, et chaque condition, en considérant conjointement les données des essais n° 3, 4 et 5 (l'échelle en ordonnée dépend du sujet).

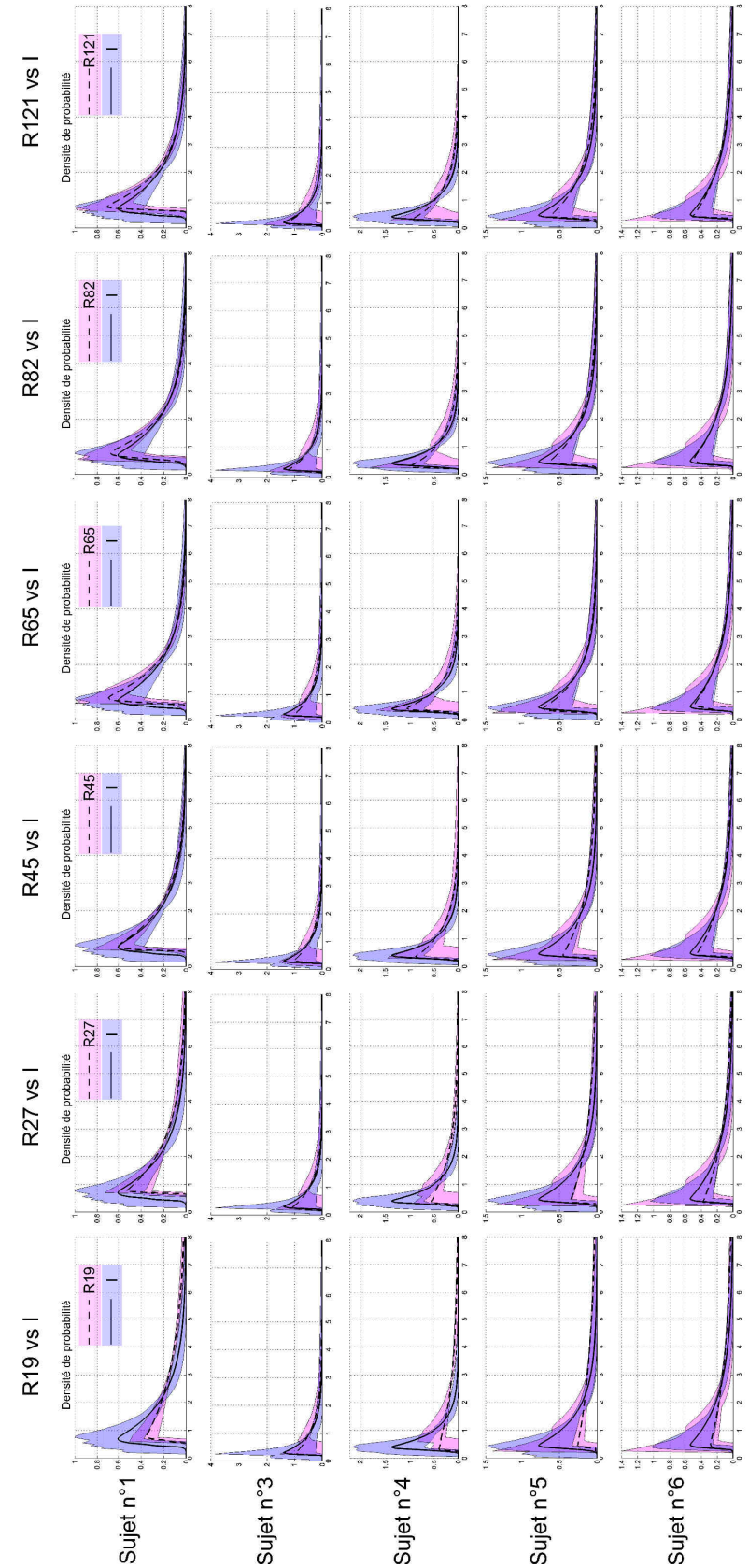


Figure 6.48 – Ajustement de distributions ex-gaussiennes sur les temps de réponse normalisés : pour chaque sujet, on considère conjointement les données des essais n° 3, 4 et 5, et on superpose les résultats obtenus pour la condition I à ceux de chaque condition R19 à R121. Les surfaces colorées correspondent à l'incertitude de l'estimation par MLE des paramètres de l'ex-gaussienne avec *bootstrapping* (intervalle de confiance à 95%), tandis que les courbes représentent les moyennes de toutes les densités de probabilité ex-gaussiennes ajustées.

Sujet n°	Condition					
	R19	R27	R45	R65	R82	R121
1	$\neq$	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
3	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
4	$\neq$	$\neq$	$\neq$	$\equiv$	$\equiv$	$\neq$
5	$\neq$	$\neq$	$\neq$	$\equiv$	$\equiv$	$\equiv$
6	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$

Table 6.2 – Résultats des tests de permutation de l'hypothèse $H_0 : F_R \equiv F_I$. On représente le résultat par le symbole $\neq$ ou $\equiv$ selon que l'hypothèse H_0 est respectivement rejetée ou non.

$\{z_1^*, \ldots, z_{m+n}^*\}_i, \quad i = 1, \ldots, r$, à partir de $\mathcal{E}$, d'où l'on tire autant de valeurs simulées $\{s_i^*\}_{i=1,\ldots,r}$ de S, de façon à estimer la statistique de test sous H_0 :

$$\mathcal{E}_{Ri}^* = \{z_1^*, \ldots, z_m^*\}_i, \quad i = 1, \ldots, r \tag{6.11}$$

$$\mathcal{E}_{Ii}^* = \{z_{m+1}^*, \ldots, z_{m+n}^*\}_i \tag{6.12}$$

$$s_i^* = \overline{\mathcal{E}_{Ri}^*} - \overline{\mathcal{E}_{Ii}^*} \tag{6.13}$$

Enfin, on note λ le nombre de valeurs s_i^* supérieures à s_{obs}, et on calcule la p-valeur p^* :

$$p^* = \frac{\lambda}{r} \tag{6.14}$$

On rejette l'hypothèse H_0, c'est-à-dire qu'on conclut que les comportements correspondant aux conditions R et I sont significativement différents, si $p^* < 0.05$ ou si $p^* > 0.95$. On représente sur le tableau 6.2 les résultats de ces tests d'hypothèse.

On cherche à déterminer le nombre limite de mesures nécessaires en entrée de la technique de reconstruction proposée pour assurer une équivalence perceptive entre les HRTF reconstruites et les HRTF réellement mesurées. On a conscience de l'asymétrie des tests d'hypothèse précédents : ils offrent un résultat significatif lorsque l'hypothèse H_0 est rejetée, mais à l'inverse, s'ils nous informent que H_0 ne peut être rejetée, on ne peut pas pour autant conclure à une équivalence des distributions F_R et F_I avec la même significativité. C'est pourtant ce résultat qui nous permettrait de conclure sur le nombre limite recherché. En s'appuyant sur les résultats de l'ajustement des distributions ex-gaussiennes par MLE (cf. Fig. 6.48), on observe néanmoins une grande similitude entre les distributions des temps de réponse obtenus, et on concluera donc abusivement à une transparence de la reconstruction des HRTF pour un nombre de mesures donné si l'hypothèse H_0 n'est pas rejetée entre la condition de test correspondante et la condition I. En ce sens, il apparaît que le sujet n°3 montre des performances idéales quelles que soient les HRTF utilisées,

résultat conforté par le fait que ce sujet est particulièrement performant par rapport aux autres. Pour le sujet n°6, au moins 27 mesures sont nécessaires, pour le sujet n°1 il faut 45 mesures, et enfin il faut 65 mesures pour le sujet n°5. Notons que pour le sujet n°4, le comportement selon la condition R121 apparaît significativement différent de celui observé selon la condition de référence I. L'observation du paramètre τ, ainsi que celle des distributions ex-gaussiennes confirment ce résultat (cf. Fig. 6.47 et 6.48). Le sujet n°4 est particulièrement rapide quand les sources sont diffusées avec ses HRTF individuelles, et l'intervalle de confiance est restreint. Pour la condition R121, le paramètre τ est légèrement plus élevé, et l'intervalle de confiance est là aussi resserré, ce qui montre que les comportements en conditions R121 et I sont effectivement différents. La distribution ex-gaussienne ajustée sur la distribution des temps de réponse semble elle aussi révéler un allongement du temps nécessaire pour localiser les sources. Ce résultat est troublant, car tant en termes de spectre d'amplitude qu'en termes d'ITD, les différences objectives entre le jeu de HRTF de la condition R121 et les HRTF individuelles sont plus faibles que pour toutes les autres conditions de test, parmi lesquelles certaines offrent une reconstruction transparente pour ce sujet (R65 et R82) (cf. Fig. 6.27 et 6.28). Bien que, comme pour les autres sujets, l'inflexion du paramètre τ n'apparaisse seulement à partir de la condition R45 ou R65, on ne peut nier que pour le sujet n°4 des différences de spatialisation sont perçues assez rapidement, même avec le jeu de HRTF le plus fidèle. Le nombre minimal de mesures nécessaires pour reconstruire convenablement les HRTF dépend donc de l'individu.

6.4.4 Résultats : HRTF non-individuelles

Les conditions de contrôle non-individuelles NI1, NI2 et NI3 ont été introduites dans cette évaluation perceptive pour répondre à plusieurs interrogations. D'abord, dans la mesure où cette méthodologie d'évaluation est nouvelle, on ne sait pas si elle permet une discrimination suffisante entre différentes conditions de spatialisation. Les HRTF non-individuelles offrant souvent une spatialisation de moins bonne qualité, leur introduction dans le test permet donc d'estimer le contraste de notre outil de mesure. Par ailleurs, il s'agit d'évaluer la qualité des HRTF reconstruites avec la technique proposée par rapport à celle de HRTF non-individuelles issues d'une base de données.

Observations préliminaires

On représente figures 6.49, 6.50, 6.51, 6.52, 6.53, 6.54, 6.55, 6.56 et 6.57 les trajectoires observées et les profils de vitesse selon les conditions NI1, NI2 et NI3 propres à chacun des sujets n°1, 4 et 5, pour 3 des 35 directions de test, et les essais 1, 3 et 5. On peut comparer ces comportements avec ceux observés pour les mêmes

sujets, et les mêmes directions pour les conditions R19 à R121 et I, et représentés en 6.4.3 et en Annexe K. A première vue, la qualité de la spatialisation semble être généralement moins satisfaisante dans ces conditions que dans les conditions de test. Cela se traduit parfois par des trajectoires plus erratiques (cf. Fig. 6.50, 6.52 et 6.55), et par des profils de vitesse à plusieurs pics, qui dénotent une certaine hésitation, ou peut-être l'existence de confusions avant/arrière. On voit notamment figure 6.52 que le sujet n°1 réussit sans problème à se tourner dans l'azimut de la source, puis cherche au hasard, par des mouvements de tête de haut en bas et de bas en haut, à obtenir la validation, ce qui montre une perception très mauvaise de l'élévation. Néanmoins, des trajectoires obliques, proches de l'optimum, sont parfois obtenues (cf. Fig. 6.51 et 6.56), ce qui suggère que le percept de localisation est dans certains cas net et correct.

On représente sur le tableau 6.3 le nombre de dépassements du temps imparti pour chaque sujet, en fonction de l'essai et de la condition. La plupart des dépassements apparaissent au premier essai, et sont en nombre bien plus important que dans les conditions de test. Le fait que ces dépassements se raréfient avec le temps dénote probablement l'existence d'un effet d'apprentissage. Rappelons que chronologiquement les sessions au cours desquelles ces conditions de contrôle ont été testées se sont déroulées tout à la fin du test. Il paraît donc raisonnable de considérer que les sujets maîtrisaient déjà parfaitement la technique de report de localisation, et donc que l'apprentissage observé est le fruit d'une adaptation aux HRTF elles-mêmes.

Analyse

L'analyse des conditions NI1, NI2 et NI3 est menée selon la méthode décrite précédemment. On représente figure 6.58 le paramètre τ des distributions ex-gaussiennes ajustées par MLE avec *bootstrapping* aux distributions des temps de réponse normalisés, pour les sujets n°1, 4 et 5 et toutes les conditions de test. Globalement, les conditions NI1, NI2 et NI3 sont associées à des temps de réponse plus longs que les conditions de test R19 à R121, mais il existe des variabilités de comportement d'un individu à l'autre. On représente figure 6.59, pour chaque essai, la moyenne sur les 3 sujets des τ moyens, en fonction des conditions de contrôle. Pour les conditions NI1 et NI2, l'amélioration des performances apparaîtrait dès le deuxième essai, avec une relative stabilité pour les essais suivants. Les performances face à la condition NI3 ne semblent pas s'améliorer au cours du temps. On pourrait donc mener l'analyse conjointement sur les essais n°2 à 5, mais on préfèrera comme précédemment ne considérer que les 3 derniers essais, afin que les analyses soient en tous points comparables. Le paramètre τ obtenu sur ces échantillons est représenté figure 6.60 pour les 3 sujets et toutes les conditions. Pour le sujet n°1, les trois conditions de contrôle

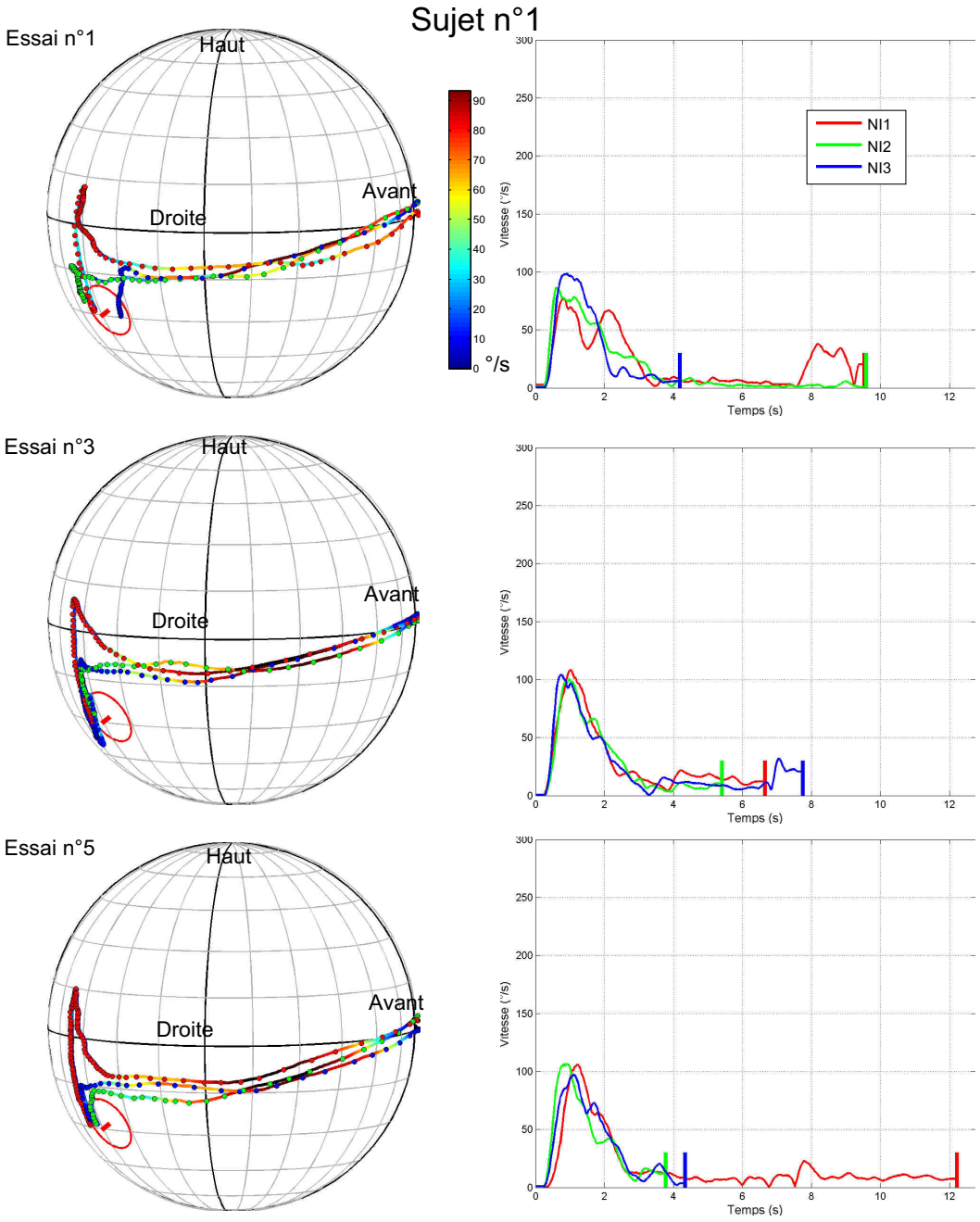


Figure 6.49 – A gauche : trajectoires adoptées par l'axe médian du sujet n°1 pour la direction n°1 (azimut 229°, élévation -28.15°, système polaire vertical), pour les essais 1, 3 et 5, et pour les conditions de contrôle non-individuelles NI1, NI2 et NI3. La vitesse angulaire est codée en couleur le long de ces trajectoires. La direction de la source est matérialisée par un trait rouge, et le cône de la validation qui l'entoure par un cercle rouge. A droite : vitesse angulaire correspondante. Les traits verticaux à la fin de chaque courbe marquent l'instant de la validation. Un trait de couleur noire indique un dépassement du temps imparti (30 s).

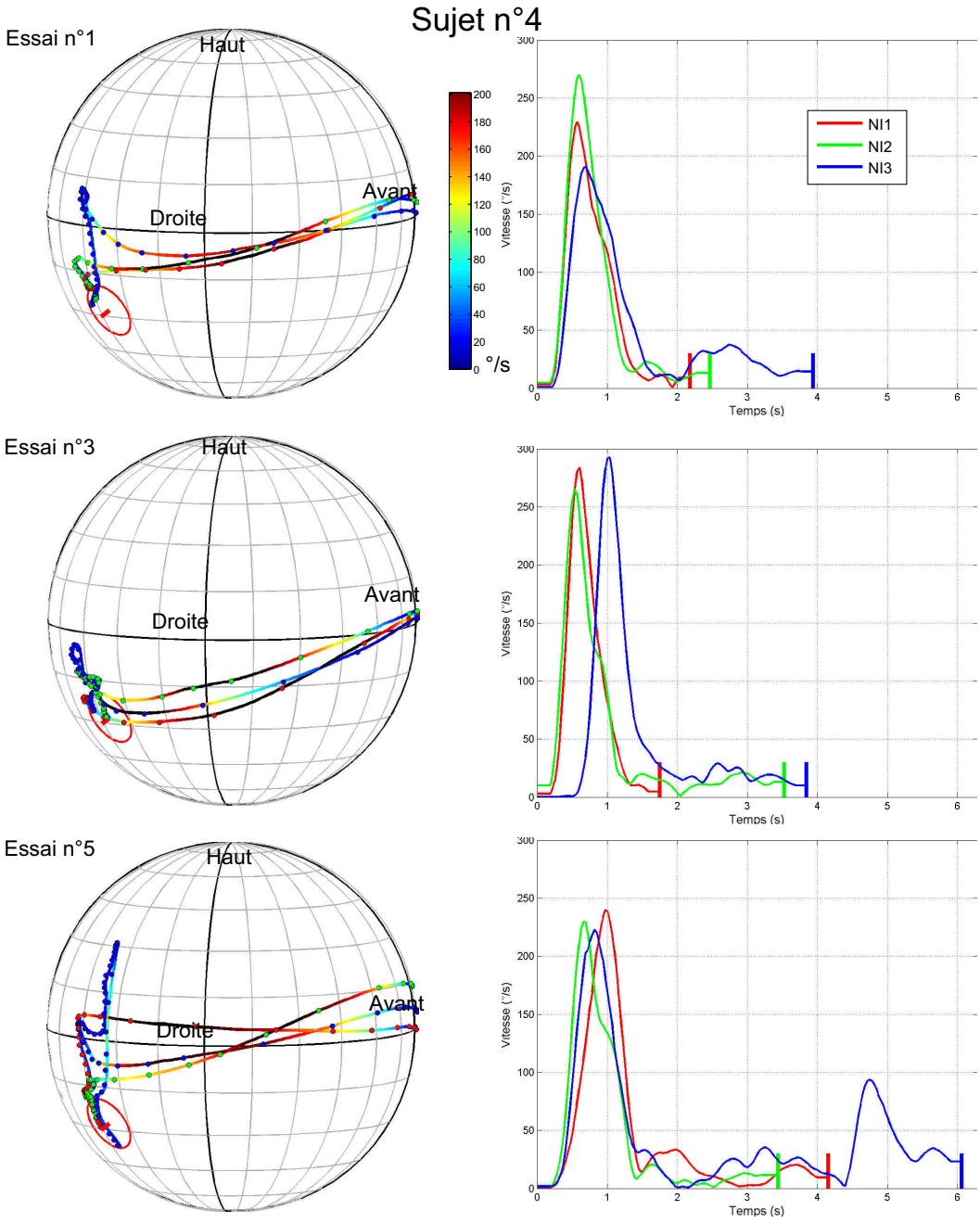


Figure 6.50 – Comportement du sujet n°4 pour la direction n°1 (azimut 229°, élévation -28.15°, système polaire vertical). Voir figure 6.49 pour les détails.

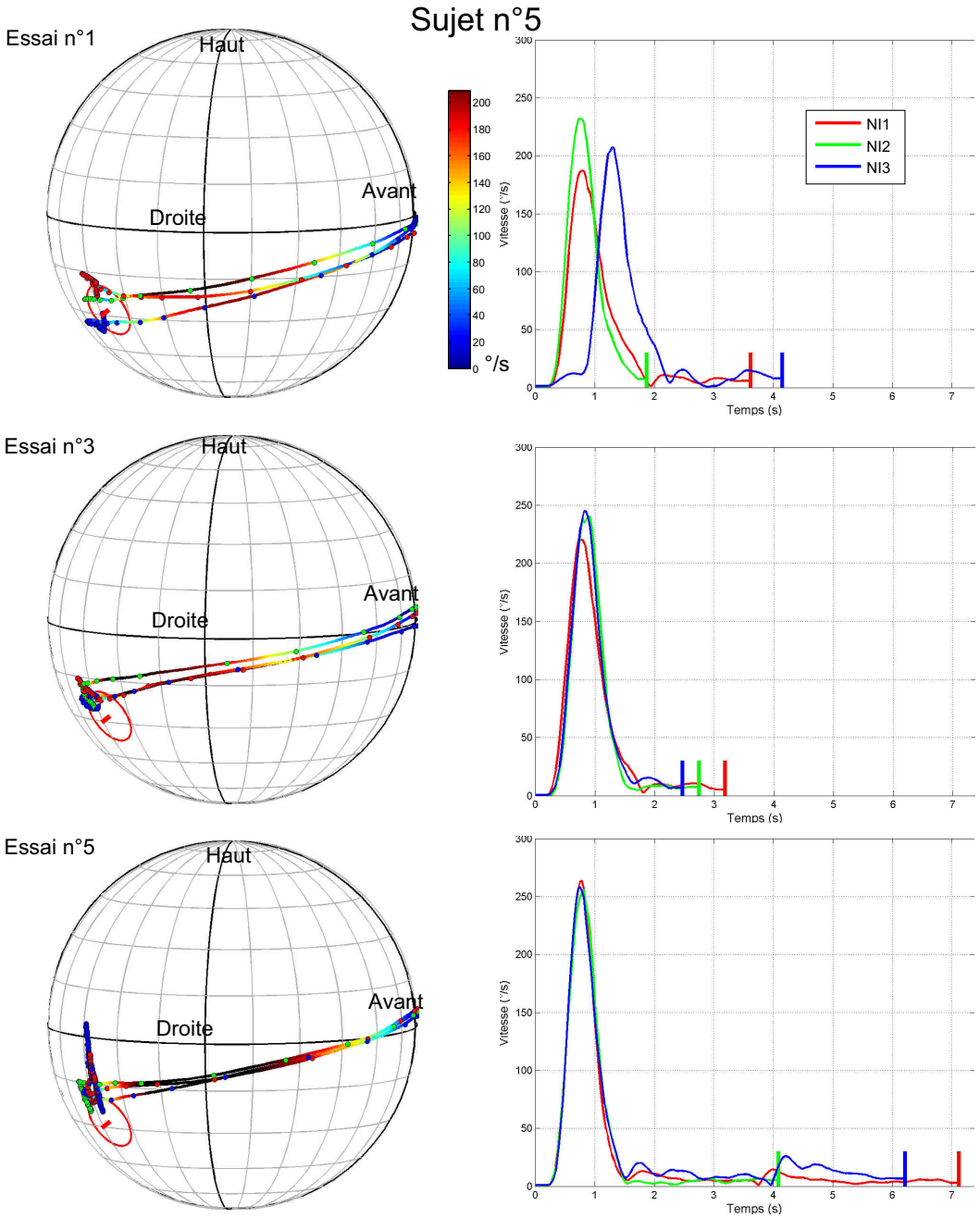


Figure 6.51 – Comportement du sujet n°5 pour la direction n°1 (azimut 229°, élévation -28.15°, système polaire vertical). Voir figure 6.49 pour les détails.

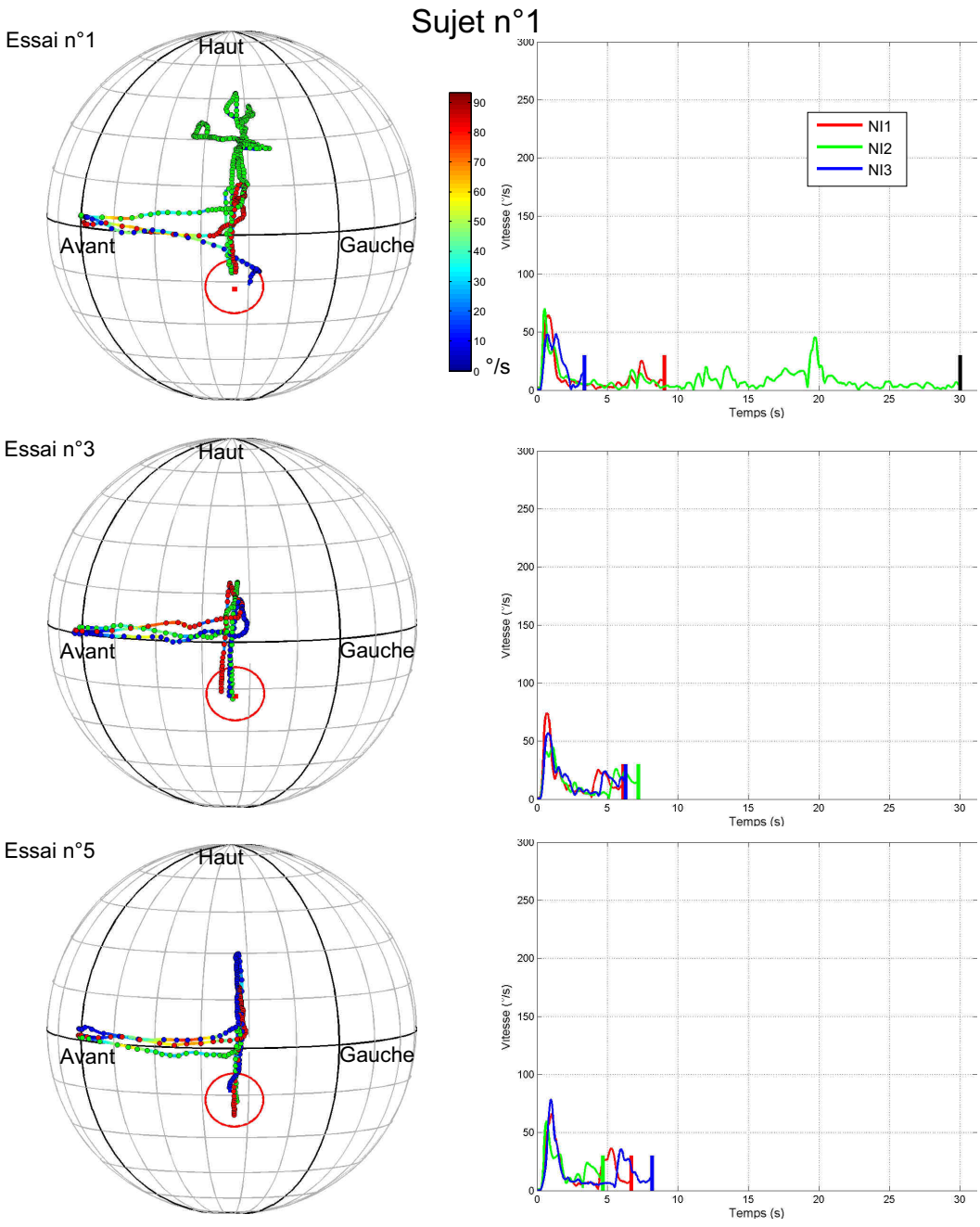


Figure 6.52 – Comportement du sujet n°1 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.49 pour les détails.

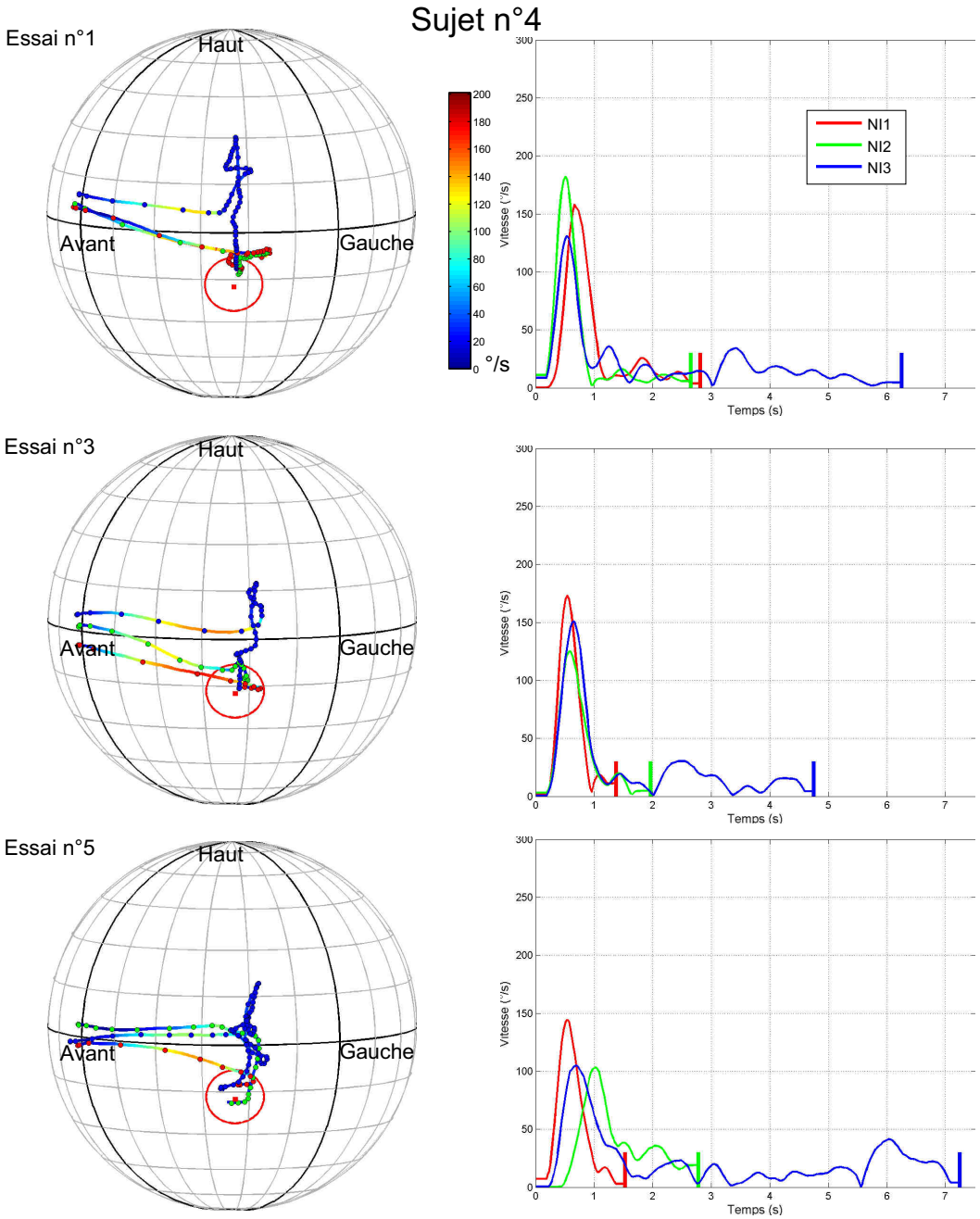


Figure 6.53 – Comportement du sujet n°4 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.49 pour les détails.

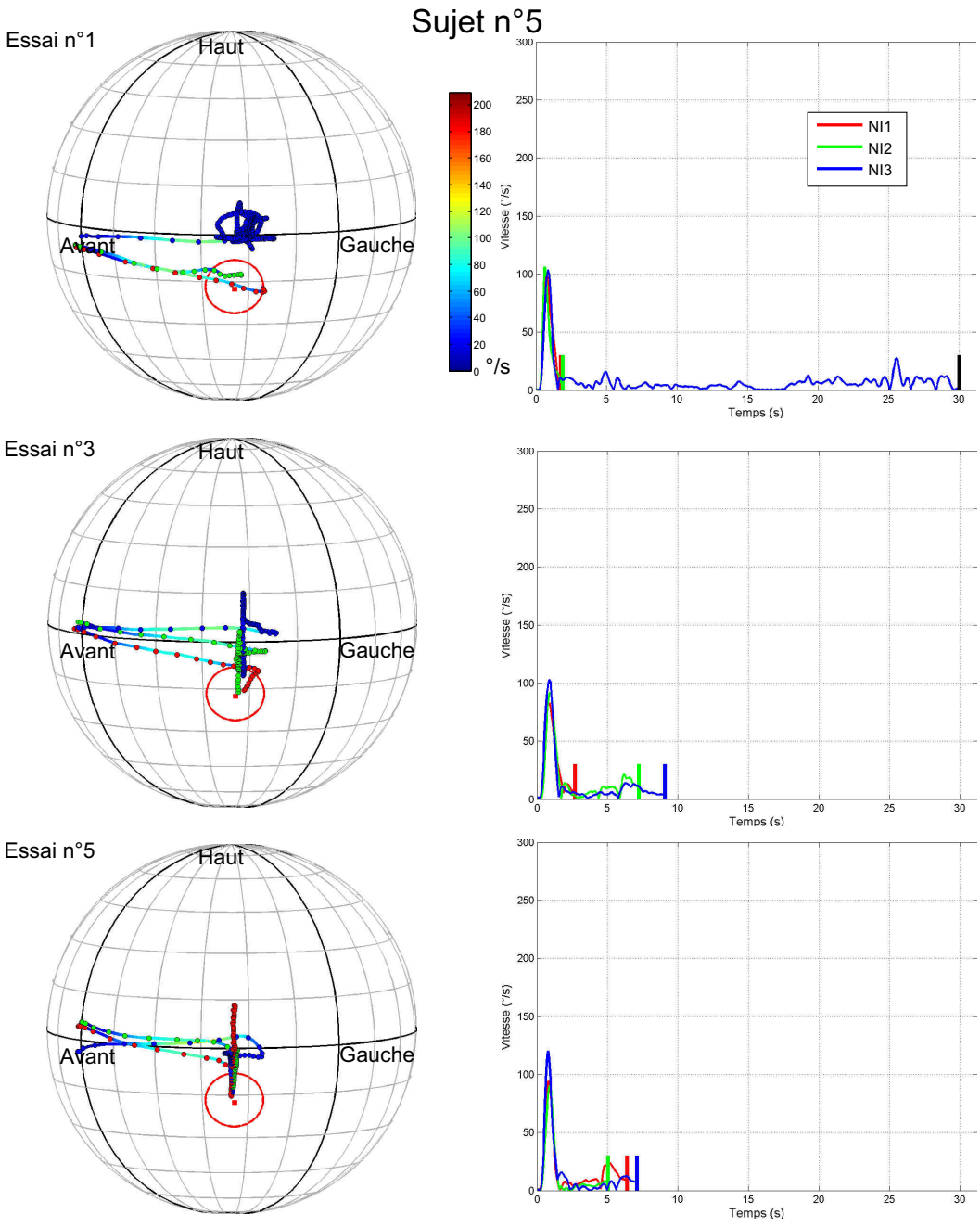


Figure 6.54 – Comportement du sujet n°5 pour la direction n°7 (azimut 55.4°, élévation -16.9°, système polaire vertical). Voir figure 6.49 pour les détails.

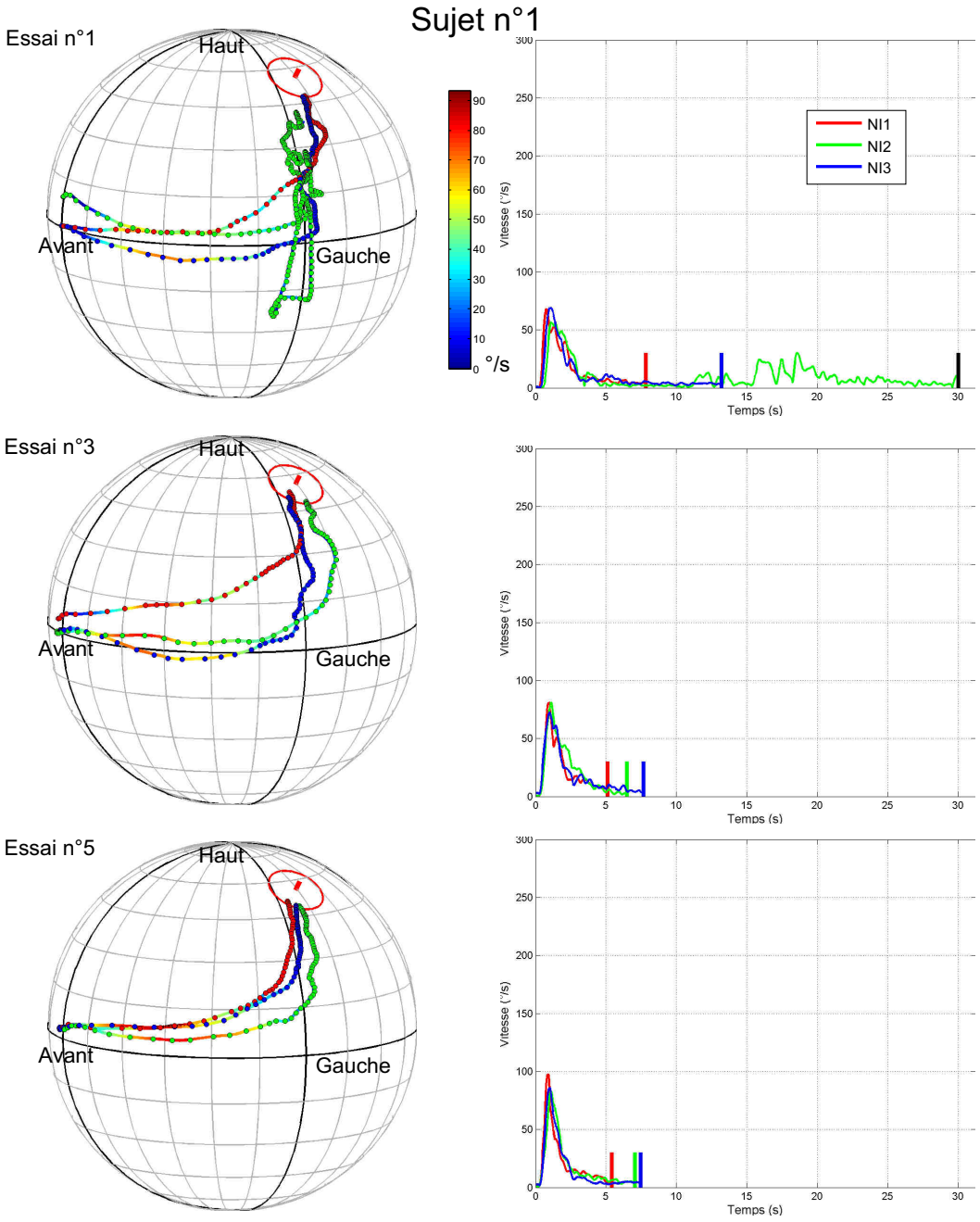


Figure 6.55 – Comportement du sujet n°1 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.49 pour les détails.

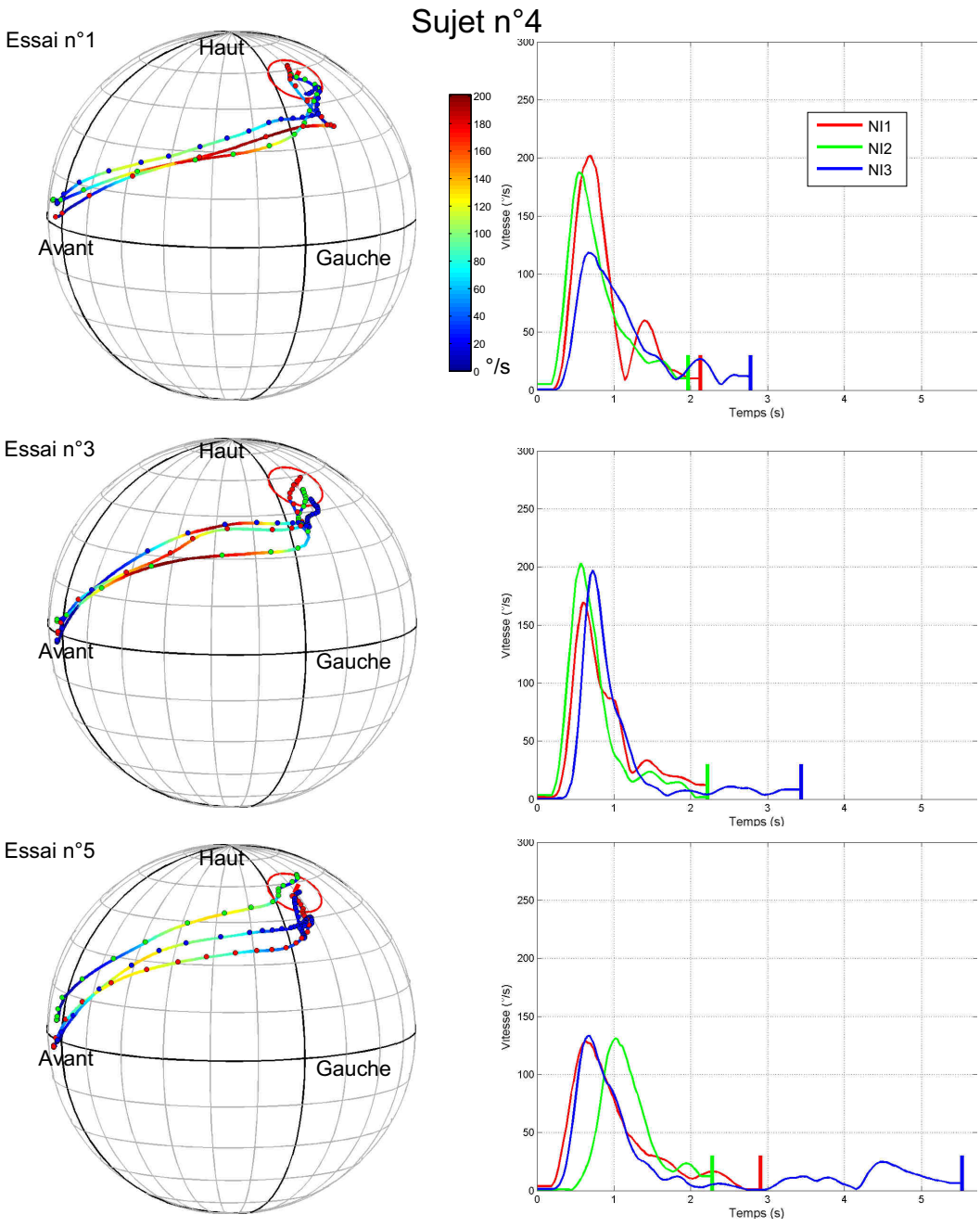


Figure 6.56 – Comportement du sujet n°4 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.49 pour les détails.

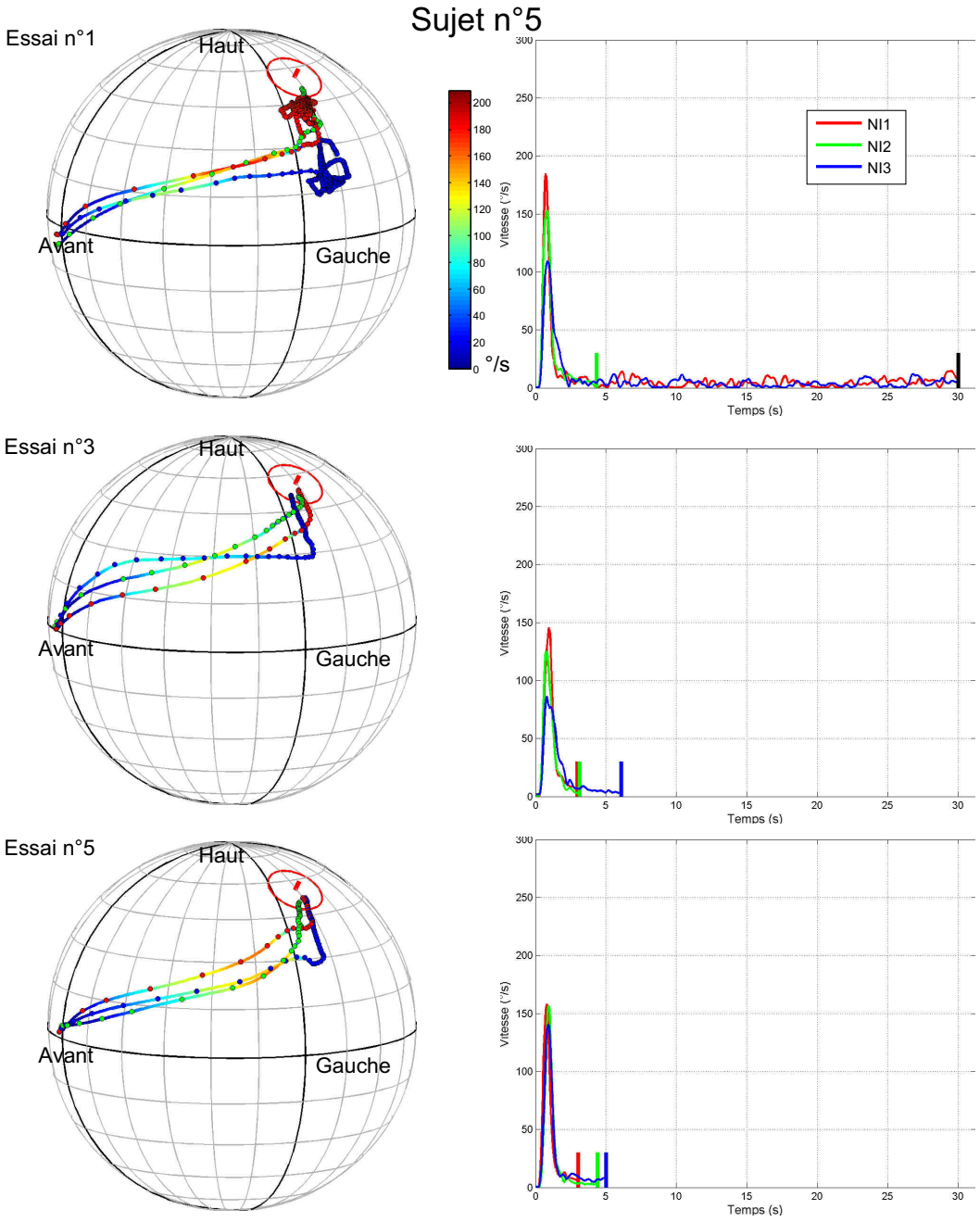


Figure 6.57 – Comportement du sujet n°5 pour la direction n°35 (azimut 105°, élévation 56.25°, système polaire vertical). Voir figure 6.49 pour les détails.

Sujet n°	Essai n°	Condition		
		NI1	NI2	NI3
1	1	0	10	2
	2	0	0	0
	3	1	0	0
	4	0	0	0
	5	0	0	1
4	1	0	0	1
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0	0	0
5	1	10	2	0
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0	0	0

Table 6.3 – Nombre de dépassements du temps imparti (30 secondes) en fonction du sujet, de l'essai et de la condition de test.

montrent des performances bien moindres que celles observées avec les conditions de test. Pour le sujet n°4, le paramètre τ est croissant avec l'ISSD, et sa valeur semble assez basse pour la condition NI1. Enfin pour le sujet n°5, les performances ne sont clairement dégradées que pour la condition NI3.

On représente figure 6.61 les distributions ex-gaussiennes ajustées sur les distributions des temps de réponse normalisés, en considérant conjointement les données des essais n° 3, 4 et 5. On y compare la condition I à chacune des conditions R19 à R121 et NI1 à NI3. De plus, une série de tests d'hypothèse par permutation permet selon la technique décrite précédemment de conclure sur la significativité des différences observées. Si l'on note de façon générale NI la condition de contrôle, il s'agit de tester l'hypothèse d'identité :

$$H_0 : F_{NI} \equiv F_I \tag{6.15}$$

On représente conjointement sur le tableau 6.4 les résultats de ces tests pour les conditions de contrôle avec ceux menés pour les conditions de test pour les trois sujets concernés. Comme le laissait penser l'observation du paramètre τ, aucun des jeux de HRTF non-individuelles n'offre une spatialisation satisfaisante pour le sujet n°1. Le sujet n°4 obtient de très bonnes performances dans les conditions NI1, de même que le sujet n°5 dans les conditions NI1 et NI2. C'est ce que confirme l'observation des distributions ex-gaussiennes ajustées sur les données correspondantes (cf. Fig. 6.61).

Les appréciations libres exprimées par les sujets sur les différentes qualités de spatialisation sont instructives. D'abord, tous les sujets ont perçu d'emblée des différences nettes de la spatialisation dans les conditions NI1, NI2 et NI3 par rapport à toutes les conditions précédemment testées (R19 à R121 et I). Le sujet n°1 a jugé les 3 conditions NI1, NI2 et NI3 très difficiles, au sens où la localisation était extrêmement biaisée. Ce sujet a admis avoir accompli les tâches de localisation en s'adaptant à cette perception biaisée, mais sans jamais s'approprier les indices de localisation fournis pas les HRTF non-individuelles. Les mêmes commentaires ont été laissés par les sujets n°4 et 5 pour la condition NI3. Bien que l'analyse quantitative montre que de hautes performances sont obtenues selon la condition NI1 pour le sujet n°4, et selon les conditions NI1 et NI2 pour le sujet n°5, quelques critiques ont été émises sur la spatialisation. Le sujet n°5 a indiqué que de fortes colorations affectaient de façon non naturelle le spectre des stimulis. De plus, un léger biais de localisation a été observé tout au long de l'expérience, mais il était suffisamment faible pour que le sujet s'en accomode et réussisse malgré tout la tâche de localisation. Ce même sujet a enfin indiqué avoir rencontré de très furtifs problèmes de confusions avant/arrière et d'externalisation avec chacune des conditions NI1, NI2 et NI3, aux premiers instants de la diffusion des stimuli. Le sujet n°4 a noté selon la condition NI1 que, bien que la spatialisation était en général satisfaisante, une "distorsion de l'espace" apparaissait

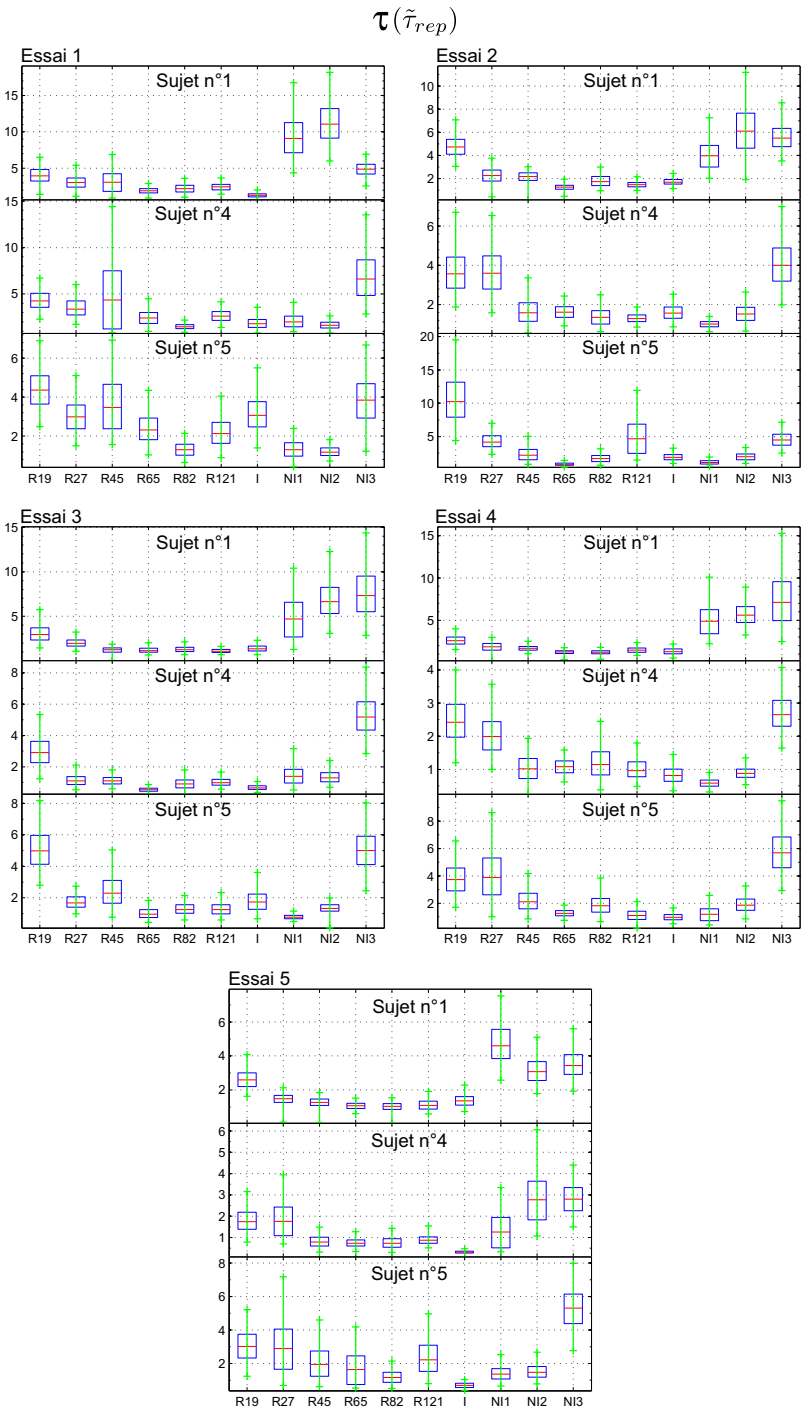


Figure 6.58 – Paramètre τ de la distribution ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les 3 sujets considérés, chaque condition, et chaque essai (l'échelle en ordonnée dépend du sujet et de l'essai).

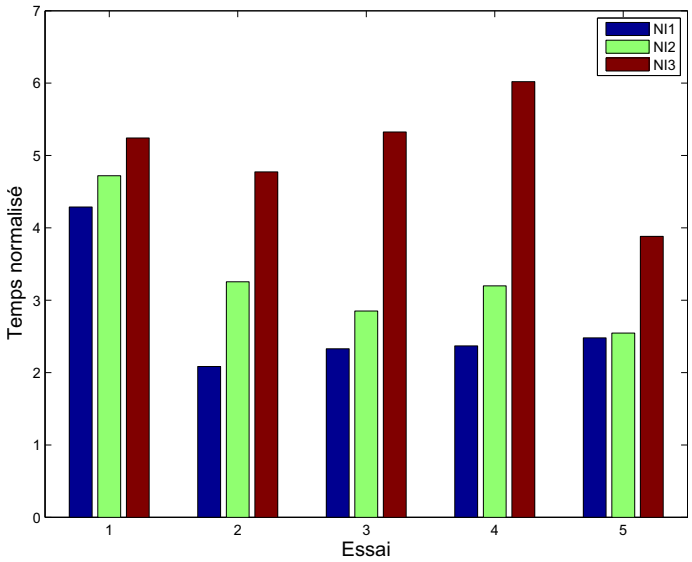


Figure 6.59 – Illustration du phénomène d'apprentissage pour les conditions de contrôle : pour chaque essai et chaque condition les paramètres τ moyens sont moyennés sur les 3 sujets considérés.

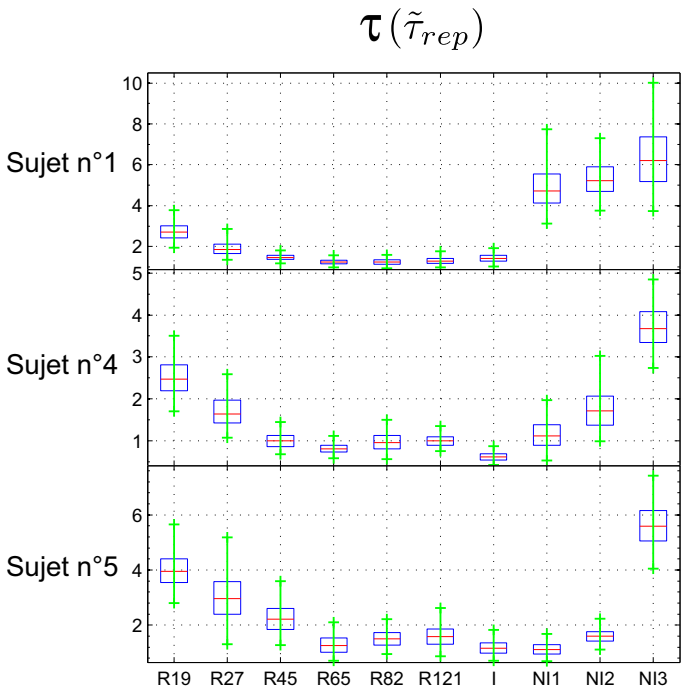


Figure 6.60 – Paramètre τ de la distribution ex-gaussienne ajustée à la distribution des temps de réponse normalisés, pour les différents sujets, chaque condition (les données des essais 3, 4 et 5 sont fusionnées, et l'échelle en ordonnée dépend du sujet).

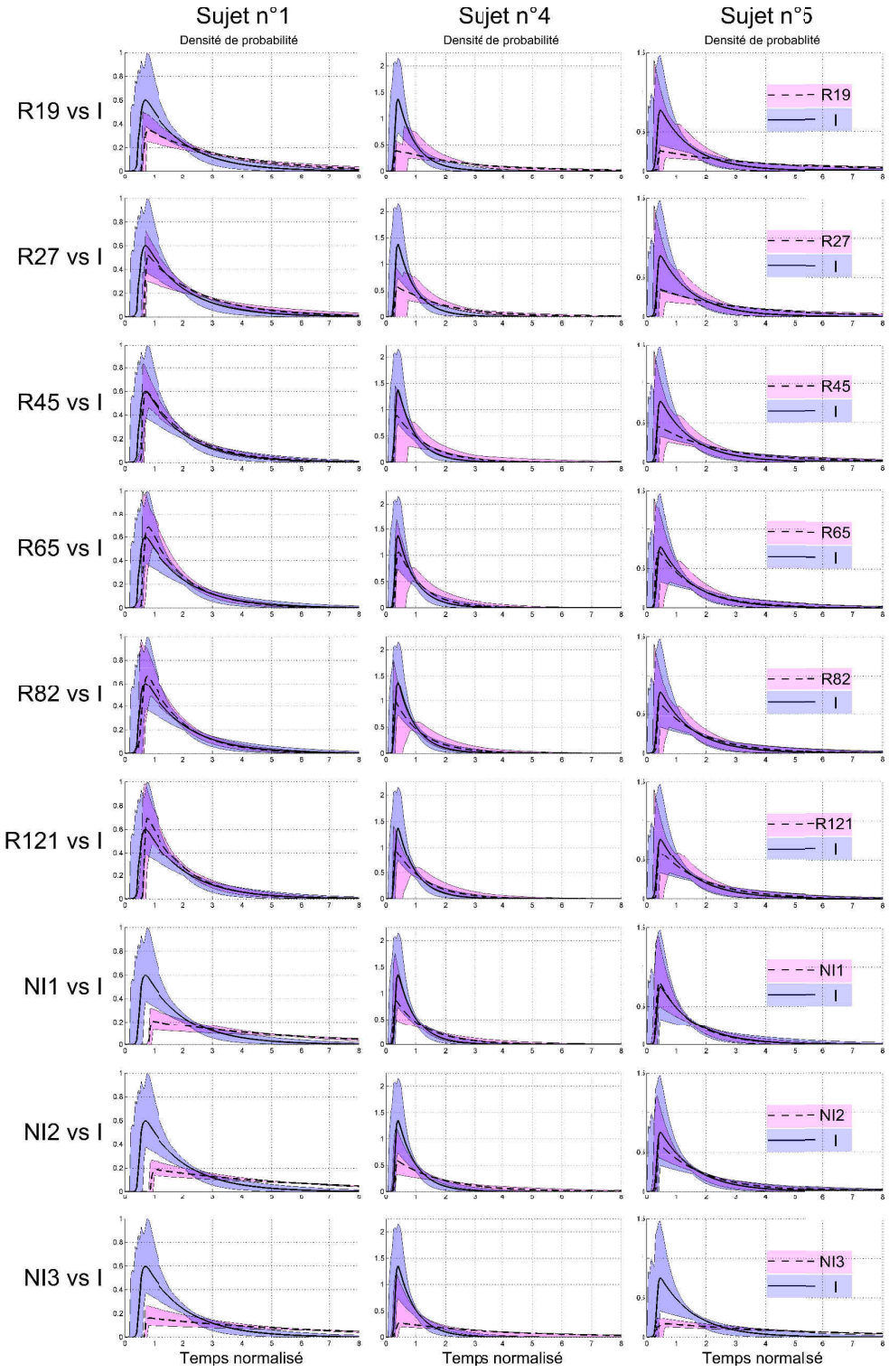


Figure 6.61 – Ajustement de distributions ex-gaussiennes sur les temps de réponse normalisés : pour les sujets n°1, 4 et 5, on considère conjointement les données des essais n° 3, 4 et 5, et on superpose les résultats obtenus pour la condition I à ceux de chaque condition R19 à R121 et NI1 à NI3. Les surfaces colorées correspondent à l'incertitude de l'estimation par MLE des paramètres de l'ex-gaussienne avec *bootstrapping* (intervalle de confiance à 95%), tandis que les courbes représentent les moyennes de toutes les densités de probabilité ex-gaussiennes ajustées.

Sujet n°	Condition								
	R19	R27	R45	R65	R82	R121	NI1	NI2	NI3
1	$\neq$	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\neq$	$\neq$	$\neq$
4	$\neq$	$\neq$	$\neq$	$\equiv$	$\equiv$	$\neq$	$\equiv$	$\neq$	$\neq$
5	$\neq$	$\neq$	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\neq$

Table 6.4 – Résultats des tests de permutation des hypothèses $H_0 : F_R \equiv F_I$ ou $H_0 : F_{NI} \equiv F_I$. On représente le résultat par le symbole $\neq$ ou $\equiv$ selon que l'hypothèse H_0 est respectivement rejetée ou non.

pendant les mouvements rapides de la tête. L'évolution dynamique des indices de localisation est donc perçue comme médiocre. Cette observation est probablement à mettre sur le compte des indices temporels, que le système auditif est le plus apte à analyser de manière dynamique. On remarque en effet figure 6.28 que globalement l'ITD dans la condition NI1 est particulièrement mal adaptée au sujet n°4, car pour près de 50% des 965 directions de l'espace, la différence observée par rapport à l'ITD individuelle est au-dessus de la JND.

6.4.5 Discussion

Les conditions de contrôle ont été introduites pour s'assurer de la capacité de la méthodologie proposée à évaluer la qualité de spatialisation offerte par un jeu de HRTF donné. Les HRTF de la condition NI3 peuvent être considérées comme une ancre de basse qualité, car elles engendrent une spatialisation que les sujets jugent eux-mêmes comme très mauvaise. Les outils d'analyse proposés identifient clairement cette condition comme significativement différente de la condition individuelle I, ce qui dénote un contraste satisfaisant. Il semble néanmoins essentiel de considérer les appréciations des sujets en complément de cette analyse quantitative des performances. En effet, certains défauts de spatialisation ont été perçus sans que cela n'affecte la capacité des sujets à accomplir la tâche de localisation, comme s'ils avaient réussi à compenser un certain handicap de perception. D'après toutes ces observations, on peut tirer plusieurs conclusions.

D'abord, les conditions R19 et R27 sont insatisfaisantes, car les sources virtuelles ont été systématiquement jugées comme diffuses, et même globalement, l'espace a été perçu comme flou. L'étendue spatiale des sources est ici un attribut d'autant plus saillant que les sujets disposaient d'indices dynamiques pour en juger. On peut probablement relier cette perception diffuse à un manque de granularité du jeu de HRTF utilisé. En effet, les HRTF des conditions R19 et R27 sont reconstruites à partir d'un nombre très limité de mesures, ce qui se traduit inévitablement par

un lissage spatial prononcé des données. Ainsi, lors d'un mouvement de la tête, l'évolution perçue des indices de localisation peut paraître trop douce, ce qui se traduit par une localisation floue et un manque de réalisme de la scène sonore.

Une amélioration des performances a été observée au cours du temps. Plusieurs composantes sont à considérer dans cet effet d'apprentissage. D'abord, il y a eu un apprentissage moteur : les sujets ont appris à viser avec le visage, mais aussi à stabiliser leur mouvement dès qu'ils percevaient les sources comme frontales. De plus, on pourrait penser que la rétroaction que constitue la validation automatique peut avoir favorisé l'apprentissage des indices de localisation disponibles au niveau central. Néanmoins, l'alternance aléatoire des différentes jeux de HRTF au cours du test constituait un obstacle important, car une telle adaptation, basée sur la plasticité du système auditif, nécessite généralement une utilisation prolongée d'un jeu de HRTF [93]. C'est donc plus probablement un apprentissage à un niveau plus conscient qui a pu intervenir, grâce à une familiarisation avec le spectre. Les sujets n°4 et 5 ont affirmé avoir dans un premier temps perdu en partie leurs repères spatiaux en conditions NI1 et NI2. La localisation était suffisamment correcte pour les mener rapidement dans le voisinage de la source, mais selon ces sujets les derniers mouvements menant à la validation étaient non pas guidés par le percept spatial de la source, mais plutôt par une identification consciente de l'évolution dynamique du spectre de la source autour de la direction frontale. Cet aspect n'a été signalé face aux HRTF reconstruites que pour les conditions R19 et R27. Ces remarques suggèrent que les stratégies mises en place par les sujets dans leur accomplissement de la tâche ont pu masquer l'existence d'artefacts de la spatialisation en conditions non-individuelles.

L'analyse des résultats en conditions non-individuelles est une nouvelle validation de l'ISSD comme mesure efficace de la dissimilarité entre deux jeux de HRTF. On note en effet pour chaque sujet une corrélation entre l'augmentation de l'ISSD et l'allongement moyen des temps de réponse (cf. Fig. 6.27 et 6.60). Cependant, une même valeur d'ISSD a un impact différent d'un sujet à l'autre. En effet, la condition NI2 présente pour les sujets n°4 et 5 une même valeur d'ISSD, mais pourtant les performances correspondantes sont équivalentes à la condition individuelle pour l'un, et significativement différente de cette condition pour l'autre. L'ISSD ne permet donc pas de déterminer une limite stricte à partir de laquelle les artefacts de localisation deviennent plus qu'anecdotiques.

L'analyse particulière des conditions NI3 pour les sujets n°4 et 5 est également instructive. Pour ces deux sujets, cette condition correspondait au jeu de HRTF du sujet n°1. Les résultats très mauvais observés sont sans équivoque : ces HRTF ne conviennent pas à ces sujets, et ce n'est pas étonnant au regard de la distance objective observée. Le sujet n°1, comme tous les autres, avait passé préalablement

à cette expérience une série de tests de localisation plus classiques, réalisés dans le cadre d'une étude qui visait à valider les données de la base d'Orange Labs [197]. Ce sujet s'est distingué par de très bonnes performances de localisation, et pourrait en ce sens être classé dans la catégorie des bons localisateurs. Notre expérience est donc une occasion de plus remettre en question le vague principe énoncé par Wenzel *et al.* [265], et évoqué en 4.2.3, selon lequel il suffirait d'utiliser les HRTF d'un bon localisateur pour obtenir une spatialisation peu dégradée par rapport à celle offerte par des HRTF individuelles. Les indices de localisation ont beau être particulièrement saillants, et donc facilement interprétables par le système auditif du bon localisateur lui-même, ils peuvent être absolument inadaptés pour un sujet quelconque. Il est donc sans fondement de choisir des HRTF non-individuelles sur la seule base des performances de localisation de leur propriétaire.

Enfin, il aurait probablement été préférable de réaliser ce test de localisation dans l'obscurité, ou en bandant les yeux des participants. En effet, deux des sujets ont relevé une difficulté à percevoir les sources à distance lointaine quand elles étaient poisitionnés là où des objets (table, écran, dispositif de *head-tracking*) occupaient l'espace. Il s'agit certainement du phénomène appelé *proximity-image effect* [72], évoqué au chapitre 1 : en l'absence d'indices visuels matérialisant la source virtuelle, sa distance apparente est déterminée par la distance de l'objet à portée de vue le plus proche susceptible d'être à l'origine du son perçu. La distance n'était cependant pas un attribut d'intérêt dans cette évaluation, et selon les commentaires des sujets eux-mêmes, ce phénomène n'a pas perturbé leur perception de la direction.

On peut finalement conclure que généralement la technique de reconstruction proposée nécessite entre 45 et 65 directions de mesure pour assurer une transparence perceptive de la reconstruction. De plus, l'intérêt de l'individualisation est confirmé, car l'utilisation de HRTF non-individuelles offre généralement une spatialisation de qualité moindre que celle obtenue avec les HRTF reconstruites selon la technique proposée, dès lors que le nombre de mesures individuelles est suffisant.

6.5 Conclusion

Nous avons proposé une technique d'individualisation qui tire parti de l'analyse d'une base de données, afin de reconstruire, à partir d'un nombre réduit de mesures individuelles, les HRTF d'un nouvel auditeur dans les directions intermédiaires. L'évaluation objective a permis de montrer pour divers critères que la reconstruction des données selon cette méthode est plus fidèle que selon les techniques aveugles de l'état de l'art. L'évaluation subjective en synthèse binaurale dynamique permet de confirmer ces résultats. Il apparaît possible que seules 45 à 65 mesures soient nécessaires en entrée de la technique de reconstruction pour obtenir une spatialisation

équivalente à celle offerte par les HRTF individuelles mesurées sur un échantillonnage fin. Ce résultat a été obtenu selon un protocole nouveau, qui offre des conditions d'écoute plus écologiques que ce que permet le cadre plus classique de la synthèse binaurale statique. On sait que les indices dynamiques disponibles facilitent l'obtention d'une bonne spatialisation sonore. C'est pourquoi le nombre limite de 45 à 65 mesures nécessaires pour notre technique ne peut être rigoureusement comparé au résultat de Carlile *et al.*, qui avançaient une limite de 150 mesures avec leur technique, mais avec des conditions d'évaluation plus défavorables. Néanmoins, il demeure que l'évaluation objective révèle la supériorité de notre technique, ce qui permet de prédire de meilleures performances de spatialisation pour un même nombre de mesures.

Afin d'améliorer la technique de recontruction proposée, il semblerait utile d'étudier pour un nombre donné de mesures de HRTF, si des directions optimales se dégagent, c'est-à-dire si le fait de choisir tel jeu de directions de mesure plutôt que tel autre permet une reconstruction plus fidèle des données. On pourrait par exemple concentrer les mesures dans les zones de l'espace où le système auditif est le plus apte à discriminer les indices de deux directions voisines. Un résultat récent vient par exemple de confirmer ce que laissaient supposer les études sur les performances de localisation : les capacités du système auditif à détecter des changements des indices spectraux sont plus faibles pour des sources très élevées [89]. Ce phénomène prend probablement sa source dans le fait que les indices spectraux eux-mêmes connaissent de faibles variations spatiales dans cette zone de l'espace (cf figures 3.8 et 3.9). Cela suggère que la résolution de l'échantillonnage de mesure pourrait être plus grossière au-delà d'une certaine élévation limite, au profit d'une résolution plus fine sur le reste de la sphère. Un obstacle à dépasser reste le fait que notre technique inclut une décomposition des données en harmoniques sphériques, et que pour cela l'échantillonnage de mesure doit couvrir la sphère de façon homogène. Si l'on conserve cette homogénéité, il reste un degré de liberté sur le choix des directions de mesure. On pourrait imaginer que les HRTF dans certaines directions portent en elles davantage d'informations sur la forme globale des SFRS, et ainsi que leur mesure favoriserait la reconstruction des données. Il faudrait cependant dégager des constantes indépendantes de l'individu, de sorte que ces directions soient optimales pour un sujet quelconque. Or, un des principes fondateurs de la technique proposée est précisément le fait que, d'un sujet à l'autre, les mêmes phénomènes peuvent être observés avec un décalage spatial. Nos hypothèses de travail semblent donc tout à fait contradictoires avec l'idée qu'un unique jeu de directions de mesures pourrait être optimal de façon universelle. C'est aussi une des conclusions de l'étude de Busson [38] décrite en 4.2.6.

La mise en place d'une évaluation subjective a été l'occasion de proposer un nouveau protocole. On a pu montrer que nos hypothèses de départ étaient pertinentes : le temps de réponse des sujets est un bon indicateur pour discriminer différentes

qualités de spatialisation. En outre, on a confirmé l'importance de s'appuyer sur le ressenti des sujets : cela nous a permis de mettre à jour des imperfections invisibles dans une analyse quantitative des temps de réponse, notamment des colorations non naturelles, des biais systématiques de localisation et des distorsions de l'espace sonore. Des tests complémentaires, permettant d'inciter les sujets à déceler ces défauts, pourraient enrichir le protocole proposé. D'abord, les stimuli naturels - musique ou parole - permettent de détecter plus facilement des problèmes de détimbrage. De tels stimuli pourraient être utilisés brièvement dans une seule position de l'espace, et seuls les mouvements propres du sujet lui suffiraient pour évaluer d'éventuels phénomènes de coloration. Les biais de localisation pourraient être révélés par une comparaison multimodale : un point lumineux pourrait matérialiser la position de la source virtuelle et permettre à l'auditeur d'évaluer la discordance éventuelle entre les deux modalités perceptives.

On peut interpréter comme suit le phénomène de distorsion de l'espace évoqué par un des sujets. D'après les informations que l'expérimentateur lui a données, l'auditeur fait l'hypothèse d'une source virtuelle fixe dans le référentiel du laboratoire. Au début de la diffusion du stimulus, un percept spatial de cette source se forme, et l'auditeur engage un mouvement dans sa direction. Il a conscience que tous les changements acoustiques perçus ne seront le fruit que de ses mouvements propres, et d'après le percept spatial initial, il est même en mesure de prévoir la relation entre un mouvement donné et les changements acoustiques associés. Une rupture de la localisation intervient donc dès que les changements acoustiques de la source, perçus à la fin d'une série de mouvements propres de l'auditeur, ne semblent pas pouvoir être compensés par la même série de mouvements, mais opérée en sens inverse. Selon Poincaré [203], c'est notamment cette notion de compensation qui permet d'expliquer comment, par une exploration active, et par un apprentissage des liens entre nos sensations musculaires et les changements externes associés, on appréhende les principes géométriques du monde qui nous entoure. La perception d'une distorsion de l'espace apparaît donc probablement dès que le percept spatial de la source évolue de façon incohérente au cours des mouvements de l'auditeur. La diffusion prolongée des sources sonores ainsi que la mise en œuvre dynamique de la synthèse binaurale permet aux auditeurs de déceler facilement ce type de défaut. On pourrait imaginer une étape de l'évaluation perceptive qui en favorise le jugement. Il suffirait de générer dans une même scène sonore plusieurs sources à des positions cardinales connues du sujet, et de les diffuser de façon entrelacée. Ainsi, l'auditeur pourrait non seulement détecter des distorsions spatiales d'après l'évolution dynamique de chacune des sources, mais il pourrait en outre juger la cohérence globale de la scène par l'analyse des relations spatiales entre les sources, qui doivent évoluer de façon rigide pour que le réalisme de la scène soit maintenu.

Van Gogh, "Autoportrait à l'oreille bandée", 1889

Conclusion

L'objectif principal de cette thèse était de répondre à une question concrète : proposer des solutions simples pour offrir à tout nouvel auditeur des HRTF adaptées, en vue de créer des VAS dotés d'une spatialisation proche d'une écoute naturelle. Néanmoins, puisque la synthèse binaurale repose sur une illusion fragile, dont les fondements psychoacoustiques sont complexes, nous avons consacré la première partie de nos travaux à la synthèse des résultats de la littérature nécessaires à la compréhension du traitement par le système auditif des indices spectraux de la localisation. C'est l'observation minutieuse des HRTF, collectées dans différentes bases de données, qui nous a ensuite permis de dégager les idées fondamentales qui ont guidé de nos investigations : malgré des différences apparentes marquées, il existe des similarités entre les HRTF, potentiellement masquées par deux sources morphologiques de variabilité d'un individu à l'autre, que sont la taille et l'orientation des pavillons. Nous avons développé des outils appropriés à la recherche de ces similarités cachées. L'autre principe fondamental sur lequel s'appuient les solutions techniques proposées est le fait que pour obtenir les HRTF adaptées à un nouvel auditeur, on peut avantageusement tirer parti d'une base de données de HRTF mesurées sur de nombreux individus.

La première solution développée vise à adapter, pour un nouvel auditeur, les HRTF d'un autre individu. Les transformations à appliquer à ce jeu de HRTF - *scaling* fréquentiel et rotation du système de coordonnées - sont contrôlées par le résultat d'une comparaison morphologique entre les pavillons des deux sujets. Cette méthode est l'extension au cas de l'humain d'une idée préalablement proposée pour la gerbille de Mongolie. Nos contributions principales sont les suivantes :

- Adaptation d'un algorithme d'alignement de surfaces en 3D (ICP), avec incorporation de l'homothétie comme degré de liberté complémentaire ;
- Acquisition en 3D des morphologies des 6 sujets impliqués dans l'expérience, au moyen d'un scanner laser portable ;
- Détermination de la portion pertinente de la surface des pavillons, à retenir lors de la comparaison morphologique ;
- Etablissement du lien entre morphologie et transformation des HRTF.

La seconde solution permet de reconstruire les HRTF d'un nouvel auditeur pour une direction quelconque de l'espace, à partir d'un nombre réduit de HRTF individuelles mesurées. L'analyse d'une base de données permet de dégager des prototypes, utilisés comme informations *a priori* dans le processus de reconstruction. Nos contributions sont les suivantes :

- Conception et mise au point d'un modèle de reconstruction de HRTF basé sur l'exploitation d'une base de données et un processus de reconnaissance de formes ;
- Utilisation de l'intercorrélation normalisée comme mesure de similarité ;
- Utilisation d'un algorithme de classification spectrale normalisée pour élire les prototypes de SFRS.

Les deux solutions proposées ont été évaluées, et on a démontré qu'elles atteignaient des performances supérieures à celles des techniques comparables de l'état de l'art.

Il conviendrait à l'avenir de poursuivre l'élaboration de la solution basée sur l'adaptation morphologique de HRTF individuelles. D'abord une évaluation subjective de ses performances est nécessaire pour parfaire sa validation. De plus, la morphologie en 3D d'un nouvel auditeur pourrait être obtenue selon la technique développée par Dellepiane *et. al* [61], pour laquelle il suffit de quelques photographies. Enfin, dans une version alternative de la méthode, les paramètres de transformation des HRTF pourraient être guidés par un ajustement psychophysique.

L'évaluation subjective de la seconde technique a été menée selon un protocole novateur en synthèse binaurale dynamique, permettant des conditions d'écoute écologiques. Au lieu de s'intéresser à la précision angulaire avec laquelle un auditeur peut localiser une source virtuelle, on a considéré le temps nécessaire pour atteindre ces sources avec une précision donnée. On a pu montrer que ce temps de réponse est un indicateur de la qualité de la spatialisation. L'analyse des résultats des tests de localisation a reposé conjointement sur l'utilisation d'outils statistiques appropriés, et sur la prise en considération des impressions ressenties par les sujets. Plusieurs pistes d'amélioration ont été proposées pour une utilisation future de ce protocole d'évaluation.

Il reste que l'évaluation de la qualité des HRTF, au sens de leur capacité à leurrer le système auditif pour lui donner l'illusion d'un espace sonore, est une question encore ouverte, et qui de toute façon n'a de sens que pour un individu donné. C'est bien l'adéquation d'un jeu de HRTF à un auditeur qu'il s'agit d'évaluer. Les tests de localisation sont lourds et complexes à mettre en œuvre, et il conviendrait donc de continuer à chercher des solutions alternatives. Au lieu d'évaluer la direction perçue de sources sonores virtuelles, on pourrait par exemple tirer parti des illusions de localisation induites, en champ libre, par des stimuli de spectre contrôlé. Les expériences

menées avec des stimuli de bande étroite ont en effet montré que l'évolution spatiale des illusions engendrées révèle des informations pertinentes sur le décodage spatial opéré par le système auditif d'un individu. On pourrait avantageusement utiliser ces informations individuelles pour caractériser la pertinence des indices de localisation fournis par un jeu de HRTF.

Annexes

Annexe A

Distribution de Kent

La distribution de Kent, ou distribution de Fisher-Bingham à 5 paramètres est une distribution de probabilité qui est l'équivalent sur la sphère unitaire S^2 d'une distribution normale bivariée. La densité de probabilité $f(\mathbf{x})$ de la distribution de Kent est donnée par la relation :

$$f(\mathbf{x}) = \frac{1}{\mathrm{c}(\kappa,\beta)}\exp\{\kappa\gamma_1 \cdot \mathbf{x} + \beta[(\gamma_2 \cdot \mathbf{x})^2 - (\gamma_3 \cdot \mathbf{x})^2]\}$$

where $\mathbf{x}$, est un vecteur de dimension 3, et $\mathrm{c}(\kappa,\beta)$, est une constante de normalisation. Le paramètreκ, $(\kappa > 0)$ détermine la concentration ou l'étalement de la distribution, tandis que β, (où $0 \leq 2\beta < \kappa$) détermine l'ellipticité des contours d'équiprobabilité. Plus κ, et β, sont élevés, plus la distribution est respectivement concentrée et elliptique. Le vecteur γ_1 est la direction moyenne, et les vecteurs γ_2 et γ_3, sont repsectivement le petit et le grand axe de l'ellipse. γ_2 et γ_3 déterminent l'orientation spatiale sur la sphère des contours d'équiprobabilité, tandis que γ_1 détermine leur centre commun. On représente figure A.1 les réalisations de trois différentes variables aléatoires suivant des distributions de Kent deparamètres différents.

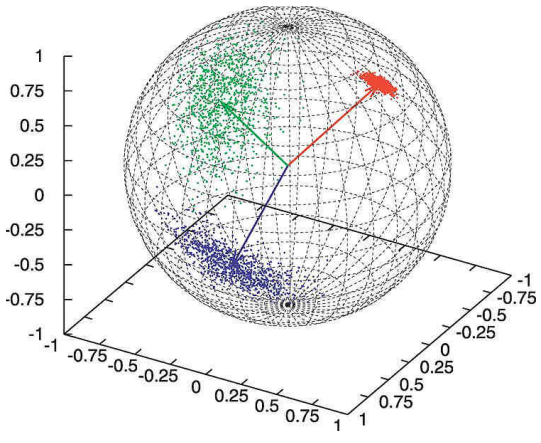


Figure A.1 – Représentation des réalisations de 3 variables aléatoires suivant des distributions de Kent de paramètres différents (un jeu de paramètres par couleur). Les directions moyennes sont matérialisées par des flèches. Le paramètre κ est le plus élevé pour le jeu de points représentés en rouge.(D'après [279]).

Annexe B

Interpolation par spline de type plaque mince sur la sphère (STPS)

La technique d'interpolation par spline de type plaque mince sur la sphère résoud le problème suivant [261] : trouver $u_{\lambda,m} \in \mathcal{H}_m(S^2)$, qui minimise

$$\frac{1}{n}\sum_{j=1}^{n}(u(\chi_i) - z_i)^2 + \lambda J_m(u) \tag{B.1}$$

où

$$J_m(u) = \begin{cases} \int_0^{2\pi}\int_0^{\pi}(\Delta^{m/2}u(\vartheta,\varphi))^2 \sin\vartheta \,\mathrm{d}\vartheta \,\mathrm{d}\varphi & m \text{ pair} \\ \int_0^{2\pi}\int_0^{\pi}\left\{\frac{(\partial(\Delta^{(m-1)/2}u)/\partial\varphi)^2}{\sin^2\vartheta} + (\partial(\Delta^{(m-1)/2}u)/\partial\vartheta)^2\right\}\sin\vartheta \,\mathrm{d}\vartheta \,\mathrm{d}\varphi & m \text{ impair} \end{cases} \tag{B.2}$$

et $\mathcal{H}_m$ est un noyau reproduisant (espace de Sobolev) sur la sphère S^2, Δ est le Laplacien sur la sphère, $\{\chi_i = (\vartheta_i, \varphi_i)\}_{i=1,\ldots,n}$ sont les points de la sphère, représentés par leurs coordonnées sphériques [1], $\{z_i\}_{i=1,\ldots,n}$ sont les valeurs mesurées en ces points. Pour résoudre le problème, la fonction u est décomposée en harmoniques sphériques. Le Laplacien en coordonnées sphériques vaut :

$$\Delta u = \frac{\partial^2 u/\partial\varphi^2}{\sin^2\vartheta} + \frac{\partial(\sin\vartheta\,\partial u/\partial\vartheta)/\partial\vartheta}{\sin\vartheta} \tag{B.3}$$

Les harmoniques sphériques forment un ensemble complet de fonctions propres du Laplacien :

$$\Delta Y_\nu^k = -\nu(\nu+1)Y_\nu^k, \quad k = -\nu,\ldots,\nu, \quad \nu = 0,1,\ldots \tag{B.4}$$

1. L'angle ϑ est la colatitude, et l'angle φ est la longitude. Ils s'expriment en fonction des coordonnées dans le système polaire-vertical selon les relations : $\vartheta = \pi/2 - \phi_{pv}$ et $\varphi = \theta_{pv}$ (cf. Figure 1).

Ils sont définis par la relation :

$$Y_\nu^k(\vartheta,\varphi) = \sqrt{\frac{2\nu+1}{4\pi}\frac{(\nu-k)!}{(\nu+k)!}}\mathcal{P}_\nu^k(\cos\vartheta)e^{ik\varphi} \tag{B.5}$$

où $\mathcal{P}_\nu^k(x)$ est le polynôme de Legendre, tel que :

$$\mathcal{P}_\nu^k(x) = \frac{(-1)^k}{2^\nu\nu!}(1-x^2)^{k/2}\frac{\mathrm{d}^{\nu+k}}{\mathrm{d}x^{\nu+k}}(x^2-1)^\nu \tag{B.6}$$

Soit $\mathcal{H}_m^0(S^2)$ le sous ensemble des fonctions de carré intégrable $\mathcal{L}_2(S^2)$, qui présentent une décomposition sous la forme :

$$u(\chi) = \sum_{\nu=1}^{\infty}\sum_{k=-\nu}^{\nu} u_{\nu k}Y_\nu^k(\chi), \tag{B.7}$$

$$\text{où} \qquad u_{\nu k} = \int_{S^2} u(\chi)Y_\nu^k(\chi)\,\mathrm{d}\chi \tag{B.8}$$

$$\text{et} \qquad \sum_{\nu=1}^{\infty}\sum_{k=-\nu}^{\nu}\frac{u_{\nu k}^2}{\lambda_{\nu k}} < \infty, \quad \lambda_{\nu k}^{-1} = [\nu(\nu+1)]^m \tag{B.9}$$

Les fonctions u de $\mathcal{H}_m^0(S^2)$ sont telles que :

$$\int_{S^2} u(\chi)\,\mathrm{d}\chi = 0 \tag{B.10}$$

Il apparaît que $\mathcal{H}_m^0(S^2)$ est un espace de Hilbert, dont la norme est définie, pour tout $m \geq 0$, par :

$$\| u \|_m^2 = \sum_{\nu=1}^{\infty}\sum_{k=-\nu}^{\nu}\frac{u_{\nu k}^2}{\lambda_{\nu k}} \tag{B.11}$$

On définit $K_m(\chi,\chi')$, où $(\chi,\chi') \in S^2 \times S^2$, pour $m > 1$, par :

$$K_m(\chi,\chi') = \sum_{\nu=1}^{\infty}\sum_{k=-\nu}^{\nu}\lambda_{\nu k}Y_\nu^k(\chi)Y_\nu^k(\chi') \tag{B.12}$$

$$= \frac{1}{4\pi}\sum_{\nu=1}^{\infty}\frac{2\nu+1}{\nu^m(\nu+1)^m}\mathcal{P}_\nu(\cos(\gamma(\chi,\chi'))) \tag{B.13}$$

où $\gamma(\chi,\chi')$ est l'angle sphérique entre les deux points χ et χ'. Il s'ensuit que pour $m > 1$, K_m est le noyau reproduisant de $\mathcal{H}_m^0(S^2)$, relatif au produit scalaire induit par la norme $J_m^{1/2}(\cdot)$ [261] :

$$J_m(u) = \sum_{\nu=1}^{\infty}\sum_{k=-\nu}^{\nu}\frac{u_{\nu k}^2}{\lambda_{\nu k}} \tag{B.14}$$

La solution au problème de minimisation B.1 est donnée par les relations [261] :

$$u_{m,\lambda}(\chi) = \sum_{i=1}^{n} c_i K(\chi,\chi') + d \tag{B.15}$$

où $\mathbf{c} = [c_1, \ldots, c_n]^T$ et d sont tels que :

$$(\mathbf{K}_n + n\lambda\mathbf{I})\mathbf{c} + d\mathbf{F} = \mathbf{z} \tag{B.16}$$

$$\mathbf{F}^T\mathbf{c} = 0 \tag{B.17}$$

$$\mathbf{F} = [1, \ldots, 1]^T \tag{B.18}$$

$$(\mathbf{K}_n)_{ij} = K(\chi, \chi') \tag{B.19}$$

$$\mathbf{z} = [z_1, \ldots, z_n]^T \tag{B.20}$$

Il n'existe de pas de forme approchée pour le calcul de $K(\chi, \chi')$. On résoud le problème en s'intéressant au noyau reproduisant relatif à un produit scalaire induit par une norme $Q_m(u)$ topologiquement équivalente à $J_m(u)$:

$$Q_m(u) = \sum_{\nu=1}^{\infty} \sum_{k=-\nu}^{\nu} \frac{u_{\nu k}}{\xi_{\nu k}} \tag{B.21}$$

où

$$\xi_{\nu k} = \left[\left(\nu + \frac{1}{2}\right)(\nu+1)(\nu+2)\ldots(\nu+2m-1)\right]^{-1} \tag{B.22}$$

Le noyau $R_m(\chi, \chi')$ associé est :

$$R_m(\chi, \chi') = \sum_{\nu=1}^{\infty} \sum_{k=-\nu}^{\nu} \xi_{\nu k} Y_\nu^k(\chi) Y_\nu^k(\chi') \tag{B.23}$$

$$= \frac{1}{2\pi} \sum_{\nu=1}^{\infty} \frac{\mathcal{P}_\nu(\cos(\gamma(\chi, \chi')))}{(\nu+1)(\nu+2)\ldots(\nu+2m-1)} \tag{B.24}$$

Si l'on note $\varrho = \cos(\gamma(\chi, \chi'))$, une expression approchée de R_m est :

$$R_m(\chi, \chi') = \frac{1}{2\pi}\left[\frac{1}{(2m-2)!} q_{2m-2}(\varrho) - \frac{1}{(2m-1)!}\right] \tag{B.25}$$

où

$$q_m(\varrho) = \int_0^1 (1-h)(1-2h\varrho+h^2)^{-1/2} \mathrm{d}h, \quad m = 0, 1, \ldots \tag{B.26}$$

Cette dernière intégrale peut être calculée pour obtenir l'expression algébrique de $q_m(\varrho)$, $m > 1$. En particulier on a :

$$q_1(\varrho) = 2\ln\left(1 + \sqrt{\frac{2}{1-\varrho}}\right)\frac{1-\varrho}{2} - 2\sqrt{\frac{1-\varrho}{2}} + 1 \tag{B.27}$$

$$\begin{aligned} q_2(\varrho) = {} & \tfrac{1}{2}\left\{\ln\left(1+\sqrt{\tfrac{2}{1-\varrho}}\right)\left[12\left(\tfrac{1-\varrho}{2}\right)^2 - 4\left(\tfrac{1-\varrho}{2}\right)\right]\right\} \\ & + \tfrac{1}{2}\left\{12\left(\tfrac{1-\varrho}{2}\right)^{3/2} + 6\left(\tfrac{1-\varrho}{2}\right) + 1\right\} \end{aligned} \tag{B.28}$$

Le problème à résoudre devient finalement :

$$u_{m,\lambda}(\chi) = \sum_{i=1}^{n} c_i R(\chi, \chi') + d \tag{B.29}$$

où $\mathbf{c} = [c_1, \ldots, c_n]^T$ et d sont tels que :

$$\begin{aligned}
(\mathbf{R}_n + n\lambda\mathbf{I})\mathbf{c} + d\mathbf{F} &= \mathbf{z} && \text{(B.30)}\\
\mathbf{F}^T\mathbf{c} &= 0 && \text{(B.31)}\\
\mathbf{F} &= [1, \ldots, 1]^T && \text{(B.32)}\\
(\mathbf{R}_n)_{ij} &= R(\chi_i, \chi_j) && \text{(B.33)}\\
\mathbf{z} &= [z_1, \ldots, z_n]^T && \text{(B.34)}
\end{aligned}$$

Les paramètres d et $\mathbf{c}$ sont estimés simplement d'après les relations B.30 à B.33, en choisissant les points χ_i et χ_j parmi les points de mesure. La relation B.29 permet d'estimer la valeur de la fonction $u_{m,\lambda}$ en un point quelconque χ, les points $\{\chi_i\}_{i=1,\ldots,n}$ étant les points de mesure. La variable λ peut être utilisée comme un paramètre de lissage, le lissage étant d'autant plus prononcé que λ est élevé. Pour $\lambda = 0$, on obtient une interpolation : la fonction $u_{m,\lambda}$ passe rigoureusement par les valeurs mesurées, dans les directions de mesure.

Annexe C

Estimation des paramètres de transformation optimaux de l'ICP

On décrit ici les développements mathématiques, issus de [251], menant à l'expression des paramètres de transformation optimaux, calculés à chaque étape de l'algorithme ICP utilisé en 5.3.1. Pour développer le théorème menant aux expressions 5.7, 5.8 et 5.9, un lemme est nécessaire : il permet l'évaluation des paramètres de rotation dans un problème aux moindres carrés.

C.1 Lemme

Soient $\mathbf{A}$ et $\mathbf{B}$ des matrices $m \times n$, $\mathbf{R}$ une matrice de rotation $m \times m$, et $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ la décomposition en valeurs singulières de $\mathbf{A}\mathbf{B}^T$ (où $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, et $\mathbf{\Lambda}$ est diagonale). Alors le minimum de $\| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2$ en fonction de $\mathbf{R}$ est donné par la relation suivante :

$$\min_{\mathbf{R}} \| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2 = \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi}) \tag{C.1}$$

où

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{A}\mathbf{B}^T) \geq 0 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{A}\mathbf{B}^T) < 0 \end{cases} \tag{C.2}$$

Quand le rang de $\mathbf{A}\mathbf{B}^T$ est supérieur à $m - 1$, la rotation optimale permettant d'atteindre le minimum est déterminé par la relation :

$$\mathbf{R} = \mathbf{U}\mathbf{\Xi}\mathbf{V}^T \tag{C.3}$$

avec en particulier, quand $\det(\mathbf{A}\mathbf{B}^T) = 0$

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = 1 \\ \text{diag}(1, 1, \ldots, -1) & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = -1 \end{cases} \tag{C.4}$$

Démonstration

Soit la fonction F définie selon la relation :

$$F = \| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2 + \text{tr}(\mathbf{L}(\mathbf{R}\mathbf{R}^T - \mathbf{I})) + g\{\det(\mathbf{R}) - 1\} \tag{C.5}$$

où g est un multiplicateur de Lagrange, et $\mathbf{L}$ une matrice symétrique de multiplicateurs de Lagrange. Les deuxième et troisième termes de F représentent respectivement le fait que la matrice $\mathbf{R}$ doit être orthogonale et représenter une rotation. La différentiation partielle de F par rapport à $\mathbf{R}$, $\mathbf{L}$, et g mène au système d'équations suivant :

$$\frac{\partial F}{\partial \mathbf{R}} = -2\mathbf{A}\mathbf{B}^T + 2\mathbf{R}\mathbf{B}\mathbf{B}^T + 2\mathbf{R}\mathbf{L} + g\mathbf{R} = 0 \tag{C.6}$$

$$\frac{\partial F}{\partial \mathbf{L}} = \mathbf{R}^T\mathbf{R} - \mathbf{I} = 0 \tag{C.7}$$

$$\frac{\partial F}{\partial g} = \det(\mathbf{R}) - 1 = 0 \tag{C.8}$$

où l'on a utilisé la relation :

$$\frac{\partial}{\partial \mathbf{R}}\det(\mathbf{R}) = \text{adj}(\mathbf{R}^T) = \det(\mathbf{R}^T)(\mathbf{R}^T)^{-1} = \mathbf{R} \tag{C.9}$$

valable car $\mathbf{R}$ est une rotation. De la relation C.6, il vient :

$$\mathbf{R}\mathbf{L}' = \mathbf{A}\mathbf{B}^T, \quad \text{où} \quad \mathbf{L}' = \mathbf{B}\mathbf{B}^T + \mathbf{L} + \frac{1}{2}g\mathbf{I} \tag{C.10}$$

$\mathbf{L}'$ étant symétrique, on obtient en transposant C.10 de chaque côté :

$$\mathbf{L}'\mathbf{R}^T = \mathbf{B}\mathbf{A}^T \tag{C.11}$$

En multipliant chaque côté de C.10 respectivement par chaque côté de C.11, on obtient C.12, car $\mathbf{R}^T\mathbf{R} = \mathbf{I}$.

$$\mathbf{L}'^2 = \mathbf{B}\mathbf{A}^T\mathbf{A}\mathbf{B}^T = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T \tag{C.12}$$

Puisque $\mathbf{L}'$ et $\mathbf{L}'^2$ commutent ($\mathbf{L}'\mathbf{L}'^2 = \mathbf{L}'^2\mathbf{L}'$), chacune de ces matrices peut être diagonalisée par la même matrice orthogonale. On peut donc écrire :

$$\mathbf{L}' = \mathbf{V}\mathbf{\Lambda}\mathbf{\Xi}\mathbf{V}^T \tag{C.13}$$

où $\mathbf{\Xi} = \text{diag}(\xi_i), \quad \xi_i = 1 \quad \text{ou} - 1$. D'après C.13, il vient :

$$\begin{aligned} \det(\mathbf{L}') &= \det(\mathbf{V}\mathbf{\Lambda}\mathbf{\Xi}\mathbf{V}^T) \\ &= \det(\mathbf{V})\det(\mathbf{\Lambda})\det(\mathbf{\Xi})\det(\mathbf{V}^T) \\ &= \det(\mathbf{\Lambda})\det(\mathbf{\Xi}). \end{aligned} \tag{C.14}$$

Par ailleurs, d'après C.10 :

$$\begin{aligned} \det(\mathbf{L}') &= \det(\mathbf{R}^T\mathbf{A}\mathbf{B}^T) \\ &= \det(\mathbf{R}^T)\det(\mathbf{A}\mathbf{B}^T) \\ &= \det(\mathbf{A}\mathbf{B}^T). \end{aligned} \tag{C.15}$$

Donc :

$$\det(\mathbf{\Lambda})\det(\mathbf{\Xi}) = \det(\mathbf{A}\mathbf{B}^T) \tag{C.16}$$

Puisque les valeurs singulières sont positives ou nulles, $\det(\mathbf{\Lambda}) = \lambda_1\lambda_2\ldots\lambda_m \geq 0$. Donc nécessairement $\det(\mathbf{\Xi}) = 1$ quand $\det(\mathbf{A}\mathbf{B}^T) > 0$, et $\det(\mathbf{\Xi}) = -1$ quand $\det(\mathbf{A}\mathbf{B}^T) < 0$.

Les extrema de $\| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2$ sont donc obtenus comme suit. D'après C.10, il vient :

$$\begin{aligned} \| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2 &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\text{tr}(\mathbf{A}\mathbf{B}^T\mathbf{R}^T) \\ &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\text{tr}(\mathbf{R}^T\mathbf{A}\mathbf{B}^T) \\ &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\text{tr}(\mathbf{L}'). \end{aligned} \tag{C.17}$$

En substituant C.13 dans C.17, on obtient :

$$\begin{aligned} \| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2 &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\text{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{\Xi}\mathbf{V}^T) \\ &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2\,\text{tr}(\mathbf{\Lambda}\mathbf{\Xi}) \\ &= \| \mathbf{A} \|^2 + \| \mathbf{B} \|^2 - 2(\lambda_1\xi_1 + \lambda_2\xi_2 + \ldots\lambda_m\xi_m). \end{aligned} \tag{C.18}$$

Le minimum de $\| \mathbf{A} - \mathbf{R}\mathbf{B} \|^2$ est donc atteint quand $\xi_1 = \xi_2 = \ldots = \xi_m = 1$ si $\det(\mathbf{A}\mathbf{B}^T) \geq 0$, et quand $\xi_1 = \xi_2 = \ldots = \xi_{m-1} = 1$ et $\xi_m = -1$ si $\det(\mathbf{A}\mathbf{B}^T) < 0$. Il reste à déterminer la rotation $\mathbf{R}$ qui permet d'atteindre ce minimum. Quand le rang de $\mathbf{A}\mathbf{B}^T$ est égal à m, alors $\mathbf{L}'$ est non singulière, et donc son inverse $\mathbf{L}'^{-1}$ est tel que $\mathbf{L}'^{-1} = (\mathbf{V}\mathbf{\Lambda}\mathbf{\Xi}\mathbf{V}^T)^{-1} = \mathbf{V}\mathbf{\Xi}^{-1}\mathbf{\Lambda}^{-1}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{\Xi}\mathbf{V}^T$ (car $\mathbf{\Xi}^{-1} = \mathbf{\Xi}$, et $\mathbf{\Xi}\mathbf{\Lambda}^{-1} = \mathbf{\Lambda}^{-1}\mathbf{\Xi}$). Donc, d'après C.10, il vient :

$$\mathbf{R} = \mathbf{A}\mathbf{B}^T\mathbf{L}'^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{\Xi}\mathbf{V}^T = \mathbf{U}\mathbf{\Xi}\mathbf{V}^T. \tag{C.19}$$

Finalement, quand $\text{rg}(\mathbf{A}\mathbf{B}^T) = m - 1$, d'après C.10 et C.13 :

$$\mathbf{R}\mathbf{V}\mathbf{\Lambda}\mathbf{\Xi}\mathbf{V}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T. \tag{C.20}$$

En multipliant chaque côté de C.20 par $\mathbf{V}$ à droite, et en utilisant la relation $\mathbf{\Lambda\Xi} = \mathbf{\Lambda}$ (puisque $\lambda_m = 0$ et $\xi_1 = \xi_2 = \ldots \xi_{m-1} = 1$), il vient :

$$\mathbf{RV\Lambda} = \mathbf{U\Lambda}. \tag{C.21}$$

Si l'on définit la matrice orthogonale $\mathbf{Q}$ telle que $\mathbf{Q} = \mathbf{U}^T\mathbf{RV}$, il vient :

$$\mathbf{Q\Lambda} = \mathbf{\Lambda} \tag{C.22}$$

Soient $\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}$ les vecteurs colonnes de $\mathbf{Q}$ ($\mathbf{Q} = [\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}]$). Les relations suivantes sont obtenues par comparaison des deux côtés de la relation C.22 :

$$\lambda_i \mathbf{q}_i = \lambda_i \mathbf{e}_i \quad 1 \leq i \leq m-1 \tag{C.23}$$

d'où

$$\mathbf{q}_i = \mathbf{e}_i \quad 1 \leq i \leq m-1 \tag{C.24}$$

où $\mathbf{e}_i$ est le vecteur colonne unitaire dont le i ème élément est égal à 1. Le dernier vecteur colonne $\mathbf{q}_m$ de $\mathbf{Q}$ est orthogonal à tous les autres vecteurs $\mathbf{q}_i$ ($1 \leq i \leq m-1$), car $\mathbf{Q}$ est une matrice orthogonale. Donc il vient :

$$\mathbf{q}_m = \mathbf{e}_m \quad \text{ou} \quad \mathbf{q}_m = -\mathbf{e}_m. \tag{C.25}$$

Par ailleurs :

$$\begin{aligned} \det(\mathbf{Q}) &= \det(\mathbf{U}^T)\det(\mathbf{R})\det(\mathbf{V}) \\ &= \det(\mathbf{U})\det(\mathbf{V}). \end{aligned} \tag{C.26}$$

Donc $\det(\mathbf{Q}) = 1$ si $\det(\mathbf{U})\det(\mathbf{V}) = 1$, et $\det(\mathbf{Q}) = -1$ si $\det(\mathbf{U})\det(\mathbf{V}) = -1$. Finalement, on obtient :

$$\begin{aligned} \mathbf{R} &= \mathbf{UQV}^T \\ &= \mathbf{U\Xi V}^T \end{aligned} \tag{C.27}$$

où

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = 1 \\ \text{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = -1 \end{cases} \tag{C.28}$$

C.2 Théorème

D'après le lemme précédent découle ce thérorème. Soient $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ et $\mathbf{Y} = \{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}\}$ deux ensembles de points en correspondance dans un espace à m dimensions. Soit e^2 l'erreur quadratique moyenne entre ces deux ensembles de points en fonction des paramètres de la similitude (rotation $\mathbf{R}$, translation $\mathbf{t}$, homothétie de facteur s) :

$$e^2(\mathbf{R}, \mathbf{t}, s) = \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{y_i} - (s\mathbf{R}\mathbf{x_i} + t) \|^2 \tag{C.29}$$

La valeur minimale ε^2 de cette erreur est donnée par la relation :

$$\varepsilon^2 = \sigma_y^2 - \frac{\mathrm{tr}(\mathbf{\Lambda\Xi})^2}{\sigma_x^2} \tag{C.30}$$

où :

$$\mu_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i} \tag{C.31}$$

$$\mu_{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y_i} \tag{C.32}$$

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{x_i} - \mu_{\mathbf{x}} \|^2 \tag{C.33}$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{y_i} - \mu_{\mathbf{y}} \|^2 \tag{C.34}$$

$$\mathbf{\Sigma_{xy}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y_i} - \mu_{\mathbf{y}})(\mathbf{x_i} - \mu_{\mathbf{x}})^T \tag{C.35}$$

$$\tag{C.36}$$

$\mathbf{U\Lambda V}^T$ est la décomposition en valeurs singulières de $\mathbf{\Sigma_{xy}}$ ($\mathbf{\Lambda} = \mathrm{diag}(\lambda_i), \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$), et

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{\Sigma_{xy}}) \geq 0 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{\Sigma_{xy}}) < 0 \end{cases} \tag{C.37}$$

$\mathbf{\Sigma_{xy}}$ est la matrice de covariance de $\mathbf{X}$ et $\mathbf{Y}$, $\mu_{\mathbf{x}}$ et $\mu_{\mathbf{y}}$ sont les moyennes respectives de $\mathbf{X}$ et $\mathbf{Y}$, et σ_x^2 et σ_y^2 sont les variances respectives autour de ces moyennes. Quand $\mathrm{rg}(\mathbf{\Sigma_{xy}}) \geq m - 1$, les paramètres optimaux de la transformation sont déterminés par les relations suivantes :

$$\mathbf{R} = \mathbf{U\Xi V}^T \tag{C.38}$$

$$\mathbf{t} = \mu_y - s\mathbf{R}\mu_x \tag{C.39}$$

$$s = \frac{1}{\sigma_x^2} \mathrm{tr}(\mathbf{\Lambda\Xi}) \tag{C.40}$$

$$\tag{C.41}$$

où $\mathbf{\Xi}$ dans l'équation C.38 doit être choisie telle que

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = 1 \\ \text{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = -1 \end{cases} \tag{C.42}$$

quand $\text{rg}(\mathbf{\Sigma_{xy}}) = m - 1$.

Démonstration

On représente les ensembles de points $\mathbf{X}$ et $\mathbf{Y}$ sous forme de matrices $m \times n$: respectivement $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]$, $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}]$. L'erreur e^2 est reformulée comme suit :

$$e^2(\mathbf{R}, \mathbf{t}, s) = \frac{1}{n} \parallel \mathbf{Y} - s\mathbf{RX} - \mathbf{t}\mathbf{h}^T \parallel^2 \tag{C.43}$$

où

$$h = (1, 1, \ldots, 1)^T \tag{C.44}$$

On introduit une matrice de normalisation $\mathbf{K}$ de dimensions $n \times n$: $\mathbf{K} = \mathbf{I} - (1/n)\mathbf{h}\mathbf{h}^T$ ($\mathbf{K}^2 = \mathbf{K}^T = \mathbf{K}$). Les équations C.33, C.34 et C.35 deviennent :

$$\sigma_x^2 = \frac{1}{n} \parallel \mathbf{XK} \parallel^2 \tag{C.45}$$

$$\sigma_y^2 = \frac{1}{n} \parallel \mathbf{YK} \parallel^2 \tag{C.46}$$

$$\mathbf{\Sigma_{xy}} = \frac{1}{n}\mathbf{YKX}^T \tag{C.47}$$

On utilise de plus les équations :

$$\mathbf{X} = \mathbf{XK} + \frac{1}{n}\mathbf{X}\mathbf{h}\mathbf{h}^T \tag{C.48}$$

$$\mathbf{Y} = \mathbf{YK} + \frac{1}{n}\mathbf{Y}\mathbf{h}\mathbf{h}^T \tag{C.49}$$

L'erreur e^2 devient alors :

$$e^2(\mathbf{R}, \mathbf{t}, s) = \frac{1}{n} \parallel \mathbf{YK} + \frac{1}{n}\mathbf{Y}\mathbf{h}\mathbf{h}^T - s\mathbf{RXK} - \frac{s}{n}\mathbf{RX}\mathbf{h}\mathbf{h}^T - \mathbf{t}\mathbf{h}^T \parallel^2 \tag{C.50}$$

$$= \frac{1}{n} \parallel \mathbf{YK} - s\mathbf{RXK} + \left(\frac{1}{n}\mathbf{Y}\mathbf{h} - \frac{s}{n}\mathbf{RX}\mathbf{h} - \mathbf{t}\right)\mathbf{h}^T \parallel^2 \tag{C.51}$$

$$= \frac{1}{n} \parallel \mathbf{YK} - s\mathbf{RXK} - \mathbf{t}'\mathbf{h}^T \parallel^2 \tag{C.52}$$

$$= \frac{1}{n}\{\parallel \mathbf{YK} - s\mathbf{RXK} \parallel^2 + \parallel \mathbf{t}'\mathbf{h}^T \parallel^2 - 2\text{tr}(\mathbf{K}(\mathbf{Y}^T - s\mathbf{X}^T\mathbf{R}^T)\mathbf{t}'\mathbf{h}^T)\} \tag{C.53}$$

où

$$\mathbf{t}' = -\frac{1}{n}\mathbf{Y}\mathbf{h} + \frac{s}{n}\mathbf{RX}\mathbf{h} + \mathbf{t} \tag{C.54}$$

En utilisant les relations suivantes :

$$\begin{aligned} \mathrm{tr}(\mathbf{K}(\mathbf{Y}^T - s\mathbf{X}^T\mathbf{R}^T)\mathbf{t}'\mathbf{h}^T) &= \mathrm{tr}(\mathbf{h}^T(\mathbf{I} - \mathbf{h}\mathbf{h}^T)(\mathbf{Y}^T - s\mathbf{X}^T\mathbf{R}^T)\mathbf{t}') \\ &= \mathrm{tr}((\mathbf{h}^T - \mathbf{h}^T)(\mathbf{Y}^T - s\mathbf{X}^T\mathbf{R}^T)\mathbf{t}') \\ &= 0 \qquad \text{(C.55)} \\ \| \mathbf{t}'\mathbf{h}^T \|^2 &= n \| \mathbf{t}' \|^2 \qquad \text{(C.56)} \end{aligned}$$

Il vient alors :

$$e^2(\mathbf{R}, \mathbf{t}, s) = \frac{1}{n} \| \mathbf{YK} - s\mathbf{RXK} \|^2 + \| \mathbf{t}' \|^2 \qquad \text{(C.57)}$$

D'après cette relation, $\mathbf{t}'$ doit être égal à 0 pour minimiser e^2, soit :

$$\mathbf{t} = \frac{1}{n}\mathbf{Yh} - \frac{s}{n}\mathbf{RXh} = \mu_y - s\mathbf{R}\mu_x \qquad \text{(C.58)}$$

De plus, si $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ est la décomposition en valeurs singulières (SVD) de $\mathbf{\Sigma}_{xy} = \frac{1}{n}\mathbf{YKX}^T$, alors $sn\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ est la SVD de $\mathbf{YK}(s\mathbf{XK}^T) = s\mathbf{YKK}^T\mathbf{X}^T = s\mathbf{YX}^T$. Donc la valeur minimale $\varepsilon^2(s)$ de $\frac{1}{n} \| \mathbf{YK} - s\mathbf{RXK} \|^2$ est donnée en fonction de $\mathbf{R}$ par le lemme présenté en C.1 :

$$\begin{aligned} \varepsilon^2(s) &= \frac{1}{n}\{\| \mathbf{YK} \|^2 + \| s\mathbf{XK} \|^2 - 2\,\mathrm{tr}(sn\mathbf{\Lambda}\mathbf{\Xi})\} \qquad \text{(C.59)} \\ &= \sigma_y^2 + s^2\sigma_x^2 - 2\,s\,\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi}) \qquad \text{(C.60)} \end{aligned}$$

où

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{\Sigma_{xy}}) \geq 0 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{\Sigma_{xy}}) < 0 \end{cases} \qquad \text{(C.61)}$$

Du même lemme découle

$$\mathbf{R} = \mathbf{U}\mathbf{\Xi}\mathbf{V}^T \qquad \text{(C.62)}$$

si $\mathrm{rg}(\Sigma_{xy}) \geq m - 1$, et où $\mathbf{\Xi}$ est tel que :

$$\mathbf{\Xi} = \begin{cases} \mathbf{I} & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = 1 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{si} \quad \det(\mathbf{U})\det(\mathbf{V}) = -1 \end{cases} \qquad \text{(C.63)}$$

si $\mathrm{rg}(\Sigma_{xy}) = m - 1$. Finalement, $\varepsilon^2(s)$ étant une frome quadratique de s, sa valeur minimale est obtenue pour :

$$s = \frac{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi})}{\sigma_x^2} \qquad \text{(C.64)}$$

et la valeur minimale atteinte est :

$$\begin{aligned} \varepsilon^2 &= \sigma_y^2 + \left\{\frac{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi})}{\sigma_x^2}\right\}^2 \sigma_x^2 - 2\left\{\frac{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi})}{\sigma_x^2}\right\}\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi}) \\ &= \sigma_y^2 + \frac{\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Xi})^2}{\sigma_x^2} \qquad \text{(C.65)} \end{aligned}$$

ce qui conclut la démonstration du théorème.

Annexe D

k-dimensional tree

D.1 Description

Un *kd-tree* (ou *k-dimensional-tree*) est une structure de partionnement spatial de données, destiné à l'organisation de points appartenant à un espace de dimension k. Cette structure se révèle utile pour la recherche des plus proches voisins Un *kd-tree* est un arbre binaire dont chaque noeud est un point de l'espace de dimension k. Chaque noeud qui n'est pas une feuille correspond à un hyperplan divisant l'espace en deux sous-espaces. Les deux ensembles de points séparés par cet hyperplan constituent deux sous-arbres reliés par ce noeud. Les hyperplans sont parallèles aux axes du référentiel de l'espace considéré. On représente Figure D.1 un exemple de partitionnement relatif à un *kd-tree* pour un espace de dimension 3.

D.2 Construction

On présente ici la méthode canonique de construction d'un *kd-tree*. Les contraintes sont les suivantes :

- En descendant l'arbre de noeud en noeud, l'orientation spatiale des hyperplans correspondant à chaque noeud doit évoluer de manière cyclique. Par exemple la racine correspond à un hyperplan normal à x, le fils de la racine à un hyperplan normal à y, et son petit-fils à un hyperplan normal à z, puis à nouveau x, y, etc.
- A chaque étape, le point sélectionné comme noeud de l'arbre, et par lequel passe l'hyperplan séparateur, est le point médian de l'ensemble des points du sous-espace et qui restent à insérer dans la structure, relativement à la dimension normale à l'hyperplan.

Une telle construction mène à un *kd-tree* dit équilibré, dans lequel chaque feuille est approximativement à la même distance de la racine. On représente respectivement

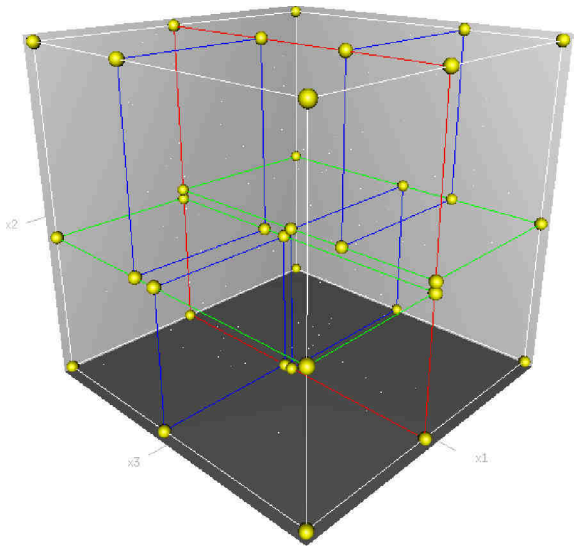


Figure D.1 – Exemple de partitionnement relatif à un *kd-tree* pour un espace de dimension 3. D'après [279].

Figure D.2 et D.3 le partitionnement d'un ensemble de points dans un espace à 2 dimensions, et le *kd-tree* correspondant.

D.3 Recherche des plus proches voisins dans un *kd-tree*

L'objectif d'un algorithme de recherche des plus proches voisins est de trouver dans l'arbre le point le plus proche d'un point quelconque de l'espace. Cette recherche peut tirer avantage des propriétés de l'arbre, afin d'éliminer des portions importantes de l'espace de recherche. Dans un *kd-tree*, la procédure est la suivante.

1. L'algorithme commence par le noeud racine, et descend de façon récursive de noeud en noeud. Chaque noeud choisi le long de cette trajectoire est celui, parmi les deux noeuds possibles, qui est le plus proche du point considéré, selon la dimension normale à l'hyperplan.
2. Quand l'algorithme atteint une feuille, ce noeud est enregistré comme meilleur candidat courant.
3. L'algorithme remonte l'arbre selon la trajectoire adoptée pour le descendre, et effectue ces étapes à chaque noeud :
   (a) Si le noeud courant est plus proche du point considéré que le meilleur candidat courant, alors il est lui-même enregistré comme le meilleur candidat courant.

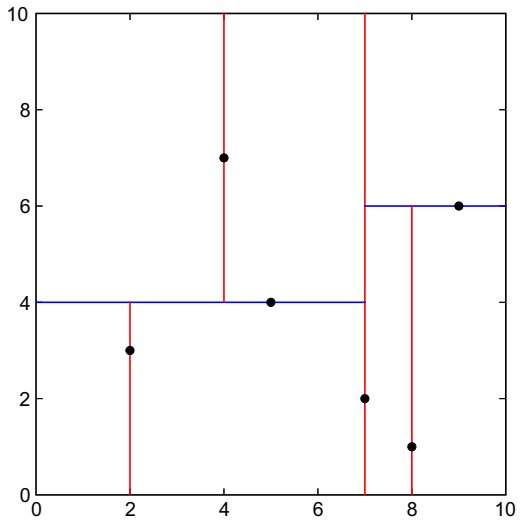


Figure D.2 – Exemple de partitionnement relatif à un k*d-tree* pour un espace de dimension 2. Les Points considérés sont de coordonnées (2,3), (5,4), (9,6), (4,7), (8,1), (7,2). D'après [279].

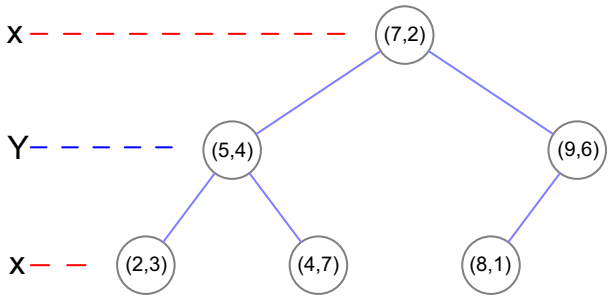


Figure D.3 – k*d-tree* associé à l'ensemble de points de la Figure D.2.

(b) L'algorithme calcule s'il peut exister des points de l'arbre plus proches de l'autre côté de l'hyperplan séparateur, en calculant l'intersection entre l'hyperplan séparateur et l'hypersphère passant par le meilleur candidat courant, et centrée sur le point considéré.

i. Si l'intersection n'est pas vide, alors il peut exister des points plus proches se situant de l'autre côté de l'hyperplan séparateur. L'algorithme passe alors à l'autre branche pour chercher un meilleur point candidat, selon la procédure récursive complète décrite jusqu'ici.

ii. Si l'intersection est vide, l'autre côté de l'hyperplan peut être éliminé de l'espace de recherche, et l'algorithme continue de remonter l'arbre.

4. Quand l'algorithme atteint le noeud racine, la recherche est terminée.

Dans le pire des cas, le temps t de recherche du plus proche voisin dans un kd-*tree* à M noeuds est donné par la relation suivante :

$$t = O(k.M^{1-\frac{1}{k}}).$$

Annexe E

Algorithme *RANdom SAmple Consensus* ou RANSAC

E.1 Objectif

L' algorithme appelé RANSAC [69], pour *RANdom SAmple Consensus* : il est fondé sur l'hypothèse qu'un jeu de données est constitué d'un ensemble de points distribués selon un certain modèle (les *inliers*), et de points marginaux, qui ne correspondent pas au modèle (les *outliers*). Les point marginaux peuvent notamment être issus d'un bruit dans l'estimation des valeurs physiques qui constituent leurs coordonnées, ou bien être le fruit d'erreurs. Etant donné un ensemble de *inliers*, même en quantité limitée et minoritaire, l'algorithme permet d'estimer les paramètres d'un modèle approprié.

E.2 Exemple

On illustre l'algorithme par l'exemple de la recherche d'une régression linéaire simple dans le jeu de données représenté Figure E.1. On suppose que l'on peut distinguer deux vatégories de points dans ces données : les *inliers*, qui suivent une relation linéaire et peuvent donc être approchés par une droite, et les *outliers*, qui ne suivent pas la même loi. Sur ce jeu de données, une solution simple est celle du problème aux moindres carrés, mais elle est généralement insatisfaisante, car la présence d'*outliers* fait passer la droite de régression loin des *inliers*. L'algorithme RANSAC est capable d'adapter le modèle seulement sur les *inliers*, à condition que la probabilité de sélectionner des sous-ensembles constitués uniquement que d'*inliers* soit suffisamment élevée. Cette condition n'est pas garantie, mais un réglage approprié des paramètres de l'algorithme permet de garder cette probabilité à un niveau élevé.

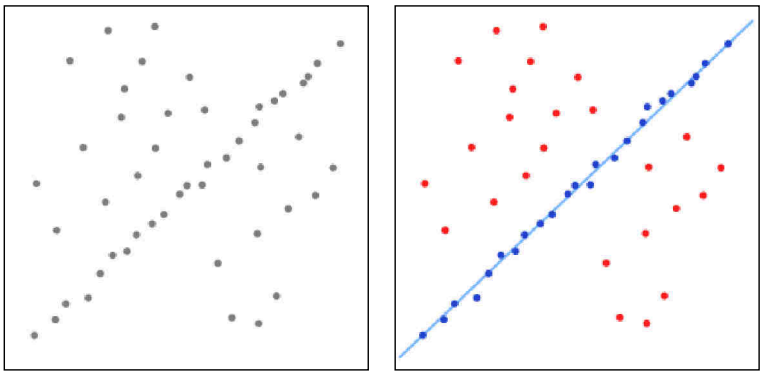


Figure E.1 – A gauche : données brutes comportant de nombreux points marginaux. A droite : résultat de l'algorithme (*inliers* en bleu, *outliers* en rouge. Les paramètres du modèle sont établis d'après les *inliers* uniquement (d'après [279]).)

E.3 Algorithme

RANSAC est un algorithme itératif non-déterministe au sens où il renvoie un résultat raisonnable avec une certaine probabilité, qui augmente avec le nombre d'itérations autorisées. En entrée de l'algorithme on considère le jeu de données observées, un modèle à adapter sur ces données ainsi que des paramètres de réglage. L'algorithme parvient à son objectif en sélectionnant itérativement un sous-ensemble aléatoire des données originales. Par hypothèse, les données sélectionnées sont des *inliers*, et cette hypothèse est testée comme suit :

1. Les paramètres du modèle sont adaptés aux *inliers* hypothétiques.
2. Les données non sélectionnées sont évaluées par rapport à ce modèle et, tous les points auxquels il convient sont ajoutés à l'ensemble des *inliers* hypothétiques.
3. Le modèle ainsi adapté est considéré comme satisfaisant si suffisamment de points viennent grossir l'ensemble des *inliers* hypothétiques.
4. Les paramètres du modèle sont réévalués d'après la totalité des *inliers* hypothétiques.
5. La validité du résultat est évaluée en estimant l'erreur prédictive du modèle par rapport aux *inliers* retenus.

Cette procédure est répétée plusieurs fois. A chaque fois, soit le modèle est rejeté car trop peu de points sont classés comme des *inliers*, soit il est retenu et caractérisé par la valeur de son erreur prédictive. On retient finalement le modèle qui offre l'erreur prédictive la plus faible.

Annexe F

Intercorrélation normalisée sur $\mathcal{L}_2(S^2)$

On considère deux fonctions f and g définies sur la sphère, et de carré intégrable : $f, g \in \mathcal{L}_2(S^2)$. L'intercorrélation normalisée $C_R(f,g)$ entre f et g, pour une rotation R, est telle que :

$$C_R(f,g) = \frac{\int_{S^2} \breve{f}(\chi)\overline{\Lambda_R(\breve{g})(\chi)\mathrm{d}\chi}}{\sqrt{\int_{S^2} |\breve{f}(\chi)|^2\mathrm{d}\chi \int_{S^2} |\breve{g}(\chi)|^2\mathrm{d}\chi}} \tag{F.1}$$

où $R \in SO(3)$,

$$\Lambda : \quad \mathcal{L}_2(S^2) \rightarrow \mathcal{L}_2(S^2) \tag{F.2}$$

$$\Lambda_R(g)(\chi) = g(R^{-1}(\chi)) \tag{F.3}$$

et $\breve{f}$ est le résultat du centrage de f autour de sa moyenne spatiale :

$$\breve{f}(\chi) = f(\chi) - \frac{1}{4\pi}\int_{S^2} f(\chi)\mathrm{d}\chi \tag{F.4}$$

On considère que ces fonctions sont de bande limitée, de largeur de bande B. On peut décomposer $\breve{f}$ et $\breve{g}$ en une combinaison linéaire finie d'harmoniques sphériques $Y_l^m(\chi)$ (le degré 0 correspondant à la moyenne n'est pas considéré).

$$\breve{f}(\chi) = \sum_{l=1}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\chi) \tag{F.5}$$

$$\breve{g}(\chi) = \sum_{l=1}^{B-1} \sum_{|m|\leq l} \hat{g}_l^m Y_l^m(\chi) \tag{F.6}$$

où les harmoniques sphériques Y_l^m sont définis pour un point de la sphère de coordonnées (ϑ, φ) [1] :

$$Y_l^m(\vartheta,\varphi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}\mathcal{P}_l^m(\cos\vartheta)e^{im\varphi} \tag{F.7}$$

où $\mathcal{P}_l^m(x)$ est le polynôme de Legendre, tel que :

$$\mathcal{P}_l^m(x) = \frac{(-1)^m}{2^l l!}(1-x^2)^{m/2}\frac{\mathrm{d}^{l+m}}{\mathrm{d}x^{l+m}}(x^2-1)^l \tag{F.8}$$

Par ailleurs, le résultat d'une rotation Λ_R d'un harmonique Y_l^m est une combinaison linéaire d'harmonique du même degré :

$$\Lambda_R(Y_l^m)(\chi) = \sum_{|k|\leq l} Y_l^k(\chi)D_{km}^{(l)}(R) \tag{F.9}$$

où $R \in SO(3)$, et $D_{km}^{(l)}$ est une fonction de Wigner-D.

$$D_{km}^{(l)}(\alpha,\beta,\gamma) = e^{-ik\alpha}d_{km}^{(l)}(\beta)e^{-ik\gamma}, \quad k,l,m \in \mathbb{N} \tag{F.10}$$

$$d_{km}^{(l)}(\beta) = \sqrt{\frac{(l+m)!(l-m)!}{(l+k)!(l-k)!}}\left(\sin\frac{\beta}{2}\right)^{m-k}\left(\cos\frac{\beta}{2}\right)^{m+k} \tag{F.11}$$

$$\times P_{l-m}^{(m-k,m+k)}(\cos(\beta)) \tag{F.12}$$

où $P_\eta^{(\mu,\nu)}(x)$ est un polynôme de Jacobi :

$$P_\eta^{(\mu,\nu)}(x) = \frac{\Gamma(\mu+\nu+1)}{\eta!\Gamma(\mu+\nu+\eta+1)}\sum_{p=0}^{\eta}\left(\begin{array}{c}\eta\\ p\end{array}\right)\frac{\Gamma(\mu+\nu+\eta+p+1)}{\Gamma(\mu+p+1)}\left(\frac{x-1}{2}\right)^p \tag{F.13}$$

où Γ est la fonction Gamma usuelle, et pour η entier :

$$\left(\begin{array}{c}x\\ \eta\end{array}\right) = \frac{\Gamma(x+1)}{\Gamma(\eta+1)\,\Gamma(x-\eta+1)} \tag{F.14}$$

En injectant F.9 dans F.1, il vient, pour le numérateur de l'intercorrélation :

$$\int_{S^2}\breve{f}(\chi)\overline{\Lambda_R(\breve{g})(\chi)}\mathrm{d}\chi$$

$$= \sum_l\sum_{|m|\leq l}\sum_{|m'|\leq l}\hat{f}_l^m\overline{\hat{g}_l^{m'}}\int_{S^2}Y_l^m(\chi)\overline{\sum_{|k|\leq l}Y_l^k(\chi)D_{km'}^{(l)}(R)}\mathrm{d}\chi \tag{F.15}$$

$$= \sum_l\sum_{|m|\leq l}\sum_{|m'|\leq l}\sum_{|k|\leq l}\hat{f}_l^m\overline{\hat{g}_l^{m'}}D_{km'}^{(l)}(R)\int_{S^2}Y_l^m(\chi)\overline{Y_l^k(\chi)}\mathrm{d}\chi \tag{F.16}$$

1. L'angle ϑ est la colatitude, et l'angle φ est la longitude. Ils s'expriment en fonction des coordonnées dans le système polaire-vertical selon les relations : $\vartheta = \pi/2 - \phi_{pv}$ et $\varphi = \theta_{pv}$ (cf. Figure 1).

La base des harmoniques sphériques étant orthogonale, l'intégrale dans F.16 est nulle, sauf si $k = m$. De plus, en utilisant les symétries des fonctions de Wigner-D, il vient :

$$\begin{aligned}
\int_{S^2} \breve{f}(\chi)\overline{\Lambda_R(\breve{g})(\chi)}\mathrm{d}\chi &= \sum_l \sum_{|m|\leq l} \sum_{|m'|\leq l} \hat{f}_l^m \overline{\hat{g}_l^{m'}}\,\overline{D^{(l)}_{mm'}(R)} && \text{(F.17)}\\
&= \sum_l \sum_{|m|\leq l} \sum_{|m'|\leq l} \hat{f}_l^m \overline{\hat{g}_l^{m'}}(-1)^{m'-m} D^{(l)}_{-m-m'}(R) && \text{(F.18)}\\
&= \sum_l \sum_{|m|\leq l} \sum_{|m'|\leq l} \hat{f}_l^{-m} \overline{\hat{g}_l^{-m'}}(-1)^{m-m'} D^{(l)}_{mm'}(R) && \text{(F.19)}
\end{aligned}$$

Introduisons la transformée de Fourier sur le groupe des rotations $SO(3)$. Soit h une fonction de carré intégrable définie sur $SO(3)$: $h \in \mathcal{L}_2(SO(3))$. La fonction h peut être décomposée sur la base des fonctions de Wigner-D :

$$h(\alpha,\beta,\gamma) = \sum_l \sum_{|m|\leq l} \sum_{|m'|\leq l} \hat{h}^l_{mm'} \overline{D^{(l)}_{mm'}(\alpha,\beta,\gamma)} \qquad \text{(F.20)}$$

Le numérateur de la fonction d'intercorrélation peut donc être calculé comme la transformée de Fourier inverse sur $SO(3)$ d'une combinaison des coefficients de la décomposition en harmoniques sphériques des fonctions f et g considérées. En pratique, on utilise $SOFT$, un package public de fonctions dédiées aux calculs de transformées de Fourier discrètes sur $SO(3)$ [124].

Le dénominateur de l'intercorrélation dans F.1 se calcule simplement :

$$\int_{S^2} |\breve{f}(\chi)|^2 \mathrm{d}\chi = \sum_{l=1}^{B-1} \sum_{|m|\leq l} (\hat{f}_l^m)^2 \qquad \text{(F.21)}$$

Finalement, on obtient rapidement l'évaluation de l'intercorrélation sur un échantillonnage discret de l'espace des rotations $R \in SO(3)$. La grille d'échantillonnage s'exprime en fonction des angles d'Euler α, β, et γ, et les pas d'échantillonnage α_{j_1}, β_k et γ_{j_2} sont inversement proportionnels à la bande B des fonctions f et g : $\alpha_{j_1} = 2\pi j_1/2B$, $\beta_k = \pi(2k+1)/4B$, $\gamma_{j_2} = 2\pi j_2/2B$, où $0 \leq j_1, j_2, k < 2B$. Notons que la rotation d'angle nul, ou transformation identité, ne fait pas partie de la grille d'échantillonnage. L'intercorrélation relative à une rotation nulle ne peut donc être calculée que de façon détournée : il s'agit au préalable de faire subir à une des deux fonctions une rotation d'angle $-\pi/4B$ autour de l'axe y. Ainsi l'échantillonnage angulaire de $SO(3)$ est-il décalé de façon à comporter le point : $\alpha = \beta = \gamma = 0$.

Annexe G

k-means

L'algorithme appelé *k-means* est un algorithme de classification, dont le but est de regrouper en k *clusters* un ensemble de n objets ($k < n$), dont les attributs forment un espace vectoriel. L'objectif de l'algorithme est de minimiser la variance totale intra-*cluster*, soit la fonction :

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \tag{G.1}$$

où l'on cherche k clusters $\{S_i\}_{i=1,\dots,k}$ et μ_i est le centroïde, ou point moyen, de l'ensemble des points $x_j \in S_i$. Dans la forme la plus courante de l'algorithme *k-means*, on utilise une recherche heuristique appelée algorithme de Lloyd, illustré Figure G.1. Cet algorithme commence par partitionner l'ensemble des points en k *clusters* initiaux, en général de façon aléatoire. Dans chaque *cluster*, le point moyen, ou centroïde, est calculé. Le partionnement suivant est construit en associant chaque point de l'ensemble au centroïde le plus proche. Les centroïdes sont à nouveau calculés pour ces nouveaux *clusters*, et l'algorithme se répète ainsi jusqu'à la convergence, obtenue quand les points ne changent plus de *cluster* (ou quand les centroïdes ne changent plus). D'autres mises en oeuvres existent pour résoudre le problème du *k-means*, mais l'algorithme de Lloyd reste populaire car il converge très rapidement en pratique, typiquement en un nombre d'itérations très inférieur au nombre de points. Cependant l'algorithme ne tombe pas systématiquement sur l'optimum global, car la solution dépend du choix initial des clusters. Il convient donc de lancer l'algorithme plusieurs fois.

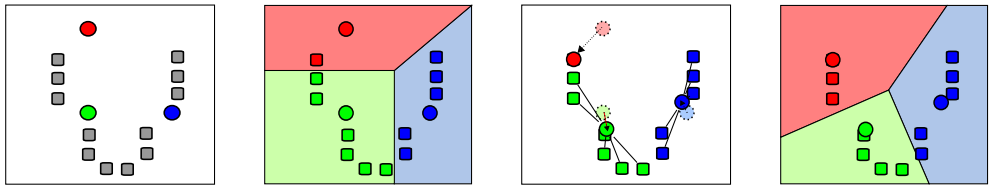


Figure G.1 – Illustration des itérations successives de l'algorithme k-*means*. On cherche ici a classer les données (carrés) en 3 *clusters*. Les centroïdes sont représentés par des cercles.

Annexe H

SFRS prototypiques

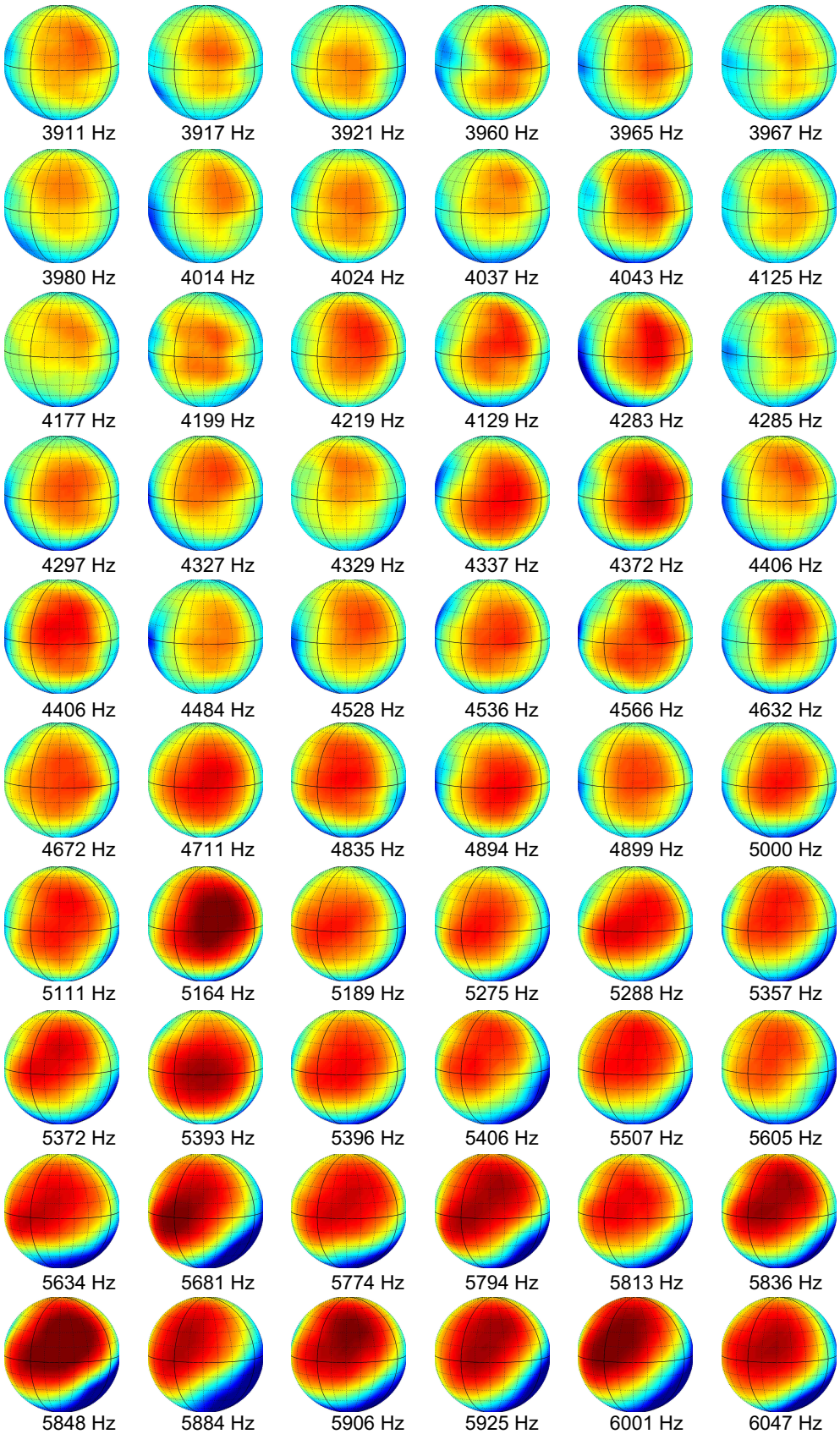


Figure H.1 – SFRS prototypiques obtenues dans l'expérience décrite en 6.3.4. Les SFRS prototypiques sont classées de gauche à droite et de haut en bas par fréquence centrale croissante du *cluster* sont elles sont issues, dont la valeur est indiquée sous chaque SFRS.

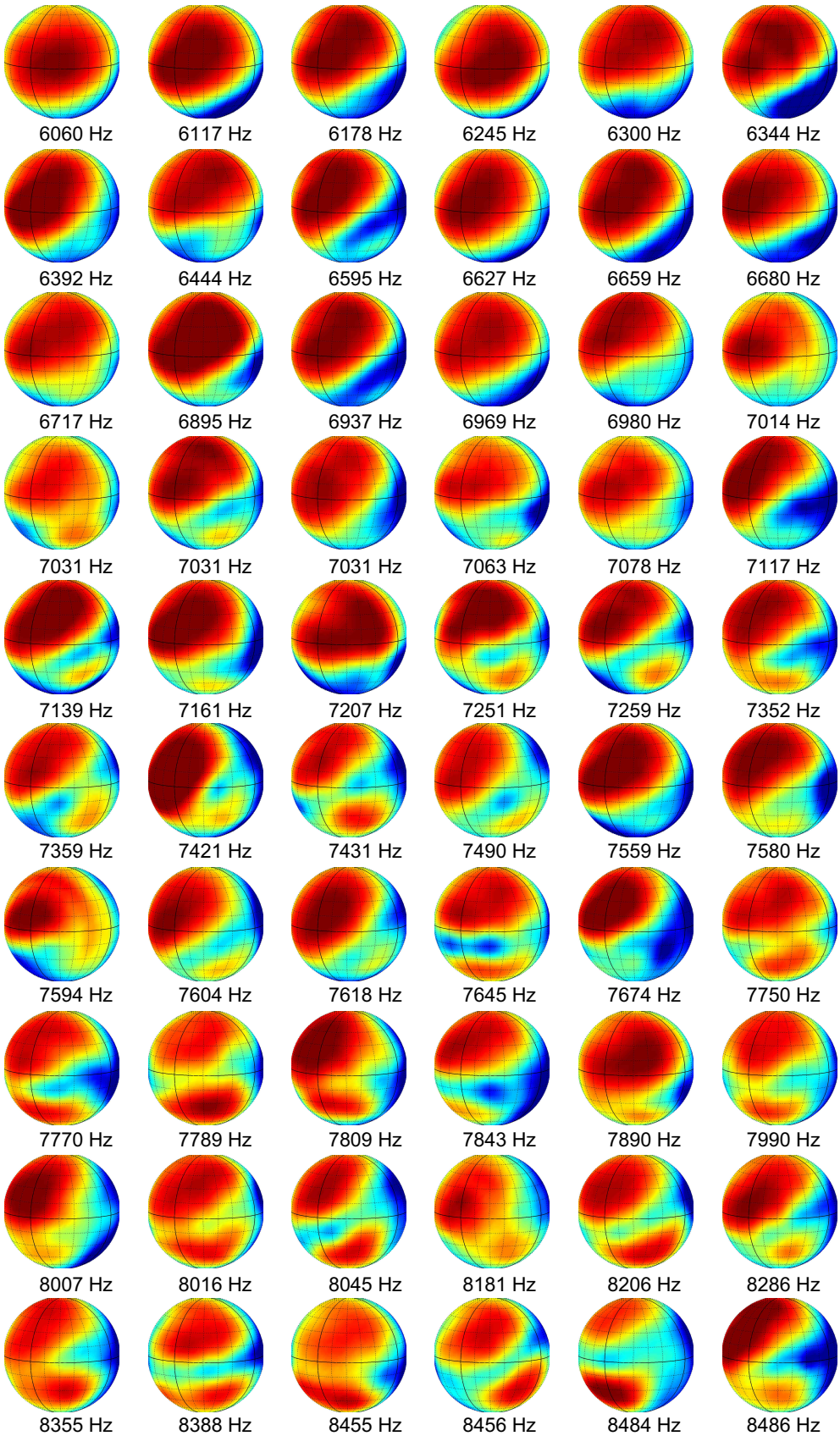


Figure H.2 – Suite de la figure H.1.

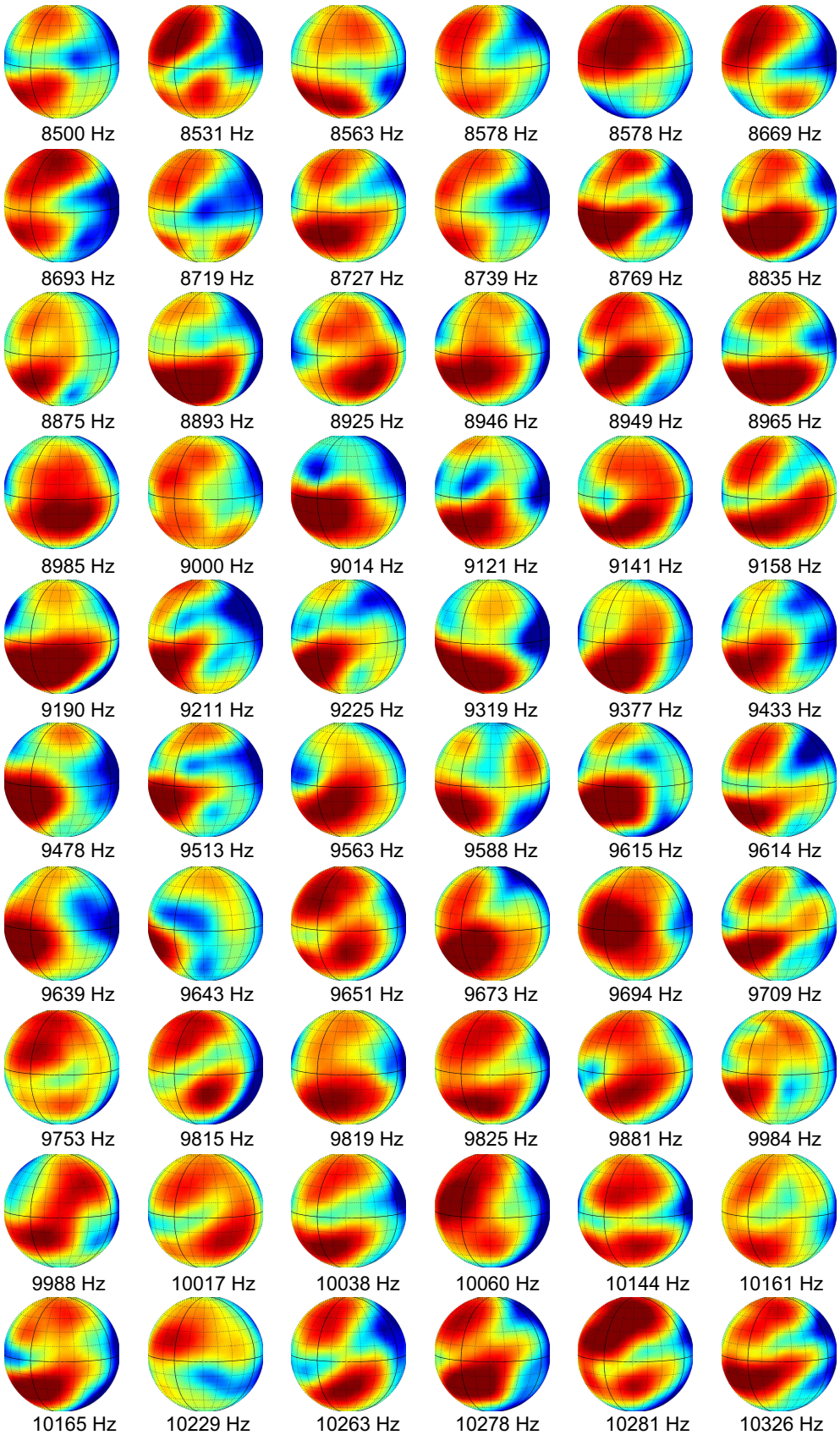


Figure H.3 – Suite de la figure H.2.

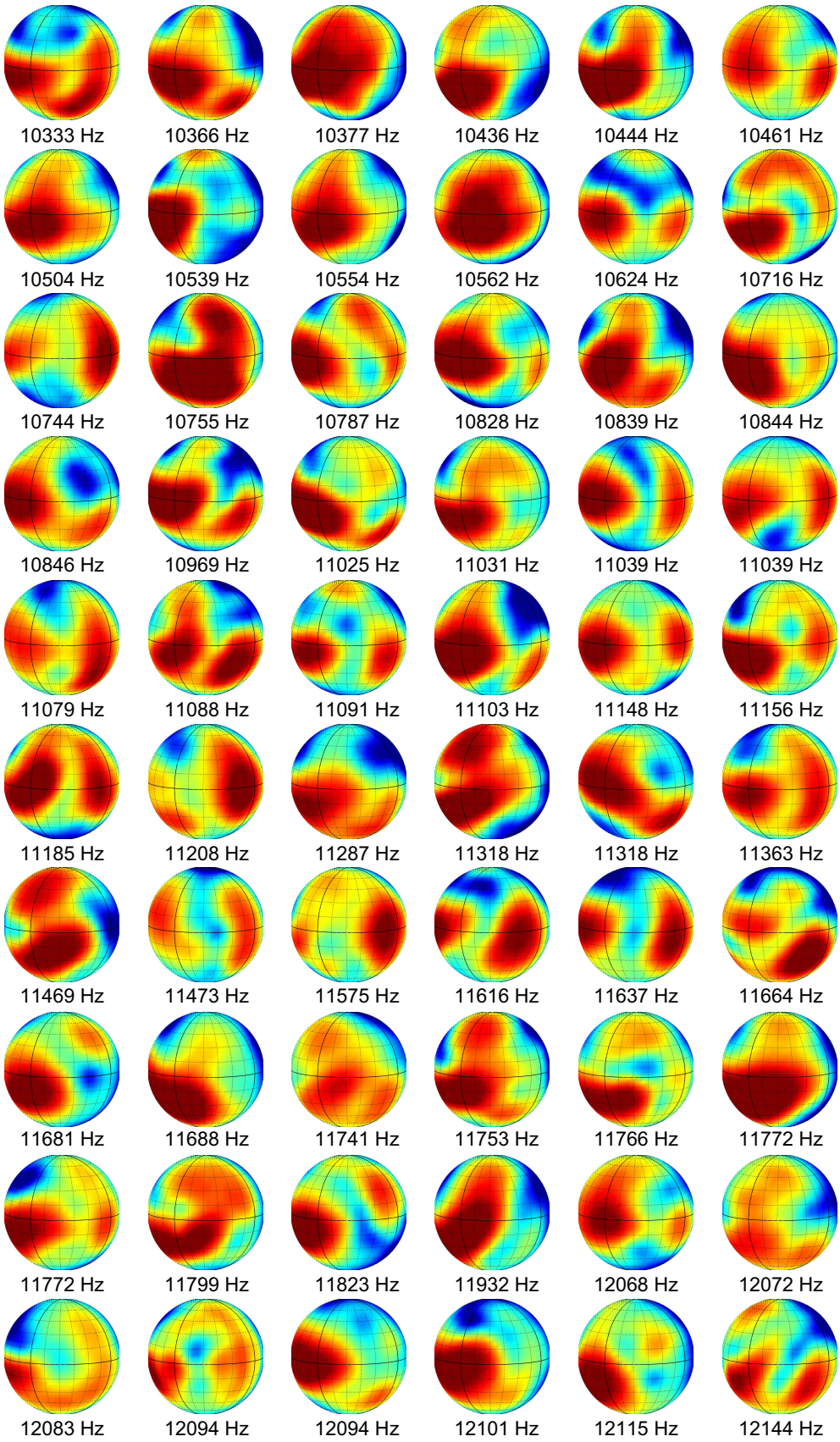


Figure H.4 – Suite de la figure H.3.

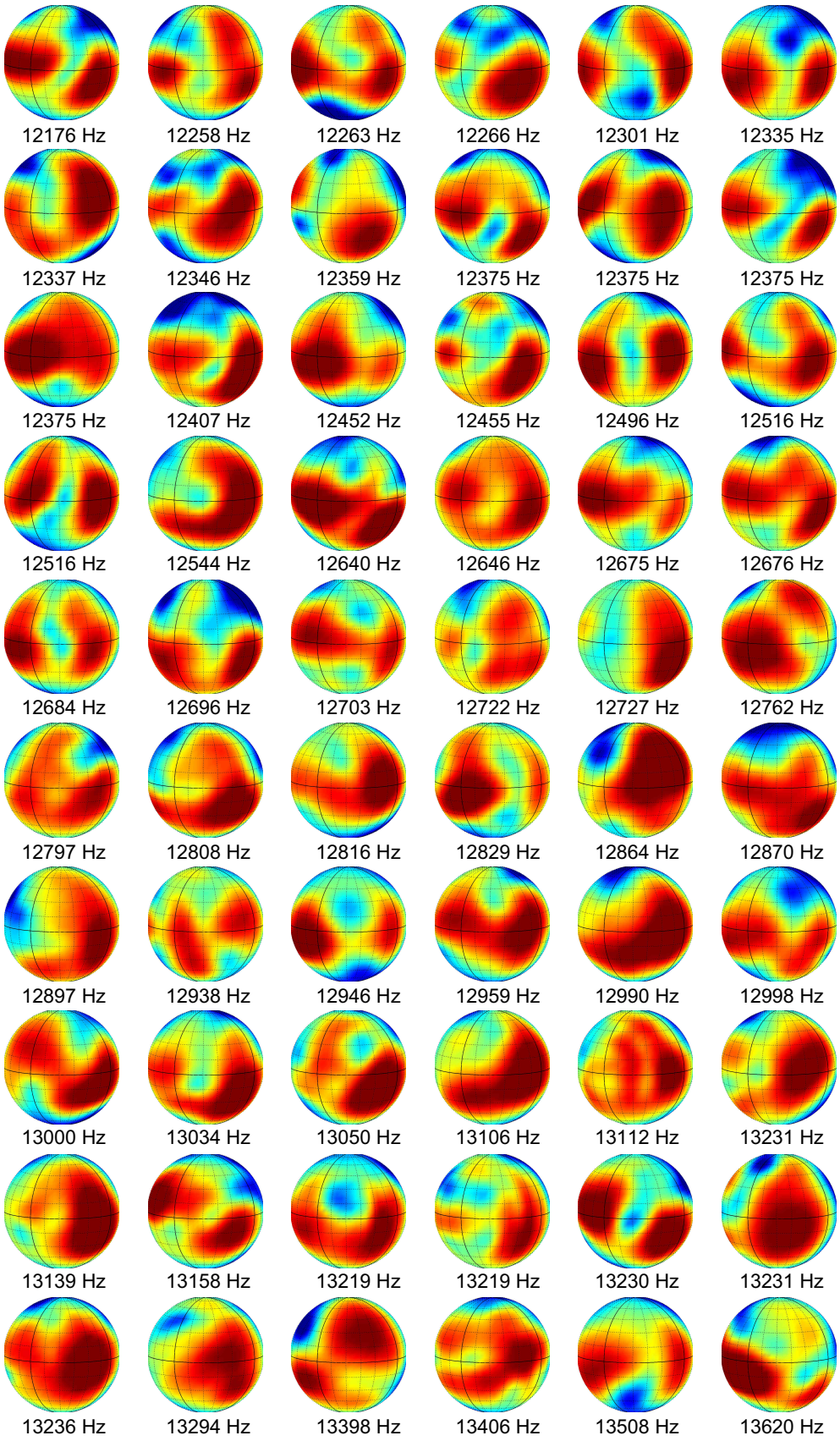


Figure H.5 – Suite de la figure H.4.

Annexe I

Analyse en Composantes Principales des HRTF

L'Analyse en Composantes Principales, ou ACP, est une procédure statistique qui fournit une représentation efficace d'un jeu de données corrélées. L'objectif de l'ACP est de réduire la dimensionnalité d'un jeu de données présentant une certaine redondance, tout en conservant la part la plus significative de la variabilité observée. C'est à partir des données elles-mêmes qu'une base de fonctions est créée. On peut avantageusement représenter les données mesurées sur cette base, car elle est organisée de façon hiérarchique : les premiers vecteurs de cette base permettent de reconstruire l'essentiel de la variabilité des données, tandis que les derniers vecteurs peuvent être plus aisément laissés de côté, sans impact sensible sur la reconstruction des données. C'est Kistler et Wightman qui ont les premiers envisagé d'analyser par ACP un ensemble de HRTF constitué de mesures effectuées sur plusieurs sujets [121]. Considérons un ensemble de M vecteurs colonnes $\mathbf{H}_i$, représentant le spectre d'amplitude en dB de M HRTF différentes, connues sur N bins fréquentiels. L'ensemble de ces vecteurs constitue la matrice $\mathbf{H}$ de dimensions $N \times M$:

$$\mathbf{H} = [\mathbf{H}_1 \dots \mathbf{H}_M] \tag{I.1}$$

On calcule le vecteur moyen $\mathbf{u}$ des HRTF mesurées, que l'on retranche de chaque colonne de $\mathbf{H}$ pour former la matrice $\mathbf{B}$:

$$\mathbf{u} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{X}_n \tag{I.2}$$

$$\mathbf{B} = [\mathbf{H}_1 - \mathbf{u} \dots \mathbf{H}_M - \mathbf{u}] \tag{I.3}$$

La matrice de covariance $\mathbf{C}$ est calculée :

$$\mathbf{C} = \frac{1}{N} \mathbf{B}\mathbf{B}^T \tag{I.4}$$

On obtient par diagonalisation de cette matrice, la décomposition suivante :

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D} \tag{I.5}$$

où $\mathbf{D}$, de dimensions $M \times M$ est diagonale constituée des valeurs propres $\mathbf{D} = diag(\lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, et $\mathbf{V}$ de dimension $M \times M$ est la matrice des vecteurs propres de $\mathbf{C}$: $\mathbf{V} = [\mathbf{V}_1 \ldots \mathbf{V}_M]$. On caractérise par l'énergie cumulative $g(l)$, la part de la variance totale exprimée par la base constituée par les l premiers vecteurs propres :

$$g(l) = \sum_{q=1}^{l} \lambda_q \tag{I.6}$$

La réduction de la dimensionnalité consiste à exprimer les données sur la base constitutée des L premiers vecteurs propres ($L < M$), où la valeur de L est choisie telle que $g(L) \geq 90\%$. Ce choix constitue une compression des données, dans la mesure où les données d'origine peuvent être reconstruites comme une somme pondérée des mêmes vecteurs propres :

$$\mathbf{W} = [\mathbf{V}_1 \ldots \mathbf{V}_L] \tag{I.7}$$

$$\mathbf{P} = \mathbf{W}^T\mathbf{B} \tag{I.8}$$

$$\hat{\mathbf{H}}_i = \mathbf{W}\mathbf{P}_i + \mathbf{u}, \quad i = 1 \ldots M \tag{I.9}$$

Les vecteurs $\hat{\mathbf{H}}_i$ sont des reconstructions des vecteurs $\mathbf{H}_i$ d'autant plus fidèles que l'on conserve un nombre élevé de vecteurs propres. En pratique, Kistler et Wightman parviennent à une reconstruction transparente perceptivement en ne conservant que 5 vecteurs propres [121]. Chen *et al.* [53] ont proposé d'obtenir pour un sujet donné les HRTF dans une direction quelconque de l'espace par interpolation STPS des L composantes des vecteurs de pondération $\mathbf{P}_i$, où $i = 1 \ldots M$, déterminés d'après les mesures M disponibles. C'est la technique retenue Carlile *et al.* pour réaliser une reconstruction des HRTF [48].

Annexe J

Consigne de test communiquée aux sujets de l'évaluation perceptive

Test de localisation en synthèse binaurale dynamique

Cher participant,

Vous allez effectuer un test de localisation en synthèse binaurale dynamique, pour une série de scènes virtuelles. Dans chaque scène, une seule source sonore virtuelle est disposée dans une direction de l'espace.

Votre objectif est de localiser cette source en pointant le regard dans sa direction. Votre réponse est validée automatiquement lorsque vous pointez le regard suffisamment longtemps dans une zone limitée autour de la direction de la source. Un signal sonore vous indique la validation, puis la scène suivante vous est présentée.

Ceci est un test de rapidité : nous mesurons le temps qui vous est nécessaire pour localiser la source. Votre tâche consiste donc à tourner donc la tête rapidement vers la direction où vous percevez la source.

Si vous ne parvenez pas à trouver la source en moins de 30 secondes, la scène s'arrête et la scène suivante est présentée. Un signal sonore vous indique alors cet échec.

Vous allez passez une session d'apprentissage puis 10 sessions de test. Chaque session est composée de 4 à 5 séries de 70 scènes. Une pause, indiquée par un signal sonore, sera marquée entre chacune de ces séries. Vous reprendrez le test en appuyant sur la barre d'espace.

Remarques et recommandations

- Dans une scène sur deux, la source est positionnée directement devant vous (par rapport à la position de la calibration).
- Si vous pensez pointer le visage près de la direction la source, mais que la validation ne se produit pas, continuez à chercher par des mouvements lents de la tête.
- Après chaque validation/échec, gardez la tête immobile jusqu'à ce que la scène suivante commence.
- Si vous rencontrez des problèmes notez svp le numéro de la scène.

Merci de votre participation, et bon courage.

Notez pour chaque série les numéros des scènes pour lesquelles vous rencontrez des problèmes, ainsi que vos impressions générales sur chaque série.

Annexe K

Evaluation perceptive : comportement des sujets n°1 et 4 en conditions R19 à R121 et I

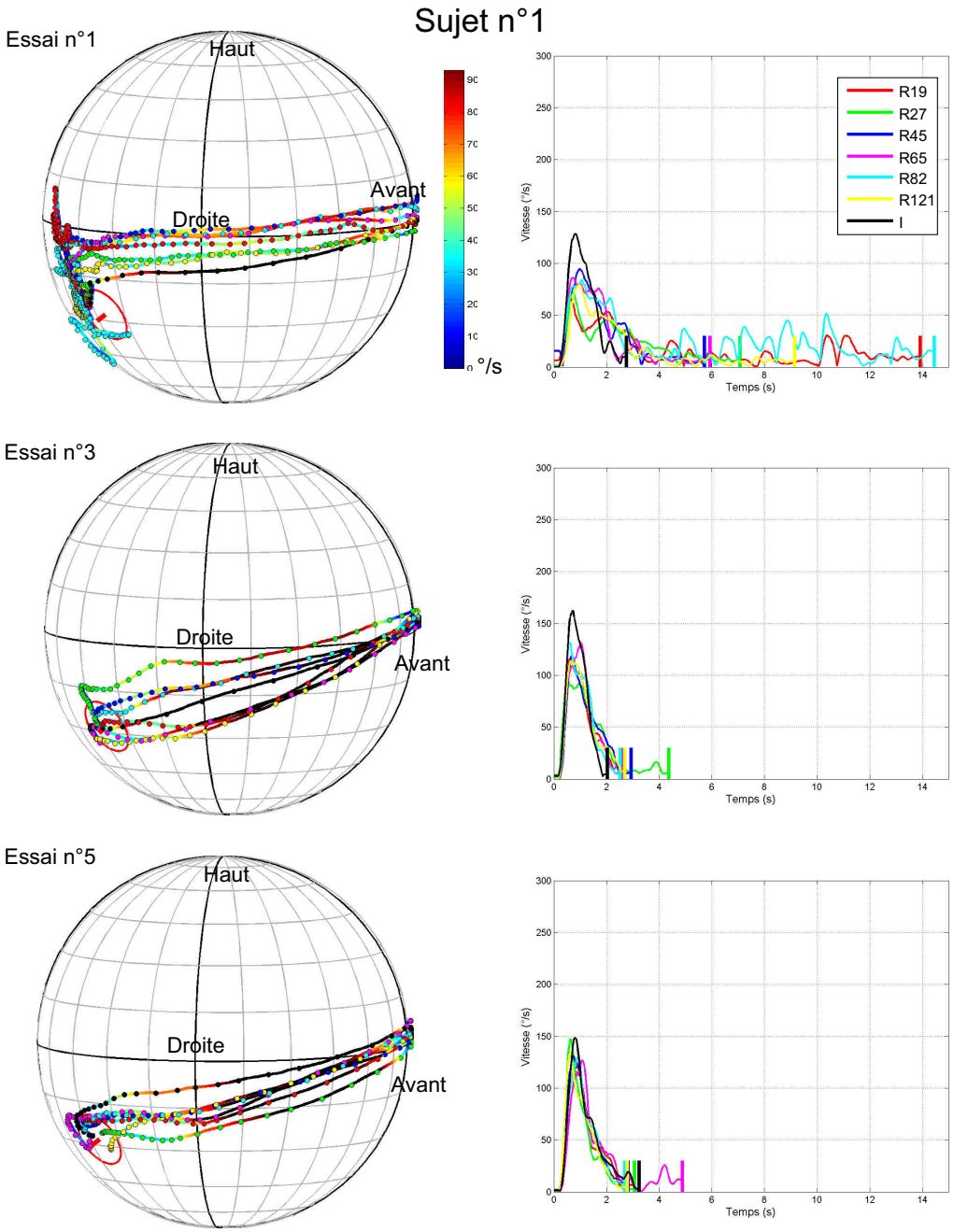


Figure K.1 – A gauche : trajectoires adoptées par l'axe médian du sujet n°1 pour la direction n°1 (azimut 229°, élévation -28.15°, système polaire-vertical), pour les essais 1, 3 et 5, et pour les conditions R19 à R121 et I. La vitesse angulaire est codée en couleur le long de ces trajectoires. La direction de la source est matérialisée par un trait rouge, et le cône de la validation qui l'entoure par un cercle rouge. A droite : vitesse angulaire correspondante. Les traits verticaux à la fin de chaque courbe marquent l'instant de la validation.

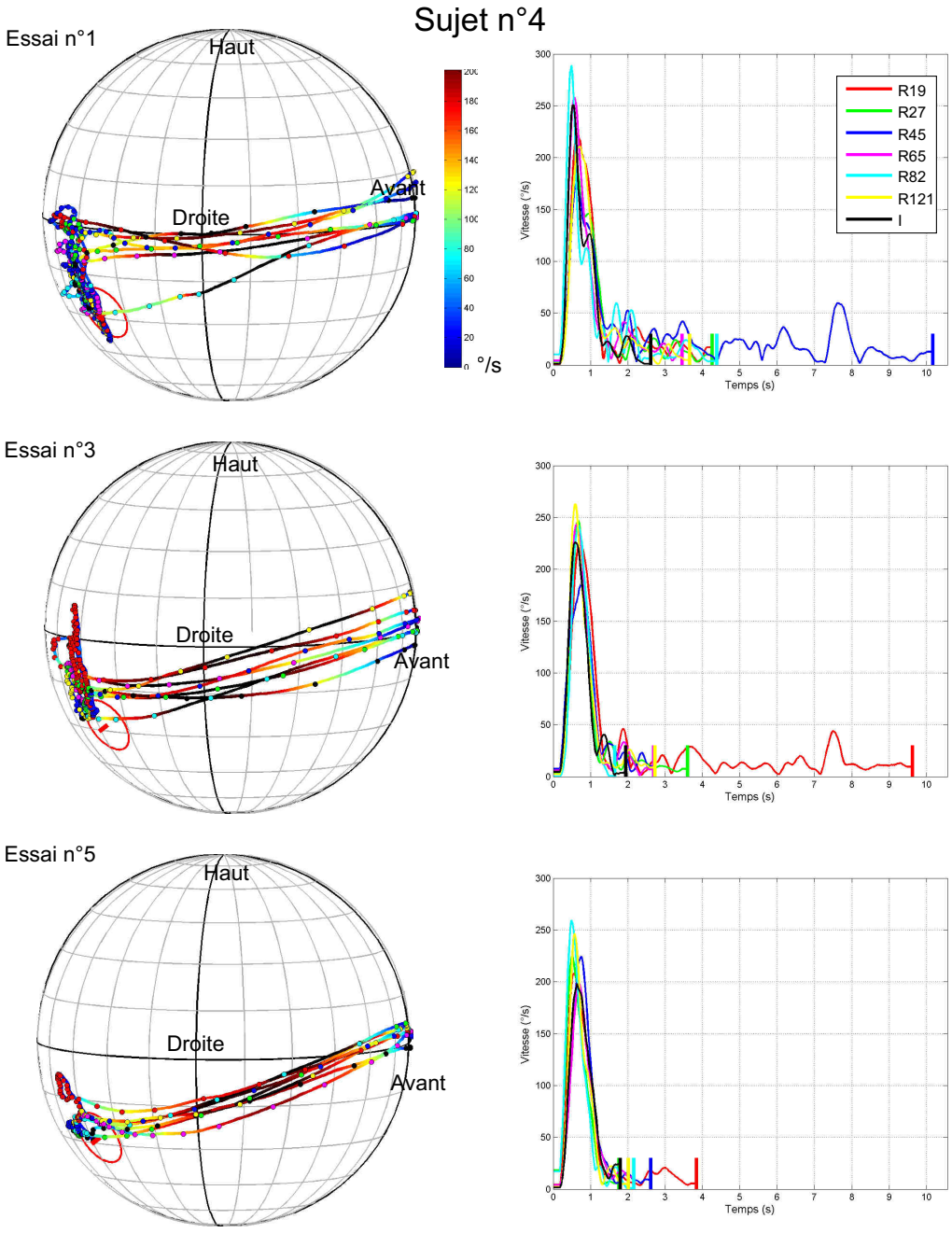


Figure K.2 – Comportement du sujet n°4 pour la direction n°1 (azimut 229°, élévation -28.15°). Voir figure K.1 pour les détails.

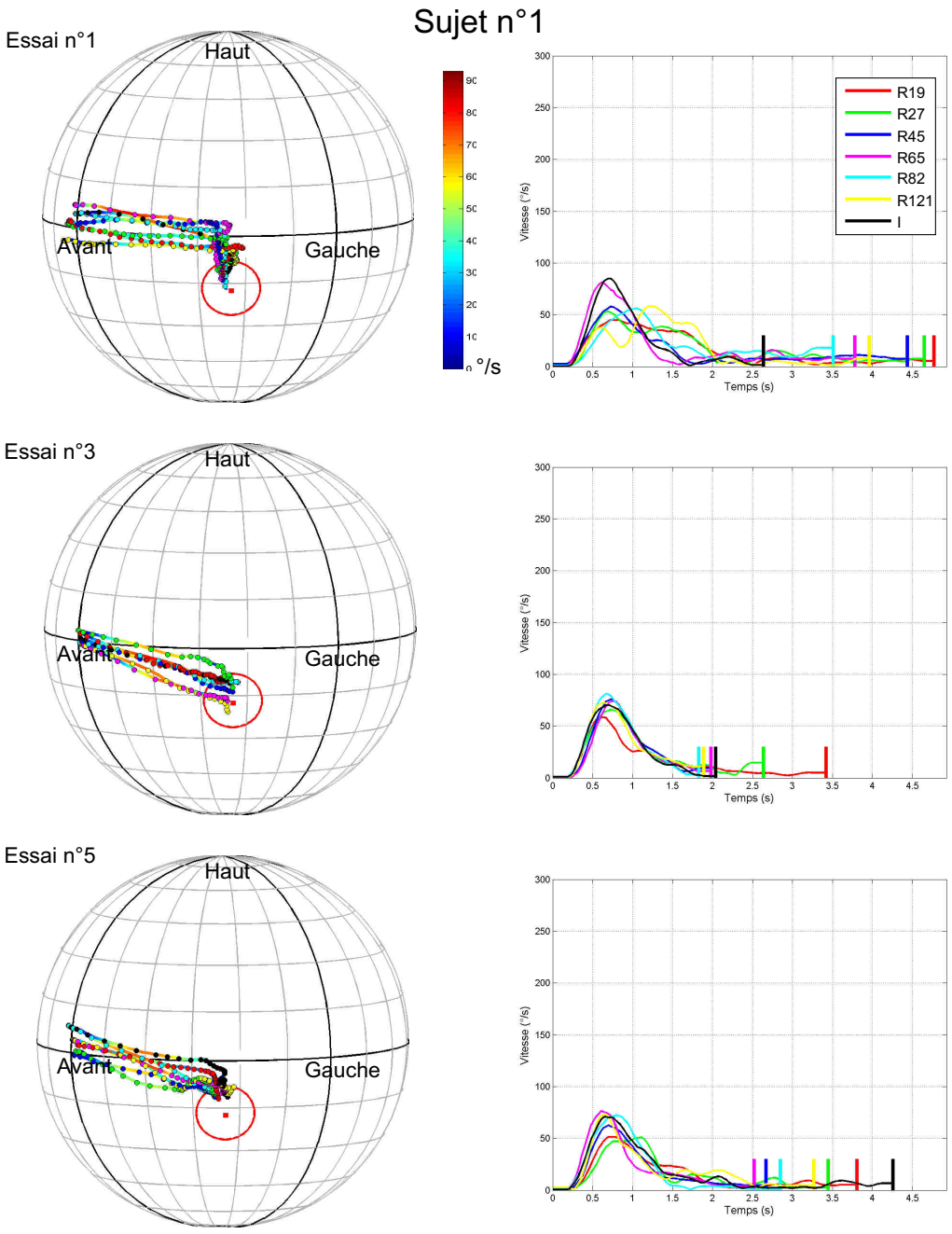


Figure K.3 – Comportement du sujet n°1 pour la direction n°7 (azimut 55.4°, élévation -16.9°). Voir figure K.1 pour les détails.

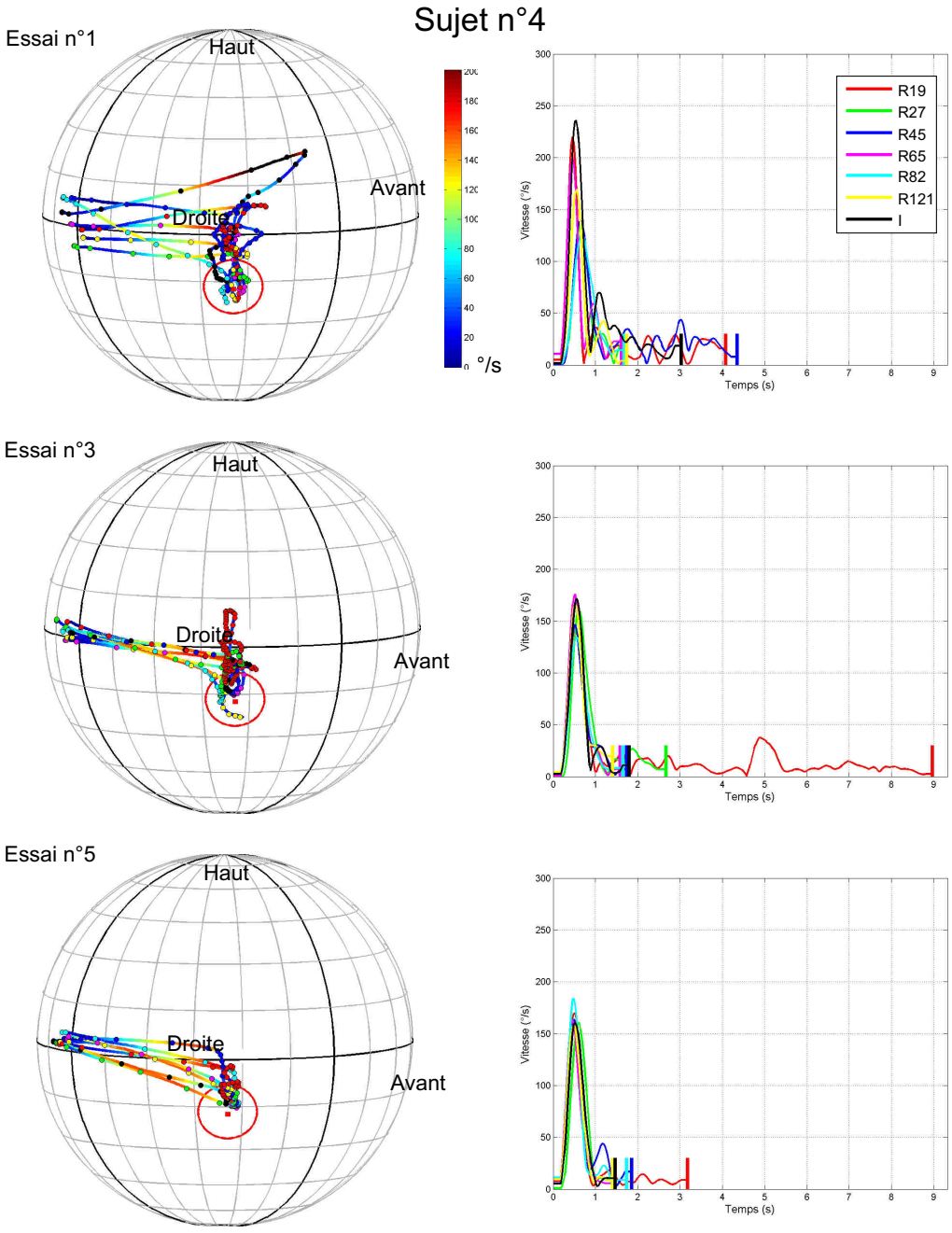


Figure K.4 – Comportement du sujet n°4 pour la direction n°7 (azimut 55.4°, élévation -16.9°). Voir figure K.1 pour les détails.

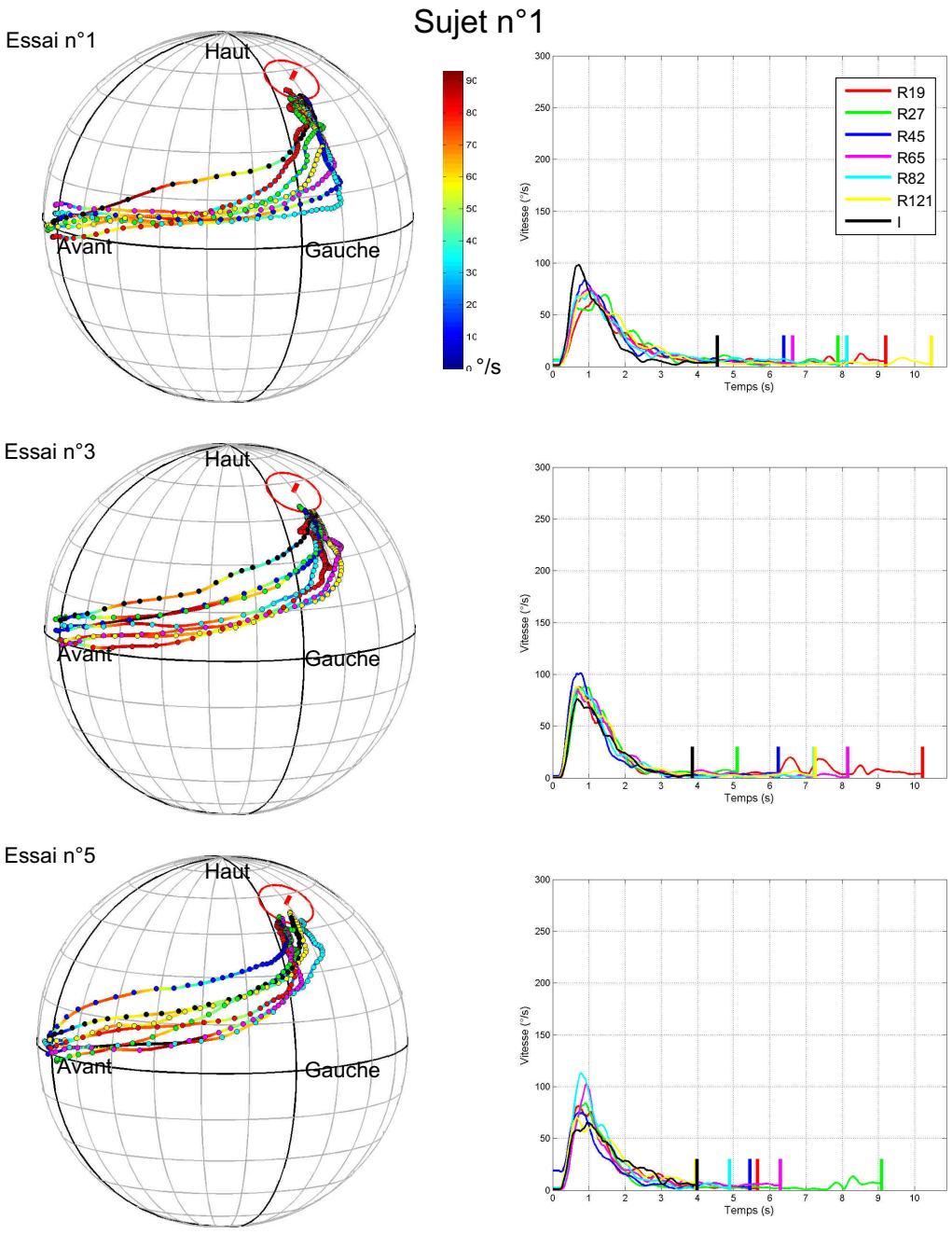


Figure K.5 – Comportement du sujet n°1 pour la direction n°35 (azimut 105°, élévation 56.25°). Voir figure K.1 pour les détails.

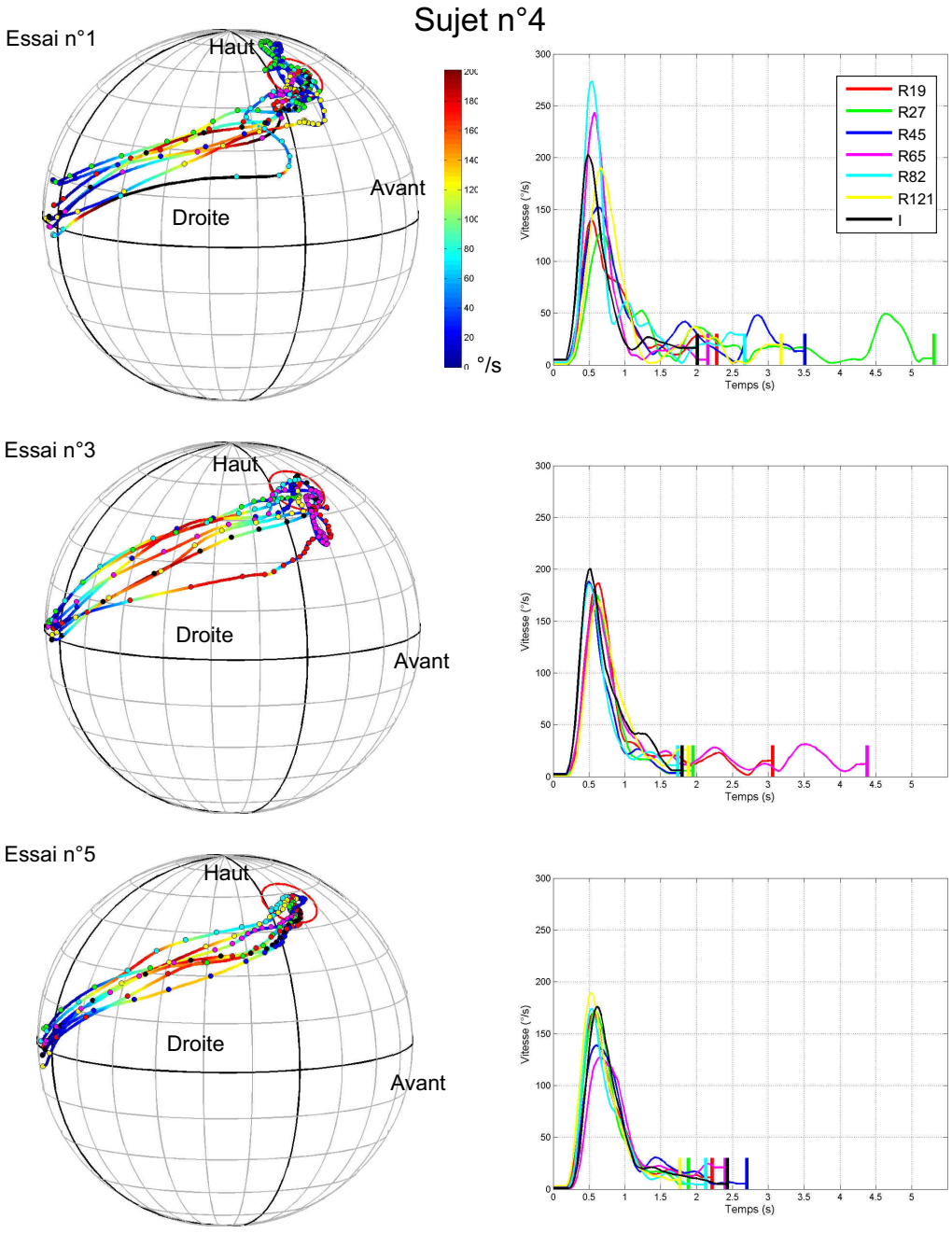


Figure K.6 – Comportement du sujet n°4 pour la direction n°35 (azimut 105°, élévation 56.25°). Voir figure K.1 pour les détails.

Annexe L

Distribution ex-gaussienne

L.1 Définition et description

La densité de probabilité ex-gaussienne f correspond à la distribution de la somme d'une variable aléatoire normale (ou gaussienne) suivant la distribution $g(x, \mu, \sigma)$ (μ est la moyenne et σ l'écart-type), et d'une variable aléatoire exponentielle de distribution $h(x, \tau)$ (τ est la moyenne). f est donc la convolution d'une fonction normale et d'une fonction exponentielle, et se caractérise par les 3 paramètres μ, σ et τ :

$$g(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{(L.1)}$$

$$h(x|\tau) = \frac{1}{\tau}\exp\left(-\frac{x}{\tau}\right) \quad \text{(L.2)}$$

$$f(x|\mu,\sigma,\tau) = \frac{1}{\tau}\exp\left(\frac{\mu}{\tau}+\frac{\sigma^2}{2\tau}-\frac{x}{\tau}\right)\Phi\left(\frac{x-\mu-\sigma^2/\tau}{\sigma}\right) \quad \text{(L.3)}$$

$$= \frac{1}{x\sqrt{2\pi}}\exp\left(-\frac{x-\mu}{\tau}+\frac{\sigma^2}{2\tau^2}\right)\int_{-\infty}^{-\frac{x-\mu}{\sigma}-\frac{\sigma}{\tau}}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y \quad \text{(L.4)}$$

où Φ est la fonction normale cumulative. Les moments centraux m_1 (moyenne), m_2 (variance) et m_3 sont définis à partir de ces paramètres selon les relations : $m_1 = \mu+\tau$, $m_2 = \sigma^2+\tau^2$ et $m_3 = 2\tau^3$. On représente Figure L.1 trois exemples de distributions ex-gaussiennes.

L.2 Ajustement par la méthode du maximum de vraisemblance (MLE)

Il n'existe pas de méthode algébrique pour estimer les paramètres d'une fonction ex-gaussienne. L'estimation des paramètres de la fonction qui correspondent le mieux à une distribution empirique doit donc être l'aide d'un algorithme d'ajustement. On

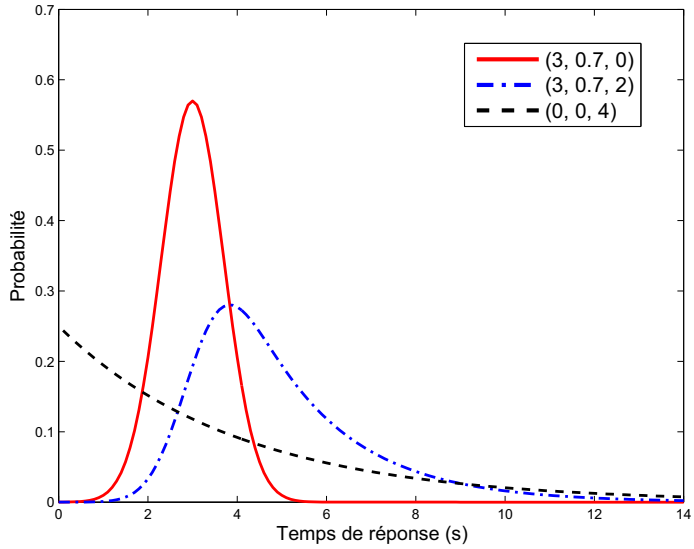


Figure L.1 – Exemples de densités de probabilité ex-gaussiennes. La légende précise les valeurs des paramètres (μ,σ,τ).

applique la méthode décrite par Lacouture *et al.* [129]. Il s'agit de considérer la fonction de vraisemblance : c'est une valeur heuristique qui représente l'ajustement ou l'adéquation entre des valeurs empiriques et certains paramètres que l'on désire estimer. La vraisemblance est d'autant plus grande que l'ajustement est bon : on cherchera donc le maximum de la vraisemblance (*Maximum Likelihood Estimation ou MLE*. Pour une densité de probabilité $f(x|\theta)$ de paramètres $\theta = [\theta_1, \theta_2, \ldots, \theta_k]$, et un échantillon de N observations $\{x_i\}_{i=1,\ldots,N}$, la vraisemblance L est définie par la relation :

$$L(\theta|X) = \prod_{i=1}^{N} f(x_i|\theta)$$

Il apparaît en pratique plus simple de s'intéresser à l'opposé du logarithme de la vraisemblance, noté $LogL$, et qu'il s'agit de minimiser pour maximiser la vraisemblance L.

$$LogL(\theta|X) = -\sum_{i=1}^{N} ln\{f(x_i|\theta)\}$$

Le minimum est atteint par la méthode Simplex, préconisée pour son efficacité et sa robustesse par Lacouture *et al.* [129]. Il faut garder à l'esprit le fait que cet ajustement n'est qu'un estimateur, dont il faut caractériser les imprécisions en termes de biais et d'écart-type. Lacouture *et al.* [129] ont mené une étude de Monte Carlo sur cet estimateur pour 2 jeux de paramètres, différentes tailles d'échantillons, et 1000 estimation. On représente les résultats Figure L.2 : le biais devient quasiment nul au delà de 100 observations, et l'écart-type décroît de façon monotone. Ces résultats

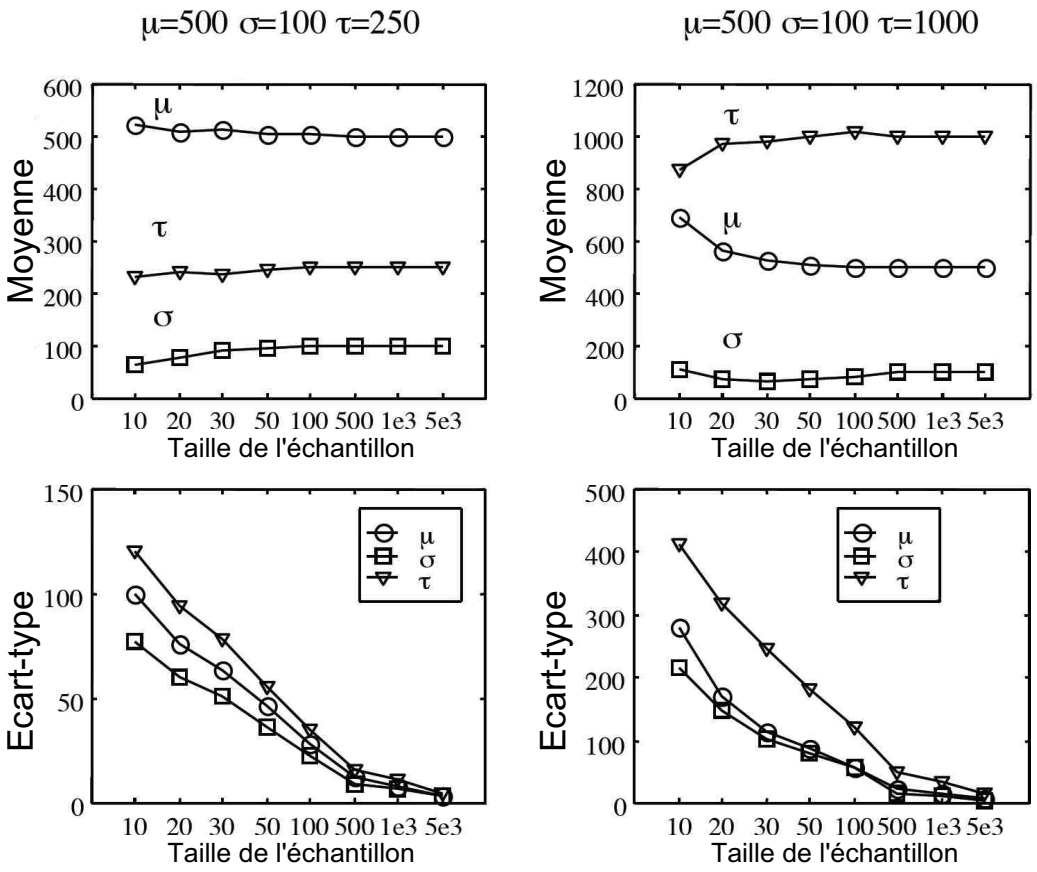


Figure L.2 – Etude de Monte Carlo : moyenne et écart-type pour 1000 réalisations de l'estimation des paramètres de la distribution ex-gaussienne par la méthode du maximum de vraisemblance, en fonction de la taille de l'échantillon. A gauche les paramètres μ,σ,τ de la distribution sont respectivement fixés à 500, 100, 250, et à droite à 500, 100, 1000 (d'après [129]).

démontrent les bonnes performances de l'estimateur choisi, mais révèlent aussi une forte dépendance avec la valeur des paramètres de la distribution.

Annexe M

Critères complémentaires pour l'analyse de l'évaluation perceptive

En complément de l'analyse de l'évaluation subjective développée au chapitre 5, plusieurs critères ont été envisagés. Leur description est proposée figure M.3.

M.1 Définition des critères

M.1.1 Temps de réaction

On définit le temps de réaction τ_{reac} comme la durée nécessaire au sujet pour engager un mouvement de tête dans la direction de la source, à partir du début du stimulus. On peut supposer que cette durée est en mesure de refléter la qualité du percept spatial initial. On l'évalue par seuillage de l'accélération angulaire. On représente figure M.1 la distribution des valeurs de τ_{reac} observées pour chacun des sujets, tous essais et conditions confondues (hormis les conditions NI1, NI2 et NI3). Les valeurs des temps de réaction sont extrêmement faibles (entre 180 ms et 400 ms), et elles ne sont donc connues qu'avec une précision relative médiocre, car la période d'échantillonnage de mesure de la trajectoire était de 30 ms. C'est pourquoi l'analyse des résultats du test perceptif ne peuvent se fonder sur ce critère.

M.1.2 Instant du maximum de la vitesse angulaire

Selon le même raisonnement que pour le temps de réaction, on peut considérer l'instant $\tau_{V_{max}}$, correspondant au maximum de la vitesse angulaire, le long de la trajectoire (cf. Fig. M.3). On représente figure M.2 la distribution de $\tau_{V_{max}}$ pour

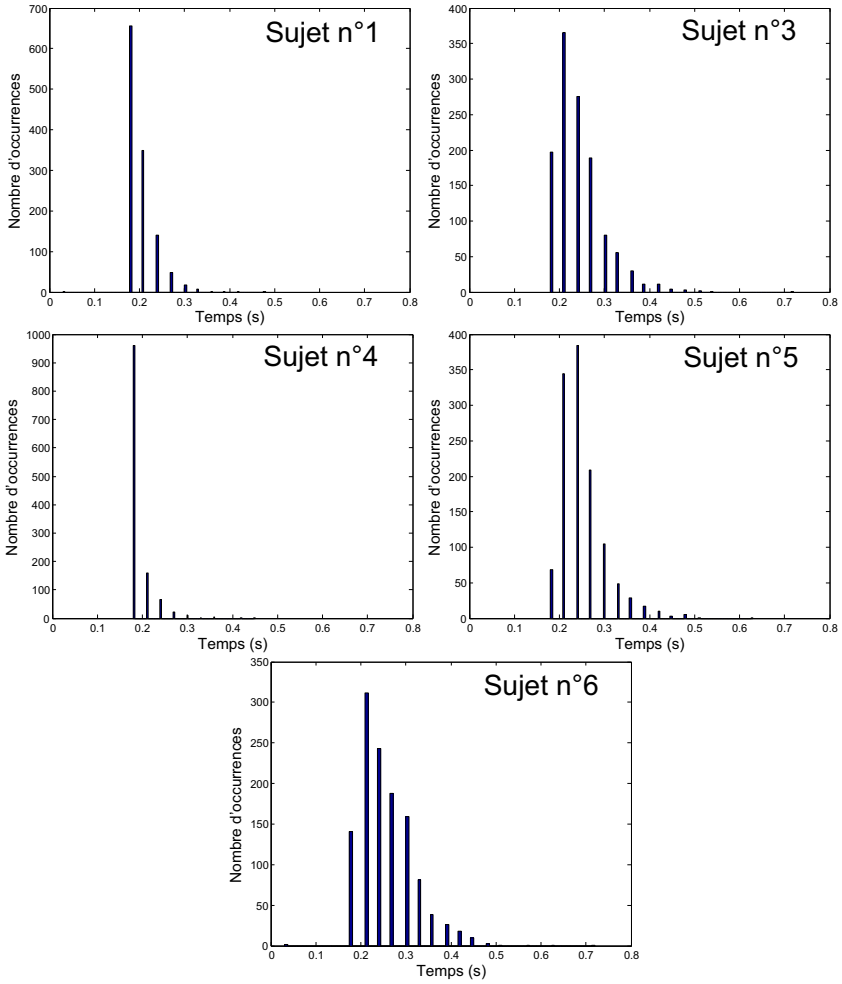


Figure M.1 – Distribution des valeurs du temps de réaction τ_{reac}, pour chaque sujet, tous essais et conditions de test confondus. La précision relative est faible, du fait de la période d'échantillonnage de 30 ms.

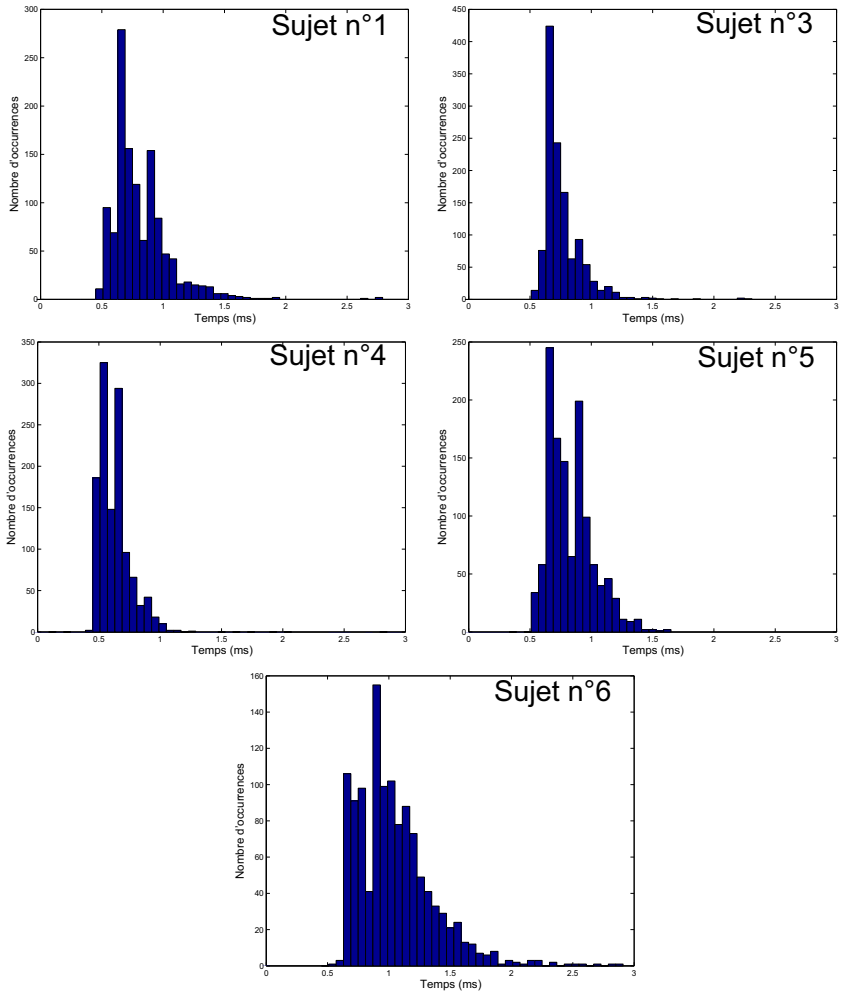


Figure M.2 – Distribution des valeurs de $\tau_{V_{max}}$, pour chaque sujet, tous essais et conditions de test confondus.

chacun des sujets. Les valeurs étant plus élevées que celles de τ_{reac}, la précision relative est acceptable, et on peut donc retenir ce critère d'analyse.

M.1.3 Temps de réponse normalisé relatif à l'azimut

On peut envisager la localisation auditive comme reposant sur deux processus, le deuxième dépendant du premier : une analyse des indices interauraux, pour déterminer la latéralisation de la source sonore, et une analyse des indices spectraux pour déterminer son élévation. Plus précisément, pour un stimulus de large bande fréquentielle, c'est essentiellement l'ITD qui participe à la formation du percept de latéralisation. De façon a évaluer séparément la qualité des indices temporels et des indices spectraux, on peut s'intéresser à la rapidité avec laquelle les sujets se tournent vers la source en termes d'azimut. On mesure donc le temps nécessaire aux sujets

pour rapprocher leur plan médian de la direction de la source. On simule *a posteriori*, les temps de réponse qui auraient été obtenus s'il avait suffi de pointer vers la source avec le plan médian, avec une précision de ±9°, et une durée nécessaire à la validation égale à 750 ms. On nomme τ_{az} le temps qui s'écoule entre la sortie du cône entourant la direction (0°,0°), et l'entrée dans la zone de la sphère comprise entre les deux plans verticaux encadrant la direction de la source à ±9° d'azimut [1]. De façon similaire à ce qui a été décrit pour $\tilde{\tau}_{rep}$ en 6.4.3, on définit le temps de réponse normalisé relatif à l'azimut $\tilde{\tau}_{az} = \tau_{az}/d_i$.

M.2 Analyse

Comme décrit au chapitre 5 pour $\tilde{\tau}_{rep}$, l'analyse des temps $\tau_{V_{max}}$ et $\tilde{\tau}_{az}$ est avantageusement réalisée en ajustant sur leurs distributions les paramètres d'une ex-gaussienne par MLE et *bootstrapping*. On représente figures M.5 M.4 les résultats de l'ajustement réalisé sur les données des essais n°3, 4 et 5 mêlés. L'observation des paramètres d'ajustement de $\tilde{\tau}_{az}$ révèlent une distribution de forme différente de celle de $\tilde{\tau}_{rep}$. En effet μ et τ sont de même ordre de grandeur et surtout σ est généralement non nul. Les paramètres d'ajustement de $\tau_{V_{max}}$ révèlent eux une prépondérance de la composante normale dans la distribution ex-gaussienne, qui est assez resserrée autour de sa moyenne. A part peut-être pour le sujet n°4, pour lequel les paramètres τ correspondant à $\tilde{\tau}_{az}$ et $\tau_{V_{max}}$ augmentent légèrement avec une diminution du nombre de HRTF mesurées, il n'apparaît pas de tendance lisible dans l'évolution des paramètres d'une condition à l'autre. Des tests d'hypothèse par permutation, similaires à ceux présentés au chapitre 5, permettent de révéler si les différences observées entre la condition I et les conditions R19 à R121 sont significatives. Les résultats sont rassemblés dans les tableaux M.1 et M.2.

Pour les sujets n°1, 3, 5, la distribution de τ_{Vmax} est équivalente entre la condition I et chacune des conditions R19 à R121. Pour le sujet n°6, le comportement se distingue seulement dans la condition R19. Le sujet n°4 montre un comportement particulier : des différences significatives sont observées pour les conditions R19, R27, R82 et R121. Cela reste inexpliqué pour les conditions R82 et R121, pour lesquelles l'erreur objective de reconstruction est moindre que pour toutes les autres conditions. En ce qui concerne $\tilde{\tau}_{az}$, aucune différence de comportement n'est observée, quelle que soit la condition, pour les sujets n°3 et 6. Pour les sujets n°1, 5 et 6, des différences de comportement sont observées là où les performances sont globalement satisfaisantes selon l'analyse de $\tilde{\tau}_{rep}$ (conditions R65 et R121 pour le sujet n°1, conditions R45,

1. L'azimut est à comprendre ici dans le système de coodonnées polaire horizontal, relatif à un repère (O, x', y', z') dont l'axe Ox' est confondu avec la direction de la source, et le plan $Ox'z'$ contient l'axe Oz du répère absolu.

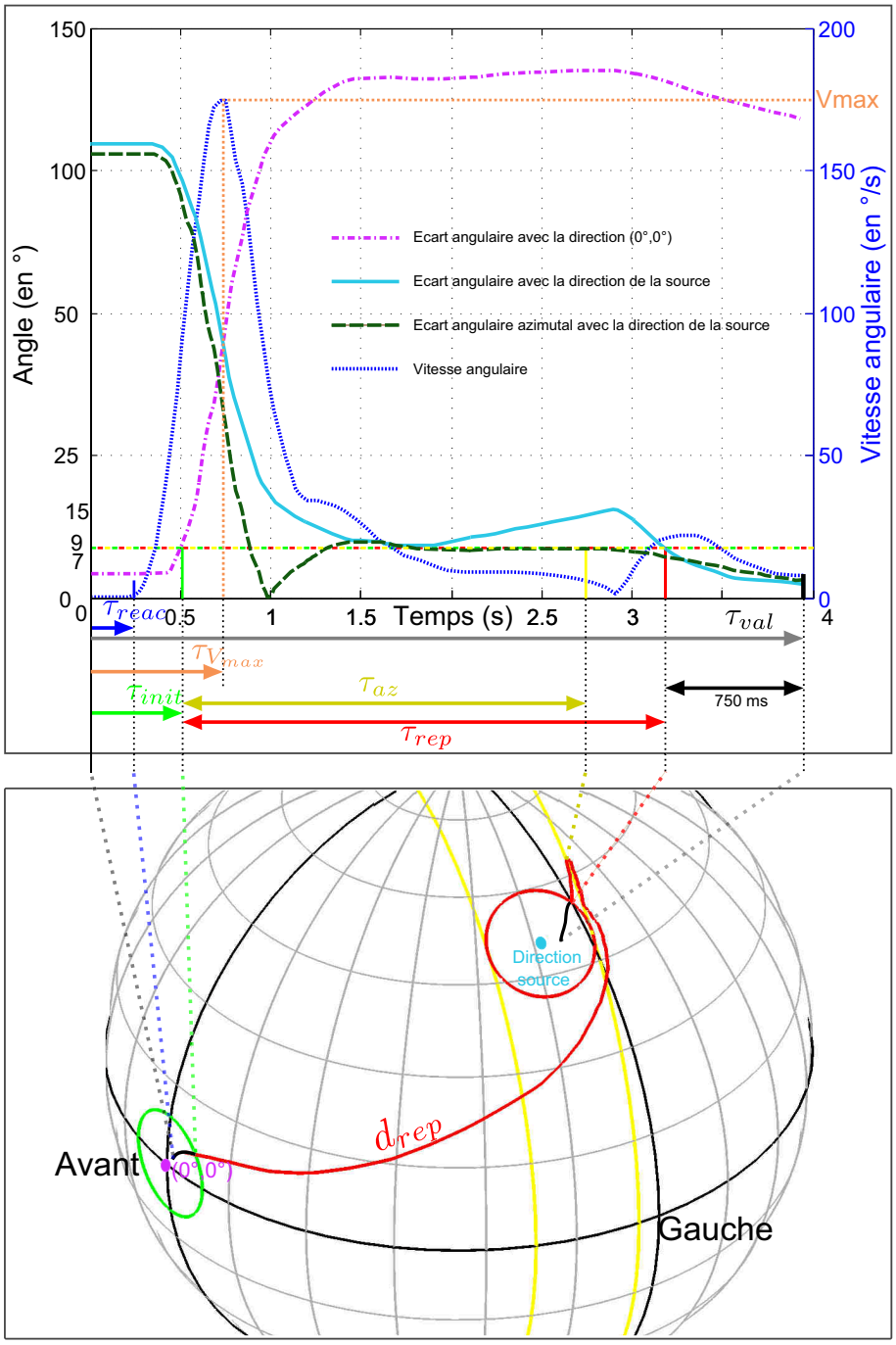


Figure M.3 – Description des quantités utilisées pour la définition des différents critères.

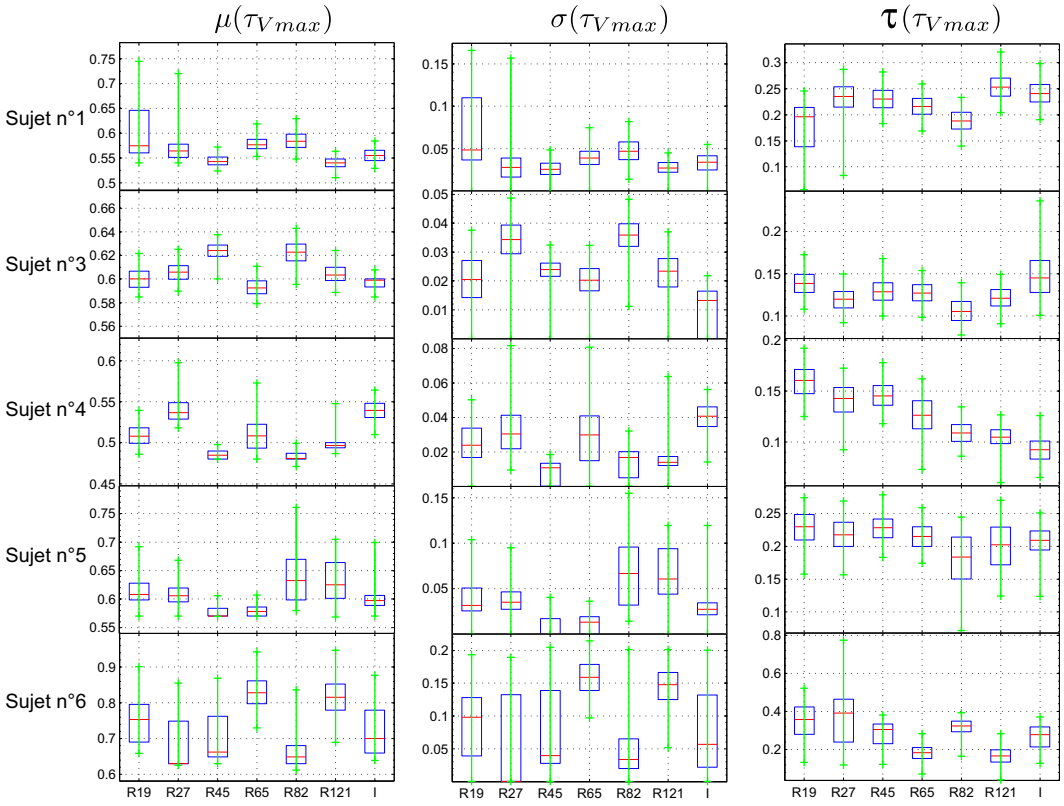


Figure M.4 – Paramètres μ, σ et τ de la fonction ex-gaussienne ajustée à la distribution de $\tau_{V_{max}}$, pour les différents sujets, et chaque condition, en considérant conjointement les données des essais n° 3, 4 et 5.

Sujet n°	Condition					
	R19	R27	R45	R65	R82	R121
1	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
3	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
4	$\neq$	$\neq$	$\equiv$	$\equiv$	$\neq$	$\neq$
5	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
6	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$

Table M.1 – Résultats des tests de permutation des hypothèses $H_0 : F_R \equiv F_I$ relatifs à $\tau_{V_{max}}$. On représente le résultat par le symbole $\neq$ ou $\equiv$ selon que l'hypothèse H_0 est respectivement rejetée ou non.

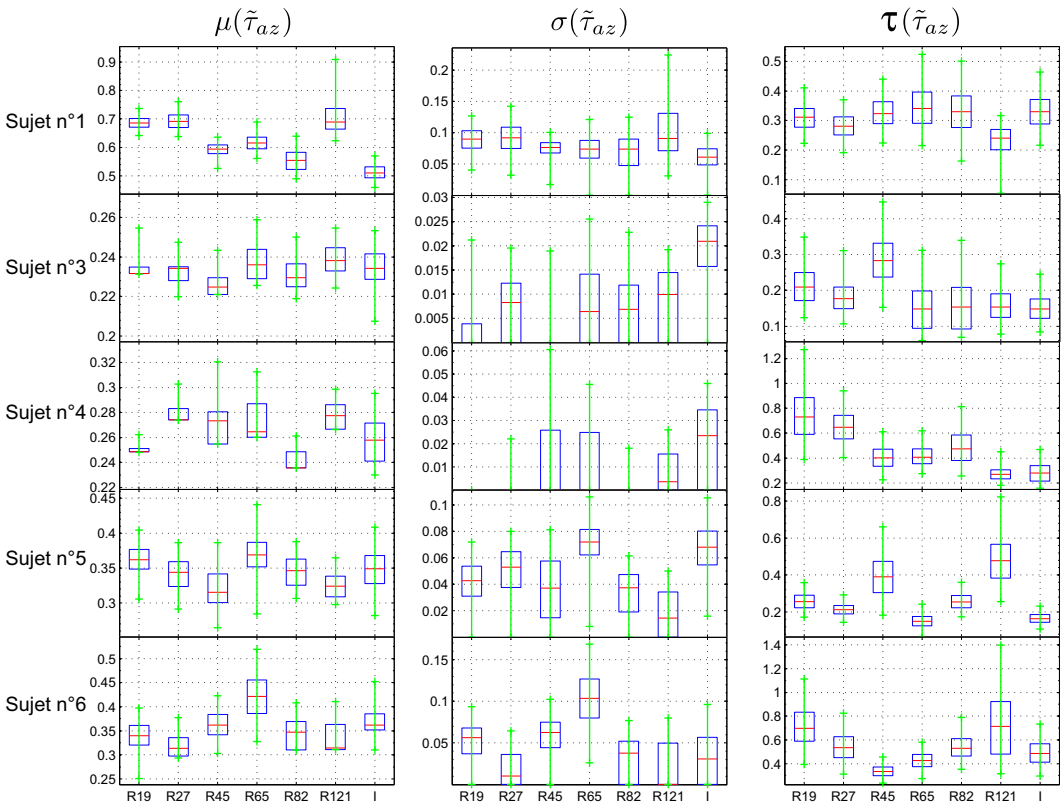


Figure M.5 – Paramètres μ, σ et τ de la fonction ex-gaussienne ajustée à la distribution de $\tilde{\tau}_{az}$, pour les différents sujets, et chaque condition, en considérant conjointement les données des essais n° 3, 4 et 5.

Sujet n°	Condition					
	R19	R27	R45	R65	R82	R121
1	$\neq$	$\neq$	$\equiv$	$\neq$	$\equiv$	$\neq$
3	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
4	$\neq$	$\neq$	$\equiv$	$\equiv$	$\equiv$	$\equiv$
5	$\neq$	$\equiv$	$\neq$	$\equiv$	$\neq$	$\neq$
6	$\equiv$	$\equiv$	$\neq$	$\equiv$	$\equiv$	$\equiv$

Table M.2 – Résultats des tests de permutation des hypothèses $H_0 : F_R \equiv F_I$ relatifs à $\tilde{\tau}_{az}$. On représente le résultat par le symbole $\neq$ ou $\equiv$ selon que l'hypothèse H_0 est respectivement rejetée ou non.

R82 et R121 pour le sujet n°5, et condition R45 pour le sujet n°6). Pour le sujet n°4, il apparaît le phénomène contraire : pour les conditions R45 et R121 qui se distinguent de la condition I en termes de temps de réponse global, le comportement est identique selon l'analyse de $\tilde{\tau}_{az}$.

Si les résultats relatifs à $\tilde{\tau}_{rep}$ sont d'une importance prépondérante, car ce sont les seuls qui traduisent dans sa globalité la tâche de localisation, les deux critères $\tilde{\tau}_{az}$ et $\tau_{V_{max}}$ ont été envisagés pour apporter des indices complémentaires sur le comportement dynamique des sujets. Malheureusement, aucune tendance nette ne ressort de leur analyse, peut-être parce que la variabilité inter-individuelle des comportements masque des informations pertinentes, ou bien parce que cette méthode d'analyse est trop bruitée. Le critère $\tilde{\tau}_{az}$ est peut-être tout simplement inadapté. En effet son introduction repose sur l'idée - peut-être simpliste - que les sujets, pour atteindre la direction de la source, s'orienteraient séquentiellement vers celle-ci d'abord en azimut, puis en élévation. Cela semble être le cas seulement quand la spatialisation est de qualité médiocre (voir notamment les trajectoires du sujet n°1 dans les conditions NI1 NI2 et NI3, figure 6.52). De plus les sujets n'avaient qu'une seule consigne : celle d'être le plus rapide possible. Ainsi, des stratégies différentes ont pu être adoptées d'un individu à l'autre, et par ailleurs, elles ont pu évoluer au cours du temps pour un sujet donné.

Publications

Audio Engineering Society

Convention Paper 7610

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

Head-Related Transfer Function reconstruction from sparse measurements considering a priori knowledge from database analysis: a pattern recognition approach

Pierre Guillon[1,2], Rozenn Nicol[1]

[1] *Laboratoire d'Acoustique de l'Université du Maine, Le Mans, France*

[2] *Orange Labs, Lannion, France*

Correspondence should be addressed to Pierre Guillon (pierre.guillon@orange-ftgroup.com)

ABSTRACT

Individualized Head-Related Transfer Functions (HRTFs) are required to achieve high quality Virtual Auditory Spaces. This study proposes to decrease the total number of measured directions in order to make acoustic measurements more comfortable. To overcome the limit of sparseness for which classical interpolation techniques fail to properly reconstruct HRTFs, additional knowledge has to be injected. Focusing on the spatial structure of HRTFs, the analysis of a large HRTF database enables to introduce spatial prototypes. After a pattern recognition process, these prototypes serve as a well-informed background for the reconstruction of any sparsely measured set of individual HRTFs. This technique shows better spatial fidelity than blind interpolation techniques.

1. INTRODUCTION

High fidelity Virtual Auditory Spaces (VAS) are efficiently created over headphones by binaural synthesis techniques, using digital filters that simulate directional-dependent acoustical phenomena between a source and the listener's eardrums [1, 2]. These filters are classically originated by Head-Related Transfer Functions (or HRTFs), defined as the linear system representation of all these phenomena. HRTFs convey all free field localization cues needed by the auditory system to perceive a

sound scene in 3D. HRTFs strongly depend on idiosyncratic morphological features (overall shape of the head, fine structure of the pinnae), which explains why the use of non-individual HRTFs leads to spatialization artifacts (in particular up-down and front-back confusions, lack of externalization)[3]. Acoustic measurements of individual HRTFs are usually achieved over the whole sphere on a dense angular sampling grid. Such measurement sessions are too long and uncomfortable for subjects, and therefore inappropriate in the context of a commercial diffusion of binaural technologies. Reducing the number of source positions can make it faster and therefore more comfortable, but then interpolation is required, in order to reconstruct HRTFs over the whole sphere. However interpolation techniques remain reliable until a certain limit of the spatial sampling coarseness : our goal is to go beyond this limit, i.e. to achieve a proper reconstruction of HRTFs for any direction from very sparse measurements.

Rayleigh's duplex theory has been extensively revisited during the last fourty years. Shaw first revealed that the pinna structure forms a series of acoustic resonators, causing a multimodal wave propagation. Since these modes are excited depending on the direction of the incident wave, the incoming sound spectrum is transformed in a direction-dependent manner. These spectral cues are dependent on the fine shape of pinna cavities. Psychophysical experiments indicate that they are needed to perceive source elevation in any sagittal plane, and to localize sound sources as out of the head. As the perceptual decoding of these cues is finely adapted to one's morphology, the use of non-individual HRTFs in VAS leads logically to artifacts. Use of individual HRTFs is therefore highly preferred. Nevertheless, the acoustical measurement of these data over a fine spatial sampling is long and uncomfortable for subjects. The following solution therefore comes naturally : to measure data for only few directions, and then to predict HRTFs everywhere via interpolation. The question of how to best accomplish interpolation was addressed by Hartung *et al.* [4]. Among several techniques, the authors proved that the one giving the best objective and subjective results is Spherical Thin Plate Splines [5] (STPS) performed over the log magnitude of HRTFs. Nevertheless, experiments were achieved with a rather fine grid (15 degrees azimuth step). Minaar *et al.* [6] tried to determine the sparseness limit of the spatial sampling needed to properly reconstruct data. They chose a linear interpolation of the minimum phase component of HRTF in time domain. A discrimination test led to the conclusion that more than 1100 directions are needed in these conditions to prevent from any audible change. Their result is a worst case limit: as the authors concede, the perceptual evaluation was very strict. Localization tests may be more convenient to evaluate HRTF reconstruction. This is precisely how Carlile *et al.* [7] addressed the problem. They performed a STPS interpolation of the coefficients weighting the PCA decomposition of HRTF spectra, as proposed by Chen *et al.* [8]. The perceptual evaluation, measured in terms of localization performances in a VAS, led to a lower limit of about 150 needed directions. Martin *et al.* [9] developed a geometrical neighbour-weighting interpolation technique, based exclusively on a regularly spaced azimuth-elevation grid. They concluded that a 20 degrees step between measured directions is sufficient, representing about 80 directions. Ajdler *et al.* [10] recently developed an angular sampling theorem, and showed that the angular Nyquist limit to avoid spatial aliasing is 4.9 degrees for HRTFs measured along horizontal circles, with a sampling rate of 44.1kHz. In order to overcome this limit, the authors proposed to work on subbands. Interpolation is performed in a complex temporal envelope domain. Accurate reconstruction is achieved on the horizontal plane with only 20 degrees spaced measurements.

These classical "blind" interpolation techniques need further information about HRTFs to properly reconstruct data from less than 80 directions. Indeed, rapid spatial variations cannot be recovered properly from a sparse sampling grid without injecting further information. Lemaire *et al.* [11] suggested to use a neural network, trained over a large HRTF database collected for many subjects. Their model achieves an optimized non-linear interpolation between measured HRTFs. Its efficiency was proved for 50 measured directions and more. However, parameterization of the model is quite difficult to handle.

We aim at supplying this lacking information, by taking advantage of a priori knowledge gathered by

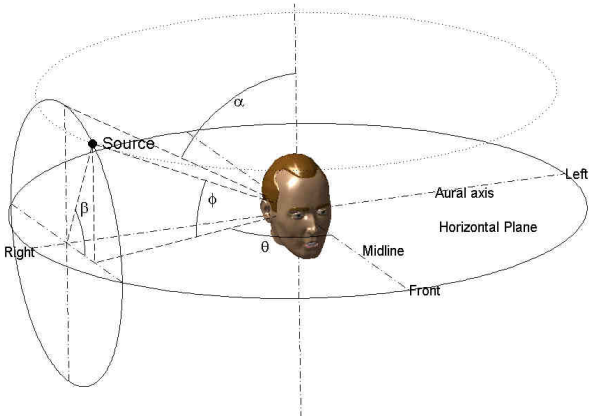


Fig. 1: Coordinate system

analyzing a large HRTF database, finely measured on many subjects. What information can be collected from this database? Focusing on the spatial structure of HRTF magnitude at a given frequency, referred to as Spatial Frequency Response Surface (SFRS) [12], great similarities are observed among subjects. However these similarities are somehow hidden, because they can occur with a frequency shift, and/or a rotation shift, due to the size differences respectively of the ear pinnae size and/or pinnae orientation between two subjects. Thus an appropriate similarity measure, i.e. which is able to point out similarity regardless frequency and/or rotation shift, should be used to analyze the SFRSs. Our similarity measure is defined as the normalized spherical cross-correlation, which is efficiently computed in the spherical harmonic domain. Using this measure, a clustering of SFRSs over the whole database is performed, so as to extract a limited number of prototypic SFRSs. These prototypes will serve as a set of well-informed high resolution background SFRSs for reconstruction. The clustering technique used is the normalized spectral clustering. Thus, our proposal is the following. Having measured HRTFs on a sparse spatial sampling for a new subject, a pattern recognition stage is carried out in order to associate each of the low resolution SFRSs to its most similar prototype. Eventually, a correction stage is performed so as to compensate for the reconstruction error.

This paper is organized as follows. In part 2, we first illustrate observations of inter-subject SFRSs similarities supporting our assumptions, and expose perceptual results. Our solution is then described in part 3, and the methodology of its assessment is detailed in part 4. Finally results are presented and discussed in part 5, before concluding.

2. FACTS AND OBSERVATIONS ABOUT HRTFS AND AUDITORY LOCALIZATION

For a particular azimuth θ and elevation ϕ (cf. Fig. 1), HRTFs are the complex frequency response functions defined as the ratio of complex sound pressure level (SPL) at the entrance of the blocked ear canal to the SPL at the center of the head when the listener is absent [15]. HRTF are ideally acquired by acoustical measurements with a great spatial resolution (about 1000 directions, Cf. Fig. 2). Such measurement sessions are very long and therefore hardly bearable for the involved subject. Using a sparse directional sampling seems to be the only way to make HRTF individual measurements conceivable. As it is widely admitted, HRTF can be replaced by their minimum phase form without causing major perceptual artifacts [16]. Doing so, temporal cues and spectrum can be treated separately : our purpose in this study is to recover the magnitude of HRTF, which represents their most critical part. We focus on the directionality evolution of HRTF magnitude for a given frequency ν, which are referred to as Spatial Frequency Response Surfaces (SFRSs) [12]. HRTFs and SFRSs are alternative views of the same data:

$$SFRS_\nu(\theta,\phi) = |HRTF(\theta,\phi)(\nu)|_{dB}$$

where ν denotes the frequency, and θ and ϕ denote azimuth and elevation (Cf. Fig. 1). Examples of SFRSs are given in Fig. 3.

2.1. Spectral cues

HRTF spectrum conveys relevant information about the location of the source: these directional encoding features are called spectral cues. Basically, while interaural cues, namely Interaural Time Difference (ITD) and Interaural Intensity Difference (IID), are decoded by the auditory system to perceive the sagittal position α of a source (Cf. Fig. 1) , spectral cues are needed to perceive its elevation β (Cf.

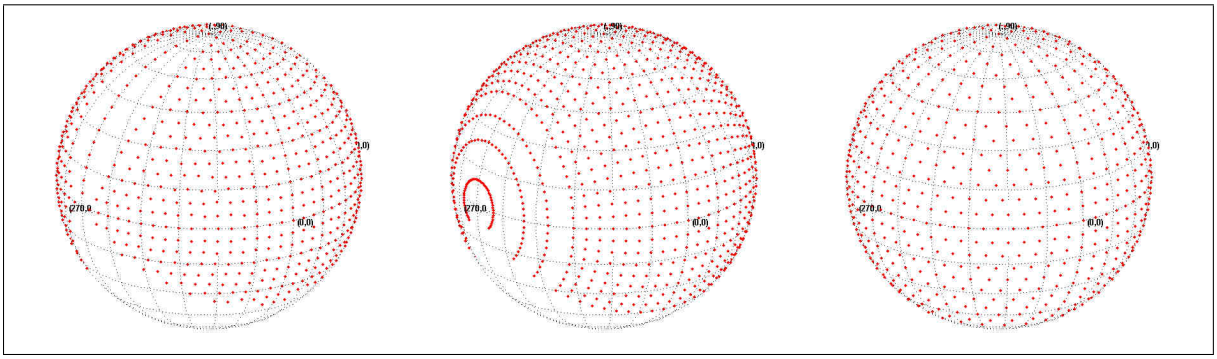


Fig. 2: Spatial sampling for HRTF measurement: Illustration for 3 databases. From left to right: E. Grassi, University of Maryland (1093 directions) [13], CIPIC, UC Davis (1250 directions) [14], Orange Labs (965 directions). Azimuth and elevation (θ, ϕ) are specified between brackets.

Fig. 1) onto the cones of confusion [17]. Spectral cues are also needed to perceive sources as "out of the head" (externalization) [18].

Many studies have addressed the problem of isolating particular features of HRTF spectrum that would be the only useful information for the auditory system [19, 20, 21]. Unfortunately, symplifying spectral cues as isolated features seems to be impossible: even if physiological processing of spectral cues remains partially unclear, it has been established that directional information is interpreted from a large portion of the spectrum [22]. More precisely, HRTF spectral shape in the whole [4 - 13 kHz] band is important [23], and it corresponds to acoustic phenomena occuring in the pinnae [24]. Other directional dependent variations of the spectrum exist at lower and higher frequencies, but they are either secondary or not interpreted [25, 26, 23]. Further studies about spectral smoothing reveal that although very rapid spectral variations can be discarded, serious losses of dynamic between spectral peaks and notches cause perception artifacts.

Although both ears contribute to the perception of the vertical angle, as a sound source shifts laterally from the median plane, the contribution of cues from the near ear increases, while conversely, that of the far ear decreases [25]. This results from a weighting of spectral information from each ear, depending on the perceived lateral angle [27]. Thus, the contralateral ear no longer contributes when the absolute lateral angle α of a sound source from the midline is larger than 30° (Cf. Fig. 1) [25].

2.2. Inter-subject similarities

Since the first acoustical measurements on ear replica [24] to the numerical simulations on mesh models [28], several comprehensive studies have observed and analyzed acoustical modes of pinnae originating spectral cues. These acoustical phenomena directly depend on the fine structure of ear pinnae, and as pinna shapes are nearly as individual as fingerprints, it was shown that HRTF spectra also are very individual [29]. Nevertheless, some very similar behaviours have been pointed out [30]: as frequency increases in the band [4kHz - 13kHz], SFRSs take successively characteristic shapes (Cf Fig. 3). First a large pole covers a wide part of the ipsilateral hemisphere. Its central direction moves backward and upward while its spatial extent becomes narrower. A second pole appears below the horizontal plane, until a bipolar shape is reached. At a stage called "breakpoint" [31], the lower pole becomes preponderant and slides continuously backward. Another pole appears in the frontal hemisphere. At higher frequencies, SFRS shapes are so rugged that no general behaviour is observed. Indeed, as frequency increases, SFRS shapes become more and more individual, and may differ from this behaviour. It can be understood considering that at long wavelengths, all pinnae resemble each other, whereas at short wavelengths, individual pinnae structures are "viewed" more precisely by incident acoustic waves. Fig. 4 depicts the trajectory followed by the SFRS spatial

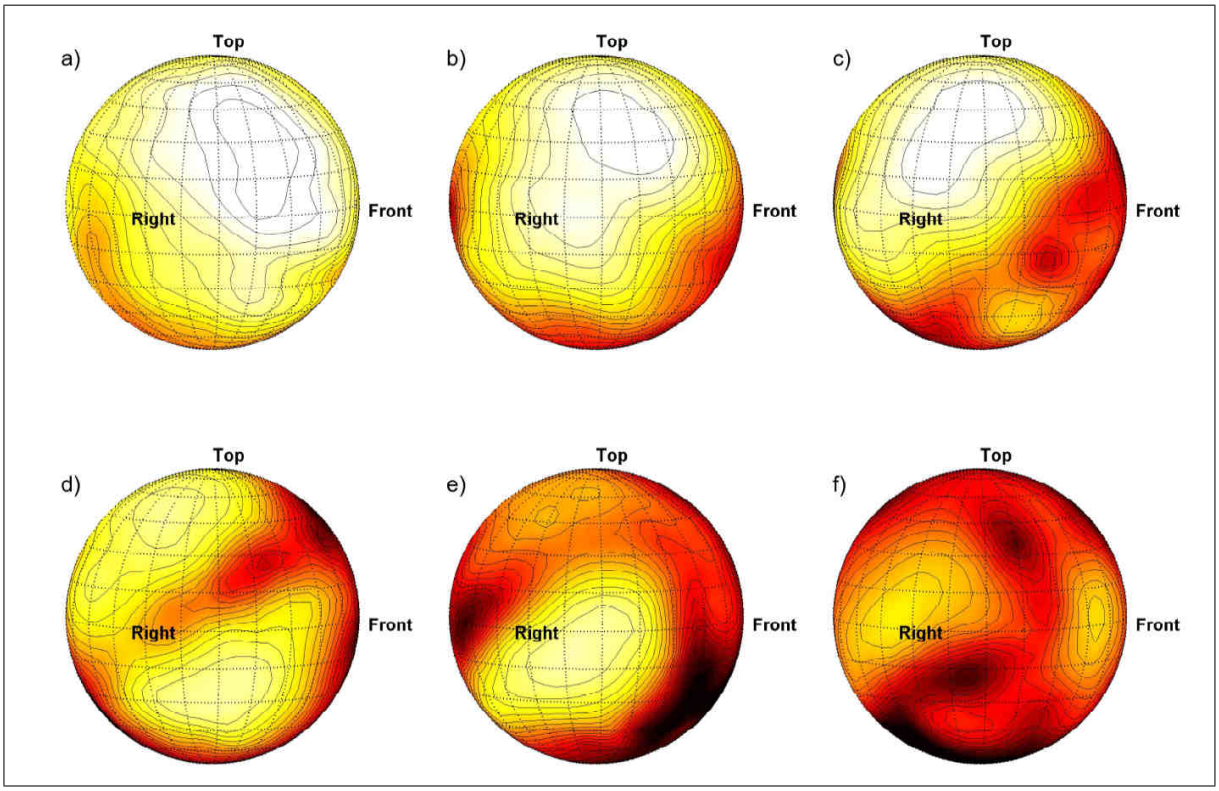


Fig. 3: Typical shapes taken by SFRSs for increasing frequency: right ear SFRSs of a subject represented over the sphere for a) 3650Hz, b) 4600Hz, c) 6000Hz, d) 7400Hz, e) 8800Hz, f) 11200Hz. (Bright zones represent high level, whereas dark zones represent low level)

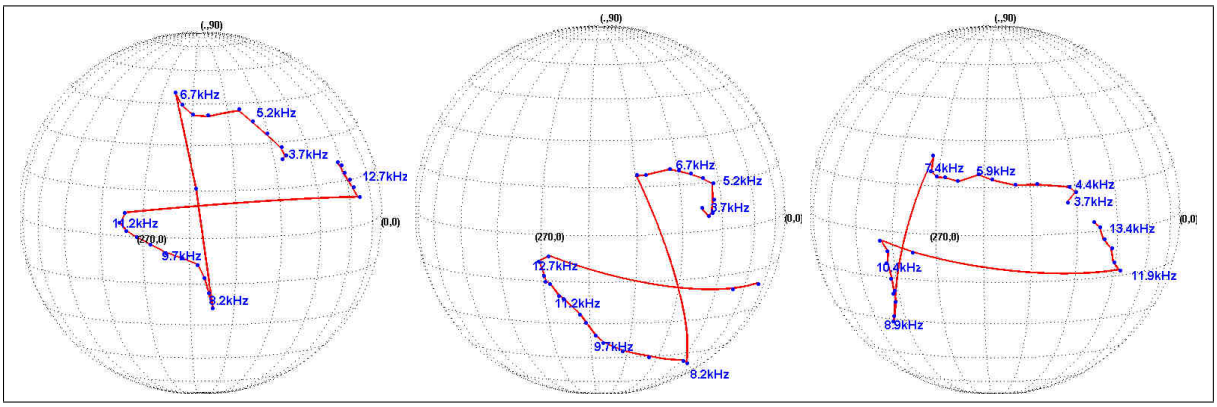


Fig. 4: When the frequency increases, the SFRS spatial maximum follows a characteristic trajectory: Illustration for 3 individuals.

maxima in function of frequency for three individuals. The inter-subject similarity is strikingly noticeable.

From subject to subject, this general evolution could be shifted along the frequency axis [30]. In other words, the particular SFRS of a subject can be very similar to the SFRS of another subject but for a different frequency. This observation is supported by the fact that a global scaling can explain part of the differences between ear pinnae [32]. Inspired by a study on Mongolian gerbils [33], we further assume that the spatial orientation of the pinnae can explain the fact that SFRSs resemble each other to a rotation shift. For example, from subject to subject, the valley orientation at the first breakpoint can take varying deviation angles from the horizontal plane. As a result, HRTFs of one individual can be derived from the HRTFs of another individual by combining frequency scaling and spatial rotation with appropriate tranformation parameters according to the morphological differences between the two individuals [34].

2.3. Concluding remarks

From these results and observations, we can make conclusions for our particular purpose.

1. Our model may only focus on the particular SFRS in the [4 - 13 kHz] band. Of course SFRSs at lower and higher frequencies have to be reconstructed, but as their spatial evolution conveys no directional information, it is assumed that a rough interpolation between the sparse measured directions should be satisfying, and keep IID unchanged.
2. As the rugged behaviour of SFRSs for extreme contralateral directions are perceptually unexploited, we can also roughly reconstruct this spatial portion of SFRSs by interpolation.
3. In the design of a similarity measure between SFRSs, a rotation shift degree of freedom must be taken into account. In the same way, comparison of SFRSs taken from different subjects must be achieved regardless of their frequency.

3. PROPOSED MODEL

Our model relies on two different processes:

- an off-line process, which consists in analyzing a database of HRTFs and organizing the information thus extracted, for purposes of future HRTF reconstruction,
- an on-line process, which is performed for any new individual in order to reconstruct his HRTFs from sparse measurement.

The off-line process is composed of the following steps:

- collecting a large database of high-resolution SFRSs measured for a given number of individuals and frequencies,
- computing the similarity measure for any pair of SFRSs extracted from the previous database,
- clustering the SFRSs according to the similarity measure: each cluster defines a prototype of SFRS, which will be used for the reconstruction process.

Then, the on-line HRTF reconstruction consists of (for each frequency):

- identifying by pattern matching the prototype which is the closest to the low-resolution SFRS of the new individual,
- adjusting this prototype in order to minimize the reconstruction error for the measured directions.

The main stages of the overall process are described in the following.

3.1. Similarity measure

In order to detect similarities between SFRSs, a similarity measure has first to be designed: we choose to use normalized cross-correlation, a known function in rotational matching problems [35, 36]. Considering two functions f and g defined on the sphere $f, g \in L^2(S^2)$, normalized cross-correlation $C_R(f,g)$ between f and g is defined for a rotation R as :

$$C_R(f,g) = \frac{\int_{S^2} \breve{f}(\omega)\overline{\Lambda_R(\breve{g})(\omega)\mathrm{d}\omega}}{\sqrt{\int_{S^2} |\breve{f}(\omega)|^2\mathrm{d}\omega \int_{S^2} |\breve{g}(\omega)|^2\mathrm{d}\omega}}$$

where $R \in SO(3)$,

$$\Lambda : \quad L^2(S^2) \rightarrow L^2(S^2)$$
$$\Lambda_R(g)(\omega) = g(R^{-1}(\omega))$$

and $\breve{f}$ results from centering f on its spatial average:

$$\breve{f}(\omega) = f(\omega) - \frac{1}{4\pi}\int_{S^2} f(\omega)\mathrm{d}\omega$$

In these equations, ω denotes the spatial coordinates. The normalized cross-correlation $C_R(f,g)$ have to be computed for an extensive range of rotations over $SO(3)$. There are two ways of computing $C_R(f,g)$: either directly in the spatial domain, which means one rotation at a time, or in the dual domain (spatial spectrum), which allows to get all the cross-correlation values at once for a full sampling of rotations over $SO(3)$ [37] (Cf. Appendix 7.2). We choose this second solution for computational efficiency. In this case, the tranformation into the dual domain of spatial coordinates relies on a Spherical Harmonics (SH) decomposition. Details about the calculation of SH decomposition can be found in Appendix 7.1. When SH are used for similarity measurement between two SFRSs, it should be kept in mind that these SFRSs have to be developed up to the same maximal degree.

The similarity measure $\Upsilon(h,h')$ between two SFRSs, h and h', is defined as the maximum value of $C_R(h,h')$ over $SO(3)$:

$$\Upsilon(h,h') = \max_{R \in SO(3)} C_R(h,h')$$

This similarity measure possesses the desired rotation shift degree of freedom. The database is fully analyzed with this measure. All SFRSs are taken regardless of their subject belonging, and each SFRS h_ν is compared with all other SFRSs $h'_{\nu'}$ within a limited range of neighbor frequencies: $\nu' \in [\nu - \nu_0/2; \nu + \nu_0/2]$. This gives us another degree of freedom along the frequency axis, enabling to detect similarities between SFRSs to a frequency shift.

3.2. Clustering

Once the similarities between all the pairs of SFRSs collected in the database are computed, an adjacency matrix $\mathbf{W}$ is then built: in each line i of $\mathbf{W}$, k entries are non-zero, corresponding to the similarity measures Υ between the ith SFRS and the k most similar SFRSs in the database (nearest neighbors). $\mathbf{W}$ is symmetrized, as needed by the following operations. Based on $\mathbf{W}$, we perform a spectral graph clustering algorithm [38]. We basically consider the whole dataset as a graph, where data are vertices, connected by edges. These edges are weighted by the value of the similarity measure between their vertices. Spectral graph clustering methods aim to find a partition of this graph such that the edges between different groups have low weight (low inter-cluster similarity), and the edges within a group have high weight (high intra-cluster similarity). This is performed considering the normalized Laplacian $\mathcal{L}$ of the graph, defined as:

$$\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$$

where $d_i = \sum_{j=1}^{n} w_{ij}$ are the entries of the diagonal degree matrix $\mathbf{D}$, n being the total number of SFRSs. Here is the algorithm :

- Compute the normalized Laplacian $\mathcal{L}$
- Compute the first κ eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\kappa$ of $\mathcal{L}$.
- Let $\mathbf{U} \in \mathbb{R}^{n \times \kappa}$ be the matrix containing the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\kappa$ of $\mathcal{L}$ as columns.
- For $i = 1, \ldots, n$, let $\mathbf{y}_i \in \mathbb{R}^\kappa$ be the vector corresponding to the ith row of U.
- Cluster the points $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^\kappa$ with the k-means algorithm [39] into clusters $\mathbf{C}_1, \ldots, \mathbf{C}_\kappa$.

The resulting SFRS clusters are $\mathbf{A}_1, \ldots, \mathbf{A}_\kappa$ with $\mathbf{A}_i = \{h_j | y_j \in \mathbf{C}_i\}$: each cluster $\mathbf{A}_i$ is composed of all SFRSs h_j whose corresponding vector $\mathbf{y}_j$ is in $\mathbf{C}_i$.

During the similarity measurement, the rotation achieving the maximum of the normalized cross-correlation is determined, but, up to now, the SFRSs are let unrotated, which means that the clusters gather homogoneous SFRSs modulo a rotation shift. An alignment process has therefore to be performed. First, a representative SFRS is chosen for each cluster: this is the SFRS maximizing the mean similarity with all members of the cluster. The alignment

process is performed by applying the optimal shift by rotation that maximizes similarity measure between each SFRS h and the representative SFRS of its cluster h_{rep}. At first guess, a good approximate is the already computed rotation R_0, argument of the maximum of $C_R(h, h_{rep})$. Nevertheless, as R_0 has been computed on a rather coarse sampling of $SO(3)$, a finer solution R_{opt} can be reached by exploring continuously $SO(3)$ with a gradient descent algorithm [36]. We initialize this algorithm with the value R_0. Details can be found in Appendix 7.3. Finally all aligned SFRSs are averaged in each cluster, and the resulting SFRSs, called prototypes, are stored.

3.3. Pattern matching

Now we suppose that low directional resolution HRTF have been measured on a new subject. We consider the corresponding SFRSs $\hbar$ collected from the whole frequency range. The aim of this template matching stage is to find among all prototypes, which is the nearest to each $\hbar$. This identification stage is carried out by computing the similarity $\Upsilon(\hbar, h_{prot})$ between each $\hbar$ and every prototype h_{prot}. An alignment of each low resolution $\hbar$ is then performed with its associated prototype, by rotating the latter. The contralateral part of the prototype is discarded: data for these directions are reconstructed by STPS interpolation of measured data. We note h_{pm} the resulting SFRS: this is the first step of the high resolution HRTF reconstruction.

As previously, the computation of the similarity between measured SFRSs and prototypes is based on SH decomposition. However it should be realized that the maximum degree of the SH decomposition of a given SFRS is limited by the total number of measurement directions (Cf. Appendix 7.1). To some extent, this maximum degree can be considered as the equivalent to the Nyquist frequency for spatial spectrum. Thus, as spatial sampling becomes sparser, this limit decreases. Therefore SH decompositions are more likely to suffer from spatial aliasing problems, and our model is more likely to perform badly.

3.4. Correction

The first step of pattern matching aims at recovering a SFRS with a proper shape (h_{pm}), but it may fail recovering exactly the measured data. We therefore perform a final correction process whose aim is to cancel the known error between h_{pm} and $\hbar$ in the measured directions. In order not to alter the gained SFRS shape, a smooth error function h_{err} is defined and computed by STPS interpolation over the high resolution final sampling. We finally have:

$$h_{rec} = h_{pm} - h_{err}$$

4. MODEL SETUP AND ASSESSMENT METHODOLOGY

Our model is set up on a HRTF database resulting from the merging of four different databases (CIPIC [14], IRCAM Listen [40], E. Grassi University of Maryland [13], and Orange Labs private database). SFRSs from a setup database of 101 subjects are compared and clustered, to get the prototypes. The HRTF reconstruction is assessed for 8 subjects whose data are not included in the setup database. Twelve different spatial sampling grids are used, from about 130 to 20 directions, which represent the sparse measurement. The performances of our technique are compared to STPS interpolation of SFRSs (as proposed in [4]), and also to the method proposed in [7], i.e. STPS interpolation of the weights of PCA decomposition.

4.1. Setup database

As original HRTFs from the merged database were measured with different sampling frequencies, we first interpolate HRTFs on a common frequency scale. Furthermore, as they were measured on different spatial sampling grids, we then interpolate SFRSs with STPS, to a common target grid (octahedron projection method [41], 1602 directions). Only 37 regularly spaced SFRSs on the band [3.5 kHz - 13 kHz] are retained. HRTFs for directions in the lower hemispherical cap, that technically could not be measured, are also generated in this interpolation stage. Although the obtained data are irrelevant, this is needed to constrain SH decomposition, and avoid blowup of the approximated function. SH decomposition of the SFRSs is obtained up to degree 30. So as to make all left and right ear SFRSs comparable, we simply symmetrize all left ear SFRSs relative to the median plane.

4.2. Sparse sampling

So as to investigate the effect of sparsity on reconstruction methods, several data samplings are constituted by selecting subsets of the available data. In order to keep the same spatial homogeneity between samplings, selected directions from the original samplings are the closest to the solutions of the sphere covering problem [42] for a given total number of directions. We test 12 sparse samplings between about 20 and 130 directions. As done on the setup database, lower cap directions are artificially added to the test samplings, and left ear SFRSs are symmetrized relative to the median plane. SH decompositions are evaluated up to the maximal degree enabled by the number of available of directions.

4.3. Model settings

During the clustering stage of the proposed model, comparisons over the whole database are performed between SFRS corresponding to nearby frequency, by setting parameter $\nu_0 = 3kHz$. Only the first $k = 40$ nearest neighbours are retained to compose the adjacency matrix $\mathbf{W}$. $\kappa = 300$ groups are constituted by spectral graphs clustering. In order to reconstruct HRTFs over the whole audible band, HRTF data measured below 3.5 kHz and above 13 kHz are interpolated with STPS and combined with the data reconstructed by our model.

The new model is compared with two methods: blind interpolation by STPS and PCA-based reconstruction as proposed by Carlile *et al.* [7]. For the latter, the basis of eigenvectors is generated from the principal component analysis of HRTFs measured on 6 subjects (different from the tested subjects). Eight eigenvectors are retained, since the eigenvalues analysis shows that they account for 95% of the total variance.

4.4. Objective evaluation

Three objective criteria are used to quantify and compare the performances of each method of HRTF reconstruction.

4.4.1. Root mean square error

First, the classical root mean square error, referred to as ϵ_{rmse}, is used. It should be noticed that the error is expressed in dB, since it is computed from the frequency-by-frequency squared difference between log-magnitudes of original (h_0) and reconstructed (h_{rec}) HRTFs. The squared difference is then averaged over all frequencies. Eventually, the square root of the result is averaged over all test directions:

$$\epsilon_{rmse} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[h_0(\nu_i,\omega_j) - h_{rec}(\nu_i,\omega_j)\right]^2}$$

where M is the number of test directions and N the total number of frequency points.

4.4.2. Maximal ipsilateral error over 1/3 octave bands

An error measure proposed by Langendijk *et al.* [43] and which takes into account coarsely the limited frequency resolution of the auditory system, is also considered. The log-magnitude spectra from 200 Hz to 16kHz is processed through a 1/3 octave filterbank, which leads to $\tilde{h}_0$ and $\tilde{h}_{rec}$, before computing the maximum difference ϵ_{max} (expressed in dB) achieved among all bands between original and reconstructed data:

$$\epsilon_{max} = \max_i \left[\tilde{h}_0(\nu_i,\omega_j) - \tilde{h}_{rec}(\nu_i,\omega_j)\right]$$

4.4.3. Covert Peak Areas reconstruction

As the design of our model is focused on the spatial structure of HRTFs, we introduce new criteria enabling to quantify the spatial fidelity of the reconstruction, based on perceptual considerations. Localization experiments with narrow band stimuli have drawn the attention to the spatial maxima of SFRSs, suggesting that they are potential spectral cues [44], in competition with frequency maxima or minima of HRTFs such as the peaks and notches of spectrum magnitude. Indeed these spatial maxima, each of which is associated with a given frequency, define a spatial mapping of frequencies, which was pointed out by Butler *et al.* [45] with the concept of spatial referents of stimulus frequencies. This concept is supported by the observation of systematic illusory mislocalizations when listening monaurally narrow band stimuli: perceived direction is only lead by frequency content and not by actual source location. To that extent stimulus frequency has a referent in space. These behavioral results were related by the authors to the spatial structure of HRTF

magnitude in narrow frequency bands (SFRS for instance). Indeed a good correlation was found between perceived source direction of a narrow band stimulus with a given central frequency, and a spatial zone called Cover Peak Area (CPA), defined as the surface on the sphere for which HRTF magnitude is close to the spatial maximum for this frequency [45]. Based on these results, a physiologically-motivated model has been proposed to explain localization of sounds with any spectral shape and bandwidth [44].

Therefore, in order to give further insight into the performances of HRTF reconstruction, we propose two new criteria paying attention to the accuracy of CPA reconstruction. CPAs are defined here as the surface for which HRTF magnitude is above the maximum gain minus 1.5dB. Their spatial position and extent for a series of frequencies are assessed. We consider CPAs for 28 frequency bins equally spaced in the band 3.5kHz - 13kHz. Two quantities are introduced: the angle θ_{CPA} between the directions of original and reconstructed CPA centroids, and the surface percentage $\cap_{CPA}$ of original CPA covered by the reconstructed CPA.

5. RESULTS AND DISCUSSION

Fig. 5 plots the root mean square error ϵ_{rmse} in function of the number of measured directions, in the band [3.5kHz - 13kHz]. Our model achieves the lowest error. In comparison with STPS interpolation and PCA-based method, the improvement in the HRTF recontruction is all the better as the spatial sampling of HRTF measurement is coarser. For instance, if the lowest error obtained by PCA-based method occurring for 130 measured directions is considered as a reference, it is observed that the same error level is reached by our model for less than 60 directions, which means a decrease by a factor greater than 2. In addition, it should remarked how STPS interpolation denotes good performances, close to our model for a number of measured directions greater than 80, but rapidly diverges for coarser sampling. Nevertheless it is striking that STPS interpolation achieves better reconstruction than PCA-based method.

Langendijk *et al.* [43] showed that for the maximal ipsilateral error ϵ_{max}, the value of 2.5 dB corresponds to the maximum threshold to prevent from

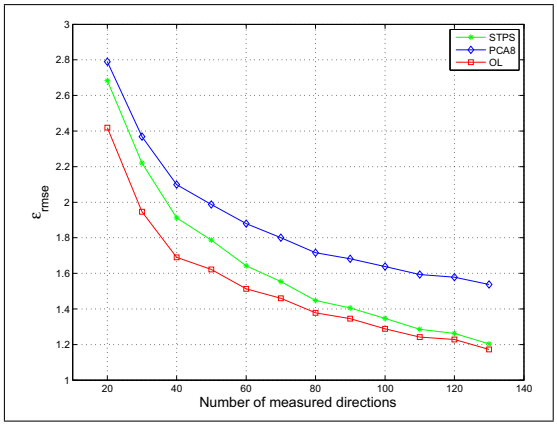


Fig. 5: Root mean square error ϵ_{rmse} in function of the number of measured directions: Comparison of the three methods of HRTF reconstruction, namely: STPS interpolation (STPS), PCA-based method with 8 eigenvectors (PCA8), the proposed model (OL). Results are averaged over all 8 tested subjects, both ears.

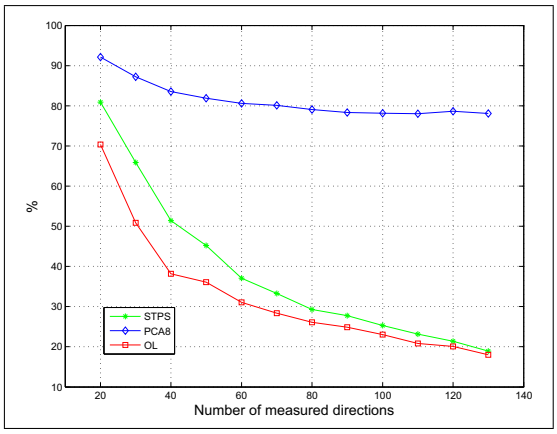


Fig. 6: Maximal ipsilateral error over 1/3 octave bands ϵ_{max} in function of the number of measured directions: Percentage of error values greater than 2.5 dB. Comparison of the three methods of HRTF reconstruction (see Fig. 5 for more details). Percentage is computed over all directions and subjects.

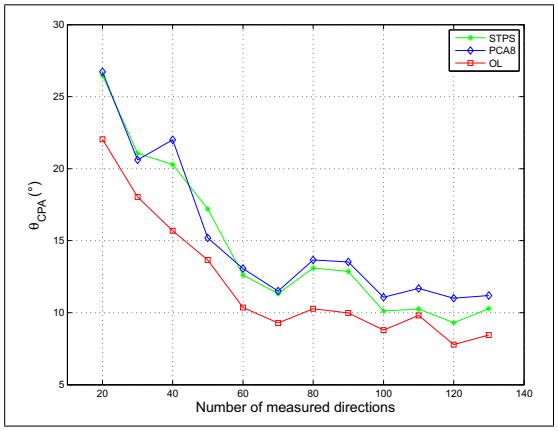


Fig. 7: Average angle θ_{CPA} between the directions of original and reconstructed CPA centroids in function of the number of measured directions: Comparison of the three methods of HRTF reconstruction (see Fig. 5 for more details). Average is computed over all frequencies and subjects.

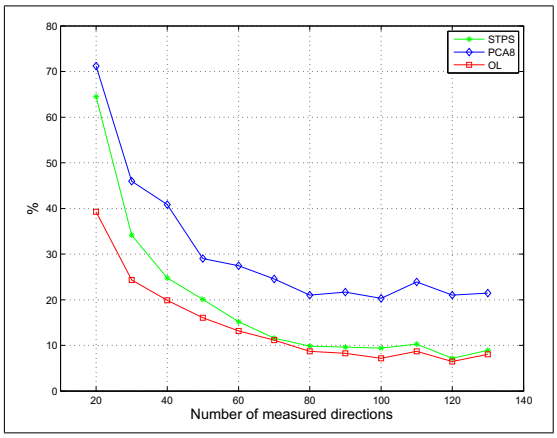


Fig. 8: Percentage of angle values θ_{CPA} greater than 10° in function of the number of measured directions: Comparison of the three methods of HRTF reconstruction (see Fig. 5 for more details). Percentage is computed over all frequencies and subjects.

localization mismatch. Therefore the percentage of cases for which ϵ_{max} is greater than 2.5 dB over all the directions and all the subjects was computed and is depicted in Fig. 6. According to this criterion, PCA-based method exhibits markedly worse performances than the other modellings: whatever the spatial sampling is, the percentage remains greater than 75%. In comparison, our model denotes a percentage lower than 20% for 130 measured directions and keeps a value lower than 30% as soon as more than 60 directions are measured. When the spatial sampling is coarse (i.e. less than 60 directions), our model is 10 points better than STPS interpolation.

As for the CPA reconstruction, first, the angle θ_{CPA} between the directions of original and reconstructed CPA centroids is plotted in function of the number of measured directions in Fig. 7. It can be seen that our model achieves a centroid mismatch noticeably lower than the two other reconstruction methods. Improvement within $2-5°$ is observed. These values correspond to the average over all directions and all subjects. The results are confirmed when examining the data distribution before averaging. Fig. 8 gives the percentage of angle values

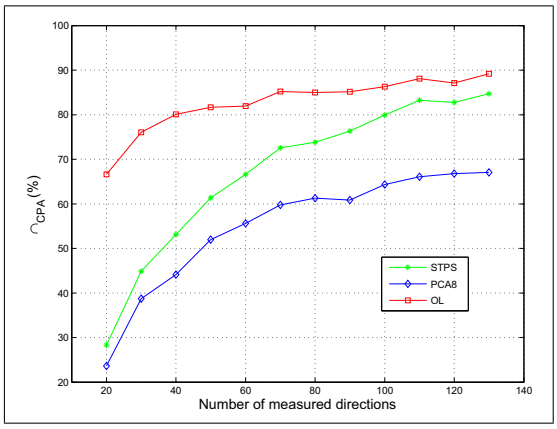


Fig. 9: Average surface percentage $\cap_{CPA}$ of original CPA covered by the reconstructed CPA in function of the number of measured directions: Comparison of the three methods of HRTF reconstruction (see Fig. 5 for more details). Average is computed over all frequencies and subjects.

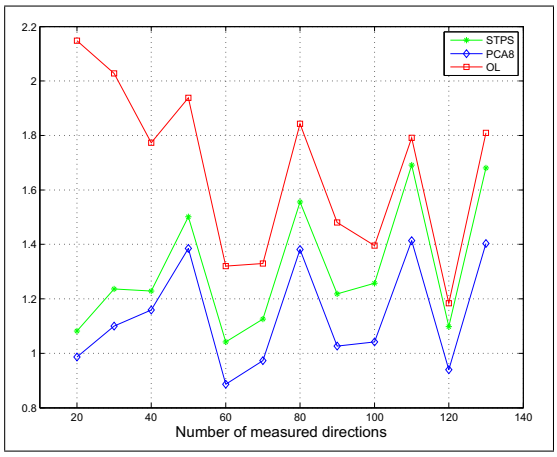


Fig. 10: Average ratio between the original CPA surface and the reconstructed one in function of the number of measured directions: Comparison of the three methods of HRTF reconstruction (see Fig. 5 for more details). Average is computed over all frequencies and subjects.

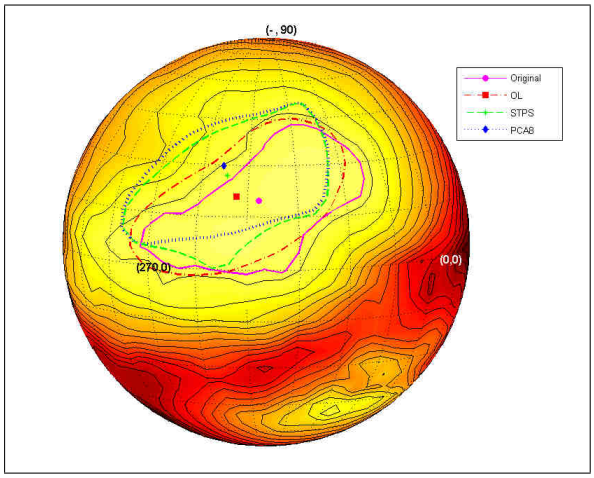


Fig. 11: Illustration of CPA reconstruction (40 measured directions): SFRS corresponding to the right ear of one individual taken from the Orange Labs database at 7 kHz.

θ_{CPA} greater than 10°[1]: our model shows a percentage 10% lower than PCA-based method. At the very best, i.e. for 130 measured directions, the PCA-based method achieves a percentage of 20 %, which is reached by our model for only 40 directions. The number of measured directions is thus decreased by a factor of more than 3. Since a CPA is described both by its direction and its surface, we have also to pay attention to how the CPA surface is reconstructed. Fig. 9 and 10 respectively depict the surface percentage $\cap_{CPA}$ of original CPA covered by the reconstructed CPA and the ratio between the original CPA surface and the reconstructed one. The CPA surface reconstructed by our model overlaps the original one at [80-90%], even for only 40 measured directions. In comparison, the percentage performed by PCA-based method decreases from about 65% down to 45% when the number of measured directions is decreased from 130 to 40. As previously, STPS interpolation exhibits in between behaviour: close to our model for fine sampling and close to PCA-based method when the spatial sampling gets coarser. However these results should be further analyzed in the light of Fig. 10, which shows that our model tends to overestimate the CPA surface by a factor [1.8-2], whereas for the other methods the factor is within [1-1.4]. This partly explains why the reconstructed CPA overlaps so well the original one. CPA reconstruction is illustrated in Fig. 11 for the three methods in comparison with the original SFRS. Moreover it is too early to draw final conclusions from this assessment in terms of CPA reconstruction, since further link with auditory perception is needed to interpret these quantitative criteria as spatialization artifacts.

6. CONCLUSION

This paper proposes a new model of HRTF reconstruction from sparse measurement for binaural individualization purposes. The specificity of the model relies on two main ideas: first it is focussed on the spatial structure of HRTF, i.e. SFRS, second the reconstruction is based on a set of SFRS prototypes which are built up from a large database of HRTF including more than 100 individuals. The perfor-

[1] The threshold of 10° was chosen as a value representative of the magnitude of MAA (Minimum Audible Angle) both for azimuth and elevation discrimination [46, 47].

mance of the proposed model was assessed by various criteria, investigating on the one hand the global reconstruction error and on the other hand the reconstruction of CPAs, as a potential candidates for spectral cues. It is shown that, in comparison with blind interpolation (STPS) and PCA-based method [7], our model achieves signifivantly better reconstruction. The next step is to examine whether this improvement is confirmed by psychoacoustic experiments and to what extent the reconstructed HRTF performs binaural rendering as convincing as the original ones.

7. APPENDICES

7.1. Computation of the Spherical Harmonics decomposition of band limited functions

Let f be a function defined on the sphere $f \in L^2(S^2)$. Considering that this function is band-limited of bandwidth B, we can decompose f as a finite discrete sum of spherical harmonics $Y_l^m(\omega)$.

$$f(\omega) = \sum_{l=0}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\omega)$$

Coefficients $\hat{f}_l^m$ of the decomposition are computed knowing values of f for M non-uniformly sampled directions on the sphere $(\omega_i)_{i=1,\ldots,M}$:

$$f(\omega_i) = \sum_{l=0}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\omega_i)$$

which can be written under a matrix form as:

$$\begin{aligned} \mathbf{f} &= \mathbf{Y}.\mathbf{C} \\ \mathbf{Y} &= \{Y_l^m(\omega_i)\}_{M\times B^2} \\ \mathbf{C} &= \{\hat{f}_l^m\}_{B^2\times 1} \\ \mathbf{f} &= \{f(\omega_i)\}_{M\times 1} \end{aligned}$$

Spherical Harmonics decomposition consists in finding the vector $\mathbf{C}$ which minimizes the error between the known data $f(\omega_i)$ and these estimated by $\mathbf{Y}.\mathbf{C}$:

$$\epsilon = (\mathbf{f} - \mathbf{Y}.\mathbf{C})^2$$

which is solved by a generalized inverse (or pseudo-inverse) procedure:

$$\mathbf{C} = (\mathbf{Y}^t.\mathbf{Y})^{-1}.\mathbf{Y}^t.\mathbf{f}$$

if $M \geq B^2$ (overdetermined problem).

7.2. Normalized cross-correlation

Considering two functions f and g defined on the sphere $f, g \in L^2(S^2)$, normalized cross-correlation $C_R(f,g)$ between f and g is defined for a rotation R as :

$$C_R(f,g) = \frac{\int_{S^2} \breve{f}(\omega)\overline{\Lambda_R(\breve{g})(\omega)\mathrm{d}\omega}}{\sqrt{\int_{S^2} |\breve{f}(\omega)|^2\mathrm{d}\omega \int_{S^2} |\breve{g}(\omega)|^2\mathrm{d}\omega}}$$

where $R \in SO(3)$,

$$\begin{aligned} \Lambda : \quad & L^2(S^2) \rightarrow L^2(S^2) \\ & \Lambda_R(g)(\omega) = g(R^{-1}(\omega)) \end{aligned}$$

and $\breve{f}$ results from centering f on its spatial average.

$$\breve{f}(\omega) = f(\omega) - \frac{1}{4\pi}\int_{S^2} f(\omega)\mathrm{d}\omega$$

Considering that these functions are band-limited of bandwidth B, we can decompose $\breve{f}$ and $\breve{g}$ as finite discrete sums of spherical harmonics $Y_l^m(\omega)$ (degree 0 representing mean value, it is discarded from the sum).

$$\breve{f}(\omega) = \sum_{l=1}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m Y_l^m(\omega)$$

$$\breve{g}(\omega) = \sum_{l=1}^{B-1} \sum_{|m|\leq l} \hat{g}_l^m Y_l^m(\omega)$$

It is known that a rotation of a given spherical harmonic expresses as a linear combination of harmonics of the same degree:

$$\Lambda_R(Y_l^m)(\omega) = \sum_{|k|\leq l} Y_l^k(\omega) D_{km}^{(l)}(R)$$

where $D_{km}^{(l)}$ are the Wigner-D functions. It follows, concerning the numerator of $C_R(f,g)$ that

$$\int_{S^2} \breve{f}(\omega)\Lambda_R(\breve{g})(\omega)\mathrm{d}\omega = \sum_l \sum_{|m|\leq l} \sum_{|m'|\leq l} \hat{f}_l^{-m}\overline{\hat{g}_l^{-m'}}(-1)^{m-m'} D_{mm'}^{(l)}(R)$$

This happens to be an inverse $SO(3)$-Fourier transform, whose computing is easily achieved using a public package [48]. Evaluation is performed all at once for a discrete sampling of rotations R over $SO(3)$. The sampling grid is expressed in terms of Euler angles α, β, γ, and the grid step for each angle is inversely proportional to the band B. Normalized cross-correlation denominator expresses simply:

$$\int_{S^2} |\breve{f}(\omega)|^2 \mathrm{d}\omega \ = \sum_{l=1}^{B-1} \sum_{|m|\leq l} \hat{f}_l^m \ ^2$$

7.3. Gradient descent over $SO(3)$

Gradient descent has been formalized on the $SO(3)$ rotation group by Stein *et al.*[49] and Chirikjian *et al.* [36]. Here are the details of the algorithm. We seek to minimize a function $f(\Gamma)$. The gradient descent procedure aims at finding the direction which reduces the value. In $SO(3)$, infinitesimal motions are captured by the basis elements of the Lie algebra of $SO(3)$:

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} X_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} X_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Directional derivatives are defined as:

$$X_i^R f(\Gamma) = \frac{d}{dt}(f(\Gamma \circ e^{tX_i}))|_{t=0}$$

Approximate gradients can be used : replace the right derivatives $X_i^R f(\Gamma), i = (1,2,3)$ with the finite difference approximations x_i :

$$x_i = \frac{f(\Gamma \circ e^{tX_i}) - f(\Gamma \circ e^{-tX_i})}{2t}$$

The collection of these directional derivatives is a gradient vector pointing in the direction of the steepest ascent. Therefore, it may be used to update the current element as :

$$\begin{aligned} \Gamma &\leftarrow \Gamma \circ exp\{-\epsilon \sum_{i=1}^{3} X_i [X_i^R f(\Gamma)]\} \\ &\approx \Gamma \circ exp \left\{ -\epsilon \begin{pmatrix} 0 & -x_1 & x_2 \\ x_1 & 0 & -x_3 \\ -x_2 & x_3 & 0 \end{pmatrix} \right\} \end{aligned}$$

where ϵ is a small step size. This process will lead to the nearest local minimum of the function. Ideally it must be restarted from several different values of Γ. Using this algorithm enables to converge precisely to the solution without computing $f(\Gamma)$ over a whole fine grid over $SO(3)$. .

8. REFERENCES

[1] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening I : Stimulus synthesis. *J. Acoust. Soc. Am.*, 85:868–878, 1989.

[2] F. L. Wightman and D. J. Kistler. Headphone simulation of free field-listening II : Psychophysical validation. *J. Acoust. Soc. Am.*, 85(2):868–878, 1989.

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94:111–123, 1993.

[4] K. Hartung, J. Braasch, and S.J. Sterbing. Comparison of different interpolation methods for the interpolation of head-related transfer functions. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction. Rovaniemi, Finland*, 1999.

[5] G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comp.*, 2:5–16, 1981.

[6] P. Minaar, J. Plogsties, and Christensen Flemming. Directional resolution of Head-related Transfer Functions required in binaural synthesis. *J. Audio Eng. Soc.*, 53(10):919–929, 2005.

[7] S. Carlile, C. Jin, and V. van Raad. Continuous virtual auditory space using HRTF interpolation : Acoustic and psychophysical errors. In *International Symposium on Multimedia Information Processing*, Sydney, NSW, Australia, 2000.

[8] J. Chen, B.D. Van Veen, and K.E Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *J. Acous. Soc. Am.*, 97(1):439–452, 1995.

[9] R. Martin and K. McAnally. Interpolation of head-related transfer functions. Technical Report DSTO-RR-0323, Australian Government - Department of Defence, February 2007.

[10] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli. Sound field analysis along a circle and its applications to HRTFs interpolation. *J. Audio Eng. Soc.*, 56(3):156–175, 2008.

[11] V. Lemaire, F. Clérot, S. Busson, R. Nicol, and V. Choqueuse. Individualized HRTFs from few measurements: a statistical learning approach. In *Proceedings of International Joint Conference on Neural Networks*, 2005.

[12] C.L. Cheng and G.H. Wakefield. A tool for volumetric visualization and sonification of Head Related Transfer Functions(HRTFs). In *International Conference on Auditory Display 2000, Atlanta, GA*, 2000.

[13] http://www.isr.umd.edu/labs/nsl/.

[14] http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm.

[15] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36:171–218, 1992.

[16] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647, 1992.

[17] M. Morimoto and Aokata H. Localization cues of sound sources in the upper hemisphere. *J. Acous. Soc. Jpn.*, 5(3):165–173, 1984.

[18] W.M. Hartmann and A. Wittenberg. On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688, 1996.

[19] P. J. Bloom. Creating source elevation illusions by spectral manipulations. *J. Audio Eng. Soc.*, 25(9):561–565, 1977.

[20] R.A. Butler and A.D. Musicant. Binaural localization: Influence of stimulus frequency and the linkage to covert peak areas. *Hear. Res.*, 67:220–229, 1993.

[21] A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396:747–749, 1998.

[22] P. M. Hofman and A. J. Van Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biol. Cybern.*, 86:305–316, 2002.

[23] E. H. A. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *J. Acous. Soc. Am.*, 112(4):1583–1596, 2002.

[24] E.A.G. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *J. Acous. Soc. Am.*, 44(1):240–249, 1968.

[25] M. Morimoto. The contribution of two ears to the perception of vertical angle in sagittal planes. *J. Acous. Soc. Am.*, 109:1596–1603, 2001.

[26] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Appoximating the head-related transfer function using simple geometric models of the head and torso. *J. Acous. Soc. Am.*, 112(5):2053–2064, 2002.

[27] E.A. MacPherson. Binaural weighting of monaural spectral cues for sound localization. *J. Acoust. Soc. Am.*, 121:3677–3688, 2007.

[28] Y. Kahana and P. A. Nelson. Spatial acousic mode shapes of the human pinna. In *Audio. Eng. Soc. 109th Convention*, Los Angeles, California, USA, 2000.

[29] H. Møller, M.F. Sørensen, D. Hammerhøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J. Audio Eng. Soc.*, 43(5):300–321, 1995.

[30] J.C. Middlebrooks, J.C. Makous, and D.M. Green. Directional sensivity of sound pressure levels in the human ear canal. *J. Acous. Soc. Am.*, 86(1):89–108, 1989.

[31] J. A. Burlingame and R. A. Butler. The effects of attenuation of frequency segments on binaural localization of sound. *Percept. Psychophysics*, 60(8):1374–1383, 1998.

[32] J.C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acous. Soc. Am.*, 106(3):1480–1492, 1999.

[33] K. Maki and S. Furukawa. Reducing individual differences in the external-ear transfer functions of the mongolian gerbil. *J. Acous. Soc. Am.*, 118(4):2392–2404, 2005.

[34] P. Guillon, T. Guignard, and R. Nicol. Head-Related Transfer Function customization by frequency scaling and rotation shift based on a new morphological matching method. In *125th AES Convention*, October 2008.

[35] K. Sorgi, L. Daniilidis. Normalized cross-correlation for spherical images. In T. Pajdla and J. Matas, editors, *ECCV 2004, LNCS 3022*, page 603616, 2004.

[36] G S. Chirikjian, P.T. Kim, J.-Y. Koo, and C.H. Lee. Rotational matching problems. *International Journal of Computational Intellignece and Applications*, 4(4):401–416, 2004.

[37] P. J. Kostelec and D. N. Rockmore. FFTs on the Rotation Group. *Santa Fe Institute Working Papers Series*, 2003.

[38] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[39] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[40] http://recherche.ircam.fr/equipes/salles/listen/.

[41] A.W. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98(5):2542–2553, 1995.

[42] RH Hardin, NJA Sloane, and WD Smith. Spherical coverings. *Electronic: http://www.research. att. com/njas/coverings/index. html, May*, 1997.

[43] E. H. A. Langendijk and A. W. Bronkhorst. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *J. Acoust. Soc. Am.*, 107(1):528–537, 2000.

[44] Craig T. Jin. *Spectral Analysis and Resolving Spatial Ambiguities in Human Sound Localization*. PhD Thesis, University of Sydney, 2001.

[45] R.A. Butler. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter Spatial referents of stimulus frequencies: their role in sound localization, pages 99–115. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

[46] A. W. Mills. On the minimum audible angle. *J. Acoust. Soc. Am.*, 30:237–248, 1958.

[47] S. R. Oldfield and S. P. A. Parker. Acuity of sound localisation: a topography of auditory space. I. normal hearing conditions. *Perception*, 13:581–600, 1984.

[48] P.J. Kostelec and D.N. Rockmore. SOFT: SO (3) Fourier Transforms. *http://www.cs.dartmouth.edu/ geelong/soft/.*

[49] D. Stein, E.R. Scheinerman, and G.S. Chirikjian. Mathematical models of binary spherical-motion encoders. *Mechatronics, IEEE/ASME Transactions on*, 8(2):234–244, 2003.

Audio Engineering Society

Convention Paper 7550

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

Head-Related Transfer Function customization by frequency scaling and rotation shift based on a new morphological matching method

Pierre Guillon[1,2], Thomas Guignard [2], Rozenn Nicol [2]

[1] *Laboratoire d'Acoustique de l'Université du Maine, Le Mans, France*

[2] *Orange Labs, Lannion, France*

Correspondence should be addressed to Pierre Guillon (pierre.guillon@orange-ftgroup.com)

ABSTRACT

Head-Related Transfer Functions (HRTFs) individualization is required to achieve high quality Virtual Auditory Spaces. An alternative to acoustic measurements is the customization of non-individual HRTFs. To transform HRTF data, we propose a combination of frequency scaling and rotation shift, whose parameters are predicted by a new morphological matching method. For six subjects, mesh models of head and pinnae are acquired, and differences in size and orientation of the pinnae are evaluated with a modified Iterative Closest Point (ICP) algorithm. Optimal HRTF transformations are computed in parallel. A relatively good correlation between morphological and transformation parameters is found and allows to predict the customization parameters from the registration of pinna shapes. The resulting model achieves better customization than frequency scaling only, which shows that adding the rotation degree of freedom improves HRTF individualization.

1. INTRODUCTION

High fidelity Virtual Auditory Spaces (VAS) can be achieved over headphones via binaural synthesis techniques. VAS filters are efficiently designed using HRTFs (Head-Related Transfer Functions) which embed all the acoustical phenomena occurring on the path between a source at a given position in space and the listener's eardrums. As these phenomena highly depend on the listener's morphology, HRTFs contain idiosyncratic features. Therefore the

use of individual HRTFs is preferred. Since acoustic measurements are long, uncomfortable for subjects, technically difficult and expensive, HRTF individualization remains an open issue.

Morphology-based methods have been proposed. BEM can solve the whole acoustic problem numerically [1, 2, 3]: having acquired a 3D mesh model of the listener's head and torso, HRTFs are generated for any direction in space. BEM efficiency is well established, but its computational cost remains high. Easier techniques follow a different philosophy: their aim is to customize HRTFs, i.e. to select and optionally transform non-individual HRTFs from a database, so that they fit with a new listener, only knowing some morphological parameters. Zotkin *et al.*[4, 5] proposed to simply select non-individual HRTFs using a set of 8 dimensions measured on a photography of each listener's pinna. Middlebrooks [6, 7] developed a technique based on the fact that inter-subject differences in size of pinnae lead to a shift of acoustical phenomena along the frequency axis. The author's technique, called frequency scaling, consists in reducing inter-subject differences by an homothetic transformation of HRTF spectral profiles. The optimal scaling factor, acting globally for all directions, can be predicted by morphology. Maki *et al.*[8] further proved on the Mongolian gerbil that a difference in spatial orientation of the pinnae could be compensated by a global rotation shift of the HRTF coordinate system. Furthermore, it was proved that the combination of these two degrees of freedom enabled to achieve better results than taking rotations only or frequency scaling only. Our objective is to investigate the extension of this method to human subjects, as suggested by the authors. The essence of the technique is its ability to predict data transformations to be performed on a set of HRTFs from a database, knowing the morphology differences between its owner and the new listener.

From this previous work, the present paper proposes a new model of HRTF individualization. Individualized HRTFs are obtained by transforming a set of measured HRTFs available from a database. The novelty relies on the transformation process which combines both coordinate rotation and frequency scaling, as described in [8]. What's more the HRTF transformation is controlled by the parameters of the morphological matching of the ear pinna between the new individual and the HRTF's owner.

Building this individualization model raises two main questions. First, concerning the morphological aspect, it should be found an appropriate method to compare and match ear pinnae of two different subjects, in order to jointly get differences in scale and orientation. Second, in the same way but considering now HRTF data, appropriate tools have to be defined to compare and match HRTF data, in order to find the optimal frequency scaling factor and global rotation shift to be performed on a set of HRTF data to minimize its differences with those of another subject.

The first question is solved by designing a registration process of ear pinnae. A key issue is that the registration matches shapes beyond their possible differences in size and spatial orientation, i.e. it includes scaling and rotation in order to focus on the intrinsic shape and relative volumes of pinnae cavities, playing a crucial role in spectral cues genesis. As proposed in the study of ear canal and concha shapes, we choose the Iterative Closest Point (ICP) algorithm [9], to which we further add a scaling degree of freedom. The registration parameters sum up the differences in orientation and size between the two pinnae shapes. The algorithm input is a 3D mesh model of the listeners' morphology consisting of head and pinnae surfaces, which are acquired with a laser scanner.

As for the second question, HRTF data matching is achieved by transformation combining frequency scaling factor and global rotation by using a gradient descent algorithm. More precisely, for a given series of frequency scaling factors, a gradient descent algorithm is applied over the $SO(3)$ group of rotations. The optimal parameters are the couple of coordinate rotation factor and frequency scaling factor, which minimizes the Inter-Subject Spectral Difference, a HRTF similarity measure proposed by Middlebrooks [6].

The methods used for the morphological and HRTF matching are described in Section 2. Then Section 3 presents results obtained for the private database of Orange Labs. It is examined to what extent the morphological and HRTF data optimal transformations correlate with each other. It is also shown how

to derive the optimal parameters of HRTF transformation from the registration parameters of pinnae morphology. The individualization performances of the proposed model is assessed.

2. METHODS

2.1. Pinnae registration

In order to quantify differences in size and orientation between the pinnae of two subjects, we propose to use a registration algorithm, acting on 3D mesh models of the pinnae surfaces. We choose Iterative Closest Point (ICP), as it was proved efficient to compare earcanals and conchae shapes [10]. The aim of ICP is to align two shapes defined as point clouds, by iteratively establish a pointwise correspondence between the two sets of points, and find the global geometric transformation that reduces to minimum the total distance between these associated points. The geometric transformation is usually defined as combination of a rotation and a translation. This would be sufficient to compensate for orientation differences, but not for size. We therefore add a scaling degree of freedom, as formalized by Umeyama *et al.*[11] and Du *et al.*[12].

Given two point sets describing the pinnae surfaces in $\mathbb{R}^3$, let $M = \{m_i\}_{i=1}^{N_m}$ be the pinna shape of a new subject, and let $P = \{p_i\}_{i=1}^{N_p}$ be the pinna shape of a subject of a database, whose HRTFs are known. To register between these two point sets is to find a similarity transformation, with which P is in the best alignment with M. Mathematically, the problem is to find the rotation $\mathbf{R}$, the translation $\vec{t}$ and the scale s solution of the following LS problem:

$$\min_{s,\mathbf{R},\vec{t}} \left(\sum_{i=1}^{N_p} \|(s\mathbf{R}\vec{p}_i + \vec{t}) - \vec{m}_j\|_2^2\right)$$

$$s.t.\ \mathbf{R}^t\mathbf{R} = \mathbf{I}_3, \det(\mathbf{R}) = 1$$

In each iteration, two steps are included.

Step 1, build up the set of correspondences c by the current similarity transformation $(s_k, \mathbf{R}_k, \vec{t}_k)$:

$$c(i) = \underset{j\in(1,2,\ldots,N_m)}{\arg\min} \left(\|(s_k\mathbf{R}_k\vec{p}_i + \vec{t}_k) - \vec{m}_j\|_2^2\right)$$

Step 2, compute the new similarity transformation $(s^*, \mathbf{R}^*, \vec{t}^*)$:

$$(s^*, \mathbf{R}^*, \vec{t}^*) = \underset{(s,\mathbf{R},\vec{t})}{\arg\min} \left(\sum_{i=1}^{N_p} \|s\mathbf{R}(s_k\mathbf{R}_k\vec{p}_i + \vec{t}_k) + \vec{t} - \vec{m}_{c(i)}\|_2^2\right)$$

Update $s_{k+1}, \mathbf{R}_{k+1}, \vec{t}_{k+1}$

$$s_{k+1} = s^*s_k, \mathbf{R}_{k+1} = \mathbf{R}^*\mathbf{R}_k, \vec{t}_{k+1} = s^*\mathbf{R}^*\vec{t}_k + \vec{t}^*$$

Details leading to the expressions of s^*, $\mathbf{R}^*$ and $\vec{t}^*$ can be found in [12]. Here are the results:

Let $\overrightarrow{p_i} \triangleq s_k\,\mathbf{R}_k\,\vec{p}_i + \vec{t}_k$, $\vec{q}_i \triangleq \overrightarrow{p_i} - \frac{1}{Np}\sum_{j=1}^{N_p}\overrightarrow{p_j}$ and $\vec{n}_i \triangleq \vec{m}_{c(i)} - \frac{1}{Np}\sum_{j=1}^{N_p}\vec{m}_{c(j)}$.

Calculate 3×3 matrix $\mathbf{H}$ and its SVD:

$$\mathbf{H} \triangleq \frac{1}{Np}\sum_{i=1}^{N_p}\vec{q}_i\vec{n}_i^T, \mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$$

Finally, the transformation parameters at each iteration express as:

$$\begin{aligned}
\mathbf{R}^* &= \mathbf{V}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & det(\mathbf{V}\mathbf{U}^T) \end{pmatrix}\mathbf{U}^T \\
s^* &= \sum_{i=1}^{N_p}\vec{n}_i^T\mathbf{R}^*\vec{q}_i / \sum_{i=1}^{N_p}\vec{q}_i^T\vec{q}_i \\
\vec{t}^* &= \frac{1}{Np}\sum_{i=1}^{N_p}\vec{m}_{c(i)} - \frac{1}{Np}\sum_{i=1}^{N_p}s^*\mathbf{R}^*\vec{p}_i'
\end{aligned}$$

Iteration is stopped when distance between shapes no longer evolves. Final registration parameters sum up the differences in orientation and size between the two pinnae shapes. For each facet of M, facet error is defined as the average distance between its vertices and P. The total registration error is computed as follows : facet errors are weighted by their surface and summed, and the result is divided by the total mesh surface. All facets lying on the edges of the mesh model are discarded during this operation.

2.2. Reducing differences between HRTF datasets

It was proved that differences between HRTF spectral profiles of two subjects can be reduced by an

homothetic transformation along the frequency axis [6]. Maki *et al.* further introduced the rotation of the coordinate system as a complementary degree of freedom to customize HRTFs, and showed the efficiency of these two combined transformations to reduce inter-subject differences between HRTFs of Mongolian gerbils. We aim at applying the same transformations to customize non-individual HRTFs for a new human subject. We adopt here the denomination proposed in [6, 8].

From the HRTFs, we extract the Directional Transfer Functions (DTFs)[13], which are the directional components of HRTFs. As we only work here with log-magnitude spectra, we calculate DTFs at each direction by substracting from the HRTFs the spectrum averaged for all measured directions. Furthermore, DTFs pass through a bank of rectangular bandpass filters, whose bandwidths are the same as the critical bandwidths, in order to take into account the limited frequency resolution of the auditory system.

Inter-Subject Spectral Difference (ISSD) is used as a global indicator of difference between the DTFs of a pair of subjects. ISSD is calculated as follows. For each direction, the dB amplitude of the DTF of one subject is substracted frequency-by-frequency from those of the other subject. The variance of the difference is calculated for the frequency range 3.5kHz - 13kHz., which is the frequency range where perceptually relevant direction specific colorations are generated by ear pinnae. This variance, expressed in dB^2, is called direction-specific ISSD. ISSD is finally calculated as the average of all direction-specific ISSDs over the whole available directions.

The Optimal Scale Factor (OSF)[6] is defined as the frequency scaling factor common across all directions, and applied on one set of DTFs, required to minimize the ISSD for a given pair of subjects. As we here work on a linear frequency scale, OSF is expressed as a linear homothetic factor. Similarly, the Optimal Coordinate Rotation (OCR)[8] is the amount of rotation of the coordinate system for one of the ears needed to minimize ISSD.

We aim at simultaneously optimizing rotation Γ and frequency scaling α. First, DTF data are considered independently for each frequency bin: the resulting directivity surfaces are referred to as Spatial Frequency Response Surfaces (SFRSs) [14]. SFRSs are decomposed in Spherical Harmonics, in order to get a functional representation [15]. So as to regularize the obtained decomposition, the missing data at lower directions are interpolated using STPS. For a given frequency scale factor, a gradient descent algorithm enables to converge to the coordinate rotation minimizing ISSD. This algorithm is performed for a series of frequency scaling factors. Finally, OSF and OSR are respectively the combined scaling factor α and rotation of the coordinate system Γ that globally minimize ISSD.

For each combined rotation Γ and frequency scaling factor α, ISSD is calculated as follows. For computational time considerations, instead of rotating each SFRS by Γ, the spatial sampling grid onto which HRTFs have been measured is rotated by the inverse rotation Γ^{-1}, and SFRSs are then evaluated on the resulting grid. SFRSs are reorganized along the frequency axis according to an homothety of factor α. So as to calculate ISSD frequency-by-frequency and direction-by-direction, spline interpolation is carried out over DTFs in frequency.

Gradient descent was formalized on the $SO(3)$ rotation group by Stein *et al.*[16] and Chirikjian *et al.* [17]. Here are the details of the algorithm. We seek to minimize a function $f_\alpha(\Gamma)$. In our case, f_α is the continuous function describing ISSD. The gradient descent procedure aims at finding the direction which reduces the value. In $SO(3)$, infinitesimal motions are captured by the basis elements of the Lie algebra of $SO(3)$:

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} X_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} X_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Directional derivatives are defined as:

$$X_i^R f_\alpha(\Gamma) = \frac{d}{dt}(f_\alpha(\Gamma \circ e^{tX_i}))|_{t=0}$$

Approximate gradients can be used. The right derivatives $X_i^R f_\alpha(\Gamma), i = (1, 2, 3)$ are replaced with the finite difference approximations x_i :

$$x_i = \frac{f_\alpha(\Gamma \circ e^{tX_i}) - f_\alpha(\Gamma \circ e^{-tX_i})}{2t}$$

The collection of these directional derivatives is a gradient vector pointing in the direction of the steepest ascent. Therefore, it may be used to update the

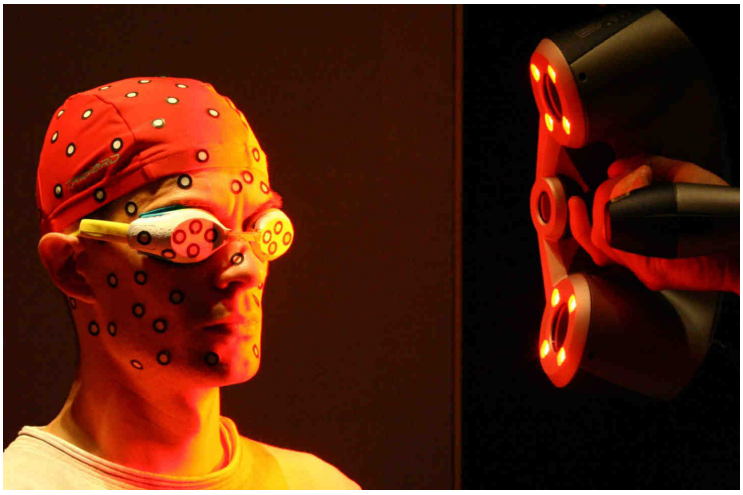


Fig. 1: Laser scanning procedure with Handyscan^{TM}

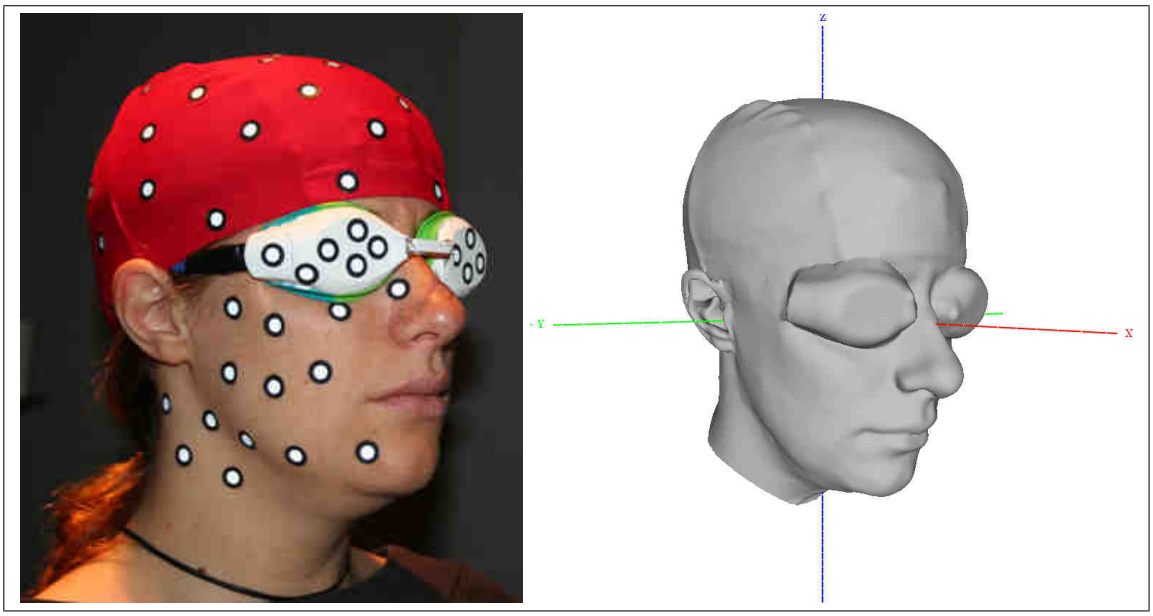
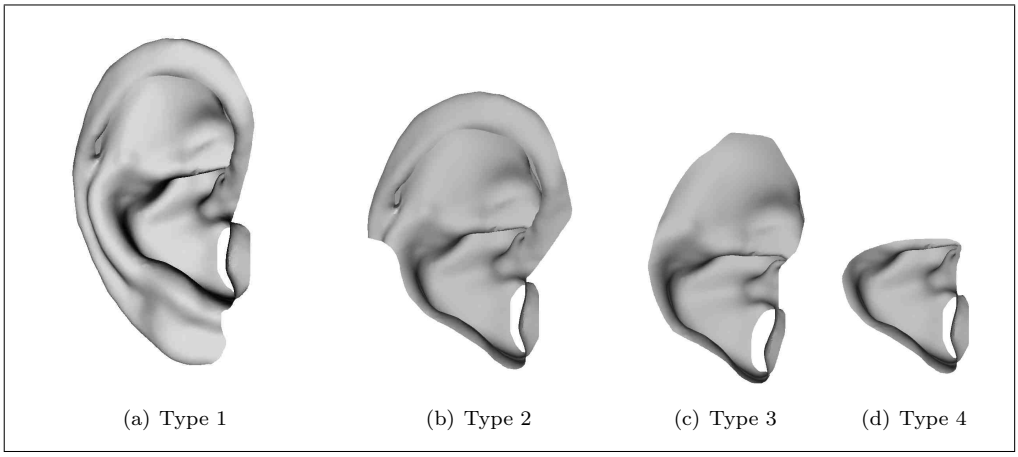


Fig. 2: 3D mesh obtained from the scanning of a subject's head and pinnae



(a) Type 1 (b) Type 2 (c) Type 3 (d) Type 4

Fig. 3: These four types of mesh are tested for morphological registration. Registration between two pinnae are performed with equivalent type.

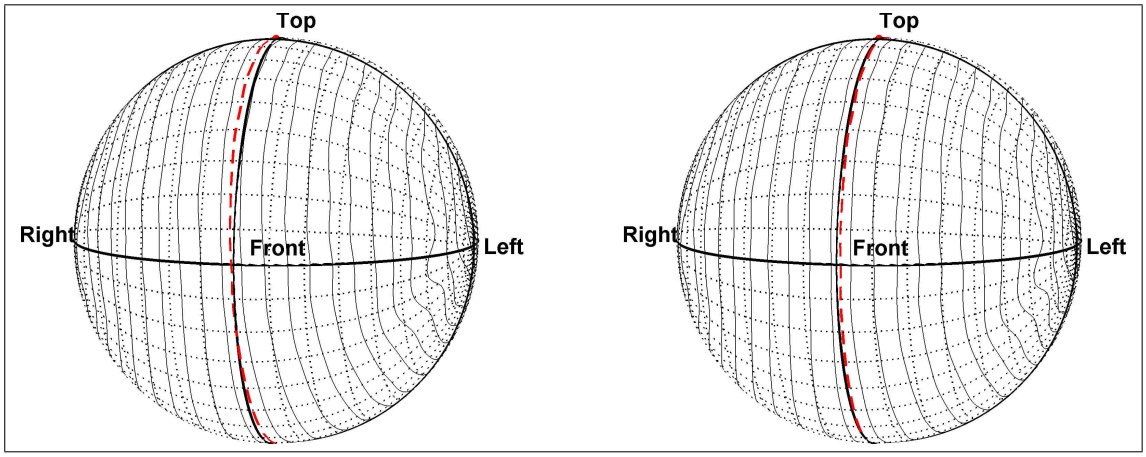


Fig. 4: Iso-ITD contours plotted over the sphere. Left: before misalignment correction. Right: after correction. The iso-0μs contour (dashed red) is closer to the median plane after correction.

current element as :

$$\Gamma \leftarrow \Gamma \circ exp\{-\epsilon \sum_{i=1}^{3} X_i[X_i^R f_\alpha(\Gamma)]\}$$
$$\approx \Gamma \circ exp\left\{-\epsilon \begin{pmatrix} 0 & -x_1 & x_2 \\ x_1 & 0 & -x_3 \\ -x_2 & x_3 & 0 \end{pmatrix}\right\}$$

where ϵ is a small step size. This process leads to the nearest local minimum of the function. Therefore it must be restarted from several initial values of Γ. Using this algorithm enables to converge precisely to the solution without computing ISSDs over a whole fine grid over $SO(3)$.

2.3. Experimental setup

2.3.1. 3D mesh acquisition

Head and pinnae surfaces of our 6 subjects are acquired with a Handyscan^{TM} laser scanner (cf. Fig. 1). As hair cannot be acquired by the scanner, subjects wear a swimming hat. They also wear darkened swimming glasses to protect their eyes from the laser beam. In order to get the shadowed parts of the pinnae, pinna molds are performed with impression polymer. Their surfaces are also acquired, and associated to global head scans. Head frame of reference is created as follows. First, interaural axis is defined as the line crossing each tragus summit. The center of the head is the middle of the segment delimited by these two points. A point on the nose finally defines the horizontal axis from the center of the head. This point is determined with a photography of subjects head profiles, when keeping their sight horizontal. As a result, we know the exact position and orientation of pinnae surfaces in the head frame, the one corresponding to HRTF measurements. So as to compare a maximum of pinnae meshes, all left pinnae models are symmetrized relative to the median plane. Therefore, with 12 ears and corresponding HRTFs, a total of 60 pairs were studied. Four different pinnae mesh types are tested: from the first to the last, pinnae elements are progressively cut out, ending with the concha surface only (cf. Fig. 3).

2.3.2. HRTF data

HRTFs used in our experiments belong to Orange labs private database [18]. See [19] for details about the measurements conditions. During the measurement procedure, position and orientation of the subject's head were permanently tracked, so that a correction of the position of the loudspeaker could automatically compensate for small movements from the initial pose. Nevertheless, this initial pose could deviate from the optimal alignment with the frame of the measurement system. Such a global rotation shift has to be compensated before any comparison between HRTF data. As proposed by Maki *et al.* [8], misalignment is evaluated by analyzing Interautal Time Difference (ITD). It is assumed that the function on the sphere describing ITD is coarsely antisymmetric relative to the median plane. Therefore finding the misalignment rotation is equivalent to finding the rotation shift of the coordinate system that minimizes the symmetric part (even part) of the ITD function relative to the vertical plane including the midline. ITD is first evaluated at each measured direction by cross-correlation between left and right Head Related Impulse Responses (HRIRs). Values for unmeasured directions on the lower hemispherical cap are obtained by interpolation (Spherical Thin Plate Spline (STPS)[20]). A decomposition of ITD onto Spherical Harmonics (SH) is then performed, by pseudo-inverse procedure. Finally, a gradient descent algorithm is used to find the rotation minimizing the total energy of the even part of ITD function, captured by the real part of the coefficients of the SH decomposition. This misalignment is evaluated for each of the subjects. Roll, pitch and yaw angles of these rotations were found respectively in the ranges $[-3.75°; 2.2°]$, $[-0.17°; 0.14°]$, and $[-1.86°; 1.43°]$, showing that misalignment is very small. Nevertheless, these rotations are carried out on each HRTF dataset before evaluating OCR and OSF. Fig. 4 illustrates ITD before and after misalignment correction. All HRTFs corresponding to left ears are symmetrized relative to the median plane, as done on pinnae meshes.

3. RESULTS AND DISCUSSION

3.1. Pinnae models registration

Visual inspection of ICP results shows that pinnae models are successfully registered for all mesh types. It seems that concha surfaces greatly influence the result of the orientation and scaling. Fig. 5 gives an

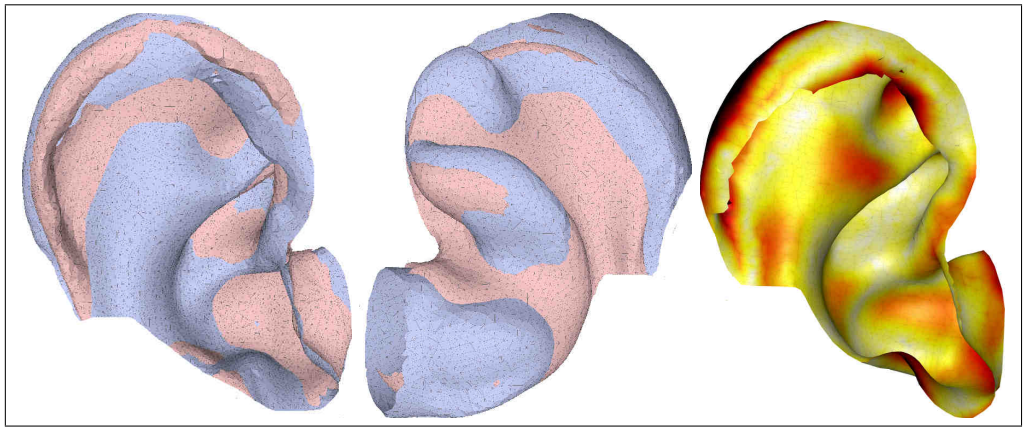


Fig. 5: Left and Center : registration result of two pinnae meshes of type 2 (one in pink and the other in blue). Left: front view. Center: back view. Right: color-coded facet registration error (low in bright zones, high in dark zones)

example of a successfull registration between type 2 pinna meshes. Shapes interleave properly, denoting good matching between pinnae elements.

3.2. ISSD reduction

In order to prove the usefullness of the rotation degree of freedom, we optimize ISSD reduction between all pairs for three conditions : classical frequency scaling without rotation shift, rotation shift without frequency scaling, and frequency scaling with rotation shift. Resulting ISSDs are plotted in Fig. 6, in addition to ISSD obtained without processing. The solution combining rotation and frequency scaling leads for all pairs to the lowest ISSD, especially when the ISSD prior to any transformation is high. Fig. 7 illustrates the corresponding DTF transformation for one particular pair and one given direction. It can be observed how valleys and peaks are well aligned with the proposed solution, which explains lower ISSD values.

3.3. Predicting HRTF transformations with morphological parameters

Both morphological matching and ISSD optimal reduction lead to the definition of a scaling factor (rsp. s and OSF $= \alpha$), and a rotation matrix (rsp. $\mathbf{R}$ and OCR $= \Gamma$). Each rotation matrix can be decomposed into three rotations around X, Y, and Z axis, respectly with angles roll θ, pitch ψ and yaw ϕ. We therefore denote as $s_{sig} = \alpha$, θ_{sig}, ψ_{sig}, ϕ_{sig}, and $s_{morph} = s$, θ_{morph}, ψ_{morph}, ϕ_{morph}, the set of parameters describing respectively optimal HRTF (signal) transformations and morphological registration. See Fig. 8 for angle definitions.

$$\begin{aligned} OCR &= \Gamma = \{\theta_{sig}, \psi_{sig}, \phi_{sig}\}_{XYZ} \\ \mathbf{R} &= \{\theta_{morph}, \psi_{morph}, \phi_{morph}\}_{XYZ} \end{aligned}$$

For practical reasons, logarithms of the scaling parameters s_{sig} and s_{morph} will be considered, since unitary frequency scaling corresponds to null transformation.

First, we compute for each pinna mesh type all one to one correlation coefficients between signal and morphological parameters. Type 2 (cf. Fig. 3), only discarding ear lobe, offers the greatest correlations, and as a result, was retained for the following experiments. For each signal parameter, we only retain the corresponding morphological parameter being the most correlated. Fig. 9 depicts the results. Associated parameters are plotted for all pinnae pairs: s_{sig} with s_{morph}, θ_{sig} with θ_{morph}, ϕ_{sig} with ϕ_{morph}, and more surprisingly, ψ_{sig} with θ_{morph}. We could expect a one-to-one correspondence between parameters, as everything is done to match head scan frame with HRTF coordinate system. What is the most striking is that ψ_{morph} (i.e. the pitch angle) apparently does not correlate with

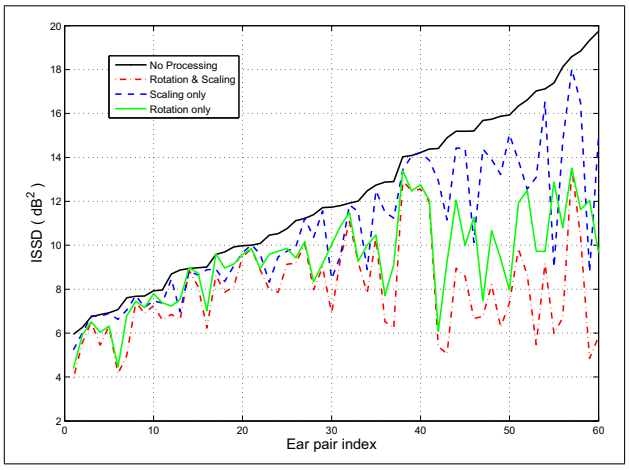


Fig. 6: ISSD after optimal transformations for every ear pair: frequency scaling and rotation (dash-dotted red), frequency scaling only (dashed blue), rotation only (cont. green), and initial ISSD (dotted black). Ear pairs are sorted by ascending initial ISSD

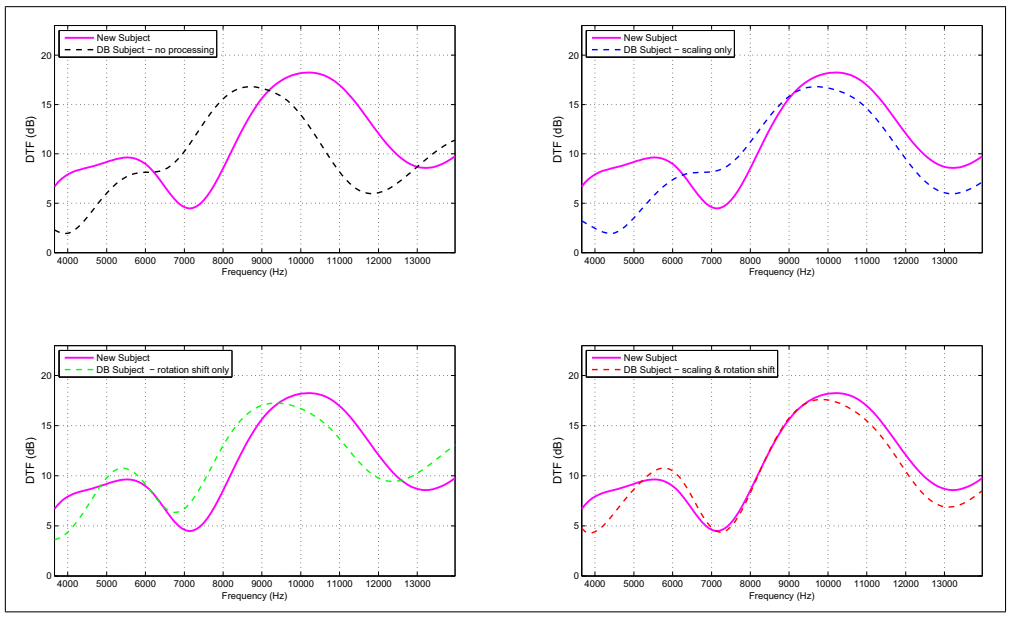


Fig. 7: Optimal transformations of DTFs for one given pair of ears, and one direction. The target DTF (cont. magenta) represents the DTF of a new subject to be approximated with the DTF from a database (dashed black, upper left corner). Three methods are compared to this unprocessed solution: scaling only (upper right corner), rotation only (lower left corner) and rotation & scaling (lower right corner)

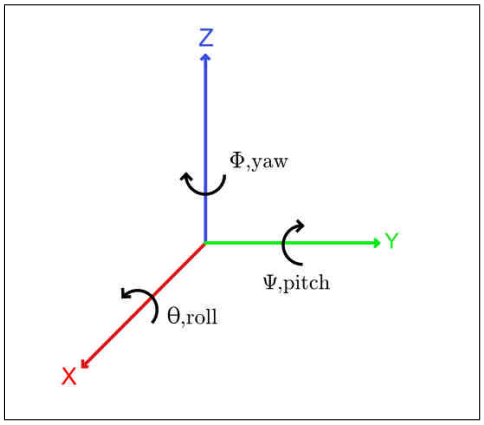


Fig. 8: Roll, pitch and yaw definition

any HRTF transformation. Indeed it is intuitively the strongest degree of freedom of pinna orientation. However it is difficult to define the null pitch reference. In our experiment, the null pitch is referred to as the position of the subject's head keeping his(her) sight horizontally. But we have to be aware that there is a possible bias of the horizontal reference between HRTF measurement and 3D mesh acquisition sessions. This deviation cannot be corrected with ITD analysis. At this point correlation coefficients values are only high between s_{sig} and s_{morph} (0.79), which is similar to what Larcher [21] obtained between concha height ratios and OSF for scaling-only technique. Moderate correlations are found for other parameter pairs : 0.64 for $(\theta_{sig}, \theta_{morph})$, 0.66 for $(\psi_{sig}, \theta_{morph})$, and 0.64 for $(\phi_{sig}, \phi_{morph})$. Though point clouds spread for angle parameters, a linear relationship may be found, if outliers are discarded. As our statistical analysis relies on 60 data points only, outliers may impact strongly on the results. Therefore outliers are discarded with a RANSAC [22] procedure, and model is then fitted with the inliers only. It can be expected that no transformation is needed to adapt HRTFs for pinnae corresponding each other with null $log_{10}(s_{morph})$, θ_{morph}, ψ_{morph}, and ϕ_{morph}. Consequently, a proper linear regression between every pair of parameters should have zero intercept. The results of both zero and non-zero intercept linear regressions are depicted in Fig. 9. As the regression with non-zero intercept achieves better results, it is clear that there is a bias between signal and morphological angles The cause of this bias is not identified. Despite these observations, we nevertheless fit the model with non-zero intercept linear regression.

Customized HRTFs are obtained by performing the rotations and scaling predicted by the fitted model. We also assess the standard frequency scaling performance, whose scaling factor was linearly predicted by s_{morph}, after a linear fitting between scaling-only OSF and s_{morph}. Performance in terms of ISSD are depicted Fig. 10, and compared to optimal results (i.e. ISSD achieved with OCR and OSF transformation). It is observed that although the model does not reach optimal performance, it is better than frequency scaling only, especially for ear pairs with high initial ISSD. Sometimes an increase of ISSD even occurs for some low initial ISSD ear pairs, and marginally for high initial ISSD ear pairs. Most of these pairs were considered as outliers by the RANSAC procedure, and it seems natural that the model fails in predicting proper signal transformations for them.

An additional parameter is expected to give further prediction about the success of the model for a given ear pair: the registration error, which, indeed, may be considered as representative of the intrinsic error between pinna morphologies once differences in size and orientation are compensated (cf. Section 2.1). It seems reasonable that, if two pinnae are badly registered, the model will fail in reducing differences between corresponding HRTF data. Fig. 11 depicts the ISSD after optimal HRTFs transformation, as a function of the registration error between pinnae shapes. Correlation coefficient is low (0.25), which contradicts our expectations. Therefore, registration error is not a relevant indicator to choose a subject in a database. Nevertheless, contrary to classical frequency scaling, the proposed model seems to perform properly even for high initial ISSD value between HRTF sets, making the initial selection of a promising subject quite useless.

A last experiment is carried out in order to understand the low correlation coefficient between ψ_{sig} and ψ_{morph} (0.32). Our assumption is that it is originated by the pitch alignment uncertainties Δ_ψ between HRTF measurement and 3D model reference frames, for each subject. The effect of such misalignments is easily evaluated by multiplying all registration rotations matrices $\mathbf{R}$ to the left and to

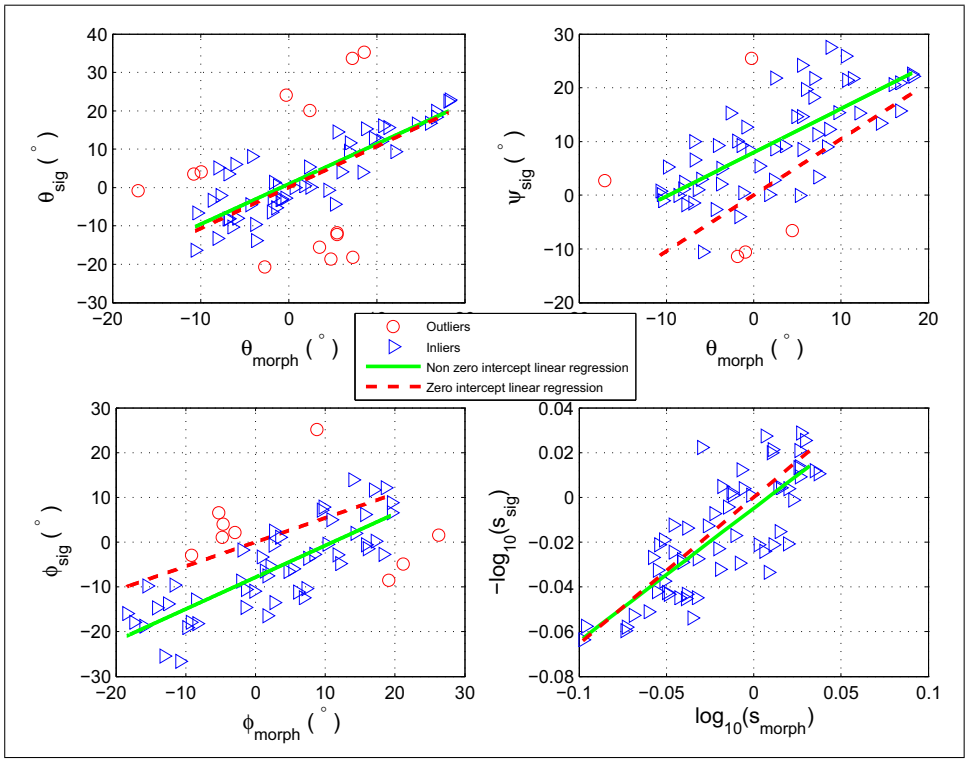


Fig. 9: Associated signal and morphological parameters. Outliers (red circles) are detected prior to fit the model with a RANSAC procedure. Two linear fittings solutions are depicted: zero intercept (dashed red lines) and non zero intercept (continuous green lines)

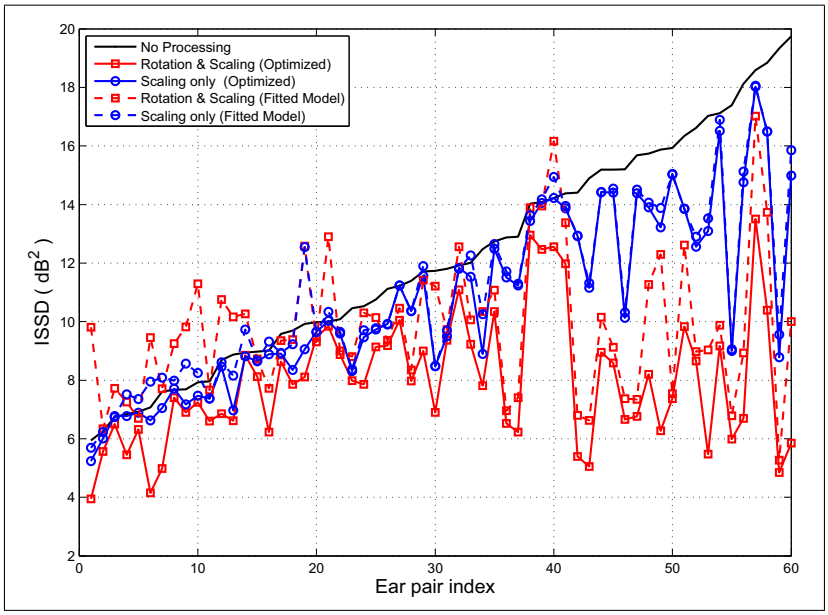


Fig. 10: ISSD after customization for every ear pair: frequency scaling and rotation (square red) and frequency scaling only (blue circles). Results from the models assessments (dashed line) may be compared to optimal values (continuous lines) and to initial ISSD (black dotted line). Ear pairs are sorted by ascending initial ISSD

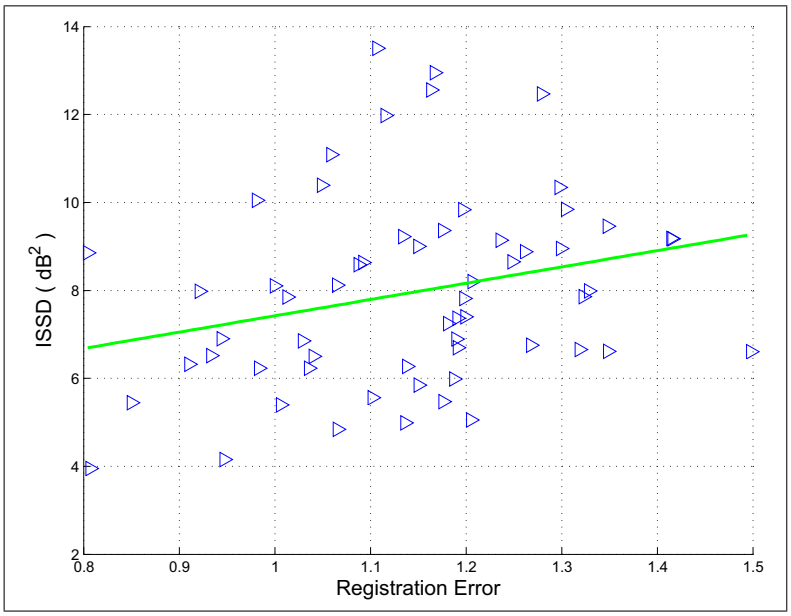


Fig. 11: ISSD obtained after optimal transformation vs. registration error. Low correlation is found.

the right by the rotation matrices simulating hypothetic Δ_ψ values for the concerned subjects. The question is: what values of Δ_ψ can lead to optimal correlations between $(\theta_{morph}, \theta_{sig})$, $(\psi_{morph}, \psi_{sig})$, and $(\phi_{morph}, \phi_{sig})$, while keeping the slope of linear fittings close to 1. Numerical optimization shows that for our 6 subjects, Δ_ψ absolute values lying in the range $[2° - 11°]$ can indeed lead to a better correlation coefficient between ψ_{morph} and ψ_{sig} (0.65), while keeping other correlation coefficients unchanged. These misalignment angle values are plausible, and therefore this experiment points out that careful alignment between signal and morphological reference frames is really crucial for a reliable interpretation of angle correlations.

4. CONCLUSION

This paper proposes a model of HRTF customization based on morphological dissimilarity taking into account differences in both pinna size and orientation. HRTF transformation thus combines frequency scaling and rotation shift. As suggested by Maki *et al.* for the Mongolian gerbil, it is shown for human HRTFs that adding the rotation degree of freedom can improve the frequency scaling customization technique, which proves that part of the differences between HRTFs is explained by differences in orientation of the pinnae. Furthermore, for the customization process, the transformation parameters (i.e. frequency scaling and spatial rotation of HRTFs) are linked with the morphological parameters (i.e. registration parameters of pinna shapes). The relation is based on a linear regression between morphological and transformation parameters, since relatively good correlation is found. The resulting model achieves better customization than frequency scaling only, especially when the dissimilarity between the initial HRTF set and the target one is high. 3D mesh acquistion by a laser scanner is maybe the most tricky step of the overall process, but recent work suggests that a proper 3D mesh of head and pinnae can be obtained from a set of 2D pictures [23]. In addition, results concerning the correspondences between signal and morphological rotation angles may be thoroughly investigated in the light of BEM simulations as proposed in [24, 25].

5. ACKNOWLEDGEMENT

We thank David Leroy, teacher at Lycée Le Dantec, Lannion, for having carried out all 3D meshes acquisition and processing.

6. REFERENCES

[1] B.F.G. Katz. *Measurement and calculation of individual head-related transfer function using a boundary element model including the measurement and effect of skin and hair impedance.* PhD Thesis, Pennsylvannia State University, USA, 1998.

[2] Y. Kahana. *Numerical modelling of the head-related transfer function.* PhD Thesis, University of Southampton, 2000.

[3] M. Otani and S. Ise. Fast calculation system specialized for head-related transfer function based on boundary element method. *J. Acous. Soc. Am.*, 119(5):2589–2598, 2006.

[4] D.N. Zotkin, R. Duraiswami, and L.S. Davis. Customizable auditory displays. In *Proc. Int. Conf. on Auditory Display, ICAD 2002*, Kyoto, Japan, 2002.

[5] D.N. Zotkin, J. Hwang, R. Duraiswami, and L.S. Davis. HRTF personalization using anthropometric measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2003*, New Patlz, NY, USA, 2003.

[6] J.C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acous. Soc. Am.*, 106(3):1480–1492, 1999.

[7] J.C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *J. Acous. Soc. Am.*, 106(3):1493–1510, 1999.

[8] K. Maki and S. Furukawa. Reducing individual differences in the external-ear transfer functions of the mongolian gerbil. *J. Acous. Soc. Am.*, 118(4):2392–2404, 2005.

[9] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE*, 14:239–256, 1992.

[10] S. Darkner, M. Vester-Christensen, R. Larsen, C. Nielsen, and R.R. Paulsen. Automated 3D Rigid Registration of Open 2D Manifolds. In *MICCAI 2006 Workshop From Statistical Atlases to Personalized Models*, 2006.

[11] S. Umeyama. Least-squares estimation of transformation parameters between twopoint patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(4):376–380, 1991.

[12] S. Du, N. Zheng, S. Ying, Q. You, and Y. Wu. An extension of the ICP algorithm considering scale factor. In *IEEE International Conference on Image Processing. ICIP 2007.*, volume 5, 2007.

[13] J.C. Middlebrooks, J.C. Makous, and D.M. Green. Directional sensivity of sound pressure levels in the human ear canal. *J. Acous. Soc. Am.*, 86(1):89–108, 1989.

[14] C.L. Cheng and G.H. Wakefield. A tool for volumetric visualization and sonification of Head Related Transfer Functions(HRTFs). In *International Conference on Auditory Display 2000, Atlanta, GA*, 2000.

[15] M.J. Evans, J.A.S. Angus, and A.I. Tew. Analyzing Head-Related Transfer Function measurements using surface spherical harmonics. *J. Acous. Soc. Am.*, 104(4):2400–2411, 1998.

[16] D. Stein, E.R. Scheinerman, and G.S. Chirikjian. Mathematical models of binary spherical-motion encoders. *Mechatronics, IEEE/ASME Transactions on*, 8(2):234–244, 2003.

[17] G S. Chirikjian, P.T. Kim, J.-Y. Koo, and C.H. Lee. Rotational matching problems. *International Journal of Computational Intellignece and Applications*, 4(4):401–416, 2004.

[18] J.-M. Pernaux. *Spatialisation du son par les techniques binaurales: application aux services de télécommunications.* Thèse de doctorat, Institut National Polytechnique de Grenobles, 2003.

[19] A.W. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98(5):2542–2553, 1995.

[20] G. Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comp.*, 2:5–16, 1981.

[21] V. Larcher. *Techniques de spatialisation du son pour la rálité virtuelle.* Thèse de doctorat, Université de Paris VI, 2001.

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

[23] M. Dellepiane, N. Pietroni, N. Tsingos, M. Asselot, and R. Scopigno. Reconstructing head models from photographs for individualized 3d-audio processing. In *Computer Graphics Forum (Special Issue - Pacific Graphics 2008 Proc.), Volume 27, Number 7, page (in press)*, 2008.

[24] Y. Iwaya and Y. Suzuki. Numerical analysis of the effects of pinna shape and position on the characteristics of head-related transfer functionsp. *J. Acoust. Soc. Am.*, 123:3297, 2008.

[25] P. Plaskota and A.B. Dobrucki. The influence of pinna position on head-related transfer function. *J. Acoust. Soc. Am.*, 123(5):3724, 2008.

Références bibliographiques

[1] Algazi (V.), Avendano (C.) et Duda (R.), « Elevation localization and head-related transfer function analysis at low frequency », *J. Acoust. Soc. Am.*, **109**(3), 2001, p. 1110–1122. (cité pages 49, 51 et 65)

[2] Algazi (V.), Duda (R.), Duraiswami (R.), Gumerov (N.) et Tang (Z.), « Appoximating the head-related transfer function using simple geometric models of the head and torso », *J. Acoust. Soc. Am.*, **112**(5), 2002, p. 2053–2064. (cité pages 50 et 51)

[3] Algazi (V.), Duda (R.) et Thompson (D.). « The CIPIC HRTF Database ». Dans *IEEE Proc. Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain House,(WASPAA'01), New Paltz, NY*, 2001. (cité page 53)

[4] Alves-Pinto (A.) et Lopez-Poveda (E. A.), « Detection of high-frequency spectral notches as a function of level », *J. Acoust. Soc. Am.*, **118**(4), 2005, p. 2458–2469. (cité page 66)

[5] Asano (F.), Suzuki (Y.) et Sone (T.), « Role of spectral cues in median plane localization », *J. Acoust. Soc. Am.*, **88**(1), 1990, p. 159–168. (cité pages 32, 36, 53 et 55)

[6] Atal (B.) et Schroeder (M.). « Apparent sound source translator ». United States Patent 3236949, 1966. (cité page 30)

[7] Batteau (D. W.). « Localization of sound - part v : Auditory perception ». Rapport technique TP-3109, U.S. Naval Ordnance Test Station Report., 1963. (cité page 68)

[8] Batteau (D. W.), « The role of pinna in human localization », *Proc. R. Soc. London*, **Ser. B168**, 1967, p. 158–180. (cité page 51)

[9] Bauck (J.), « A simple loudspeaker array and associated crosstalk canceler for improved 3D audio », *J. Audio Eng. Soc.*, **49**(1/2), 2001, p. 3–13. (cité page 30)

[10] Bauck (J. L.) et Cooper (D.). « On acoustical specification of natural stereo imaging ». Dans *Proc. 66th Convention of the Audio Eng. Soc.*, volume Preprint 1649, Los Angeles,CA, USA, 1980. (cité page 48)

[11] Behrend (O.), Dickson (B.), Clarke (E.), Jin (C.) et Carlile (S.), « Neural response to free field and virtual acoustic simulation in the inferior colliculus of the guinea pig », *J. Neurophysiol.*, **92**, 2004, p. 3014–3029. (cité page 83)

[12] Bentley (J.), « Multidimensional binary search trees used for associative searching », *Communications of the ACM*, **18**(9), 1975, p. 509–517. (cité page 123)

[13] Bernstein (L.) et Trahiotis (C.), « Enhancing sensitivity to interaural delays at high frequencies by using transposed stimuli », *J. Acoust. Soc. Am.*, **112**, 2002, p. 1026. (cité page 12)

[14] Bernstein (L.) et Trahiotis (C.), « Enhancing interaural-delay-based extents of laterality at high frequencies by using transposed stimuli », *J. Acoust. Soc. Am.*, **113**, 2003, p. 3335. (cité page 12)

[15] Bernstein (L.) et Trahiotis (C.), « Measures of extents of laterality for high-frequency "transposed" stimuli under conditions of binaural interference », *J. Acous. Soc. Am.*, **118**, 2005, p. 1626. (cité page 12)

[16] Bertillon (A.), *Identification anthropométrique. Instructions signalétiques. Nouvelle édition entièrement refondue et considérablement augmentée.* Imprimerie Administrative (de la prison), 1893. (cité page 54)

[17] Besl (P.) et McKay (N.), « A method for registration of 3-d shapes », *IEEE Transactions on pattern analysis and machine intelligence*, **14**, 1992, p. 239–256. (cité pages 113 et 121)

[18] Blauert (J.), « Sound localization in the median plane », *Acustica*, **22**, 1969/70, p. 205–213. (cité page 66)

[19] Blauert (J.), *Spatial Hearing*. MIT Press, Cambridge, MA, 1997. (cité pages 15, 22, 80 et 82)

[20] Blommer (A.) et Wakefield (G. H.), « Pole-zero approximations for head-related transfer functions using a logarithmic error criterion », *IEEE Trans. Speech. and Audio Process.*, **5**(3), 1997, p. 278–287. (cité page 98)

[21] Bloom (P. J.), « Creating source elevation illusions by spectral manipulations », *J. Audio Eng. Soc.*, **25**(9), 1977, p. 561–565. (cité page 70)

[22] Bloom (P. J.), « Determination of monaural sensivity changes due to the pinna by use of minimum-audible-field measurements in the lateral vertical plane », *J. Acoust. Soc. Am.*, **61**(3), 1977, p. 820–828. (cité page 55)

[23] Blum (A.). « Etude de la plasticité du système auditif en localisation sonore. application au problème de l'individualisation en synthèse binaurale. ». Rapport de stage dea acoustique, Université d'Aix-Marseille II, 2003. (cité page 101)

[24] Blum (A.), Katz (B. F. G.) et Warusfel (O.). « Eliciting adaptation to non-individual hrtf spectral cues with multi-modal training ». Dans *Proc. CFA/DAGA'04*, Strasbourg, France, 2004. (cité pages 101 et 102)

[25] Boehnke (S.) et Phillips (D.), « Azimuthal tuning of human perceptual channels for sound location », *J. Acoust. Soc. Am.*, **106**, 1999, p. 1948. (cité page 24)

[26] Box (G.) et Cox (D.), « An analysis of transformations (with discussion) », *Journal of the Royal Statistical Society, Series B*, **26**(21), 1964, p. 1–264. (cité page 209)

[27] Box (G.) et Cox (D.), « An Analysis of Transformations Revisited (Rebutted) », *Journal of the American Statistical Association*, **77**, 1982, p. 209–210. (cité page 209)

[28] Breebaart (J.) et Kohlrausch (A.). « The perceptual (ir) relevance of HRTF magnitude and phase spectra ». Dans *Proc. 110th Convention of the Audio Eng. Soc.*, 2001. (cité pages 36 et 37)

[29] Bronkhorst (A.), « Localization of real and virtual sound sources », *J. Acoust. Soc. Am.*, **98**(5), 1995, p. 2542–2553. (cité pages 24, 32, 83, 117 et 173)

[30] Bronkhorst (A.). « Effects of stimulus properties on auditory distance perception in rooms ». Dans Breebart (D.), Houtsma (A.), Kohlrausch (A.), Prijs (V.) et Schoonhoven (R.), éditeurs, *Proc. of the 12th International Symposium on Hearing (ISH) : Physiological and Psychological Bases of Auditory Function*, p. 184–191. Mierlo, The Netherlands, 2000. (cité page 33)

[31] Brown (T.). « Characterization of acoustic head-related transfer functions for nearby sources ». Mémoire de Master, Massachusetts Institute of Technology, 2001. (cité page 22)

[32] Brungart (D. S.) et Rabinowitz (W. M.), « Auditory localization of nearby sources. Head-Related Transfer Functions », *J. Acoust. Soc. Am.*, **106**(3), 1999, p. 1465–1479. (cité pages 13, 15 et 49)

[33] Brungart (D.). « Auditory parallax effects in the HRTF for nearby sources ». Dans *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, p. 171–174, 1999. (cité page 22)

[34] Brungart (D.), Durlach (N.) et Rabinowitz (W.), « Auditory localization of nearby sources. II. localization of a broadband source », *J. Acoust. Soc. Am.*, **106**(3), 1999, p. 1956–1968. (cité page 15)

[35] Brungart (D.), Simpson (B.), McKinley (R.), Kordik (A.), Dallman (R.) et Ovenshire (D.). « The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2004*, p. 6–9, 2004. (cité pages 39 et 187)

[36] Burbeck (S.) et Luce (R.), « Evidence from auditory simple reaction times for both change and level detectors. », *Percept. Psychophys.*, **32**(2), 1982, p. 117–33. (cité page 209)

[37] Burlingame (J. A.) et Butler (R. A.), « The effects of attenuation of frequency segments on binaural localization of sound », *Percept. Psychophys.*, **60**(8), 1998, p. 1374–1383. (cité page 55)

[38] Busson (S.). *Individualisation d'indices acoustiques pour la synthèse binaurale*. Thèse de doctorat, Université de la Méditerranée, Aix-Marseille II, 2006. (cité pages 12, 86, 100, 152 et 240)

[39] Butler (R. A.), « An analysis of the monaural displacement of sounds in space », *Percept. Psychophys.*, **41**(1), 1987, p. 1–7. (cité page 72)

[40] Butler (R. A.) et Belendiuk (K.), « Spectral cues utilized in the localization of sound in the median sagittal plane », *J. Acoust. Soc. Am.*, **61**(5), 1977, p. 1264–1269. (cité pages 55 et 90)

[41] Butler (R. A.), Humanski (R. A.) et Musicant (A. D.), « Binaural and monaural localization of sound in two-dimensional space », *Perception*, **19**, 1990, p. 241–256. (cité page 72)

[42] Butler (R.) et Flannery (R.), « The spatial attributes of stimulus frequency and their role in monaural localization of sound in the horizontal plane. », *Percept. Psychophys.*, **28**(5), 1980, p. 449–57. (cité page 72)

[43] Butler (R.) et Helwig (C.), « The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. », *Am. J. Otolaryngol.*, **4**(3), 1983, p. 165–73. (cité pages 66 et 72)

[44] Butler (R.), Levy (E.) et Neff (W.), « Apparent distance of sound recorded in echoic and anechoic chambers », *J. Exptl. Psychol. : Hum. Percept. Perform.*, **6**, 1980, p. 745–750. (cité page 22)

[45] Butler (R.) et Musicant (A.), « Binaural localization : Influence of stimulus frequency and the linkage to covert peak areas », *Hearing Research*, **67**, 1993, p. 220–229. (cité page 72)

[46] Campbell (R.), Doubell (T.), Nodal (F.), Schnupp (J.) et King (A. J.), « Interaural timing cues do not contribute to map of space in the superior colliculus : a virtual acoustic space study », *J. Neurophysiol.*, **95**, 2006, p. 242–254. (cité page 83)

[47] Carlile (S.), *Virtual Auditory Space : Generation and Applications*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1996. (cité pages 15, 51, 75 et 182)

[48] Carlile (S.), Jin (C.) et van Raad (V.). « Continuous virtual auditory space using HRTF interpolation : Acoustic and psychophysical errors ». Dans *International Symposium on Multimedia Information Processing*, Sydney, NSW, Australia, 2000. (cité pages 160, 173, 181, 182 et 284)

[49] Carlile (S.), Leong (P.) et Hyams (S.), « The nature and distribution of errors in sound localization by human listeners », *Hearing Research*, **114**(1-2), 1997, p. 179–196. (cité pages 22, 25, 26, 181, 182 et 184)

[50] Carlile (S.), Leong (P.), Pralong (D.) et Hyams (S.). « High fidelity virtual auditory space : An operational definition ». Dans *Proc. Simtect 96*, p. 79–84, Melbourne, 1996. (cité pages 24 et 83)

[51] Carlile (S.) et Pralong (D.), « The location-dependent nature of perceptually salient features of the human head-related transfer functions », *J. Acoust. Soc. Am.*, **95**(6), 1994, p. 3445–3459. (cité pages 53, 55 et 68)

[52] Chen (F.). « The reaction time for subjects to localize 3D sound via headphones ». Dans *Proc. 22nd Conference of the Audio Eng. Soc.*, 2002. (cité pages 182, 196 et 209)

[53] Chen (J.), Van Veen (B.) et Hecox (K.), « A spatial feature extraction and regularization model for the head-related transfer function », *J. Acoust. Soc. Am.*, **97**(1), 1995, p. 439–452. (cité pages 160 et 284)

[54] Cheng (C.) et Wakefield (G.). « A tool for volumetric visualization and sonification of Head Related Transfer Functions(HRTFs) ». Dans *Proc. Int.l Conf. on Auditory Display, ICAD 2000*, 2000. (cité page 132)

[55] Cheung (N.-M.) et Trautmann (S.), « Head-related transfer function modeling in 3-d sound systems with genetic algorithm », *J. Audio Eng. Soc.*, **46**(6), 1998, p. 531–539. (cité page 37)

[56] Chirikjian (G. S.), Kim (P.), Koo (J.-Y.) et Lee (C.), « Rotational matching problems », *International Journal of Computational Intelligence and Applications*, **4**(4), 2004, p. 401–416. (cité pages 133 et 162)

[57] CIPIC. « UC Davis CIPIC HRTF Database », 2001. (cité pages 107, 163, 165 et 166)

[58] Constan (Z.) et Hartmann (W.), « On the detection of dispersion in the head-related transfer function », *J.Acoust. Soc. Am.*, **114**, 2003, p. 998. (cité page 34)

[59] Coupin (H.), « Notre oreille », *La Nature*, **1470**, 1901, p. 138–141. (cité page 46)

[60] Darkner (S.), Vester-Christensen (M.), Larsen (R.), Nielsen (C.) et Paulsen (R.). « Automated 3D Rigid Registration of Open 2D Manifolds ». Dans *MICCAI 2006 Workshop From Statistical Atlases to Personalized Models*, 2006. (cité pages 121 et 124)

[61] Dellepiane (M.), Pietroni (N.), Tsingos (N.), Asselot (M.) et Scopigno (R.). « Reconstructing head models from photographs for individualized 3D-audio processing ». Dans *Proc. Computer Graphics Forum (Special Issue - Pacific Graphics 2008*, volume 27, 2008. (cité pages 88, 152, 153 et 246)

[62] Duda (R. O.). « Elevation dependence of the interaural transfer function ». Dans *Binaural and Spatial Hearing in Real and Virtual Environments, edited by R. H. Gilkey and A. T. B. (Erlbaum, Mahwah, NJ)*, p. 49–75. 1997. (cité page 68)

[63] Duda (R. O.) et Martens (W. L.), « Range dependence of the response of a spherical head model », *J. Acoust. Soc. Am.*, **104**(5), 1998, p. 3048–3058. (cité pages 16, 49, 50 et 152)

[64] Durlach (N.), Rigopulos (A.), Pang (X.), Woods (W.), Kulkarni (A.), Colburn (H.) et Wenzel (E.), « On the externalization of auditory images »,

Presence : Teleoperators and Virtual Environments, **1**(2), 1992, p. 251–257. (cité page 41)

[65] Efron (B.) et Tibshirani (R.), « An Introduction to the Bootstrap », *NY Monographs on Statistics and Applied Probability Chapman and Hall*, **57**, 1993. (cité pages 212 et 214)

[66] Evans (M.), Angus (J.) et Tew (A.), « Analyzing Head-Related Transfer Function measurements using surface spherical harmonics », *J. Acoust. Soc. Am.*, **104**(4), 1998, p. 2400–2411. (cité page 132)

[67] Faure (J.). « Les systèmes de Head Tracking (Les systèmes de poursuite de tête) ». Rapport technique, France Télécom R&D, 2004. (cité page 39)

[68] Faure (J.). « Evaluation de la synthèse binaurale dynamique ». Rapport technique FT/DIVISION R&D/TECH/SSTP/JF/2005-63, France Télécom R&D, 2005. (cité pages 20 et 182)

[69] Fischler (M.) et Bolles (R.), « Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography », *Communications of the ACM*, **24**(6), 1981, p. 381–395. (cité pages 144 et 269)

[70] Foster (S.). « Impulse response measurement using golay codes ». Dans *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'86*, p. 929–932, 1986. (cité page 32)

[71] Foucault (M.). « Utopies et hétérotopies, 1966 ». Dans *INA Mémoire Vive, 2003*. (cité page 8)

[72] Gardner (M.), « Proximity image effect in sound localization », *J. Acoust. Soc. Am.*, **43**, 1968, p. 163. (cité pages 26 et 239)

[73] Gardner (M.), « Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space », *J. Acoust. Soc. Am.*, **45**, 1969, p. 47. (cité page 24)

[74] Gardner (M.) et Gardner (R.), « Problem of localization in the median plane : effect of pinnae cavity occlusion », *J. Acoust. Soc. Am.*, **53**(2), 1973, p. 400–408. (cité page 64)

[75] Glasberg (B. R.) et Moore (B. C. J.), « Derivation of auditory filter shapes from notched-noise data », *Hearing Research*, **47**, 1990, p. 103–138. (cité page 35)

[76] Grassi (E.), Tulsi (J.) et Shamma (S.). « Measurement of head-related transfer functions based on the empirical transfer function estimate ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2003*, p. 119–122, Boston, MA, USA, 2003. (cité pages 109, 164, 166 et 167)

[77] Gupta (N.), Ordonez (C.) et Barreto (A.). « Improved localization of virtual sound by spectral modification of hrtfs to simulate protruding pinnae ». Dans *Proc. 6th World Multiconference on Systemics, Cybernetics and Informatics*, p. III 291–296, Orlando, FL, USA, 2002. (cité page 97)

[78] Hacihabiboglu (H.), Gunel (B.) et Murtagh (F.). « Wavelet-based spectral smoothing for head-related transfer function filter design ». Dans *Proc. 22nd Conference of the Audio Eng. Soc.*, 2002. (cité page 37)

[79] Hardin (R.), Sloane (N.) et Smith (W.), « Spherical coverings », *Electronic : http ://www. research. att. com/njas/coverings/index. html, May*, 1997. (cité page 173)

[80] Hartmann (W.) et Rakerd (B.), « Auditory spectral discrimination and the localization of clicks in the sagittal plane », *J. Acoust. Soc. Am.*, **94**(4), 1993, p. 2083–2092. (cité page 65)

[81] Hartmann (W.) et Wittenberg (A.), « On the externalization of sound images », *J. Acoust. Soc. Am.*, **99**(6), 1996, p. 3678–3688. (cité pages 41 et 83)

[82] Hartung (K.), Braasch (J.) et Sterbing (S.). « Comparison of different interpolation methods for the interpolation of head-related transfer functions ». Dans *Proc. 16th Conference of the Audio Eng. Soc.*, 1999. (cité pages 160 et 172)

[83] Heathcote (A.), Popiel (S.) et Mewhort (D.), « Analysis of response time distribution : an example using the stroop task », *Psychological bulletin*, **109**(2), 1991, p. 340–347. (cité page 209)

[84] Hebrank (J.), « Pinna disparity processing : A case of mistaken identity ? », *J. Acoust. Soc. Am.*, **59**(1), 1976, p. 220–221. (cité page 68)

[85] Hebrank (J.) et Wright (D.), « Spectral cues used in the localization of sound sources on the median plane », *J. Acoust. Soc. Am.*, **56**(6), 1974, p. 1829–1834. (cité pages 55, 65 et 70)

[86] Hetherington (C.) et Tew (A. I.). « Parametrizing human pinna shape for the estimation of head-related transfer functions ». Dans *Proc. 114th Convention of the Audio. Eng. Soc.*, Amsterdam, The Netherlands, 2003. (cité page 88)

[87] Hetherington (C.), Tew (A. I.) et Tao (Y.). « Three-dimensional elliptic fourier methods for the parameterization of human pinna shape ». Dans *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP 2003*, Hong Kong, 2003. (cité pages 88 et 100)

[88] Hockley (W.), « Analysis of response time distributions in the study of cognitive processes », *J. of Exptl Psychol.. Learning, memory, and cognition*, **10**(4), 1984, p. 598–615. (cité page 209)

[89] Hoffmann (P.) et Møller (H.), « Audibility of differences in adjacent Head-Related Transfer Functions », *Acta Acustica united with Acustica*, **94**, 2008, p. 945–954. (cité page 240)

[90] Hofman (P.) et Van Opstal (A.), « Spectro-temporal factors in two-dimensional human sound localization », *J. Acoust. Soc. Am.*, **103**(5), 1998, p. 2634–2648. (cité pages 65 et 74)

[91] Hofman (P.) et Van Opstal (A.), « Bayesian reconstruction of sound localization cues from responses to random spectra », *Biol. Cybern.*, **86**, 2002, p. 305–316. (cité page 89)

[92] Hofman (P.) et Van Opstal (A.), « Binaural weighting of pinna cues in human sound localization », *Exp. Brain. Res.*, **148**, 2003, p. 458–470. (cité pages 67 et 69)

[93] Hofman (P.), Van Riswick (J.) et Van Opstal (A.), « Relearning sound localization with new ears », *Nature neuroscience*, **1**(5), 1998, p. 417–421. (cité pages 100 et 238)

[94] Hohle (R.), « Inferred components of reaction times as functions of foreperiod duration. », *J. Exptl. Psychol.*, **69**, 1965, p. 382–6. (cité page 209)

[95] Honda (A.), Shibata (H.), Gyoba (J.), Saitou (K.), Iwaya (Y.) et Suzuki (Y.), « Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game », *Applied Acoustics*, **68**(8), 2007, p. 885–896. (cité page 101)

[96] Hultsch (D.) et MacDonald (S.), « Intraindividual variability in performance as a theoretical window onto cognitive aging », *New frontiers in cognitive aging*, 2004, p. 65–88. (cité page 212)

[97] Humanski (R. A.) et Butler (R. A.), « The contribution of the near and far ear toward localization of sound in the sagittal plane », *J. Acoust. Soc. Am.*, **83**(6), 1988, p. 2300–2310. (cité page 69)

[98] Huopaniemi (J.) et Zacharov (N.), « Objective and subjective evaluation of head-related transfer function filter design », *J. Audio Eng. Soc.*, **47**(4), 1999, p. 218–239. (cité page 37)

[99] Iida (K.) et Itoh (M.). « A novel head-related transfer function model based on spectral and interaural difference cues ». Dans *Proc. 9th Western Pacific Acoustics Conference, WESPAC IX 2006*, Seoul, Korea, 2006. (cité page 70)

[100] Iida (K.), Matsuda (J.), Nakamura (K.), Itoh (M.) et Morimoto (M.). « 3-d sound reproduction through stereo inner-phones ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2002*, Kyoto, Japan, 2002. (cité page 90)

[101] Ingård (U.), « A review of the influence of meteorological conditions on sound propagation », *J. Acoust. Soc. Am.*, **25**, 1953, p. 405. (cité page 22)

[102] Inoue (N.), Nishino (T.), Itou (K.) et Takeda (K.). « Hrtf modeling using physical features ». Dans *Proc. Forum Acusticum 2005*, Budapest, Hungary, 2005. (cité page 100)

[103] IRCAM Listen HRTF database. « http ://recherche.ircam.fr/equipes/salles/ listen/ », 2002. (cité pages 168 et 172)

[104] Iwaya (Y.), « Individualization of head-related transfer functions with tournament-style listening test : Listening with other's ears », *Acoust. Sci. & Tech.*, **6**, 2006, p. 340–343. (cité pages 91 et 101)

[105] Iwaya (Y.) et Suzuki (Y.), « Numerical analysis of the effects of pinna shape and position on the characteristics of head-related transfer functions », *J. Acoust. Soc. Am.*, **123**, 2008, p. 3297. (cité pages 145 et 146)

[106] Jack (C.) et Thurlow (W.), « Effects of degree of visual association and angle of displacement on the " ventriloquism" effect », *Percept. Mot. Skills*, **37**(3), 1973, p. 967–79. (cité page 10)

[107] Janko (J. A.), Anderson (T. R.) et Gilkey (R. H.). « Using neural networks to evaluate the viability of monaural and interaural cues for sound localization ». Dans *Binaural and Spatial Hearing in Real and Virtual Environments, edited by R. H. Gilkey and A. T. B. (Erlbaum, Mahwah, NJ)*, p. 557–570. 1997. (cité page 68)

[108] Jeppesen (J.) et Moller (H.). « Cues for localization in the horizontal plane ». Dans *Proc. 118h Convention of the Audio. Eng. Soc.*, Barcelona, Spain, 2005. (cité page 67)

[109] Jin (C.), Leong (P.), Leung (J.), Corderoy (A.) et Carlile (S.). « Enabling individualized virtual auditory space using morphological measurements ». Dans *IEEE Pacific-Rim Conference on Multimedia - International Symposium on Multimedia Information Processing*, p. 220–223, 2000. (cité page 99)

[110] Jin (C. T.). *Spectral Analysis and Resolving Spatial Ambiguities in Human Sound Localization.* PhD Thesis, University of Sydney, 2001. (cité pages 65, 66, 72 et 75)

[111] Jin (C.), Corderoy (A.), Carlile (S.) et Van Schaik (A.). « Spectral cues in human sound localization ». Dans Müller (S. A. S.), Leen (T. K.) et Klaus-Robert, éditeurs, *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, p. 768–774. MIT Press, 2000. (cité pages 68 et 69)

[112] Jot (J.) et Warusfel (O.). « A real-time spatial sound processor for music and virtual reality applications ». Dans *ICMC, International Computer Music Conference*, p. 294–295, 1995. (cité page 33)

[113] Kacelnik (O.), Nodal (F. R.), Parsons (C. H.) et King (A. J.), « Training-induced plasticity of auditory localization in adult mammals », *PLoS Biology*, **4**(4), 2006, p. 627–638. (cité page 101)

[114] Kahana (Y.). *Numerical modelling of the head-related transfer function.* PhD Thesis, University of Southampton, 2000. (cité pages 88 et 89)

[115] Kahana (Y.) et Nelson (P. A.). « Spatial acousic mode shapes of the human pinna ». Dans *Proc. 109th Convention of the Audio. Eng. Soc.*, Los Angeles, California, USA, 2000. (cité page 63)

[116] Katz (B.). *Measurement and calculation of individual head-related transfer function using a boundary element model including the measurement and effect of skin and hair impedance.* PhD Thesis, Pennsylvannia State University, USA, 1998. (cité page 88)

[117] Katz (B.), « Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements », *J. Acoust. Soc. Am.*, **110**, 2001, p. 2449. (cité page 88)

[118] Kim (H.-Y.), Suzuki (Y.), Tkane (S.) et Sone (T.), « Control of auditory distance perception based on the auditory parallax model », *Applied Acoustics*, **62**, 2001, p. 245–270. (cité page 22)

[119] Kim (S.-M.) et Choi (W.), « On the externalisation of virtual sound images in headphone reproduction : A Wiener filter approach », *J. Acoust. Soc. Am.*, **117**(6), 2005, p. 3657–3665. (cité pages 41, 42, 81, 83 et 85)

[120] King (R. B.) et Oldfield (S. R.), « The impact of signal bandwidth on auditory localization : Implications for the design of three-dimensional audio displays », *Human Factors*, **39**(2), 1997, p. 287–295. (cité page 65)

[121] Kistler (D. J.) et Wightman (F. L.), « A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction », *J. Acoust. Soc. Am.*, **91**(3), 1992, p. 1637–1647. (cité pages 37, 91, 283 et 284)

[122] Klumpp (R.) et Eady (H.), « Some Measurements of Interaural Time Difference Thresholds », *J. Acoust. Soc. Am.*, **28**, 1956, p. 859. (cité page 192)

[123] Kostelec (P. J.) et Rockmore (D. N.), « FFTs on the Rotation Group », *Santa Fe Institute Working Papers Series*, 2003. (cité page 162)

[124] Kostelec (P.) et Rockmore (D.), « SOFT : SO (3) Fourier Transforms », *http ://www.cs.dartmouth.edu/ geelong/soft/.* (cité page 273)

[125] Kuhn (G. F.), « Model for the interaural time difference in the azimuthal plane », *J. Acoust. Soc. Am.*, **62**(1), 1977, p. 157–167. (cité pages 10 et 12)

[126] Kulkarni (A.) et Colburn (H. S.), « Variability in the characterization of the headphone transfer-function », *J. Acoust. Soc. Am.*, **107**(2), 2000, p. 1071–1074. (cité pages 41 et 42)

[127] Kulkarni (A.) et Colburn (H.), « Role of spectral detail in sound-source localization », *Nature*, **396**, 1998, p. 747–749. (cité pages 36 et 83)

[128] Kulkarni (A.), Isabelle (S.) et Colburn (H.), « Sensitivity of human subjects to head-related transfer-function phase spectra », *J. Acoust. Soc. Am.*, **105**, 1999, p. 2821. (cité pages 34, 35 et 188)

[129] Lacouture (Y.) et Cousineau (D.), « How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times », *Tutorials in Quantitative Methods for Psychology*, **4**(1), 2008, p. 35–45. (cité pages 214, 296 et 297)

[130] Langendijk (E. H. A.) et Bronkhorst (A. W.), « Fidelity of three-dimensional-sound reproduction using a virtual auditory display », *J. Acoust. Soc. Am.*, **107**(1), 2000, p. 528–537. (cité pages 159, 175 et 176)

[131] Langendijk (E. H. A.) et Bronkhorst (A. W.), « Contribution of spectral cues to human sound localization », *J. Acoust. Soc. Am.*, **112**(4), 2002, p. 1583–1596. (cité pages 19, 37, 65, 74 et 131)

[132] Larcher (V.). *Techniques de spatialisation du son pour la réalité virtuelle.* Thèse de doctorat, Université de Paris VI, 2001. (cité pages 15, 39, 40, 95, 106 et 111)

[133] Lee (S.), Kim (L.) et Sung (K.), « Reduction of sound localization error for non-individualized HRTF by directional weighting function », *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, **87**(6), 2004, p. 1531–1536. (cité page 97)

[134] Little (A.), Mershon (D.) et Cox (P.), « Spectral content as a cue to perceived auditory distance », *Perception*, **21**(3), 1992, p. 405–416. (cité page 22)

[135] Loomis (J.), Golledge (R.), Klatzky (R.), Speigle (J.) et Tietz (J.). « Personal guidance system for the visually impaired ». Dans *Proc. of the first annual ACM conference on Assistive technologies*, p. 85–91. ACM Press New York, NY, USA, 1994. (cité pages 40 et 41)

[136] Loomis (J.), Klatzky (R.) et Golledge (R.). *Mixed Reality : Merging Real and Virtual Worlds*, chaptitre Auditory distance perception in real, virtual, and mixed environments., p. 201–214. Ohta, Y. and Tamura, H., 1999. (cité page 41)

[137] Lopez-Poveda (E.). *The physical origin and physiological coding of pinna-based spectral cues.* PhD Thesis, Loughborough University, 1996. (cité pages 53 et 55)

[138] Lopez-Poveda (E.) et Meddis (R.), « A physical model of sound diffraction and reflections in the human concha », *J. Acoust. Soc. Am.*, **100**(5r), 1996, p. 3248–3259. (cité pages 63 et 64)

[139] Macpherson (E.), « Source spectrum recovery at different spatial locations », *J. Acoust. Soc. Am.*, **98**, 1995, p. 2946. (cité page 76)

[140] MacPherson (E.). *Spectral cue processing in the auditory localization of sounds with wideband non-flat spectra.* PhD Thesis, University of Wisconsin-Madison, USA, 1998. (cité pages 67, 69, 71, 72, 73 et 74)

[141] MacPherson (E.), « Binaural weighting of monaural spectral cues for sound localization », *J. Acoust. Soc. Am.*, **121**, 2007, p. 3677–3688. (cité page 69)

[142] Macpherson (E.). « Stimulus continuity is not necessary for the salience of dynamic sound localization cues ». Dans *157th meeting of the ASA, Portland, OR, USA*, 2009. (cité page 20)

[143] Macpherson (E.) et Kerr (D.). « The salience of dynamic sound localization cues as a function of head velocity and stimulus frequency ». Dans *APCAM, Auditory Perception Cognition and Action Meeting, 2008, Chicago, IL, USA*, 2008. (cité page 20)

[144] MacPherson (E.) et Middlebrooks (J.), « Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited », *J. Acoust. Soc. Am.*, **111**(5), 2002, p. 2219–2236. (cité pages 12, 15 et 68)

[145] MacPherson (E.) et Middlebrooks (J.), « Vertical-plane sound localization probed with ripple-spectrum noise », *J. Acoust. Soc. Am.*, **114**(1), 2003, p. 430–445. (cité pages 67 et 70)

[146] MacQueen (J. B.). « Some methods for classification and analysis of multivariate observations ». Dans *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, p. 281–297, 1967. (cité page 169)

[147] Majdak (P.), Balazs (P.) et Laback (B.), « Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions », *J. Audio Eng. Soc.*, **55**(7/8), 2007, p. 623. (cité pages 32 et 158)

[148] Maki (K.) et Furukawa (S.), « Reducing individual differences in the external-ear transfer functions of the mongolian gerbil », *J. Acoust. Soc. Am.*, **118**(4), 2005, p. 2392–2404. (cité pages 111, 113, 118, 131, 132 et 147)

[149] Martens (W. L.). « Uses and misuses of psychophysical methods in the evaluation of spatial sound reproduction ». Dans *Proc. 110th Convention of the Audio. Eng. Soc.*, Amsterdam, The Netherlands, 2001. (cité page 81)

[150] Martens (W. L.). « Rapid psychophysical calibration using bisection scaling for individualized control of source elevation in auditory display ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2002*, Kyoto, Japan, 2002. (cité pages 95 et 96)

[151] Martin (R.) et McAnally (K.). « Interpolation of head-related transfer functions ». Rapport technique DSTO-RR-0323, Australian Government - Department of Defence, February 2007. (cité page 160)

[152] Martin (R.), McAnally (K.) et Senova (M.), « Free-field equivalent localization of virtual audio », *J. Audio Eng. Soc.*, **49**(1/2), 2001, p. 14–22. (cité page 83)

[153] Martin (R.), Paterson (M.) et McAnally (K.), « Utility of monaural spectral cues is enhanced in the presence of cues to sound-source lateral angle », *Journal of the Association for Research in Otolaryngology*, **5**, 2004, p. 80–89. (cité page 69)

[154] Møller (H.), « Fundamentals of binaural technology », *Applied Acoustics*, **36**, 1992, p. 171–218. (cité pages 18, 32 et 42)

[155] Møller (H.), « Binaural technique : Do we need individual recordings ? », *J. Audio Eng. Soc.*, **44**(6), 1996, p. 451–469. (cité page 85)

[156] Møller (H.), Jensen (C.), Hammerhøi (D.) et Sørensen (M.). « Using a typical human subject for binaural recording ». Dans *Proc. 100th Convention of the Audio Eng. Soc.*, Copenhagen, Denmark, 1996. (cité page 90)

[157] Møller (H.), Jensen (C.), Hammershøi (D.) et Sørensen (M.), « Selection of a typical human subject for binaural recording », *Acta Acustica*, **82**, 1996, p. s215. (cité page 90)

[158] Møller (H.), Hammershøi (D.), Jensen (C.) et Sørensen (M. F.), « Transfer characteristics of headphones measured on human ears », *J. Audio Eng. Soc.*, **43**(4), 1995, p. 203–217. (cité pages 41 et 42)

[159] Møller (H.), Hammershøi (D.), Jensen (C.) et Sørensen (M.), « Evaluation of artificial heads in listening tests », *J. Audio Eng. Soc.*, **47**(3), 1999, p. 83–99. à trouver. (cité page 85)

[160] Møller (H.), Sørensen (M.), Hammerhøi (D.) et Jensen (C.), « Head-related transfer functions of human subjects », *J. Audio Eng. Soc.*, **43**(5), 1995, p. 300–321. (cité pages 32, 80 et 86)

[161] UC Davis CIPIC HRTF database. « http ://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm ». (cité pages 167 et 172)

[162] University of Maryland, Neural System Laboratory HRTF database. « http ://www.isr.umd.edu/Labs/NSL/ ». (cité pages 164, 166, 167 et 172)

[163] McAnally (K. I.) et Martin (R. L.), « Variability in the headphone-to-ear-canal transfer function », *J. Audio Eng. Soc.*, **50**(4), 2002, p. 263–266. (cité pages 41 et 42)

[164] Mehrgardt (S.) et Mellert (V.), « Transformation characteristics of the external human ear », *J. Acoust. Soc. Am.*, **61**(6), 1977, p. 1567–1576. (cité pages 35 et 86)

[165] Mershon (D.) et King (L.), « Intensity and reverberation as factors in the auditory perception of egocentric distance », *Percept. Psychophys.*, **18**(6), 1975, p. 409–415. (cité page 21)

[166] Mewhort (D.), Braun (J.) et Heathcote (A.), « Response time distributions and the Stroop Task : a test of the Cohen, Dunbar, and McClelland (1990) model. », *J. Exp. Psychol. Hum. Percept. Perform.*, **18**(3), 1992, p. 872–82. (cité page 209)

[167] Middlebrooks (J.), « Narrow-band sound localization related external ears acoustics », *J. Acoust. Soc. Am.*, **92**(5), 1992, p. 2607–2624. (cité pages 66, 67, 72 et 73)

[168] Middlebrooks (J.), « Spectral shape cues for sound localization », *Binaural and Spatial Hearing in Real and Virtual Environments, edited by R. H. Gilkey and A. T. B. (Erlbaum, Mahwah, NJ)*, 1997, p. 77–97. (cité page 68)

[169] Middlebrooks (J.), « Individual differences in external-ear transfer functions reduced by scaling in frequency », *J. Acoust. Soc. Am.*, **106**(3), 1999, p. 1480–1492. (cité pages 92, 95, 106, 113, 131, 135, 147 et 149)

[170] Middlebrooks (J.), « Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency », *J. Acoust. Soc. Am.*, **106**(3), 1999, p. 1493–1510. (cité pages 92, 94, 131, 149 et 175)

[171] Middlebrooks (J.) et Green (D.), « Sound localization by human listeners », *Annual Review of Psychology*, **42**, 1991, p. 135–159. (cité page 12)

[172] Middlebrooks (J.) et Green (D.), « Observations on a principal component analysis of head-related transfer functions », *J. Acoust. Soc. Am.*, **92**(1), 1992, p. 597–599. (cité page 32)

[173] Middlebrooks (J.), MacPherson (E.) et Onsan (Z.), « Psychophysical customization of directional transfer functions for virtual sound localization », *J. Acoust. Soc. Am.*, **108**(6), 2000, p. 3088–3091. (cité page 92)

[174] Middlebrooks (J.), Makous (J.) et Green (D.), « Directional sensivity of sound pressure levels in the human ear canal », *J. Acoust. Soc. Am.*, **86**(1), 1989, p. 89–108. (cité pages 53, 55, 68, 86 et 131)

[175] Middlebrooks (J.) et Pettigrew (J.), « Functional classes of neurons in primary auditory cortex of the cat distinguished by sensitivity to sound location », *J. Neurosci.*, **1**, 1981, p. 107–120. (cité page 72)

[176] Mills (A. W.), « On the minimum audible angle », *J. Acoust. Soc. Am.*, **30**, 1958, p. 237–248. (cité page 181)

[177] Mills (A. W.). *Foundations of modern auditory theory*, chaptitre Auditory localisation. New York : Academic Press, 1972. (cité page 11)

[178] Minaar (P.), Olesen (S. K.), Christensen (F.) et Møller (H.), « Localization with binaural recordings from artificial and human heads », *J. Audio Eng. Soc.*, **49**(5), 2001, p. 323–335. (cité page 85)

[179] Minaar (P.), Plogsties (J.) et Flemming (C.), « Directional resolution of Head-related Transfer Functions required in binaural synthesis », *J. Audio Eng. Soc.*, **53**(10), 2005, p. 919–929. (cité page 159)

[180] Moore (B. C. J.), Oldfield (S. R.) et Dooley (G. J.), « Detection and discrimination of spectral peaks and notches at 1 and 8 khz », *J. Acoust. Soc. Am.*, **85**(2), 1989, p. 820–835. (cité page 71)

[181] Morimoto (M.), « The contribution of two ears to the perception of vertical angle in sagittal planes », *J. Acoust. Soc. Am.*, **109**, 2001, p. 1596–1603. (cité pages 69 et 172)

[182] Morimoto (M.) et Aokata (H.), « Localization cues of sound sources in the upper hemisphere », *J. Acoust. Soc. Jpn.*, **5**(3), 1984, p. 165–173. (cité page 67)

[183] Morimoto (M.), Yairi (M.), Iida (K.) et Itoh (M.), « The role of low frequency components in median plane localization », *Acoust. Sci. & Tech.*, **24**(2), 2003, p. 76–82. (cité page 65)

[184] Mrsic-Flogel (T.), King (A.), Jenison (R.) et Schnupp (J.), « Listening through different ears alters spatial response fields in ferret primary auditory cortex », *J. Neurophysiol.*, **86**, 2001, p. 1043–1046. (cité pages 85 et 87)

[185] Mrsic-Flogel (T.), King (A.) et Schnupp (J.), « Encoding of virtual acoustic space stimuli by neurons in ferret primary auditory cortex », *J. Neurophysiol.*, **93**, 2005, p. 3489–3503. (cité pages 85 et 86)

[186] Musicant (A. D.) et Butler (R. A.), « The influence of pinnae-based spectral cues on sound localization », *J. Acoust. Soc. Am.*, **75**(4), 1984, p. 1195–1200. (cité page 69)

[187] Musicant (A. D.), Chan (J. C. K.) et Hind (J. E.), « Direction-dependent spectral properties of cat external ear : New data and cross-species comparisons », *J. Acoust. Soc. Am.*, **87**(2), 1990, p. 757–781. (cité pages 63 et 68)

[188] Nandy (D.) et Ben-Arie (J.), « Estimating the azimuth of a sound source from the binaural spectral amplitude », *IEEE Trans. Speech. and Audio Process.*, **4**(1), 1996, p. 45–55. (cité page 68)

[189] Ng (G.). *Elevation localization of single- and multiple-band noises.* Thèse de doctorat, Boston University, College of engineering, 2005. (cité page 74)

[190] Oldfield (S. R.) et Parker (S. P. A.), « Acuity of sound localisation : a topography of auditory space. I. normal hearing conditions », *Perception*, **13**, 1984, p. 581–600. (cité pages 22 et 181)

[191] Otten (J.). *Factors influencing acoustical localization.* PhD Thesis, Universität Oldenburg, Deutschland, 2001. (cité page 37)

[192] Palmer (A.) et Russell (I.), « Phase locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells », *Hearing Research*, **24**, 1986, p. 1–15. (cité page 11)

[193] Park (M.-H.), Choi (S.-I.), Kim (S.-H.) et Bae (K.-S.). « Improvement of front-back sound localization characteristics in headphone-based 3d sound generation ». Dans *Proc. Internet and Multimedia Systems, and Applications*, 2005. (cité page 97)

[194] Patterson (R. D.), Holdsworth (J.), Nimmo-Smith (I.) et Rice (P.). « SVOS final report : The auditory filterbank ». Technical Report 2341, 1988. (cité page 36)

[195] Pellieux (L.), Piedecocq (B.), Leroyer (P.), Villeneuve (F.) et Sarafian (D.). « Mesure des fonctions de transfert de tête (HRTF) ». Rapport technique IMASSA-ERMA 95-14, Centre d'Etudes et de Recherches de Medecine Aerospatiale, Département des Sciences Cognitives et Ergonomie, Bretigny sur Orge, France, 1995. (cité page 32)

[196] Pernaux (J.-M.). « Mesures de HRTF au TNO (Soesterberg) ». Rapport technique, 2001. (cité page 117)

[197] Pernaux (J.-M.). *Spatialisation du son par les techniques binaurales : application aux services de télécommunications.* Thèse de doctorat, Institut National Polytechnique de Grenobles, 2003. (cité pages 32, 33, 40, 117, 184, 188 et 239)

[198] Pernaux (J.-M.), Emerit (M.), Nicol (R.) et Daniel (J.). « Perceptual evaluation of binaural sound synthesis : the problem of reporting localization judgments ». Dans *Proc. 114th Convention of the Audio. Eng. Soc.*, Amsterdam, The Netherlands, 2003. (cité page 182)

[199] Perrett (S.) et Noble (W.), « The contribution of head motion cues to localization of low-pass noise », *Percept. Psychophys.*, **59**(7), 1997, p. 1018–1026. (cité page 20)

[200] Perrett (S.) et Noble (W.), « The effect of head rotations on vertical plane sound localization », *J. Acoust. Soc. Am.*, **102**(4), 1997, p. 2325–2332. (cité page 20)

[201] Phillips (D. P.), Calford (M. B.), Pettigrew (J. D.), Aitkin (L. M.) et Semple (M. N.), « Directionality of sound pressure at the cat's pinna », *Hearing Research*, **8**(13-28), 1982. (cité page 72)

[202] Plaskota (P.) et Dobrucki (A.), « The influence of pinna position on head-related transfer function », *J. Acoust. Soc. Am.*, **123**(5), 2008, p. 3724. (cité page 145)

[203] Poincaré (H.), Rougier (L.) et Monnier (P.), *La valeur de la science*. Flammarion Paris, 1905. (cité pages 182 et 241)

[204] Poon (P. W. F.) et Brugge (J. F.), « Sensitivity of auditory nerve fibers to spectral notches », *J. Neurophysiol.*, **70**(2), 1993, p. 655–666. (cité page 71)

[205] Pralong (D.) et Carlile (S.), « Measuring the human head-related transfer functions : A novel method for the construction and calibration of a miniature "in-ear" recording system », *J. Acoust. Soc. Am.*, **95**, 1994, p. 3435. (cité page 32)

[206] Pralong (D.) et Carlile (S.), « The role of individualized headphone calibration for the generation of high fidelity virtual auditory space », *J. Acoust. Soc. Am.*, **100**(6), 1996, p. 3785–3793. (cité pages 41 et 42)

[207] Purves (D.), Augustine (G.), Fitzpatrick (D.), Coquery (J.) et Roucoux (A.), *Neurosciences Neurosciences & cognition*. De Boeck université, 1999. (cité page 84)

[208] Qian (J.) et Eddins (D.), « The role of spectral modulation cues in virtual sound localization », *J. Acoust. Soc. Am.*, **123**, 2008, p. 302. (cité page 67)

[209] Rabinowitz (W. M.), Maxwell (J.), Shao (Y.) et Wei (M.), « Sound localization cues for a magnified head : Implications from sound diffraction about a rigid sphere », *Presence*, **2**(125-129), 1993. (cité page 48)

[210] Rakerd (B.), « Identification and localization of sound sources in the median sagittal plane », *J. Acoust. Soc. Am.*, **106**(5), 1999, p. 2812–2820. (cité pages 69 et 76)

[211] Ratcliff (R.), « A theory of memory retrieval », *Psychological Review*, **85**(2), 1978, p. 59–108. (cité page 209)

[212] Ratcliff (R.) et Murdock (B.), « Retrieval processes in recognition memory », *Psychological Review*, **83**(3), 1976, p. 190–214. (cité page 209)

[213] Rauschecker (J.), Tian (B.) et Hauser (M.), « Processing of complex sounds in the macaque nonprimary auditory cortex », *Science*, **268**(5207), 1995, p. 111–114. (cité page 75)

[214] Raykar (V. C.) et Duraiswami (R.), « Extracting the frequencies of the pinna spectral notches in measured head related impulse responses », *J. Acoust. Soc. Am.*, **118**(1), 2005, p. 364–374. (cité page 99)

[215] Raykar (V. C.), Duraiswami (R.), Davis (L.) et Yegnanarayana (B.). « Extracting significant features from the hrtf ». Dans *Proc. 2003 International Conference on Auditory Display, ICAD 2003*, p. 115–118, Boston, MA, USA, 2003. (cité page 99)

[216] Rayleigh (L.), « On our perception of sound direction », *Philosophical Magazine*, **13**, 1907, p. 214–232. (cité pages 10 et 22)

[217] Riederer (K. A.). « Repeatability analysis of head-related transfer function measurement ». Dans *Proc. 105th Convention of the Audio. Eng. Soc.*, San Francisco, CA, USA, 1998. (cité pages 41 et 42)

[218] Rodriguez (S. G.) et Ramirez (M. A.). « Extracting and modeling approximated pinna-based transfer functions from hrtf data ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2005*, Limerick, Ireland, 2005. (cité page 99)

[219] Rodriguez (S. G.) et Ramirez (M. A.). « Hrtf individualization by solving the least squares problem ». Dans *Proc. 118th Convention of the Audio Eng. Soc.*, Barcelona, Spain, 2005. (cité page 99)

[220] Rodriguez (S. G.) et Ramirez (M. A.). « Linear relationships between spectral characteristics and anthropometry of the external ear ». Dans *Proc.*

Int. Conf. on Auditory Display, ICAD 2005, p. 336–339, Limerick, Ireland, 2005. (cité page 99)

[221] Roffler (S. K.) et Butler (R. A.), « Factors that influence the localization of sound in the vertical plane », *J. Acoust. Soc. Am.*, **43**(6), 1967, p. 1255–1259. (cité page 64)

[222] Rogers (M. E.) et Butler (R. A.), « The linkage between stimulus frequency and covert peak areas as it relates to monaural localization », *Percept. Psychophys.*, **52**(5), 1992, p. 536–546. (cité page 72)

[223] Romblom (D.) et Cook (B.). « Near-field compensation for HRTF processing ». Dans *Proc. 125th Convention of the Audio Eng. Soc.*, October 2008. (cité page 34)

[224] Romigh (G.) et Brungart (D.), « Real-virtual equivalent auditory localization with head motion. », *J. Acoust. Soc. of Am.*, **125**, 2009, p. 2690. (cité pages 81 et 83)

[225] Roth (G.), Kochhar (R.) et Hind (J.), « Interaural time differences : Implications regarding the neurophysiology of sound localization », *J. Acoust. Soc. Am.*, **68**, 1980, p. 1643. (cité page 11)

[226] Runkle (P.), Yendiki (A.) et Wakefield (G. H.). « Active sensory tuning for immersive spatialized audio ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2000*, Atlanta, Georgia, USA, 2000. (cité pages 97 et 99)

[227] Sandvad (J.). « Dynamic aspects of auditory virtual environments ». Dans *Proc. 100th Convention of the Audio Eng. Soc.*, Copenhagen, Denmark, 1996. (cité page 39)

[228] Schmiedek (F.), Oberauer (K.), Wilhelm (O.), Suss (H.) et Wittmann (W.), « Individual Differences in Components of Reaction Time Distributions and Their Relations to Working Memory and Intelligence », *Journal of Experiemental psychology : General*, **136**(3), 2007, p. 414. (cité page 212)

[229] Searle (C. L.), Braida (L. D.), Cuddy (D. R.) et Davis (M. F.), « Binaural pinna disparity : another localization cue », *J. Acoust. Soc. Am.*, **57**(2), 1975, p. 448–455. (cité page 68)

[230] Seeber (B. U.) et Fastl (H.). « Subjective selection of non-individual head-related transfer function ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2003*, p. 259–262, Boston, MA, USA, 2003. (cité page 90)

[231] Shaw (E.), « Earcanal pressure generated by a free sound field », *J. Acoust. Soc. Am.*, **39**(465-470), 1966. (cité page 86)

[232] Shaw (E.), « Acoustic response of external ear with progressive wave source », *J. Acoust. Soc. Am.*, **51**, 1972, p. 150. (cité page 56)

[233] Shaw (E.). *Handbook of sensory physiology*, chaptitre The external ear, p. 455–490. Springer-Verlag, 1974. (cité pages 11 et 14)

[234] Shaw (E.). « The external ear : new knowledge ». Dans *Earmolds and associated problems, Proceedings of the Seventh Danavox Symposium*, p. 24–50, 1975. (cité pages 52, 56 et 64)

[235] Shaw (E.). « Acoustical features of the human ear ». Dans *Binaural and Spatial Hearing in Real and Virtual Environments*, p. 25–47. R. H. Gilkey and A. T. B. (Erlbaum, Mahwah, NJ), 1997. (cité pages 56 et 63)

[236] Shaw (E.) et Teranishi (R.), « Sound pressure generated in an external-ear replica and real human ears by a nearby point source », *J. Acoust. Soc. Am.*, **44**(1), 1968, p. 240–249. (cité pages 52, 53, 55, 56 et 63)

[237] Shi (J.) et Malik (J.), « Normalized Cuts and Image Segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, p. 888–905. (cité pages 169 et 170)

[238] Shimada (S.), Hayashi (N.) et Hayashi (S.), « A clustering method for sound localization transfer functions », *J. Audio Eng. Soc.*, **42**(7/8), 1994, p. 577–584. (cité page 91)

[239] Shinn-Cunningham (B.). « Learning Reverberation : Considerations for Spatial Auditory Displays ». Dans *Proc. Int. Conf. on Auditory Displays, ICAD 2000*, p. 126–134, 2000. (cité page 34)

[240] Shinn-Cunningham (B.), Santarelli (S.) et Kopco (N.), « Tori of confusion : Binaural localization cues for sources within reach of a listener », *J. Acoust. Soc. Am.*, **107**, 2000, p. 1627. (cité page 22)

[241] Silzle (A.). « Selection and tuning of HRTFs ». Dans *Proc. 112th Convention of the Audio Eng. Soc.*, Munich, Germany, 2002. (cité pages 97 et 99)

[242] Sorgi (L.) et Daniilidis (K.), « Normalized cross-correlation for spherical images », *Lecture Notes in Computer Science*, 2004, p. 542–553. (cité page 162)

[243] Stan (G.), Embrechts (J.) et Archambeau (D.), « Comparison of different impulse response measurement techniques », *J. Audio Eng. Soc.*, **50**(4), 2002, p. 249–262. (cité page 32)

[244] Stein (D.), Scheinerman (E.) et Chirikjian (G.), « Mathematical models of binary spherical-motion encoders », *IEEE ASME Transactions on Mechatronics*, **8**(2), 2003, p. 234–244. (cité page 133)

[245] Sterbing (S.), Hartung (K.) et Hoffman (K.-P.), « Spatial tuning to virtual sounds in the inferior colliculus of the guinea pig », *J. Neurophysiol.*, **90**, 2003, p. 2648–2659. (cité page 85)

[246] Tan (C.-J.) et Gan (W.-S.), « User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems », *Electronic Letters*, **34**(25), 1998, p. 2387–2389. (cité pages 97 et 98)

[247] Tao (Y.), Tew (A. I.) et Porter (S. J.). « The differential pressure synthesis method for estimating acoustic pressure on human heads ». Dans *Proc. 112th Convention of the Audio Eng. Soc.*, volume Preprint 5594, Munich, Germany, 2002. (cité page 88)

[248] Teranishi (R.) et Shaw (E. A. G.), « External-ear acoustic models with simple geometry », *J. Acoust. Soc. Am.*, **44**(1), 1967, p. 257–263. (cité pages 53 et 64)

[249] Theile (G.), « On the standardization of the frequency response of high-quality studio headphones », *J. Audio Eng. Soc.*, **34**(12), 1986. (cité page 42)

[250] Torres (J. C. B.), Petraglia (M. R.) et Tenenbaum (R. A.). « Low-order modeling and grouping of hrtfs for auralization using wavelet transforms ». Dans *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, 2004. (cité page 37)

[251] Umeyama (S.), « Least-squares estimation of transformation parameters between twopoint patterns », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **13**(4), 1991, p. 376–380. (cité pages 122 et 257)

[252] Van Opstal (A. J.) et Van Esch (T.), « Estimating spectral cues underlying human sound localization », *NAG-journaal*, **168**, 2003, p. 1–10. (cité page 73)

[253] Van Wanrooij (M.). *Monaural adaptative mechanisms in Human sound localization*. Thèse de doctorat, Radboud University Nijmegen, 2006. (cité page 101)

[254] Van Wanrooij (M.) et Van Opstal (A. J.), « Sound localization under perturbed binaural hearing », *J. Neurophysiol.*, **97**, 2006, p. 715–726. (cité page 83)

[255] Van Wanrooij (M.) et Van Opstal (A.), « Contribution of head shadow and pina cues to chronic monaural sound localization », *J. Neurosci.*, **24**(17), 2004, p. 4163–4171. (cité page 67)

[256] Viemeister (N.) et Wakefield (G.), « Temporal integration and multiple looks », *J. Acoust. Soc. Am.*, **90**, 1991, p. 858. (cité page 65)

[257] Vliegen (J.) et Van Opstal (A. J.), « The influence of duration and level on human sound localization », *J. Acoust. Soc. Am.*, **115**(4), 2004, p. 1705–1713. (cité pages 65 et 66)

[258] Völk (F.), Heinemann (F.) et Fastl (H.), « Externalization in binaural synthesis : effects of recording environment and measurement procedure. », *J. Acoust. Soc. Am.*, **123**(5), 2008. (cité page 85)

[259] von Bekesy (G.), *Experiments in hearing*. McGraw-Hill Book Company, 1960. (cité page 21)

[260] von Luxburg (U.), « A tutorial on spectral clustering », *Statistics and Computing*, **17**(4), 2007, p. 395–416. (cité page 168)

[261] Wahba (G.), « Spline interpolation and smoothing on the sphere », *SIAM J. Sci. Stat. Comp.*, **2**, 1981, p. 5–16. (cité pages 253 et 254)

[262] Wallach (H.), « On sound localization », *J. Acoust. Soc. Am.*, **10**, 1939, p. 270–274. (cité page 20)

[263] Wallach (H.), « The role of head movements and vestibular and visual cues in sound localization », *J. Exptl. Psychol.*, **27**, 1940, p. 339–346. (cité page 20)

[264] Watkins (A. J.), « Psychoacoustical aspects of synthesized vertical locale cues », *J. Acoust. Soc. Am.*, **63**(4), 1978, p. 1152–1165. (cité page 70)

[265] Wenzel (E.), Wightman (F.), Kistler (D.) et Foster (S.), « Acoustic origins of individual differences in sound localization behavior », *J. Acoust. Soc. Am.*, **84**, 1988, p. S79. (cité pages 83, 90 et 239)

[266] Wenzel (E. M.), Arruda (M.), Kistler (D. J.) et Wightman (F. L.), « Localization using non-individualized head-related transfer functions », *J. Acoust. Soc. Am.*, **94**, 1993, p. 111–123. (cité pages 83 et 85)

[267] Wenzel (E.), « Localization in virtual acoustic displays », *Presence : Teleoperators and Virtual Environments*, **1**(1), 1992, p. 80–107. (cité page 41)

[268] Wenzel (E.). « Effects of increasing system latency on localization of virtual sounds ». Dans *Proc. 16th Conference of the Audio Eng. Soc.*, Rovaniemi, Finland, 1998. (cité page 39)

[269] Wenzel (E.), Wightman (F.) et Foster (S.). « A virtual display system for conveying three-dimensional acoustic information ». Dans *Proc. Human Factors Soc. 32nd Ann. Meeting*, 1988. (cité page 90)

[270] Wightman (F.), Kistler (D.) et Zahorik (P.). *Usability Evaluation and Interface Design : Cognitive Engineering,Intelligent Agents and Virtual Reality*, chaptitre Issues and non-issues in the production of high-resolution auditory virtual environments, p. 604–608. Smith, M.J. and Salvendy, G. and Harris, D. and Koubek, R.J., 2001. (cité page 42)

[271] Wightman (F.) et Kistler (D. J.). « Individual differences in human sound localization behavior ». Dans *131st Meeting Acoustical Society of America*, p. 2470, 1996. (cité page 90)

[272] Wightman (F.) et Kistler (D.), « Headphone simulation of free-field listening I : Stimulus synthesis », *J. Acoust. Soc. Am.*, **85**, 1989, p. 868–878. (cité pages 41, 83, 86 et 118)

[273] Wightman (F.) et Kistler (D.), « Headphone simulation of free field-listening II : Psychophysical validation », *J. Acoust. Soc. Am.*, **85**(2), 1989, p. 868–878. (cité pages 24, 41, 81, 83, 90 et 181)

[274] Wightman (F.) et Kistler (D.), « The dominant role of low-frequency interaural time differences in sound localization », *J. Acoust. Soc. Am.*, **91**, 1992, p. 1648–1661. (cité page 34)

[275] Wightman (F.) et Kistler (D.). « Multidimensional scaling analysis of head-related transfer functions ». Dans *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1993. (cité pages 83 et 91)

[276] Wightman (F.) et Kistler (D.), « Resolution of front-back ambiguity in spatial hearing by listener and source movement », *J. Acoust. Soc. Am.*, **105**(5), 1999, p. 2841–2853. (cité page 20)

[277] Wightman (F.) et Kistler (D.), « Measurement and Validation of Human HRTFs for Use in Hearing Research », *Acta Acustica united with Acustica*, **91**(3), 2005, p. 429–439. (cité page 42)

[278] Woodworth (R.) et Schlosberg (H.), *Experimental Psychology*. Methuew, 1954. (cité page 12)

[279] www.wikipedia.com (W.). (cité pages 252, 266, 267 et 270)

[280] Yairi (S.), Iwaya (Y.) et Suzuki (Y.), « Influence of large system latency of virtual display on behavior of head movement in sound localization task », *Acta Acustica united with Acustica*, **94**, 2008, p. 1016–1023. (cité pages 182 et 209)

[281] Zahorik (P.), « Assessing auditory distance perception using virtual acoustics », *J. Acoust. Soc. Am.*, **111**, 2002, p. 1832. (cité pages 21 et 24)

[282] Zahorik (P.). « Auditory display of sound source distance ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2002*, p. 2–5, 2002. (cité pages 24 et 26)

[283] Zahorik (P.), Kistler (D.) et Wightman (F.). « Sound localization in varying virtual acoustic environments ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 1994*, p. 179–186, 1994. (cité page 33)

[284] Zahorik (P.), Tam (C.), Wang (K.), Bangayan (P.) et Sundareswaran (V.). « Localization accuracy in 3-d sound displays : The role of visual-feedback training ». Dans *Proc. Advanced Displays and Interactive Displays ARL Fed. Lab. 5th Ann. Symp.*, p. 17–22, College Park, MD, 2001. (cité page 101)

[285] Zakarauskas (P.) et Cynader (M. S.), « A computational theory of spectral cue localization », *J. Acoust. Soc. Am.*, **94**, 1994, p. 1323–1331. (cité page 74)

[286] Zhang (M.), Tan (K.-C.) et Er (M. H.), « Three-dimensional sound synthesis based on Head-Related Transfer Functions », *J. Audio Eng. Soc.*, **46**(10), 1998, p. 836–844. (cité page 97)

[287] Zotkin (D.), Duraiswami (R.) et Davis (L.). « Customizable auditory displays ». Dans *Proc. Int. Conf. on Auditory Display, ICAD 2002*, Kyoto, Japan, 2002. (cité pages 91 et 99)

[288] Zotkin (D.), Hwang (J.), Duraiswami (R.) et Davis (L.). « HRTF personalization using anthropometric measurements ». Dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2003*, New Patlz, NY, USA, 2003. (cité pages 91 et 99)

[289] Zwislocki (J.) et Feldman (R.), « Just noticeable differences in dichotic phase », *J. Acoust. Soc. Am.*, **28**, 1956, p. 860. (cité page 11)

Résumé

Le travail de thèse qui est rapporté dans le présent document a porté sur le problème de l'individualisation des HRTF pour la synthèse binaurale. Les HRTF sont les filtres linéaires, chacun associé à une direction de l'espace, qui portent en eux l'expression de tous les indices physiques de localisation nécessaires pour une perception de l'espace par le système auditif. La synthèse binaurale utilise avantageusement ces filtres pour sculpter les signaux à présenter aux tympans de l'auditeur, afin de lui procurer l'illusion d'une scène sonore réaliste. Les HRTF étant très liées à la morphologie de la tête et des pavillons, la spatialisation n'est correctement assurée que si ces filtres sont bien adaptés à l'auditeur. Cependant, la mesure exhaustive des HRTF est coûteuse et inconfortable, et il s'agit donc de développer des moyens alternatifs pour les obtenir : c'est le problème de l'individualisation. On se focalise sur les indices spectraux de la localisation auditive, c'est-à-dire les colorations du spectre à dépendance directionnelle, qui constituent la part des HRTF la plus complexe et la plus variable d'un individu à l'autre.

Le constat fondateur de nos investigations est le suivant : bien que les HRTF présentent des caractéristiques intrinsèquement individuelles, on peut dégager des évolutions spatio-fréquentielles de leur spectre d'amplitude, communes d'un individu à l'autre, mais susceptibles d'être masquées par deux sources importantes de variabilité, que sont la taille et l'orientation des pavillons. Nous proposons des outils permettant de dépasser ces différences apparentes, afin de se focaliser sur ce qui est vraiment spécifique à chaque individu. Deux solutions techniques d'individualisation des HRTF sont développées en utilisant avantageusement la diversité des comportements offerte par les HRTF d'une base de données.

La première solution proposée permet d'adapter, pour un nouvel auditeur, les HRTF d'un autre individu issues d'une base de données, en leur appliquant des transformations guidées par une comparaison morphologique entre les pavillons des deux sujets. Les hypothèses de travail et les outils proposés pour mettre en œuvre la technique sont validés objectivement grâce aux données recueillies sur 6 sujets, et on montre que la méthode d'adaptation proposée dépasse les performances de l'état de l'art.

La seconde solution permet de reconstruire les HRTF d'un nouvel auditeur pour une direction quelconque de l'espace à partir d'un nombre réduit de HRTF individuelles mesurées. La technique proposée est basée sur une base de données constituée des HRTF mesurées finement sur une centaine de sujets, à partir desquelles on génère des prototypes. La reconstruction des HRTF repose sur un processus de reconnaissance de formes entre les HRTF individuelles mesurées et ces prototypes. Une validation objective montre que, selon différents critères, les performances de reconstruction de la technique proposée dépassent celles de l'état de l'art. Ces résultats sont confirmés par une évaluation subjective, menée selon un protocole novateur en synthèse binaurale dynamique.

Abstract

This Ph.D. thesis deals with the problem of Head-Related Transfer Functions (HRTFs) individualization, in the context of binaural synthesis. HRTFs embed all the acoustical phenomena occurring on the path between a source at a given position in space and the listener's eardrums. As these linear filters convey all free field localization cues needed by the auditory system to perceive a 3D sound scene, HRTF can be used to sculpt the signals to be reproduced over headphones in order to create convincing spatialized auditory displays : this is the aim of binaural synthesis. HRTFs strongly depend on idiosyncratic morphological features (overall shape of the head, fine structure of the pinnae), and as a result, the use of non-individual HRTFs often leads to perceptual artifacts. Unfortunately, exhaustive acoustic measurements of individual HRTFs are long and uncomfortable for subjects, and it is therefore expected to develop alternative techniques to obtain customized HRTFs : this is the problem of individualization. As they represent the most complex and the most individual part of HRTFs, our study focusses on the colorations induced by pinna filtering, known as spectral cues.

The founding assumption of our work is the following : although HRTFs contain intrinsically individual features, common spatio-frequential behaviours can be found from subject to subject. Such similarities may be hidden by the existence of two morphological sources of variability, being the size and orientation of ear pinnae. We develop tools whose aim is to go beyond apparent differences, and to focus on what is really specific of each individual. We propose two technical solutions for HRTF individualization, based on the use of a HRTF database.

The first solution uses a 3D model-based morphological matching of pinnae shapes, to properly adapt existing non-individual HRTFs from a database, so that they fit to a new listener. To transform HRTF data, we propose a combination of frequency scaling and rotation shift, whose parameters are predicted by the result of the morphological comparison. The method is designed on the basis of data acquired from six subjects, and it is shown objectively that a better customization is achieved compared to the state-of-the-art technique.

The second solution aims at reconstructing HRTF for any direction, from only sparse individual HRTF measurements. In order to overcome the performance of classical blind interpolation techniques, additional knowledge is injected in the reconstruction process : HRTF prototypes are first extracted from the analysis of a large HRTF database, and serve as a well-informed background in a pattern recognition process. An objective assessment shows that, compared to previously developped techniques, HRTF reconstruction achieves a better spatial fidelity with the proposed method. Finally, this result is confirmed by a subjective evaluation based on a new protocol.