École nationale supérieure Louis-Lumière
Promotion Son 2016

Optimisation du traitement binaural interactif pour le mixage orienté objet.

Mémoire de fin d'étude
réalisé par François Salmon

Directeurs externes : Jean-Christophe Messonnier, Matthieu Aussal
Directeur interne : Laurent Millot
Rapporteur : Alan Blum

Je tiens à remercier les personnes suivantes
pour leur encadrement et leur soutien :

Jean-Christophe Messonnier
Matthieu Aussal
Laurent Millot
François Alouges
Etienne Hendrickx
Victorine Cesbron
Le 36/38
Le Fautoir

Résumé

Le mixage orienté objet offre la possibilité de fournir un encodage spatial permettant de s'adapter à la multiplication des dispositifs de diffusion. Cette exigence de versatilité peut nécessiter l'usage de plusieurs dispositifs de contrôle pour assurer la compatibilité d'un mixage. Le rendu binaural interactif (assisté d'un *head-tracker*) pourrait constituer une aide utile pour les professionnels et un nouveau support d'écoute pour les particuliers étant donné son coût et sa facilité de mise en œuvre. Or, cette technologie présente des colorations spectrales notables dues à la reproduction d'indices spatiaux et pâtit de sa comparaison courante avec la stéréophonie au casque. Il a donc été entrepris de mettre en place des méthodes d'optimisation du traitement binaural afin de réaliser un meilleur compromis entre la sensation d'espace et la fidélité du timbre. Dans le but d'évaluer la pertinence des traitements réalisés et d'identifier des axes de développement futurs, des tests perceptifs en lien avec la capacité d'externalisation et la perception du timbre sont présentés.

Mots-clés

binaural, HRTF, mixage orienté objet, monitoring, timbre, coloration, externalisation, traitement du signal

Abstract

Object-based audio mixing can provide a spatial encoding for audio contents to suit various multichannel setups. This versatility requirement may involve several monitoring devices during a post-production to ensure the compatibility of a mix. Head-tracked binaural processing could then be helpful for professionals and be a new device for individuals to listen to spatialized contents given its low cost and its ease of implementation. However, this technology provides significant spectral coloration due to direction-dependent features of the processing and suffers from its current comparison to the reproduction of a stereophonic signal through headphones. Therefore, different methods for designing modified filters are proposed to optimize the binaural processing and achieve a better balance between the externalization of sounds and the timbral coloration. In order to assess the accuracy of such treatments and to identify future paths of development, perceptual tests related to the externalisation and timbre perception are presented.

Keywords

binaural, HRTF, object-based audio, monitoring, timbre, coloration, externalization, signal processing

Table des matières

Introduction Générale

Les changements d'habitudes des consommateurs initient souvent des développements technologiques. Ces dernières années, les habitudes d'écoute ont évolué de façon radicale. Bien que l'écoute mobile ne soit pas nouvelle, la miniaturisation de la technologie ainsi que la connectivité généralisée ont multiplié les possibilités d'écoute. Il est possible d'écouter un même média sur un téléphone portable, une tablette, à travers des haut-parleurs de salon ou dans une voiture avec des systèmes de restitution toujours plus performants. Les conditions d'écoutes très différentes s'éloignent cependant énormément d'un cas idéal de diffusion où des haut-parleurs seraient disposés de manière standard dans une acoustique contrôlée. Aussi, devant la multiplication des configurations multicanales de diffusion dans les salles de cinéma, de concert ou encore sur les équipements mobiles, plusieurs constructeurs (Dolby, DTS, etc.) ont développé des formats pour diffuser le signal audio en trois dimensions spatiales. A terme, ces formats seront accessibles aux particuliers. D'ici là, les ingénieurs du son ont besoin d'outils pour se former et réaliser des mixages en son multicanal. Les évolutions technologiques permettent de nouvelles perspectives notamment par l'écoute au casque. Les techniques binaurales utilisent les filtrages correspondant à ceux introduits par la tête et les pavillons des oreilles en fonction de la provenance du son. On peut, grâce à ces filtrages, donner à entendre au sujet un espace sonore perçu en dehors de sa tête, dans l'espace ambiant et ceci directement à partir d'un casque.

Le rendu binaural pourrait constituer une solution de mixage pour les professionnels qui désirent s'équiper sans investir dans un système complet sur enceintes afin de s'adapter aux mutations professionnelles à venir. Ce moyen de restitution présente également un intérêt économique pour les particuliers désireux de ressentir un espace sonore environnant tout en ayant un rendu proche de ceux qu'ils côtoient habituellement : la stéréophonie sur enceintes ou au casque.

Dans l'état actuel des avancées technologiques, le binaural ne peut rivaliser avec une écoute réelle sur enceintes en termes de perception de l'espace, précision de la localisation ou stabilité de l'image. D'autre part, sur le même support d'écoute (le casque), un signal binaural en comparaison au signal stéréophonique présente de fortes colorations spectrales notables.

Avec le mixage orienté objet, un même contenu audio est destiné à être diffusé sur plusieurs systèmes de restitution. De tels dispositifs de reproduction doivent permettre une restitution aussi fidèle que possible des contenus sonores mixés par un ingénieur du son. Le binaural doit lui aussi assurer une fidélité vis-à-vis du mixage en procurant *une* sensation d'espace. Cette technologie est considérée ici seulement comme un moyen de reproduction sonore et non comme la technologie permettant de reproduire notre sensation réelle de l'espace. Elle présente simplement une puissance de restitution étant donné son coût et sa facilité de mise en œuvre.

Dans le premier chapitre de ce mémoire, on s'intéresse à la question de l'intégration d'un contrôle de mixage binaural interactif (assisté d'un *head-tracker*) dans le cadre d'une production en mixage orienté objet. Dans le second chapitre, on étudie dans quelle mesure on peut améliorer le rendu binaural afin de minimiser la coloration spectrale introduite par le filtrage binaural. Le dernier chapitre consiste en une confrontation entre les traitements effectués avec les jugements d'ingénieur du son afin d'en déduire des axes de développement futurs.

CHAPITRE 1

Le mixage orienté objet

La production en mixage orienté objet est une production pensée pour une restitution sur plusieurs systèmes différents grâce à un codage audio par objets et donc indépendant de tout système de reproduction. On observe une diversification de l'offre relative à la diffusion sonore et l'intérêt de ce type de mixage est de répondre aux choix d'écoute de l'auditeur de manière cohérente en s'adaptant au système de reproduction. La production par objets vise à définir un cadre flexible afin de restituer une production avec le maximum de cohérence possible sur une grande variété de systèmes de diffusion. La restitution doit respecter la production originale, ce qui peut être très variable d'un système à un autre ou d'une production à une autre. Le mixage orienté objet permet d'autre part d'être plus efficace. Si on prend l'exemple du cinéma, le son d'un film à grand budget doit être actuellement mixé pour plusieurs standards de diffusion. On peut imaginer limiter le nombre de mixages successifs en adoptant une approche de production par objet.

Outre les formes communes que sont la stéréophonie et le 5.1, d'autres formes en lien avec la mobilité ou la spatialisation du son ont émergé. On peut citer l'écoute au casque qui s'est grandement développée avec l'arrivée des *smartphones*. Ou encore la spatialisation au cinéma, en plein essor avec le développement de formats audio tels que le Dolby Atmos,

l'Auro3D ou le DTS :X. Des systèmes interactifs tels que l'Oculus Rift proposent également une autre forme de restitution du son dans le domaine de la réalité augmentée et du jeu vidéo.

Ces formats récents de description de sessions audio 3D permettent d'ajouter des métadonnées au contenu audio afin de décrire une scène sonore indépendamment du système de reproduction. Ces informations rassemblent notamment les coordonnées polaires pour chacun des objets sonores. On peut ainsi envisager d'obtenir un enregistrement à la fois diffusable sur un système stéréophonique et une installation à 24 canaux. La scène sonore obtenue est alors plus ou moins définie spatialement selon les limites du système considéré.

Après l'encodage, il s'agit de chercher à optimiser la répartition spatiale du son afin de conserver une cohérence spatiale satisfaisante pour chaque système de reproduction. L'enjeu des industriels est d'adapter un contenu encodé dans un certain format de son 3D à une installation sonore donnée en fonction de ses contraintes spatiales. Mais plutôt que de réaliser un mixage spécifique à chaque reproduction on envisage alors un format global compatible. Une post-production en mixage orienté objet doit alors prendre en considération cette exigence de versatilité qui peut influer sur toute la chaine de production, de la prise de son au dispositif d'écoute.

Ce chapitre ne porte pas sur les différents standards et formats de diffusion, mais présente les enjeux liés au format objet en s'appuyant sur l'exemple du format ADM, gratuit et open-source ; contrairement aux autres formats bien obscurs cités précédemment. Nous verrons que la non-standardisation de l'écoute contraint à prendre des précautions s'agissant des objets sonores à utiliser ainsi que de leur position dans l'espace et qu'elle implique l'utilisation de plusieurs dispositifs d'écoute. Nous testerons l'hypothèse selon laquelle le binaural interactif pourrait constituer un candidat utile et flexible afin de contrôler le mixage et l'enregistrement dans le cadre d'une production par objets.

1.1 L'exemple de l'Audio Definition Model

Le format ADM fut élaboré sous l'égide de l'Union Européenne de Radio Télévision (UER-EBU) [1]. Gratuit et ouvert, il est un bon candidat pour unifier les technologies issues de DTS, Dolby et Auro Technologies. C'est une extension du *Broadcast Wave File* (BWF) auquel il ajoute des meta-données de position. Le contenu audio qu'il décrit peut être défini en terme de canaux, de scènes ou d'objets.

Les objets sonores représentent un son comme un élément distinct auquel on ajoute une information de position fonction du temps. Ces objets peuvent alors être reproduits au plus près de l'azimuth et l'élévation spécifiés pour un système de diffusion.

Une description par scènes audio est quant à elle, une représentation d'un champ sonore selon plusieurs sources, l'Ambisonics, ou le HOA, étant la technique la plus répandue pour ce genre de description. On obtient un encodage du son par des composantes nécessaires à la création de champ sonore. Tous les sons, directs et réverbérés sont corrélés et arrivent en un point d'écoute donné. De tels encodages sont totalement indépendants de la répartition des enceintes dans l'espace. L'Ambisonic n'est pas la seule technique capable de définir une scène puisque plusieurs objets sonores peuvent constituer une scène audio, bien que celle-ci ne représente pas forcément une scène sonore réelle.

Enfin, lorsque l'on encode le son sous forme de canaux discrets, chacun d'entre eux est alors assigné à une position fixe dans l'espace. A partir de l'ensemble du mixage, on peut retrouver des formats classiques tels que la stéréophonie, le 5.1 ou encore le 22.2 NHK. Notons que les canaux peuvent être considérés comme des objets sonores à une position fixe et que les canaux Ambisonic peuvent être considérés comme des objets. De même, si un objet sonore n'est pas en mouvement et que sa position correspond à celle d'un canal standard, il peut être encodé comme tel. Ainsi, les trois approches peuvent être combinées dans l'encodage de l'audio selon les besoins de l'utilisateur. C'est pourquoi on parle plus généralement d'*Object Based Audio* pour désigner la matière première du mixage orienté objet.

Cette approche est différente de l'encodage en stéréophonie dans la mesure où une scène stéréophonique écoutée au casque ne respecte pas la même disposition des sources qu'une scène stéréophonique sur enceintes. La position des sources sonores réelles est différente : 90° dans le premier cas, 30° dans le second. Si nous voulons écouter la même scène sonore au casque et sur deux enceintes, et ainsi garder la même position pour les sources, la synthèse binaurale doit être employée en filtrant chaque source par le filtre binaural correspondant. En l'occurrence ± 30 ° en azimuth pour virtualiser un contenu stéréophonique.

Des industriels tels que Dolby, DTS ou Fraunhofer, ont travaillé sur des technologies de rendu permettant d'adapter un contenu audio ADM à un système de reproduction donné. L'ADM s'ajoute ainsi à la liste des formats audio 3D développés pour le cinéma qui adoptent une approche d'encodage similaire. Quels que soient les constructeurs, ces formats combinent

des méthodes de mixage basées sur les canaux discrets avec un module de mixage dynamique manipulant des objets sonores.

Le système de reproduction utilisé pour le monitoring pendant la prise de son et le mixage peut alors être un ensemble d'enceintes ou encore un moteur de rendu binaural. Plusieurs méthodes sont envisageables mais certaines conditions doivent être prises en compte afin d'assurer la versatilité du mixage.

1.2 L'enjeu de la compatibilité

Un enregistrement encodé en mixage orienté objet peut être diffusé sur tout dispositif multicanal discret, en synthèse par front d'ondes (WFS) ou encore au travers d'un moteur binaural. L'azimuth et l'élévation de chaque objet seront alors réglés aussi proches que possible de l'azimuth et de l'élévation définis dans l'encodage ADM. Certains systèmes de restitution seront ainsi plus ou moins performants en termes de precision de localisation, posséderont une zone d'écoute plus ou moins large ou permettront un plus grand démasquage des sources. L'intérêt d'une telle approche réside dans le fait de nécessiter un unique mixage optimisé pour différents systèmes.

L'ADM ne décrit que les informations spatiales et non comment traiter les objets sonores lors de la diffusion sur un système donné. En effet, les formats orientés objets ne précisent pas forcément la méthode de rendu. Des problèmes de débit peuvent également apparaître en raison du nombre conséquent d'objets pouvant figurer dans un mix objet. Ce sont les technologies développées par les industriels qui gèrent les réductions de débits et prennent en charge le traitement destiné à adapter un contenu orienté objet à un système de reproduction considéré. Chaque solution est *a priori* différente et l'algorithme de rendu (VBAP, DBAP, WFS, *etc...*) peut rester inconnu selon la technologie de décodage utilisée.

On peut alors se demander comment concilier versatilité et qualité du résultat. Considérons le cas de la stéréophonie qui s'est largement généralisée à tous les systèmes de restitution communs. Cette généralisation a permis de travailler avec précision sur la profondeur, le relief ou encore les effets de masquage. La non standardisation de l'écoute remet en cause un savoir faire. Si le système d'écoute n'est plus fixe, comment espérer atteindre un niveau de qualité similaire ? Ces propos peuvent être nuancés dans la mesure où la standardisation de la stéréophonie n'est que partielle. En effet, son utilisation chez l'auditeur, dans son salon ou sa voiture, n'est pas normée en terme de disposition des enceintes, de qualité des haut-parleurs

ou d'acoustique. Le passage d'une régie LEDE à une écoute domestique (supposée respecter à peu près la disposition stéréophonique standard) correspond à une augmentation d'environ 3 dB du niveau de réverbération et change donc notre perception des plans sonores [2]. Les conditions de post-production et de restitution sont effectivement bien différentes. Au cinéma, une grande variété de salles et de dispositifs d'écoute introduisent également une importante variabilité dans la restitution. On peut alors estimer que l'introduction d'une variabilité due au décodage du mixage orienté objet est minime par rapport à la variété des dispositifs.

De plus, certains systèmes d'enceintes 3D bénéficient d'une meilleure définition spatiale et d'une plus grande qualité en termes de timbre que d'autres. Ainsi, même si des avancées technologiques peuvent prétendre minimiser ces différences, les différences fondamentales entre un système binaural et une diffusion à 24 canaux persisteront certainement.

Se pose alors la question du *mastering*, cruciale pour la diffusion. Ce processus consiste à traiter une production sonore pour en faire un ensemble homogène adapté à un support de diffusion. Tous les outils nécessaires n'étant pas encore fréquemment utilisés actuellement, cette étape reste encore à préciser. La norme MPEG-H-Audio inclut néanmoins des mécanismes de contrôle de sonie, de dynamique et de vérification du true peak selon les standards ITU-R BS.1770-3 et EBU R128.

Le *monitoring*

Le *monitoring* désigne l'ensemble des éléments servant à l'écoute du mixage ou de l'enregistrement. Afin de garantir la plus grande compatibilité possible entre différents ensembles de diffusion, une solution possible est d'utiliser un nombre important de systèmes en post-production. Le *monitoring* peut alors s'effectuer en utilisant un dispositif 5.1, stéréophonique, binaural interactif, *etc...* et se poursuivre en vérifiant la compatibilité sur différents systèmes en fin de post-production. Ces systèmes peuvent introduire des phénomènes de masquage plus ou moins importants, notamment la réverbération qui est plus ou moins spatialisée donc plus ou moins masquante. Par exemple, en passant d'un mode très masquant comme la stéréophonie sur enceintes au binaural interactif, nous sommes en présence de deux extrêmes. Le binaural permet en effet une description précise de l'espace sonore autour de l'auditeur grâce à des méthodes d'interpolation avancées et en stéréophonie la spatialisation est la plus défavorablement restituée. On peut alors fixer des «bornes» de perception du mixage dans lesquelles l'auditeur va évoluer suivant le système qu'il utilise. Deux formats peuvent donc

être privilégiés pour trouver un compromis acceptable : la stéréophonie et le binaural interactif.

Au Conservatoire national supérieur de musique et de danse de Paris (CNSMDP), grâce aux outils développés en interne, il est possible de restituer un mixage en cours sur trois systèmes différents : en 5.1, 2.0 et binaural interactif. Notons que ces diffusions sont issues du même mixage effectué en binaural objet et que le rendu sur canaux discrets est réalisé en transaural.

Lorsque l'on diffuse un mixage sur différents systèmes, il se peut que la scène sonore ne soit pas fidèlement restituée. En effet, même si tous les objets sonores se situent à l'endroit spécifié par l'encodage, ils ne correspondent pas forcement à l'emplacement d'un haut-parleur.

Les images fantômes

En stéréophonie on utilise la capacité de la perception à créer des sources, ou images, fantômes issues des signaux provenant des deux enceintes. La perception sonore génère des sources dans une direction intermédiaire entre les haut parleurs où il n'y a pourtant aucune émission physique sonore, d'où la qualification de fantôme. Ces sources fantômes ont des localisations dépendantes de la position de l'auditeur vis-à-vis des enceintes et dépendent des différences existantes entre les signaux issus des haut-parleurs. Leur positionnement est fonction d'une différence d'intensité ou de temps entre ces signaux. Les différences peuvent être créées par des systèmes de prise de son ou par des traitements. Par exemple, le potentiomètre panoramique joue sur l'intensité des signaux stéréophoniques.

Selon Pliege et Theile [3], la création d'une image fantôme s'effectue en deux temps. Si on considère un son diffusé sur deux enceintes : l'auditeur perçoit alors d'abord la position des sources sonores réelles, celles des haut-parleurs, puis il situe une source fantôme située entre les sources réelles d'après les informations de localisation perçues. En fonction de la distance qui sépare les deux enceintes, la localisation de cette source résultante peut être plus ou moins précise. C'est pourquoi un système de diffusion dense tel que le 22.2 permet une localisation plus précise des sources qu'un système 5.1 car la localisation d'une source réelle est plus fiable que celle d'une image fantôme. De plus, les sources réelles gardent leur position absolue lorsque l'auditeur est en mouvement alors que les images fantômes se déplacent. Le 22.2 permet donc également une meilleure stabilité en ce sens. Ainsi, si les objets sonores

sont proches les uns des autres, la perception de l'image fantôme sera alors d'autant plus précise et stable. On aura alors tout intérêt à multiplier le nombre de sources dans les zones pour lesquelles une plus grande précision spatiale est requise. Dans le cadre d'une production musicale par exemple, les canaux sont principalement concentrés sur la scène frontale, car les instruments sont frontaux. Les canaux latéraux ou arrières sont utilisés pour l'acoustique de la salle et les réverbérations artificielles.

En mixage orienté objet, la perception du son s'effectue selon deux niveaux d'interpolation. Une image fantôme se forme si un son est défini par plusieurs objets, que plusieurs sons corrélés coexistent dans la scène ou que l'emplacement des objets ne correspond pas à ceux des haut-parleurs. La stabilité et la précision sont donc défavorisées si les objets ne se situent pas à l'emplacement des haut-parleurs. Et, plus le nombre de canaux est faible, plus il est probable que ce soit le cas.

Un mixage objet constitué d'un nombre important de pistes peut toujours être lu sur un système moins défini en transformant les objets supplémentaires en sources fantômes codées entre les canaux de diffusion, moyennant une perte de précision et de stabilité. Il faut en revanche vérifier que ce passage aux sources fantômes ne pose pas de problème d'interactions négatives entre les objets. On peut notamment constater que lorsque des objets sonores sont trop corrélés entre eux, leur mélange peut se traduire par l'introduction d'un filtrage en peigne important. De même, des problèmes de phase conséquents peuvent apparaitre lorsque l'on considère des prises de sons reposant seulement sur les différences de temps. Celles intégrant à la fois des différences de niveau ΔI et de temps ΔT sont plus robustes vis-à-vis de la création de source fantômes.

1.3 L'utilisation du binaural

Le *monitoring* d'un dispositif multicanal ou, plus généralement, la restitution d'un espace sonore au moyen d'un simple casque, présente des subtilités dues aux mécanismes physiologiques de la localisation sonore. Celle-ci se base sur de multiples indices contenus dans les fonctions de transfert qui caractérisent les transformations subies par une onde acoustique entre le point source et les deux oreilles de l'auditeur. La synthèse binaurale permet de simuler au casque l'écoute spatialisée de sources sonores situées autour de l'auditeur en filtrant le signal émis à l'aide de ces caractéristiques. Les mécanismes en jeu dans la localisation sonore sont décrits dans la section dédiée en Annexes.

Les HRTF

La synthèse binaurale a pour objectif de reproduire et de contrôler la sensation de localisation du son pour une restitution au casque. Le traitement fait intervenir deux filtres spécifiques à chaque position de source, qui prennent en compte les transformations auxquelles le son est sujet lorsqu'il rencontre le torse, la tête et les oreilles de l'auditeur. En effet, les *Head-Related Transfer Function* (HRTF), regroupent l'ensemble des déformations subies par le son depuis la source jusqu'à notre canal auditif, en modélisant le filtrage du son par notre corps. Elles sont propres à chaque individu dans la mesure où nous disposons tous de morphologies différentes, sources de déformations, déphasages, diffractions diverses du son. Les HRTF comprennent *a fortiori* les différences interaurales de niveau et de temps, les indices spectraux et de distance. Les filtres HRTF sont alors employés afin de filtrer numériquement le signal sonore pour le situer en un point de l'espace. La synthèse binaurale consiste donc à filtrer un son monophonique par les HRTF mesurées à une incidence donnée pour reproduire la sensation d'un son provenant de cette position.

On appelle moteur de rendu l'appareil qui réalise les filtrages employant les *Head-Related Impulse Response* (HRIR), le pendant temporel des HRTF. Ce moteur de rendu peut intégrer des fonctions d'individualisation et être équipé d'un *head-tracker*, dispositif de suivi des mouvements de tête. La qualité de l'expérience sonore de l'auditeur dépend alors à la fois de la qualité des signaux traités par le moteur et du traitement réalisé par le moteur.

L'individualisation

Les problèmes les plus remarquables liés à la non individualisation de l'écoute binaurale sont un changement de couleur spectrale, une modification de l'externalisation et des erreurs de localisation. L'utilisation de réponses impulsionnelles non individuelles dans la synthèse binaurale est une problématique qui n'est pas encore résolue. Un des grands défis de la technologie binaurale réside dans la mise au point d'une méthode permettant l'individualisation de l'écoute pour un auditeur spécifique ou à trouver des réponses génériques, adaptables ou non, en s'appuyant sur la plasticité de l'audition et l'apprentissage.

Pour obtenir les HRTF d'un individu, on mesure les fonctions de transfert entre le signal émis par un haut-parleur en un point d'une sphère centrée sur la tête de l'auditeur et le signal reçu par les microphones situés dans ses oreilles. L'obtention d'HRTF sur des sujets réels peut représenter plusieurs centaines à plusieurs milliers de mesures sur une seule sphère, et donc plusieurs heures pendant lesquelles le sujet ne doit pas bouger. Il paraît donc complexe

de mesurer les HRTF à différentes distances et l'augmentation de la résolution angulaire sur une seule sphère paraît limitée. La mesure d'HRTF est une approche lourde à mettre en place tant en termes de matériel que de temps pour envisager une individualisation simple de l'écoute.

Diverses méthodes statistiques ont été utilisées pour analyser les bases de données de mesures HRTF dans le but de faire émerger une structure sous-jacente dans les données. Ces approches consistent à déterminer et identifier le nombre nécessaire et suffisant de vecteurs propres dans la base pour synthétiser des spectres de HRTF individualisées par combinaison linéaire de ces vecteurs. On peut citer l'analyse en composante principale (PCA) [4], le *Locally Linear Isomap* [5] ou le *Locally Linear Embedding* [6]. Ces méthodes permettent de réduire la dimension d'une base de données à quelques vecteurs élémentaires mais n'apportent pas d'indices quant à la combinaison d'HRTF à effectuer pour l'individualisation. On peut néanmoins imaginer ajuster le poids des vecteurs à la suite de tests d'écoute.

Des chercheurs ont développé des modèles géométriques pour le torse, la tête et les oreilles. La tête et le torse peuvent par exemple être modélisés en utilisant des ellipsoïdes [7], le pavillon peut être modélisé comme un ensemble d'objets géométriques simples [8]. Les modèles géométriques sont facilement adaptables à tout auditeur en intégrant des mesures anthropométriques dans le modèle. Cependant, bien que les modèles géométriques simplifiés soient fiables à basses fréquences, ils deviennent de plus en plus inexacts à des fréquences plus élevées. En raison de l'importance des hautes fréquences pour la localisation, les modèles géométriques simplifiés ne semblent pas adaptés pour la création de HRTF individualisées.

L'utilisation de maillage 3D de tête est une approche qui semble plus prometteuse. La représentation géométrique précise de la tête sert alors de base pour la simulation numérique de la propagation acoustique en utilisant la modélisation par éléments finis [9,10]. Avec cette méthode, les HRTF peuvent être déterminées par le calcul avec une précision équivalente aux mesures acoustiques, même à des fréquences élevées. Les méthodes de modélisation tridimensionnelle de la tête par balayage ou photographie restent cependant compliquées à mettre en œuvre pour l'obtention d'un maillage ayant une résolution spatiale adaptée aux calculs. Cependant, un modèle de tête peut être créé à partir d'éléments finis et déformé par la suite selon un ensemble de mesures anthropométriques pour s'adapter à un auditeur. L'objectif consisterait alors à élaborer un système qui pourrait calculer des HRTFs individualisées sur la base de quelques images de la tête et des oreilles du sujet. Encore faut-il développer un modèle de tête correctement adaptable à tout auditeur et aussi estimer les

caractéristiques anthropomorphiques propres à partir d'images.

L'externalisation

La perception en profondeur des sons dépend, entre autres, de l'externalisation : la scène sonore se déroule-t-elle à l'intérieur de ma tête ou dans l'espace situé autour de moi et à quelle distance de moi ? L'externalisation est la capacité à placer les sources sonores en dehors de la tête de l'auditeur dans un espace sonore virtuel, en fonction des informations qui lui parviennent. Contrairement à la stéréophonie écoutée au casque, où le son perçu se situe entre nos deux oreilles, le traitement binaural fait en sorte que les sons de la scène soient situés à l'extérieur de soi comme dans la réalité immédiate. Or dans cette dernière, nous utilisons nos propres HRTF et nous sommes en mouvement dans l'espace, ce qui nous permet de mieux analyser la position des sources par rapport à nous.

La perception intra-crânienne des sons est un défaut parfois mentionné lors d'une écoute binaurale. Ce phénomène de distorsion de localisation varie d'un individu à l'autre et concerne principalement les sources frontales. Elles ne sont alors pas situées devant l'auditeur mais très proches au-dessus ou derrière celui-ci voire à l'intérieur de sa tête. Les facteurs déclenchant le défaut d'externalisation sont nombreux et il n'est pas simple de les identifier clairement. D'autant plus qu'en binaural on ne sait pas si des signaux faisant sens d'un point de vue spatial pour un sujet le font pour toute ou une partie de la population. Il est donc complexe de définir précisément le phénomène.

Parmi les causes éventuelles d'un manque ou d'une absence d'externalisation, on peut cependant citer l'utilisation de filtres HRTF non individuels, la familiarité de l'auditeur avec l'écoute binaurale, le manque d'information liée à la réverbération ou encore l'utilisation d'un moteur binaural statique ne permettant pas une perception dynamique des objets sonores.

Le *head-tracking*

Le head-tracking permet la prise en compte des mouvements de la tête de l'auditeur via l'utilisation d'une camera ou l'ajout d'un ou plusieurs capteurs placés sur le casque. Ce dispositif permet une perception accrue de la provenance des sons en termes de localisation et d'externalisation. Il a été montré que l'utilisation d'un head-tracker permet de réduire fortement les confusions avant/arrière ainsi que les confusions pour les azimuts frontaux et l'élévation [11]. L'individualisation alliée à un effet de salle et au *head-tracking* permet no-

tamment de s'approcher d'une écoute réelle en champ libre [12].

L'usage du *head-tracking* permet de plus d'extraire les informations de localisation des HRTF. En bougeant la tête, on balaye une partie des positions de l'espace faisant intervenir un certains nombre de filtres. Les mouvements de tête permettent d'appréhender les informations communes et ainsi de ne garder que les indices de localisation. Ce phénomène permet alors une décoloration partielle du signal.

L'apprentissage

Notons qu'un phénomène d'apprentissage chez l'auditeur peut cependant venir contrebalancer les défauts liés à l'absence d'individualisation. En effet, l'apprentissage peut mener à une amélioration des performances de localisation.

Les études de Theile permettent de supposer que nous possédons une base de données d'indices de localisation personnelle acquise et mémorisée tout au long de notre vie [13]. Lorsque nous percevons un son dans l'espace il est décodé suivant cette base de données afin d'en déterminer l'incidence. La plascticité du système auditif peut alors permettre d'adapter ce décodage à des indices de localisation différents par un nouvel apprentissage [14].

Sur enceintes, l'auditeur a une perception externalisée car l'espace de diffusion est un espace réel - sur les plans cognitifs, auditifs et visuels - qu'il aborde avec ses HRTF personnelles et ses mouvements de tête, ce qui n'est pas forcément le cas en binaural. Notons cependant que la stéréophonie sur enceintes nécessite un apprentissage afin de gagner en précision de localisation. En stéréophonie binaurale, on peut également considérer qu'au-delà d'une externalisation effective, un apprentissage doive être mis en place afin d'assimiler les distorsions de localisation, se familiariser avec l'outil et enrichir ainsi sa base de données perceptives.

Certains contenus nécessitent plus ou moins l'utilisation d'une écoute individualisée, du *head-tracking* ou d'un apprentissage car ils contiennent une description cognitive et dynamique de l'espace sonore différente. Une scène sonore pourra alors donner des résultats sensiblement différents avec ou sans *head-tracker* ou individualisation. On peut alors affirmer que la création de la scène sonore se partage entre les signaux diffusés et le système utilisé pour la diffusion ; de même que la création de l'image stéréophonique sur enceintes est due à la fois aux signaux utilisés et au dispositif permettant sa diffusion. Le binaural correspondrait alors à un outil qu'il faut apprendre à manier tant au niveau des contenus

que de la restitution pour être en mesure d'en exploiter les possibilités offertes.

Les limites du traitement binaural pour une production par objets

Du fait de sa capacité à restituer un espace sonore en trois dimensions spatiales autour de l'auditeur et de son aspect économique non négligeable vis-à-vis d'un système équivalent sur enceintes, le binaural semble pouvoir constituer un outil adapté au mixage orienté objet. L'emploi du binaural nécessite néanmoins d'être vigilant et de prendre certaines précautions.

L'utilisation du binaural en mixage orienté objet complexifie en effet la perception dans la mesure où les HRTF utilisées ne sont pas nécessairement individualisées et qu'il n'est pas forcément possible d'appréhender les changements d'information dûs aux mouvements de tête. Deux solutions sont alors à envisager : l'individualisation des HRTF et l'intégration d'un *head-tracker*. L'avantage du mixage orienté objet tient au fait qu'il intègre des méta-données qui permettent une flexibilité vis-à-vis du traitement binaural. L'individualisation ou le *head-tracking* peuvent s'intégrer aisément.

Lors du mixage au casque, une attention particulière doit être portée également à la réverbération. En effet, un mixage effectué avec des HRTF anéchoïques aura tendance à privilégier un niveau de réverbération élevé qui, lors d'une restitution sur haut-parleurs, sera exacerbée et pourra flouter l'image. Cet effet peut être amoindri si le moteur binaural utilisé permet la simulation d'une salle d'écoute. De plus, il a d'autant été montré que l'ajout d'un effet de salle favorise l'externalisation [15].

En binaural, la position 0° n'est pas équivalente au mélange des positions $+30°$, $-30°$. En effet, vu que le traitement binaural ne simule pas parfaitement notre perception naturelle ou que les HRTF ne sont pas nécessairement individualisées, les images fantômes ne sont pas aussi bien définies qu'en stéréophonie par exemple. Les systèmes de binauralisation ne permettent pas de traiter les sources fantômes aussi facilement que sur enceintes car la stéréophonie sur enceintes repose sur nos HRTF personnelles et la variation de l'information sonore due à nos mouvements de tête. Ce phénomène est donc d'autant plus important si l'on utilise pas nos propres HRTF et que le traitement binaural est statique. Pour une prise de son donnée, selon le nombre de canaux et des modes de codage en ΔI ou ΔT, la localisation en binaural peut poser problème. Il reste à savoir si l'on peut arriver à un niveau d'individualisation et à une performance des *head-trackers* permettant de localiser facilement ces signaux ou si l'on doit se limiter à certains types de systèmes lorsque l'on souhaite une

compatibilité avec le binaural. Néanmoins, les images fantômes entre deux objets sonores introduisent des erreurs qui peuvent être minimisées si les objets sont proches les uns des autres [16]. Il s'avère donc nécessaire d'utiliser des prises de sons comportant un nombre important d'objets sonores. Outre l'exigence de compatibilité, d'autres contraintes sont également à prendre en compte pour faire le choix d'un système de prise de son et aucun d'entre eux ne présentera la solution idéale pour tous les cas. Le coût, la visibilité du dispositif, le caractère fixe ou en mouvement de la scène, le projet esthétique, *etc...* sont autant de facteurs à prendre en compte. Une des tâches du projet de recherche collaboratif Bili * (*Binaural Listening*), lancé en 2013, a été de travailler sur les techniques d'enregistrement binaural et de post-production. Les différents acteurs ont ainsi pu identifier les systèmes d'enregistrement qui s'adaptent à une écoute binaurale [16,17] et organiser plusieurs comparaisons subjectives, dont une expérience menée par Nicol [18].

Alors que la post-production en mixage orienté objet implique un contrôle régulier sur différents systèmes de diffusion, il est difficile d'en faire autant lors d'une prise de son. Il s'agit alors de maîtriser le *monitoring* sur un seul système et d'extrapoler par expérience le comportement de la prise de son sur d'autres systèmes. Le *monitoring* au casque au travers d'un traitement binaural interactif permet notamment de vérifier rapidement la spatialisation de la scène. Le binaural étant également le cas le plus défavorable vis-à-vis de la création des images fantômes, son utilisation permet à l'ingénieur du son de vérifier comment le système principal se comportera dans les conditions de restitution les plus défavorables.

1.4 Conclusions locales

Le mixage orienté objet vise à restituer un mixage dans un format compatible à une grande variété de systèmes de reproduction. Cet encodage spatial des sources sonores a pour but de diffuser le plus fidèlement possible une production malgré la variabilité des dispositifs d'écoute. L'exigence de versatilité influe sur toute la chaine de production qui impose l'utilisation de plusieurs systèmes de diffusion afin de préciser la compatibilité du mixage. Le décodage d'un mixage constitué d'un grand nombre d'objets sonores peut s'adapter à un dispositif moins défini par la création d'images fantômes entre les canaux de diffusion. Il faut alors s'assurer que ce passage aux sources fantômes ne pose pas de problème d'interactions négatives entre les objets. L'usage de prises de sons intégrant à la fois des différences de niveau ΔI et de temps ΔT et non des différences en ΔT seules, l'utilisation d'objets décorrélés entre eux ou une description spatiale dense dans les zones significatives sont autant de

∗. http ://www.bili-project.org/

solutions permettant une robustesse vis-à-vis de ces sources fantômes, sources d'imprécision et d'instabilité.

La stéréophonie, un dispositif très masquant du fait de son faible nombre de canaux, peut représenter un système permettant d'identifier efficacement les intéractions négatives dues à l'apparition de source fantôme. Le binaural interactif quant à lui, par sa capacité à décrire de manière précise l'espace peut constituer une autre «borne» de perception en ce sens. L'individualisation de l'écoute binaurale alliée à un effet de salle et au *head-tracking* permet notamment de s'approcher d'une écoute réelle en champ libre. Il permet de répartir les sources dans les trois dimensions spatiales et ainsi favoriser le démasquage.

Au delà d'une recherche d'individualisation, la plasticité de l'audition et l'apprentissage constituent des pistes prometteuses pour pallier aux erreurs de localisation ou à l'altération du sentiment d'externalisation. L'apprentissage peut effectivement permettre de se familiariser avec l'outil afin d'assimiler les distorsions de localisation et enrichir sa base de données perceptives.

Tout système de reproduction nécessite un apprentissage minimal afin d'exploiter au mieux les possibilités offertes. Il est important d'apprendre à manier cette technologie tant au niveau de la mise en espace des contenus que des conditions de restitution. Par exemple, du fait que le binaural interactif corresponde à une simulation de la perception d'un espace sonore, l'apparition de sources fantômes n'est pas aussi précise qu'en stéréophonie, ce qui peut être minimisé si les objets concernés sont proches les uns des autres. L'usage de cet outil puissant, implique donc au préalable d'anticiper la description spatiale dès la prise de son. D'autre part, l'utilisation des HRTF mesurées en chambre anéchoïque pourra privilégier un niveau de réverbération élevé pouvant poser des problèmes de compatibilité. Un moteur binaural incluant un effet de salle peut réduire ce phénomène si la familiarité à l'outil ne permet pas son anticipation.

Afin d'appréhender les effets de masquage, les rapports de niveaux entre les sources ou le relief et la profondeur lors d'un mixage, le contenu fréquentiel des sources est une composante essentielle à prendre en compte. La coloration notable du traitement binaural semble être un frein à son usage dans cette étape de la post-production. Malgré le fait que cette coloration caractéristique soit présente afin de simuler une écoute réelle, elle est défavorable et semble peu efficace tant l'externalisation présente des problèmes dans les zones frontale et arrière de l'espace. Sans *head-tracker* cette dernière n'est d'ailleurs pas significativement perceptible

en dehors des zones situées sur les cotés. La coloration présente donc un désavantage sans justifier un réel atout. Il faut alors trouver un meilleur compromis entre la sensation d'espace et la fidélité du timbre quitte à altérer la première qui semble moins pertinente pour le mixage.

CHAPITRE 2

Quelles HRTF pour le mixage orienté objet ?

La perception de la localisation liée à l'utilisation de HRTF a été largement étudiée dans la littérature, tant au niveau des méthodes d'approximations des filtres que de l'individualisation. Néanmoins, les qualités de timbre des HRTF ont reçu relativement peu d'attention. L'objectif classique de l'égalisation spectrale de ces filtres est de prendre en compte une chaîne d'acquisition de sorte que les effets de la mesure soient annulés [19]. On tente ainsi de supprimer toutes les caractéristiques spectrales indépendantes de la direction des HRTF. Ces manipulations laissent cependant des colorations spectrales significatives. En dépit du fait d'être une composante naturelle de la propagation du son vers les oreilles d'un auditeur, celles-ci modifient significativement le timbre des contenus sonores ce qui peut poser problème pour des applications professionnelles.

L'utilité d'un monitoring au casque dans le cadre d'une production orienté objet a été mise en évidence dans le chapitre précédent. Les limites de l'usage de cette technologie et des éléments de réponse correspondants ayant été identifiées, la suite de ce mémoire s'attarde sur cette question primordiale de la coloration spectrale. Des méthodes de traitement des filtres HRTF seront explorées dans ce chapitre afin de trouver un meilleur compromis entre la sensation d'espace et la fidélité du timbre. L'ensemble de ces traitements considère l'ap-

proximation d'HRTF à phase minimale à laquelle est ajouté un retard pur. Nous invitons le lecteur à ce référer à la partie dédiée à cette question en Annexes.

Nous nous attarderons dans un premier temps à trouver des jeux de HRTF issues de campagne de mesures d'une part, et de calcul numériques d'autre part, susceptibles de constituer un bon compromis entre la sensation d'espace et la qualité du timbre. Dans ce but, des traitements exclusifs d'une partie significative de l'espace seront ensuite entrepris.

2.1 La base de données Listen

Plusieurs bases de données publiques d'HRTF mesurées sont disponibles. Le tableau 2.1 regroupe un ensemble non exhaustif accessible gratuitement. Ces bases contiennent des filtres HRIR associés aux points de mesure correspondants ainsi que des données morphologiques relatives aux individus.

Propriétaire	Nombre de mesures	Nombre de sujets
Ircam	187	51
ARI	1550	110
UC Davis - CIPIC	1250	45
RIEC	865	105
Université du Maryland	1093	7
Université de Nagoya	72	96
Université de Tohoku	464	3
MIT	710	Tête artificielle Kemar
TH Köln	16020	Tête artificielle KU100

Table 2.1 – Liste non exhaustive de bases de données publiques d'HRTF.

Il n'a pas été entrepris d'écouter l'ensemble des HRTF mentionnées car la base de données réalisée par l'Ircam en 2003 lors du projet de recherche Listen, a rapidement suscité un intérêt. En effet, en raison d'une qualité spectrale des HRTF supérieure à celles présentes dans la base ARI ou CIPIC et d'une externalisation satisfaisante pour de multiples sujets, le choix de ces HRTF s'est rapidement imposé. De plus, les travaux réalisés au Centre de Mathématiques Appliquées de l'école de Polytechnique (CMAP) en audio 3D se sont appuyés sur les filtres de l'Ircam. Il est intéressant de constater que, sans concertation préalable, le jeu d'HRTF utilisé en particulier au CMAP est également employé par les ingénieurs du

son du CNSMDP. Il s'agit de celles issues du sujet 1040, individu masculin à la chevelure courte ayant une largeur, hauteur et profondeur de tête de 15 cm, 24.5 cm et 19.3 cm respectivement. Les HRTF 1040 jouissent en effet d'une réputation quant à la qualité de leur externalisation, qui semble être efficiente pour un grand nombre d'individus. La figure 2.1 représente les contenus fréquentiels et temporels des HRTF 1040 dans le plan azimutal.

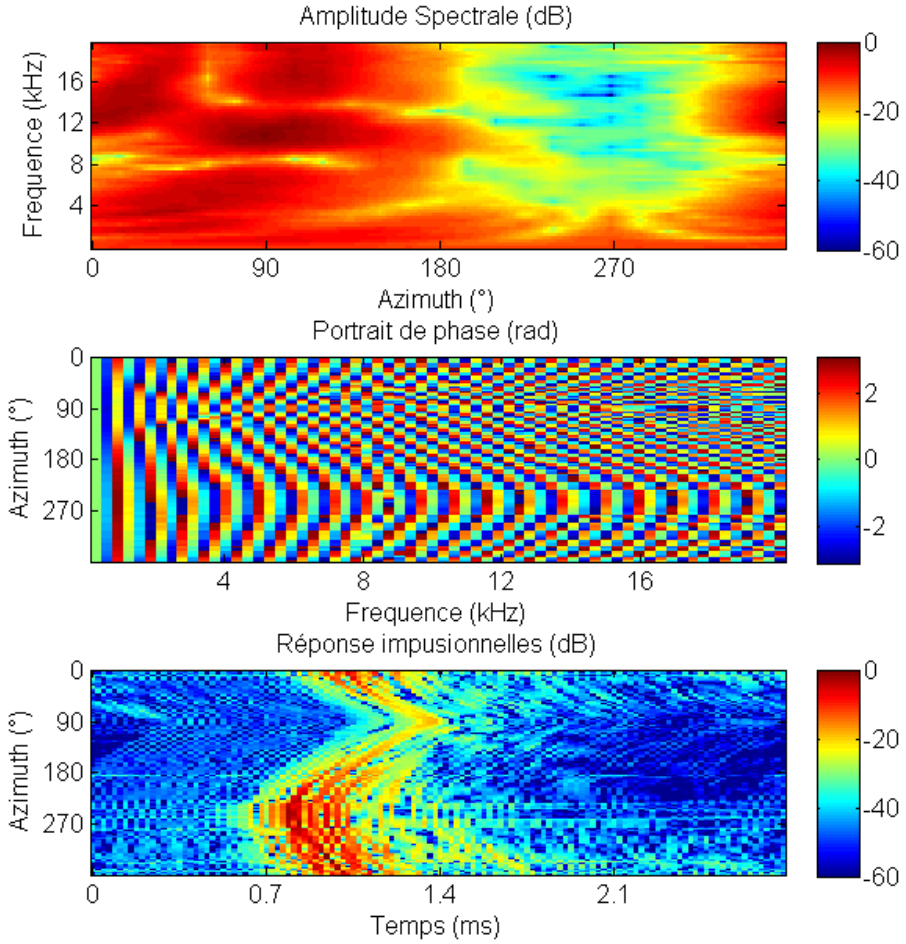


FIGURE 2.1 – Représentation spectrale et temporelle des HRTF 1040 de la base Listen dans le plan azimutal. Ces représentations emploient des interpolations et lissage effectués par le CMAP.

La campagne de mesure des HRIR de la base Listen a été effectuée dans une chambre anéchoïque de 324 m^3. Une enceinte Tannoy System 600 était disposée sur une grille métallique mobile située autour du sujet à 1.95 m de distance. Un système de suivi de tête lié au logiciel de mesure a permis de contrôler la position de la tête du sujet, de sorte que le signal

de mesure ne se déclenche uniquement que lorsque cette dernière se trouvait dans la position souhaitée. Une paire de microphones Knowles FG3329 était disposée dans les oreilles des sujets et des balayages fréquentiels logarithmiques, ou *sweeps*, de 186 ms et échantillonnés à 44100 Hz ont été employés comme stimuli. La figure 2.2 illustre la répartition des 187 points de mesures répartis dans l'espace sur une grille d'un pas angulaire de 15 ° à 1.95 m de l'auditeur. Des égalisations de l'amplitude spectrale des HRTF ont été réalisées *a posteriori* pour compenser les effets de l'amplificateur, des microphones, de l'enceinte et des conduits auditifs [19].

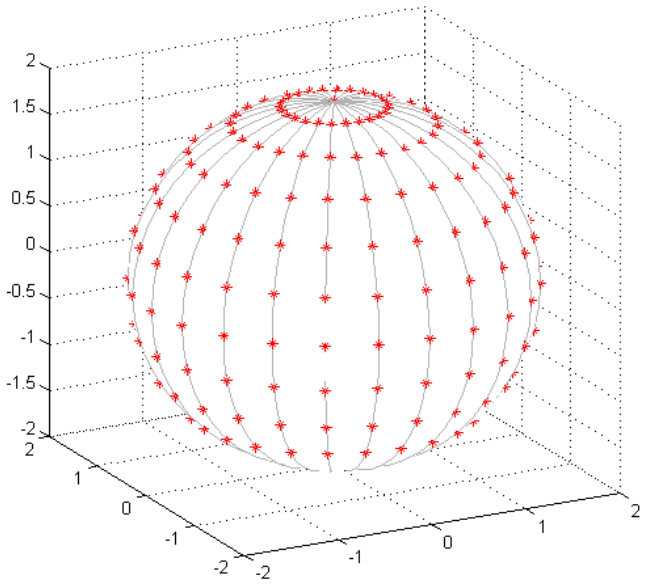


FIGURE 2.2 – Représentation des points de mesure de la base Listen.

Les filtres 1040 extraits de cette base de données ont été adoptés comme référence du fait de leur usage répandu. Différentes approches destinées à diminuer l'altération du timbre apparaissant avec le traitement binaural sont présentées dans la suite de ce document. On a envisagé dans un premier temps de trouver une alternative à l'utilisation directe de HRTF mesurées qui peuvent vraisemblablement comporter des erreurs inhérentes au procédé d'acquisition. On a alors entrepris de minimiser les accidents spectraux, en tentant de lisser l'influence de la mesure et des différences morphologiques.

La création de filtres médians

Il s'agit ici d'extraire de la base Listen des données moyennées afin d'obtenir des filtres qui confèrent une externalisation globalement satisfaisante tout en minimisant les accidents

fréquentiels. La figure 2.3 représente l'ensemble des 51 filtres de l'oreille gauche issus de tous les sujets de la base pour une source d'incidence (0 °, 0 °). On peut observer que la fréquence, l'amplitude, le nombre de pics et de creux diffèrent grandement d'un individu à l'autre. Dans le cadre de l'objectif d'extraire des indices communs à tous les sujets, nous faisons l'hypothèse que le système auditif analyse le spectre dans son intégralité [20] et non les caractéristiques locales seules des HRTF [21].

Au vu de la grande dispersion des amplitudes spectrales pouvant aller jusqu'à 30 dB, il a été choisi d'effectuer des calculs de spectres médians plutôt que d'opter pour une moyenne des données. En effet, la médiane est une donnée moins sensible aux valeurs extrêmes d'un ensemble que la moyenne. Elle représente le point milieu d'un ensemble de valeurs, qu'elle divise en deux moitiés. On peut distinguer sur la figure 2.3 que le spectre médian matérialise une tendance générale de l'évolution spectrale des mesures tout en réduisant l'écart fréquentiel maximal à 10 dB environ.

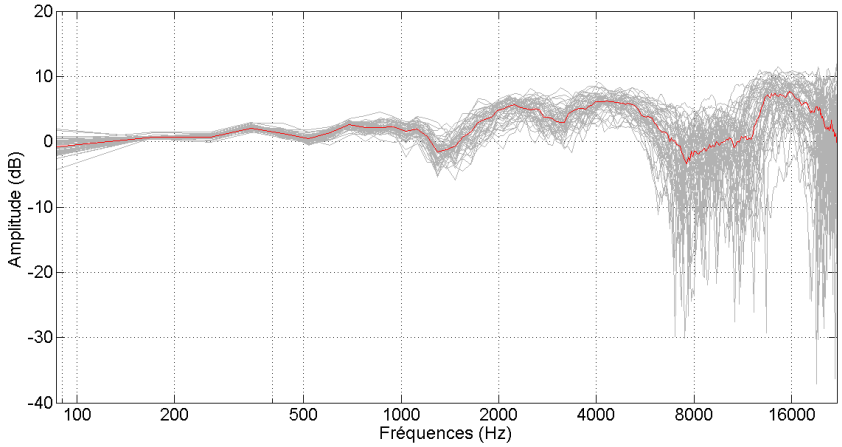


FIGURE 2.3 – Représentation de l'ensemble des spectres des HRTF de l'oreille gauche présentes dans la base Listen à la position (0 °, 0 °). En rouge figure le spectre médian calculé à partir de cet ensemble.

Après avoir considéré la composante fréquentielle des filtres, les différences interaurales de temps ont été estimées pour construire les HRIR correspondantes. La figure 2.4 représente l'ensemble des retards des 51 filtres de la base Listen pour toutes les incidences mesurées dans le plan azimutal. L'estimation des retards a été effectuée par la détection d'un seuil de montée des HRIR, après un calcul préalable de l'enveloppe de ces dernières. Une routine de vérification des résultats a permis de s'assurer de la décroissance monotone des ITD obtenus

dans la zone arrière (90 °$< \theta <$ 270 °) et de la croissance monotone dans la zone avant, selon un seuil de tolérance pré-défini de 2 échantillons (40 µs à 44100 Hz). La mise en œuvre simple de cette approche ainsi que la pertinence des résultats obtenus ont orienté le choix vers cette méthode. Notons que le calcul des retards purs a mis en évidence des résultats disparates dans la base de données. En effet, les mesures correspondant aux sujets 1002, 1007, 1034, 1044 et 1055 présentent des résultats très différents des autres et ont donc été exclues dans le calcul des filtres médians.

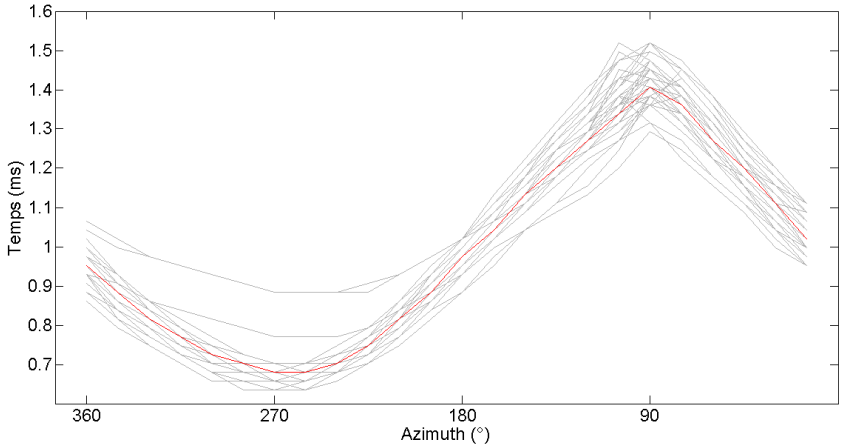


FIGURE 2.4 – Ensemble des retards purs des filtres de la base Listen ayant une élévation nulle. En rouge, retards purs médians correspondants.

Nous avons donc synthétisé des filtres HRTF médians présentant une dispersion spectrale réduite qui respectent les allures fréquentielles et temporelles des mesures. Afin de minimiser davantage les accidents spectraux l'ordre des filtres a également été réduit.

Les HRTF présentent de nombreuses irrégularités du fait du nombre important de coefficients qui les composent. La figure 2.5 illustre le contenu fréquentiel d'un filtre associé à son approximation par un polynôme d'ordre 10. Cette approximation permet de conserver les principaux pics et creux fréquentiels tout en lissant le spectre. En-deçà de l'ordre 10, le filtre résultant ne respecte plus les positions fréquentielles de ces pics et creux dans les zones contralatérales, très accidentées. On peut distinguer sur la représentation spectrale de la figure 2.5 que l'approximation à l'ordre 5, ne prend pas en compte les variations spectrales jusqu'à 6 kHz et que la fréquence centrale du creux situé aux alentours de 9 kHz est inférieure à celle de l'approximation à l'ordre 10 et à celle du filtre original. La pertinence de l'utilisation d'un ordre 10 à été avérée d'après des représentations graphiques selon diffé-

rentes directions de l'espace. Une méthode plus rigoureuse pour prouver cet usage adéquat serait de comparer les fréquences des maxima locaux entre les filtres d'origine et les filtres modifiés pour s'assurer qu'ils sont égaux à un seuil de tolérance près. L'approximation utilise la méthode des moindres carrés et les équations de Yule-Walker pour approcher la réponse en fréquence spécifiée [22]. Cette méthode permet la conception de filtres à réponse impulsionnelle infinie. Le but est de déterminer les $(P+Q+1)$ coefficients a_p et b_q, avec $p = 1, ..., P$ et $q = 0, 1, ..., Q$, permettant de minimiser ϵ, l'erreur au carré entre les modules de l'HRTF $H(\omega_k)$ et de son approximation $\hat{H}(z)$:

$$\epsilon = \sum_{k=0}^{N-1} \left(|H(\omega_k)| - |\hat{H}(z)| \right)^2 \tag{2.1}$$

où

$$\hat{H}(z) = \frac{b_0 + b_1 z^{-1} + ... + b_Q z^{-Q}}{1 + a_1 z^{-1} + ... + a_P z^{-P}} \tag{2.2}$$

avec $z = e^{j.\omega_k}$.

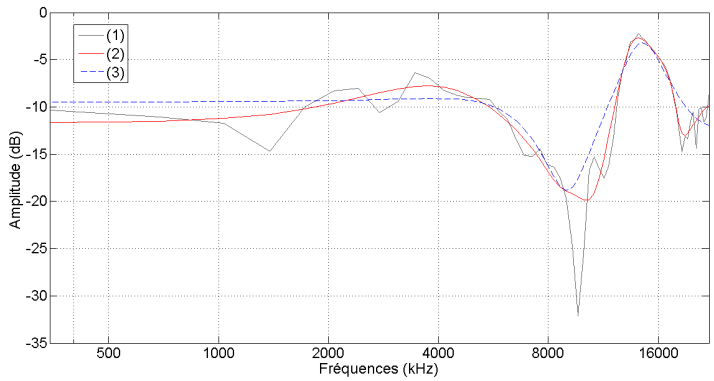


FIGURE 2.5 – Approximation spectrale d'une HRTF (1) par un polynôme d'ordre 10 (2) et d'ordre 5 (3).

La figure 2.6 met en évidence le lissage de la réponse en fréquence des HRTF médianes comparée avec la représentation spectrale de la figure 2.1. La phase n'est pas représentée du fait de l'utilisation d'un modèle de filtre à phase minimale associé à un retard pur. Cette phase étant linéaire, elle ne présente pas d'information supplémentaire pertinente. L'écoute des HRTF résultantes met en avant une plus grande présence et une meilleure définition des objets sonores virtuels. On constate en effet que les filtres médians modèrent la perte de gain dans la zone de fréquences avoisinant les 8 kHz. Le sentiment d'externalisation semble néanmoins altéré du fait d'une perception de la distance moins importante. Ces phénomènes ont

fait l'objet d'une évaluation présentée au chapitre 3 dans lequel de plus amples observations sont exposées.

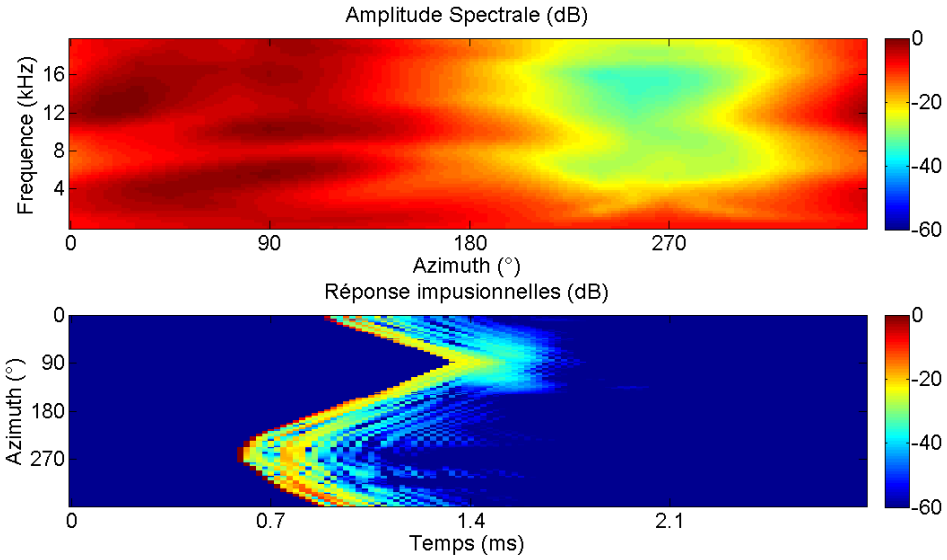


FIGURE 2.6 – Représentation spectrale et temporelle des HRTF médianes de la base Listen.

2.2 Le calcul d'HRTF

Le calcul analytique

Calculer des filtres HRTF revient à trouver des solutions au problème physique de propagation d'une onde acoustique en présence d'un obstacle, l'auditeur. La résolution d'un tel problème peut s'effectuer de manière analytique si la géométrie de la modélisation est simple. Nous avons alors entrepris de modéliser la tête d'un auditeur par une sphère et de résoudre le problème par calcul analytique. Une solution analytique peut être décrite à l'aide d'une fonction mathématique connue alors que les solutions numériques employées dans le calcul d'HRTF sont des approximations de solutions réelles en lien avec un modèle.

Duda a décrit les solutions analytiques du problème de diffraction d'onde par une sphère rigide [23]. La fonction de transfert, $H(\rho, \mu, \theta)$ correspondante est fournie par :

$$H(\rho, \mu, \theta) = -\frac{\rho}{\mu} e^{-i\mu\rho} \sum_{m=0}^{\infty} (2m+1) P_m(\cos\theta) \frac{h_m(\mu\rho)}{h'_m(\mu)} \tag{2.3}$$

avec

$$\mu = f\frac{2\pi a}{c} \tag{2.4}$$

et

$$\rho = \frac{r}{a}, \quad a < r \tag{2.5}$$

où c est la vitesse du son, r la distance de la sphère par rapport à la source, a le rayon de la sphère, h_m est la fonction de Hanckel sphérique de degré m et P_m le polynôme de Legendre de degré m. Un algorithme récursif rapide est fourni par Duda afin de calculer les solutions au problème.

Les HRTF issues de la solution à un tel problème ont été calculées pour une sphère de rayon 15 cm, rayon de la tête du sujet 1040. Les représentations temporelles et fréquentielles de ces filtres sont disponibles en figure 2.7. On y distingue l'apparition d'interférences constructives et destructives introduisant des creux de près de 30 dB pour des incidences contralatérales. En revanche, les amplitudes spectrales des sources ipsilatérales sont beaucoup moins accidentées que celles des filtres mesurés. La comparaison avec la figure 2.1, correspondant à ces derniers, met en évidence que les réflexions supplémentaires engendrées par le pavillon des oreilles, des épaules et du torse sont à l'origine des importantes déformations spectrales en amplitude et en phase. Il est intéressant de constater que les allures des amplitudes spectrales, temporelles et de la phase du modèle de Duda sont visibles de manière sous-jacente dans les représentations de la figure 2.1. La modélisation effectuée étant visiblement trop élémentaire, le calcul numérique d'HRTF a été choisi afin de prendre en compte une plus grande complexité des mécanismes en jeu.

Le calcul numérique

Un des intérêts du calcul numérique, en dehors du fait de s'affranchir de la mesure, est la possibilité d'évaluer l'impact d'une modification complexe de la morphologie sur les indices de localisation. On peut envisager d'isoler un élément en particulier, de jouer sur des paramètres de taille, de positionnement, de forme ou d'affiner la précision du modèle. De plus, ces méthodes garantissent une reproductibilité des résultats contrairement à la mesure. L'utilisation de calcul numérique pour l'obtention d'HRTF présente ici l'intérêt d'évaluer

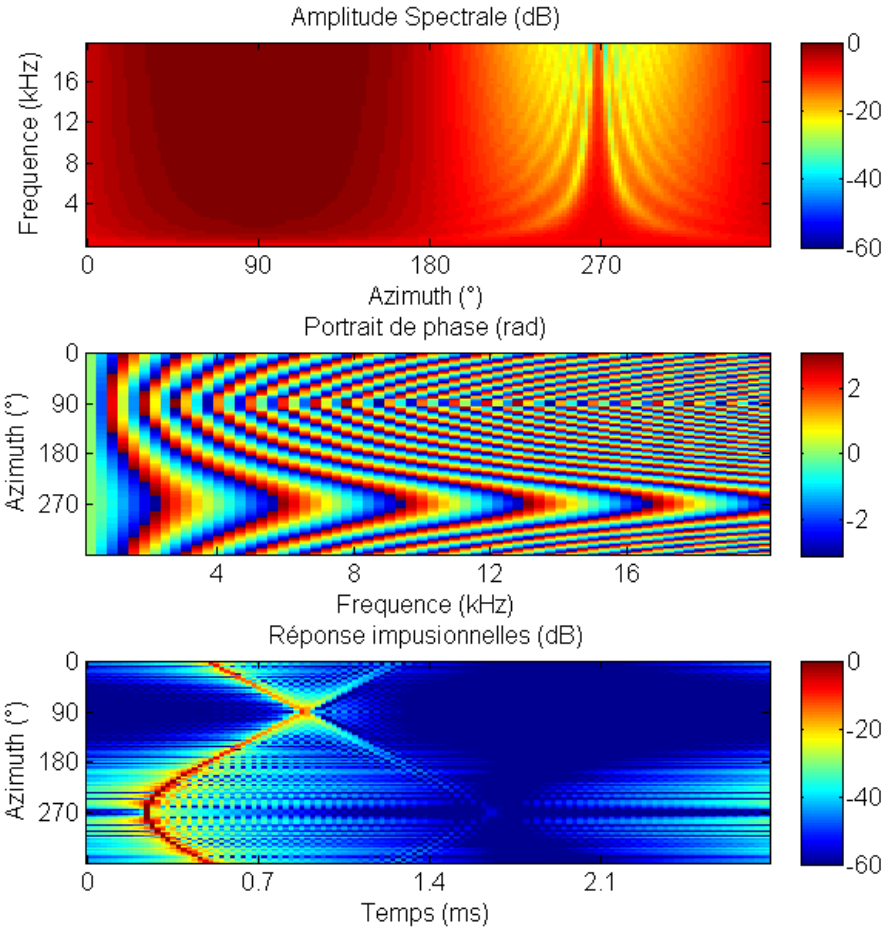


FIGURE 2.7 – Représentation spectrale et temporelle des HRTF issues d'une modélisation sphérique de la tête du sujet 1040 de la base Listen.

la modification spectrale des filtres selon différentes modélisations géométriques de la tête d'un auditeur. L'idée consiste à déterminer la complexité de la modélisation nécessaire et suffisante afin d'obtenir une externalisation satisfaisante tout en minimisant la coloration spectrale due au filtrage HRTF.

En particulier, le calcul par éléments finis permet de déterminer un champ u solution d'un ensemble d'équations aux dérivées partielles pour des conditions aux limites données, en tout point d'un domaine Ω et à tout instant t. Les conditions aux limites sont des contraintes s'exerçant sur le système considéré, soient des conditions simplifiées que l'on espère raisonnables imposées à u sur la frontière du domaine Ω.

Le calcul par élément finis est une méthode de discrétisation permettant d'approcher

une solution exacte par un champ défini par morceaux sur Ω. Il s'agit donc de morceler le domaine grâce à une géométrie approchée de celui-ci. On définit alors un domaine constitué de plusieurs sous-domaines pour lesquels on calcule des champs locaux. Par linéarisation des équations aux dérivées partielles on obtient des systèmes d'équations linéaires, plus simples à résoudre, pour chaque sous-domaine. Les conditions aux limites étant définies pour le champ global Ω, l'ensemble des systèmes d'équations linéaires des sous-domaines sont donc réunis pour résoudre le problème selon les conditions aux limites.

La solution au problème est calculée sur les nœuds du maillage ce qui nécessite une interpolation des résultats pour décrire l'ensemble du domaine Ω. Plus les points sont distants les uns des autres, plus l'approximation liée à l'interpolation risque de s'écarter des phénomènes physiques étudiés. Un compromis doit cependant être trouvé afin d'éviter des coûts de calcul trop importants dûs à des décompositions en sous-ensembles trop raffinés.

Le calcul par éléments finis de frontière, ou BEM, s'est imposé comme une alternative à l'utilisation des éléments finis en Acoustique qui permet d'une part de résoudre le problème sur une surface donnée dans un milieu de propagation infini et, d'autre part, qui ne nécessite pas la discrétisation de l'ensemble du milieu : la discrétisation des frontières est suffisante.

Dans notre cas, le domaine Ω est délimité par le volume représentant l'auditeur et la sphère de mesure qui l'entoure et le champ u à déterminer correspond au champ de pression p. L'équation d'ondes permet de déterminer le comportement de l'onde acoustique en espace libre à l'intérieur de Ω, milieu homogène, linéaire et isotrope : la variation de pression p en un point M et à l'instant t est fournie par l'équation 2.6.

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = \Delta p \tag{2.6}$$

où c est la vitesse du son et Δ l'opérateur laplacien scalaire de dérivation spatiale d'ordre 2. La linéarité de l'équation autorise une formulation fréquentielle de la propagation par transformée de Fourier. On obtient alors l'équation d'Helmholtz scalaire :

$$\Delta p + k^2 p = 0 \tag{2.7}$$

où k représente le nombre d'onde.

Différentes conditions aux limites peuvent être employées pour résoudre un tel système. Ces conditions caractérisent l'interaction de l'onde, les diffractions en particulier, avec la surface du domaine considéré ; à la fois la surface de l'auditeur et de la sphere de mesure.

Avec le calcul, l'ensemble des filtres peut être obtenu en calculant le rayonnement d'une source seulement, contrairement à la mesure. Lorsque l'on effectue une mesure d'HRTF une paire de microphones dans les conduits auditifs d'un sujet est utilisée afin de capter un signal émis selon chaque position de l'espace. Le calcul numérique permet l'inversion du système d'acquisition selon le principe de réciprocité. Ce principe repose principalement sur des propriétés d'invariances temporelles et spatiales de l'équation d'onde d'ordre deux. Cette méthode est très souvent utilisée pour résoudre les problèmes en Acoustiques comme en Mécanique. Selon le principe de réciprocité [24], la pression acoustique p_1 émise par la source S en un point M est égale à la pression acoustique p_2 produite par une source située au point M et perçue à la position de S. On peut donc placer la source acoustique directement sur le maillage de la tête, au niveau de l'oreille. La source rayonne alors sur le maillage de la sphère qui entoure l'auditeur. On récupère les données sur cette dernière et non dans le conduit auditif du sujet. Ce principe permet de réduire les calculs à effectuer. Cependant, la puissance et le temps de calcul nécessaires pour obtenir les filtres restent très importants.

Les recherches réalisées au sein du CMAP ont permis de développer d'autres méthodes de calcul performantes, en terme de rapidité et de précision. Les travaux effectués par François Alouges et Matthieu Aussal dans le cadre de la thèse de ce dernier [25], ont notamment mené à la création d'une nouvelle méthode rapide intitulée la «Décomposition Creuse en Sinus Cardinal», ou SCSD [26].

Plusieurs maillages de tête ont été réalisés en vue d'évaluer l'évolution des indices spectraux selon la géométrie plus ou moins complexe des morphologies. Une visualisation de ces maillages est disponible en Annexes. Cinq modélisations y sont répertoriées : une sphère, un ovoïde, un ovoïde muni d'oreilles, une tête et une tête associée à un torse. L'ensemble des maillages n'ayant pu faire l'objet de calculs d'HRTF, l'étude de l'évolution des indices spectraux n'a pu être réalisée. En effet, les outils à disposition n'étaient pas en mesure d'effectuer autant de calculs lors de la période dédiée à ce mémoire. Le logiciel de calcul numérique développé au CMAP, *MyBEM*, est un outil de recherche en phase de développement, qui n'est pas encore adapté à une utilisation industrielle. Les premiers résultats issus de ces calculs

ont présenté des colorations du son encore trop significatives pour s'y attarder dans le laps de temps disponible. Il est cependant intéressant de mentionner cette étude car de futurs travaux devraient reprendre cette question afin d'identifier la complexité nécessaire et suffisante d'une modélisation pour conférer une externalisation satisfaisante à un auditeur tout en minimisant la coloration spectrale. Les chercheurs du CMAP poursuivent les développements initiés par cette approche et les difficultés de mise en œuvre rencontrées sont en passe d'être levées.

2.3 Le traitement de la zone frontale

Dans le cadre d'une écoute non individualisée, l'externalisation des sources sonores est plus difficile à percevoir dans les zones frontale et arrière. Les indices spectraux contenus dans ces zones ne semblent pas suffisamment pertinents et l'utilisation du *head-tracking* associé à un apprentissage des HRTF non individuelles permettent d'améliorer l'externalisation avant et arrière, comme nous l'avons décrit dans le chapitre précédent. Aussi, pour créer un espace sonore environnant, il semblerait que les zones situées sur les cotés et à l'arrière soient principalement utilisées. Le traitement spectral exclusif de la zone frontale a alors été entrepris. Nous faisons donc l'hypothèse que l'altération des indices spectraux n'est pas forcement critique pour externaliser dans cette zone.

De plus, dans le cadre d'une production sonore, qu'elle soit stéréophonique, en 5.1 ou 11.1, les objets sonores sont principalement concentrés sur la scène frontale, dans la zone située entre les canaux gauche et droit, tandis que les canaux latéraux, zénithaux ou arrières sont utilisés pour les ambiances, l'acoustique de la salle ou les réverbérations artificielles. Par soucis de compatibilité, le mixage orienté objet nécessite également d'utiliser un grand nombre d'objets pour d'écrire au mieux une scène sonore et ainsi éviter de trop nombreuses créations d'images fantômes, sources d'imprécision et d'instabilité de la scène. Un nombre important de sons est alors susceptible de se situer dans cette zone frontale. Il est nécessaire de réduire au mieux la coloration spectrale due à l'écoute binaurale pour effectuer un mixage qui traite les timbres réels des sources, afin d'appréhender le plus objectivement possible les problèmes de masquage, les rapports de niveaux ou le placement des sources en profondeur.

D'après ces observations, différentes méthodes seront proposées pour minimiser la coloration spectrale dans la zone frontale. Dans la suite de ce chapitre, les HRTF issues du sujet 1040 de la base Listen ont été employées pour leur qualité d'externalisation constatée par une majorité des acteurs du projet Bili.

Égalisation multi-bandes

Un première approche a consisté à traiter les filtres HRTF frontaux de l'oreille en temps réel au moyen d'une interface graphique développée sous MATLAB et utilisant la bibliothèque *DSP System Toolbox*. La figure 2.8 représente les différents éléments constituant cette interface :

— les filtres HRTF gauche ou droite originaux (en rouge) ;
— les filtres correspondants modifiés (en bleu) ;
— la somme des huit filtres paramétriques (en noir) ;
— les paramètres permettant le réglage des fréquences, des gains et des facteurs de qualité de chaque filtres paramétriques ;
— un bouton permettant le déclenchement du traitement binaural ou non ;
— une glissière permettant de régler le volume du signal stéréo non binaural pour permettre une comparaison des timbres à même niveau ;

Huit filtres paramétriques ont été utilisés afin de découper le domaine fréquentiel de 62.5 Hz à 16 kHz en 8 octaves. Les gains correspondants étaient réglables sur une plage de ±10 dB et les facteurs de qualité de 0.1 à 5. Cette limite supérieure a été choisie en raison de la résolution fréquentielle minimale de 86 Hz, ce qui fournit une limite sur la précision des filtres à utiliser.

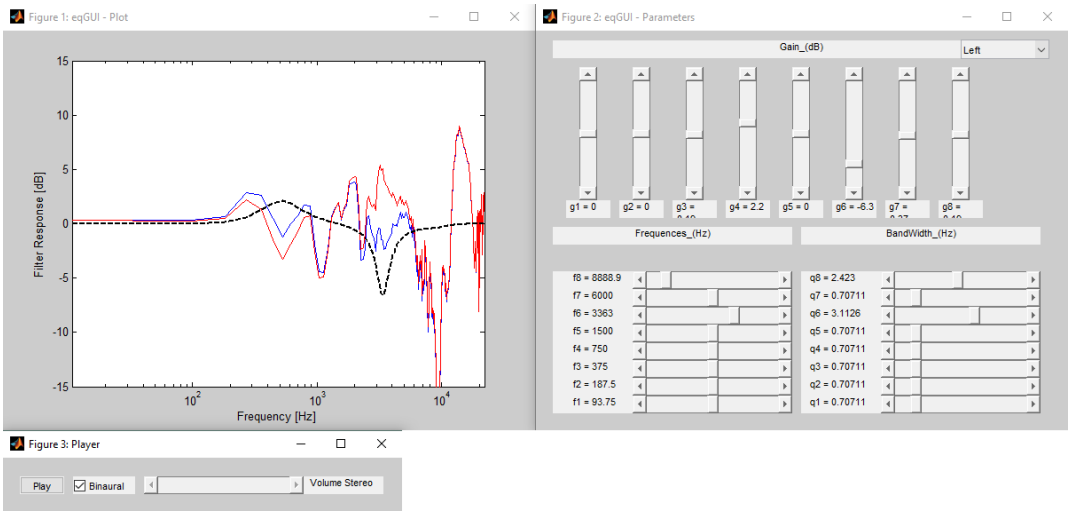


FIGURE 2.8 – Interface graphique permettant l'égalisation des filtres HRTF en temps réel.

Le même filtrage à phase minimale était effectué conjointement sur une paire de HRTF

gauche droite afin de conserver la cohérence de localisation en terme de différences intéraurales de niveau et de temps. La tâche consistait à trouver un filtrage permettant de minimiser au mieux la coloration du timbre dans la zone frontale. Plusieurs questions ont alors été soulevées durant de cette étape :

— Quelle est l'ampleur de la zone à traiter ?
— Comment traiter les filtres aux limites de cette zone ?
— Combien de filtres paramétriques doit-on utiliser pour compenser les altérations spectrales ?
— Comment maintenir une cohérence de timbre dans la zone si chaque paire de HRTF est traitée indépendamment ?

Du fait de l'aspect laborieux de la méthode, qui consiste à égaliser jusqu'à une trentaine de paires de filtres selon l'angle du fenêtrage spatial choisi, d'un nombre relativement faible de filtres paramétriques à disposition ainsi que d'une éventuelle disymétrie des filtrages dans l'espace, cette approche empirique n'a pas été menée à son terme. D'autres approches se sont révélées plus prometteuses.

L'utilisation d'une fenêtre spatiale dynamique

Afin de répondre aux questions relatives à l'ampleur de la zone à traiter ainsi qu'au traitement aux limites de cette zone, un fenêtrage spatial a été mis en place dans le moteur binaural utilisé. Cette interface, représentée figure 2.9, permet en effet d'appliquer une fenêtre spatiale au jeu de filtres de la zone frontale en temps réel selon trois paramètres. Le paramètre a correspond à l'angle à partir duquel le traitement de la zone frontal débute. Une zone de transition s'établit alors jusqu'à l'angle b correspondant à l'angle au-delà duquel les filtres HRTF sont totalement modifiés. Sur la figure 2.9, la zone bleu s'apparente à l'utilisation des filtres 1040 originaux, la zone rouge aux filtres modifiés. Dans la zone de transition une pondération des filtres originaux et des filtres modifiés est réalisée. La figure 2.10 illustre la pondération selon le paramètre k borné entre -6 et 6.

La calcul de la pondération fût réalisé de la manière suivante.

Soit une base orthonormée directe $\{\vec{x_0}, \vec{y_0}, \vec{z_0}\}$ dont le vecteur $\vec{x_0}$ est normal à la zone frontale. On définit x_a la projection sur l'axe $\vec{x_0}$ du point M_a d'angle a par rapport à $\vec{x_0}$ et situé sur la sphère de rayon r. De même, x_b la projection sur $\vec{x_0}S$ du point M_b situé sur la sphère de rayon r et d'angle b par rapport à $\vec{x_0}$. On a alors $x_a = r.\cos a$ et $x_b = r.\cos b$.

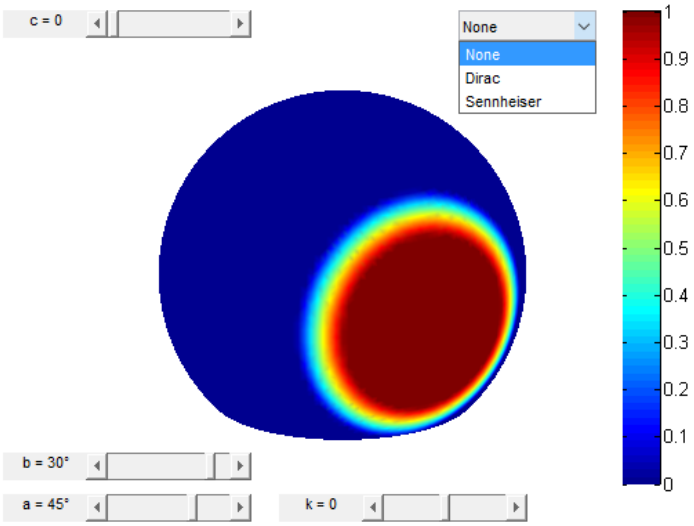


FIGURE 2.9 – Interface graphique de la fenêtre spatiale utilisée pour le traitement de la zone frontale.

Pour tous les points $M(x,y,z)$ vérifiant $x_a < x < x_b$, on a la fonction v définie sur $[x_a,x_b]$ et à valeur dans [0,1] :

$$v(x) = \frac{x - x_a}{x_b - x_a}. \tag{2.8}$$

Soit $f(v,k)$ la fonction de pondération :

$$f(v,k) = \begin{cases} \frac{1}{2}\cdot\left(\frac{\tan(2(v-0.5).\operatorname{atan}(k))}{k} + 1\right) & \text{si} \quad k < 0 \\ v & \text{si} \quad k = 0 \\ \frac{1}{2}\cdot\left(\frac{\operatorname{atan}(k(2.v-1))}{\operatorname{atan}(k)} + 1\right) & \text{si} \quad k > 0 \end{cases} \tag{2.9}$$

La pondération ainsi définie permet de jouer sur une pente plus ou moins abrupte pour la transition entre les filtres originaux, $H_O(\omega,\theta,\phi)$, et les filtres frontaux $H_F(\omega,\theta,\phi)$ situés à l'azimut θ et à l'élévation ϕ. Le module du filtre à phase minimale, $|H_T(\omega,\theta,\phi)|$, d'une

HRTF de la zone de transition s'écrit alors :

$$|H_T(\omega,\theta,\phi)| = |H_F(\omega,\theta,\phi)|^{f(v,k)}.|H_O(\omega,\theta,\phi)|^{1-f(v,k)} \tag{2.10}$$

de sorte que l'amplitude en dB de $H_T(\omega,\theta,\phi)$ est égale à la somme pondérée des amplitudes logarithmiques de $H_O(\omega,\theta,\phi)$ et $H_F(\omega,\theta,\phi)$.

Les traitements effectués étant à phase minimale, les retards purs associés sont ceux correspondant aux filtres originaux du sujet 1040. Les différences intéraurales sont donc similaires à celles contenues dans le jeu de filtres originaux 1040, seuls les indices spectraux et éventuellement les différences intéraurales de niveau changent selon la méthode employée.

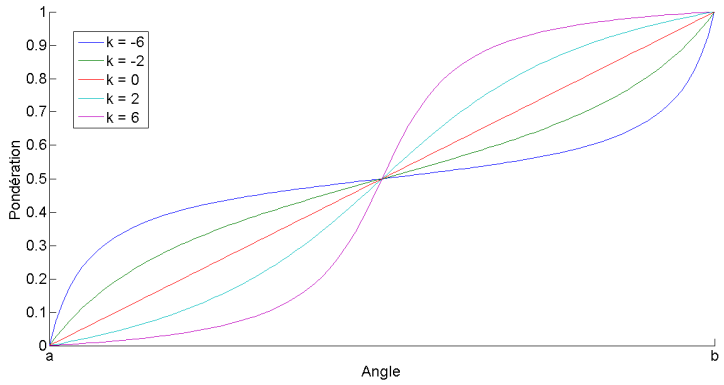


FIGURE 2.10 – Réseau de courbes de pondération du traitement des HRTF dans la zone de transition entre les angles a et b.

La suppression des indices spectraux

Après avoir mis en place la fenêtre spatiale, un premier traitement de la zone frontale a consisté à n'utiliser que les différences intéraurales de niveau et de temps en supprimant les indices spectraux contenus dans les filtres. C'est en ne conservant qu'un unique échantillon (dirac numérique) de niveau et de retard pur équivalents à ceux du filtre d'origine que l'on obtient un filtre à spectre plat respectant les ITD et ILD. L'amplitude de ce spectre est obtenue d'après la norme $L_O(\theta,\phi)$ du filtre original :

$$L_O(\theta,\phi) = \sqrt{\frac{1}{\omega_1-\omega_2}\int_{\omega 1}^{\omega 2}|H_O(\omega,\theta,\phi)|^2 d\omega}. \tag{2.11}$$

L'amplitude de l'unique échantillon utilisé est calculée de sorte que la norme du spectre modifié $L_F(\theta, \phi)$ soit égale à $L_O(\theta, \phi)$.

Comme la bande de fréquence $[\omega 1\,;\,\omega 2]$ est choisie de manière à ce qu'elle corresponde à des fréquences où l'audition humaine est la plus sensible afin que le niveau calculé possède un sens du point de vue de la perception, c'est la bande de fréquences [200 Hz ; 6 kHz] qui a été utilisée pour le calcul des amplitudes spectrales. Les courbes de Fletcher et Munson pour un niveau isosonique de 80 phones, niveau sonore adapté à la réalisation d'un mixage, ont permis d'opter pour ces bornes.

La figure 2.11 représente les échantillons calculés dans une zone frontale de 30 ° en azimut et élévation. Le contenu fréquentiel correspond à une constante dont l'amplitude est fonction de l'angle d'incidence du son. Les HRTF résultant d'une telle transformation permettent de localiser les sources sonores sans fournir d'informations de localisation permettant l'externalisation. Reste à savoir si le cerveau permet de les extrapoler grâce à l'usage d'un *head-tracker*. Ces filtres ont été soumis à l'appréciation de la qualité spectrale et de l'externalisation qu'ils apportent dans le test perceptif présenté au chapitre 3.

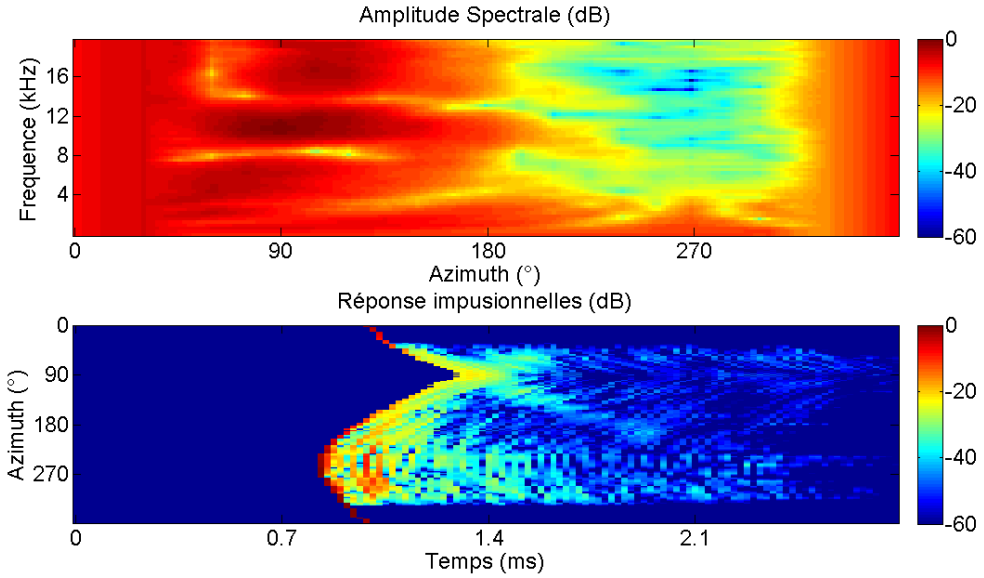


FIGURE 2.11 – Représentation spectrale et temporelle des HRTF issues d'une modélisation par un dirac numérique pour une fenêtre spatiale de 30 °.

L'exploitation des ISD

L'usage d'un unique coefficient dans le filtrage des sources sonores ne présentant pas une externalisation satisfaisante, on a envisagé de prendre en compte des indices spectraux afin d'améliorer la sensation d'externalisation dans la zone concernée. Cette approche s'appuie sur l'utilisation des différences relatives entre les filtrages gauche et droite seulement et non des spectres dans leur intégralité. Selon les études effectuées par Morimoto [27] ou Hofman [28], au-delà de la zone frontale, les indices spectraux peuvent être considérés comme monoraux : le système auditif extrait les informations spectrales issues des signaux droit et gauche de manière indépendante. Le rôle de l'oreille ipsilatérale, la plus proche de la source sonore, est en effet prédominant pour des azimuts supérieurs à 30 °. Dans la zone frontale en revanche, le système auditif extrait des informations spectrales issues des deux oreilles. La méthode exposée ici consiste à extraire seulement les différences spectrales intéraurales, ou ISD.

La figure 2.12 représente la différence intéraurale des HRTF gauche et droite pour une position de 20 ° en azimut et 0 ° en élévation. On constate des différences de niveau allant jusqu'à 26 dB entre les deux filtres HRTF. Jusqu'à 16 kHz, l'amplitude du spectre gauche est comprise entre -22 dB et -3 dB, soit 19 dB d'écart et celle du spectre droit s'étend de -29 dB à -11 dB, soit un écart de 18 dB. Nous constatons de plus l'allure accidentée du contenu fréquentiel à partir de 1 kHz.

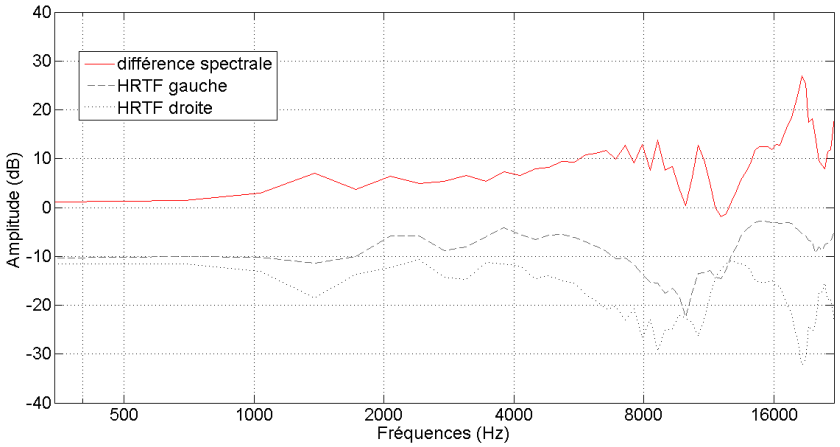


FIGURE 2.12 – Représentation des HRTF d'incidence (20 °, 0 °) gauche et droite ainsi que la différence spectrale associée.

La figure 2.13 représente les HRTF droite et gauche contenant seulement les différences

interaurales. Celles-ci sont obtenues en répartissant la moitié de la différence spectrale sur le filtre de gauche et l'autre moitié sur le filtre droit. Soient $|H_{OG}|$ et $|H_{OD}|$, les modules des filtres gauche et droit originaux, on obtient le calcul des amplitudes modifiées respectives H_{FG} et H_{FD} de la manière suivante :

$$|H_{FG}| = \frac{1}{L_M(\theta,\phi)} \cdot \sqrt{\frac{|H_{OG}(\omega,\theta,\phi)|}{|H_{OD}(\omega,\theta,\phi)|}} \tag{2.12}$$

et

$$|H_{FD}| = \frac{1}{L_M(\theta,\phi)} \cdot \sqrt{\frac{|H_{OD}(\omega,\theta,\phi)|}{|H_{OG}(\omega,\theta,\phi)|}} \tag{2.13}$$

où $L_M(\theta,\phi)$ est la moyenne de la norme $L_{OG}(\theta,\phi)$, définie grâce à l'équation 2.11, et de la norme $L_{OD}(\theta,\phi)$ des filtres gauche et droite d'origine. Il a été choisi d'utiliser la norme moyenne des deux filtres étant donné que chacune des deux HRTF modifiées ne peut être normalisée indépendamment dans la mesure où les rapports de niveaux doivent rester inchangés pour respecter la différence spectrale d'origine.

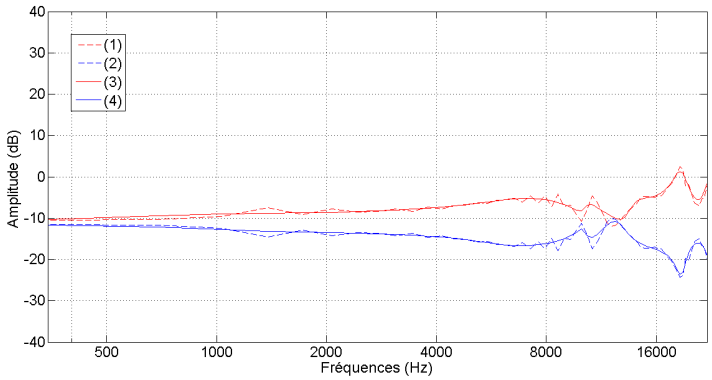


FIGURE 2.13 – La différence spectrale répartie à gauche (1) et à droite (2) ainsi que leur approximation polynomiale d'ordre 10 correspondante (3)(4).

Ainsi la différence en fréquence est respectée et les pics et les creux locaux de chacune des HRTF d'origine n'apparaissent plus. Le profil des filtres comporte alors moins d'écarts d'amplitude. On constate en effet que les amplitudes spectrales varient respectivement de 7 dB jusqu'à 16 kHz et non plus de 18 dB. La figure 2.14 illustre cette dynamique spectrale selon plusieurs incidences azimutales dans la zone frontale. Cette représentation met en évidence la réduction quasiment de moitié des dynamiques spectrales, soient des filtres

plus constants et moins accidentés fréquentiellement.

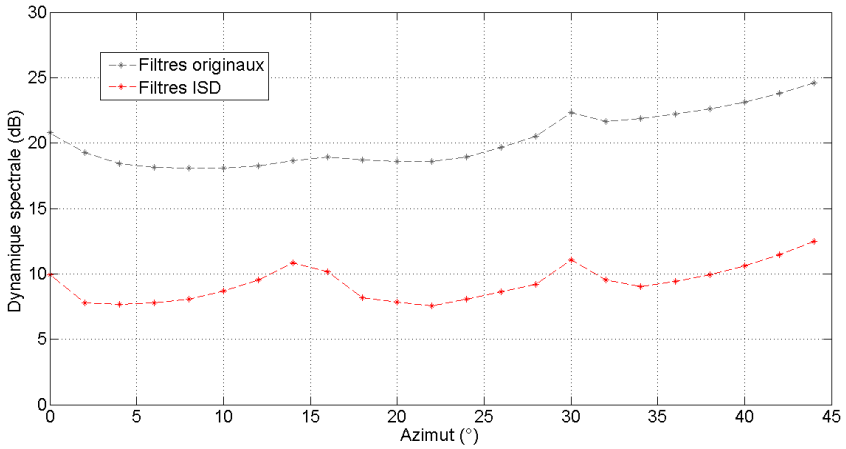


FIGURE 2.14 – Dynamique spectrale jusqu'à 16 kHz des filtres de l'oreille contralatérale en fonction de l'incidence dans le plan azimutal.

Afin de lisser un peu plus les réponses en fréquences, une approximation par un polynôme d'ordre 10 a été employée. L'approximation utilise la méthode des moindres carrés et les équations de Yule-Walker pour approcher la réponse en fréquence spécifiée [22].

La figure 2.15 représente l'allure spectrale et temporelle des filtres du plan azimutal. On peut constater que les différences spectrales semblent importantes au delà de 16 kHz. Cette amplification, pouvant aller jusqu'à 10 dB, peut s'expliquer par la diffraction et l'absorption de ces fréquences dont la longueur d'onde, de moins de 2 cm, est proportionnelle aux nombreux obstacles présents dans la morphologie d'un auditeur. La méthode présentée a tendance à créer plus de pics que de creux pour l'oreille ipsilatérale alors que les filtres HRTF sont surtout composés de creux. Ceci peut s'avérer problématique dans la mesure où les contenus audio sont relativement peu définis dans la zone de fréquences [16 kHz ; 20 kHz] et que son amplification peut avoir pour conséquence d'amplifier surtout du bruit. L'étude précédente a néanmoins exclu la bande de fréquences [16 kHz ; 20 kHz] car elle présente les dynamiques les plus importantes alors qu'elle n'est pas une zone où l'oreille est particulièrement sensible.

Afin de valider l'utilisation de cette méthode, encore faut il s'assurer de la capacité d'un auditeur à externaliser en n'utilisant que les ISD. Deux interprétations communes de l'utilité des indices spectraux considèrent que soit le système auditif analyse les creux et

les bosses des HRTF, soit il analyse le spectre incident dans son intégralité. L'approche exposée ici ne respecte pas ces deux interprétations dans la mesure où les creux et les pics sont modifiés et où l'intégralité du spectre n'est pas respectée. L'hypothèse que l'exploitation dans la zone frontale des ISD seulement, associés au *head-tracking*, permet une externalisation satisfaisante a été soumise au test présenté au chapitre 3.

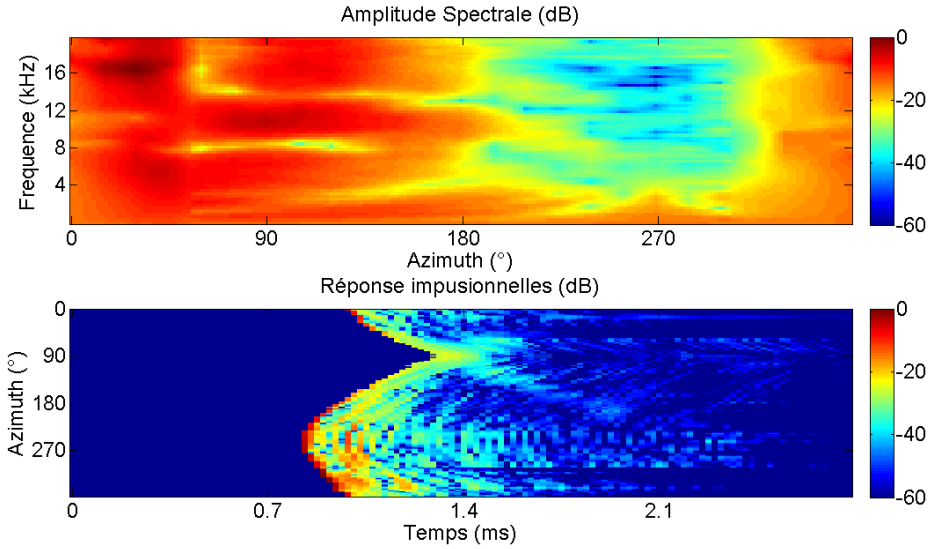


FIGURE 2.15 – Représentation spectrale et temporelle des HRTF issues du traitement des ISD pour une fenêtre spatiale de 30 °.

La méthode proposée par Sennheiser

Juha Merimaa propose une méthode développée au sein du laboratoire de recherche de Sennheiser permettant de réduire les effets indésirables de la coloration spectrale due au traitement binaural [29]. Cette méthode est basée sur la réduction de la variation spectrale de la moyenne quadratique d'une paire de HRTF tout en conservant les différences interaurales de niveau et de temps. Des tests formels d'écoute permettent de montrer que la coloration peut être réduite sans altérer les performances de localisation.

Considérons $|H_G(\omega,\theta,\phi)|$ et $|H_R(\omega,\theta,\phi)|$ les modules des filtres gauche et droite d'une paire de HRTF. Soit $H_{RMS}(\omega,\theta,\phi)$, la moyenne quadratique de ces filtres :

$$H_{RMS}(\omega,\theta,\phi) = \sqrt{\frac{|H_G(\omega,\theta,\phi)|^2 + |H_D(\omega,\theta,\phi)|^2}{2}}. \tag{2.14}$$

L'hypothèse que la variation spectrale de la moyenne quadratique est grandement responsable de la coloration est formulée sans justifications alors que ce filtre contient néanmoins des informations de localisation dont la modification pourrait entrainer une perte d'externalisation. Afin de paramétrer une plus ou moins importante modification, le coefficient c est introduit tel que :

$$|H_G^{'}(\omega,\theta,\phi)| = \Big(\frac{L_{RMS}(\theta,\phi)}{H_{RMS}(\omega,\theta,\phi)}\Big)^{1-c}.|H_G(\omega,\theta,\phi)| \tag{2.15}$$

et

$$|H_D^{'}(\omega,\theta,\phi)| = \Big(\frac{L_{RMS}(\theta,\phi)}{H_{RMS}(\omega,\theta,\phi)}\Big)^{1-c}.|H_D(\omega,\theta,\phi)| \tag{2.16}$$

où $L_{RMS}(\theta,\phi)$ est la norme du filtre H_{RMS} définie par l'équation 2.11. Son usage permet de normaliser le filtre résultant afin qu'il conserve la norme du filtre original quel que soit c. La moyenne quadratique des modules des filtres $H_D^{'}(\omega,\theta,\phi)$ et $H_G^{'}(\omega,\theta,\phi)$ s'écrit alors :

$$H_{RMS}^{'}(\omega,\theta,\phi) = \Big(\frac{L_{RMS}(\theta,\phi)}{H_{RMS}(\omega,\theta,\phi)}\Big)^{1-c}.H_{RMS}(\omega,\theta,\phi). \tag{2.17}$$

Ainsi pour $c \neq 0$, les variations spectrales de la moyenne quadratique des filtres diminuent et pour $c = 0$ le spectre résultant correspond à une constante. L'usage d'un coefficient $c < 1$ n'aplatit pas seulement la réponse en fréquence de H_{RMS} mais permet également de lisser la réponse de l'oreille ipsilatérale dans la mesure où elle contribue plus grandement à la moyenne quadratique que l'oreille contralatérale.

La figure 2.16 fait figurer plusieurs filtres HRTF de l'oreille ipsilatérale pour une incidence azimutale de 30 ° et une élévation nulle. On constate effectivement que la variation du coefficient influe grandement sur le caractère plus ou moins plat de la réponse en fréquence. Cette méthode de calcul a été employée pour traiter les filtres 1040 dans la zone frontale afin de comparer sa pertinence vis-à-vis des autres méthodes. La figure 2.17 rend compte du traitement réalisé sur une fenêtre spatiale de 30 °. On constate que les principaux accidents spectraux présents en figure 2.1 sont certes atténués mais restent présents contrairement aux cas précédemment exposés.

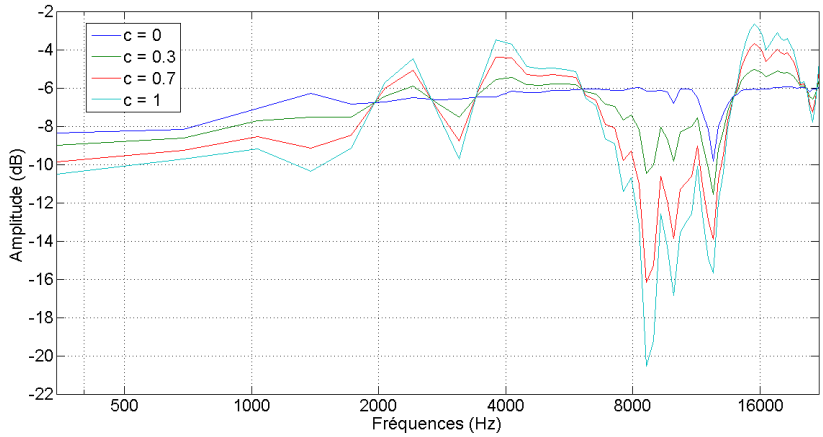


FIGURE 2.16 – Réseau de courbes correspondant à l'oreille gauche selon une minimisation plus ou moins importante de la variation de la moyenne quadratique des filtres HRTF pour une source d'incidence (30 °, 0 °).

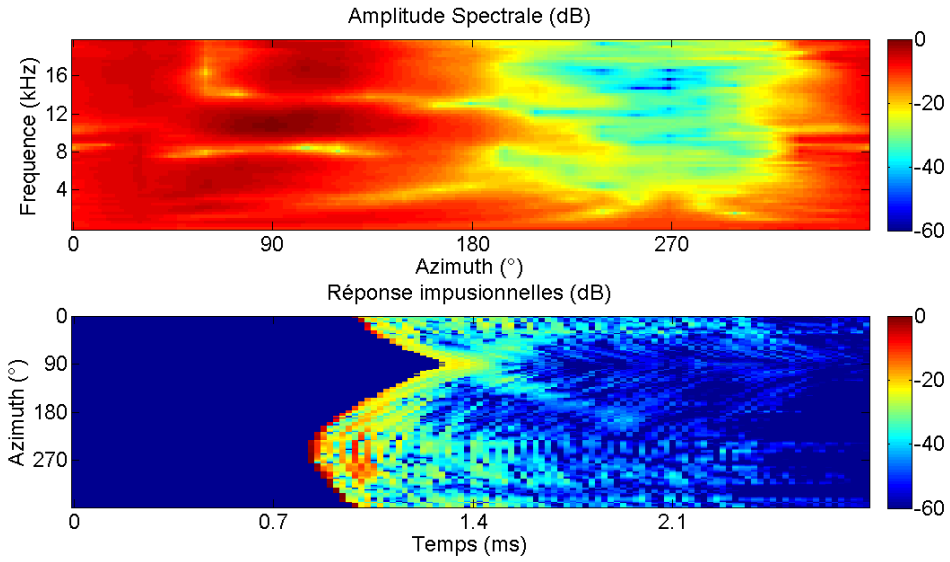


FIGURE 2.17 – Représentation spectrale et temporelle des HRTF issues de la méthode «Sennheiser» pour une fenêtre spatiale de 30 °.

La création d'un jeu de HRTF hybride

Une dernière approche a consisté à utiliser à la fois les HRTF mesurées 1040 et les filtres issus de la modélisation sphérique correspondante, détaillée en section 2.2. Dans la zone frontale, les indices spectraux ainsi que les différences intéraurales de niveau sont issues des calculs analytiques, les différences intéraurales de temps correspondent aux filtres mesurés.

Le niveau sonore des HRTF calculées a été adapté au niveau des filtres 1040 en normalisant l'ensemble des filtres par la norme $L(\theta, \phi)$ (Eq. 2.11), moyennée selon toutes les directions [29] :

$$L = \sqrt{\frac{1}{4\pi} \int \int L^2(\theta, \phi) cos(\phi) d\theta d\phi}. \tag{2.18}$$

La figure 2.18 représente les amplitudes spectrales des filtres dans une zone spatiale de 30 °. Les spectres restent compris dans ±1 dB jusqu'à 600 Hz après quoi les amplitudes des filtres s'étendent de -7 dB à 6 dB. Alors que dans le cas des spectres des HRTF mesurées, la bande de fréquence avoisinant les 8 kHz est réduite, on distingue ici une croissance monotone des amplitudes spectrales. En dehors de la sensation d'agressivité que l'on percevoir lors d'une écoute binaurale, cette amplification de la partie haute du spectre apporte une sensation de présence et de définition des sources sonores qui peuvent amoindrir la capacité d'externalisation. De plus, cette prédominance des fréquences aiguës peut se traduire par une perte de sensation de corps par la faiblesse relative des amplitudes de la bande de fréquences [125 Hz ; 250 Hz].

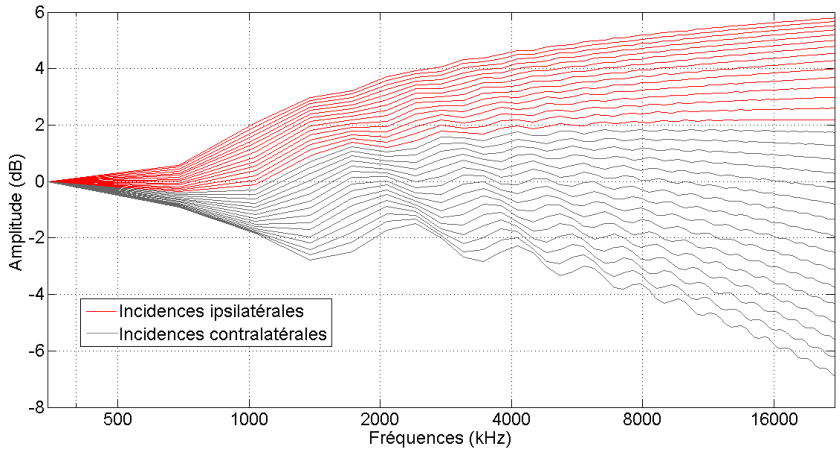


FIGURE 2.18 – Amplitudes spectrales des filtres ipsilatéraux et contralatéraux entre -30 ° et 30 ° issues des calculs analytiques de la sphère de largeur égale au rayon de la tête du sujet 1040.

2.4 Conclusions locales

Différentes approches destinées à diminuer l'altération du timbre ont été présentées dans ce chapitre. Des jeux de filtres issus de mesures ont été examinés dans le but trouver une solution au problème de la coloration spectrale du traitement binaural. Après avoir tenté de trouver une alternative à l'utilisation directe de HRTF mesurées, des traitement plus ou moins drastiques ont été effectués sur les filtres dans une zone de l'espace jugée significative. Les filtres 1040 extraits de la base de données Listen de l'Ircam ont été adoptés comme référence du fait de leur capacité d'externalisation avérée par leur usage fréquent.

L'externalisation des sources sonores est plus difficile à percevoir dans les zones frontale et arrière. Les indices spectraux contenus dans ces zones ne semblent effectivement pas suffisamment pertinents. De plus, l'utilisation du *head-tracking*, l'apprentissage ou un effet de salle peuvent palier à ce défaut d'externalisation. Les zones latérales et arrières étant plutôt dédiées à la création d'un espace environnant, le traitement spectral exclusif de la zone frontale a alors été réalisé. Afin de pouvoir juger dynamiquement de l'ampleur de la zone à traiter et de la zone de transition, un fenêtrage spatial a été intégré au moteur binaural utilisé.

En s'appuyant sur le jeu de filtres originaux 1040, seuls les indices spectraux et éventuellement les différences intéraurales de niveau ont été modifiés selon la méthode employée. Quatre méthodes de traitement de la zone frontale ont été envisagées :

— **La méthode du dirac numérique.**
Cette méthode a consisté à n'utiliser que les différences intéraurales de niveau et de temps en supprimant les indices spectraux contenus dans les filtres. Les filtres résultant présentent un spectre plat respectant les ITD et ILD des filtres originaux. Celle transformation fournit une bonne qualité spectrale mais un nombre restreint d'informations de localisation de telle sorte que l'externalisation est essentiellement due à l'usage du *head-tracker*.

— **La méthode des ISD.**
Seules les différences spectrales intéraurales, ont été extraites des filtres. La différence en fréquence entre les filtres gauche et droit est alors respectée et les pics et les creux locaux de chacune des HRTF d'origine n'apparaissent plus. Grâce à cette méthode, les dynamiques spectrales des filtres ont été significativement réduites. Les filtres présentent une allure fréquentielle plus constante et sont moins accidentés.

— **La méthode de Sennheiser.**
Cette méthode est basée sur la réduction de la variation spectrale de la moyenne quadratique d'une paire de HRTF, les différences intéraurales de temps et de niveau de HRTF originales étant conservées. Le traitement permet d'agir sur le caractère plus ou moins plat de la réponse en fréquence des filtres tout en conservant les principaux accidents spectraux des filtres originaux.

— **La méthode hybride.**
Cette méthode a consisté à utiliser à la fois des filtres mesurés et des HRTF issues d'un calcul analytique de propagation d'une onde sur une sphère. Les amplitudes spectrales de la zone frontale présentent une allure monotone. La croissance de l'amplification en fonction de la fréquence exacerbe la sensation de présence et de définition des sources sonores qui amoindrissent la capacité d'externalisation. Ceci a également pour effet une perte de sensation de corps des sources sonores.

En complément de ces traitements, une étude de l'évolution des indices spectraux a été proposée. Il serait intéressant de s'attarder sur la possibilité d'évaluer l'impact d'une modification de la morphologie sur les indices de localisation afin d'identifier une modélisation nécessaire et suffisante d'une tête ou d'isoler d'éventuels éléments physiques non pertinents pour la localisation et responsables de la coloration. De plus, des analyses relatives au réglage de la fenêtre spatiale peuvent être développées. La pondération de filtres originaux et des filtres modifiés dans la zone de transition et le réglage des angles de début et fin de traitements mériteraient d'être approfondis. Encore faut-il s'assurer du bien fondé de cette précision de paramétrage de la fenêtre pour toutes les méthodes.

Afin de juger de la pertinence des traitements réalisés et dans le but d'identifier des axes de développement vers lesquels s'orienter, des tests perceptifs ont été réalisés.

CHAPITRE 3

Tests perceptifs préliminaires

Différentes méthodes de traitement des HRTF ont été mises en place afin de trouver des filtres adaptés à une utilisation du binaural interactif dans le cadre d'une post-production orientée objet. Ces différents jeux d'HRTF ont fait l'objet de test perceptifs préliminaires pour privilégier ou écarter certaines de ces méthodes afin de faciliter l'avancée de futurs développements. La sensation d'externalisation et la fidélité du timbre des HRTF ont été jugées selon l'ampleur spatiale des traitements et la modification spectrale plus ou moins importante des HRTF. Nous présenterons dans un premier temps les protocoles des tests effectués ainsi que leurs mises en œuvre avant d'analyser les résultats obtenus.

3.1 Les protocoles de tests

Les modifications des indices spectraux exposées au chapitre 2, visant à obtenir une plus grande fidélité envers le timbre d'origine des sources, mènent nécessairement à une modification des indices de localisation. Dans le but d'évaluer la pertinence des traitements et de trouver un compromis entre ces deux aspects, le premier test a consisté à qualifier l'externalisation, le second à juger de la qualité du timbre. Neuf jeux de HRTF ont été évalués au cours de ces deux tests perceptifs :

— les filtres mesurés sur le sujet 1040 de la base Listen ;
— les filtres médians calculés d'après la base Listen ;
— les filtres dits «diracs», contenant un unique échantillon dans une fenêtre spatiale de 30 ° ;
— les filtres «diracs» dans une fenêtre spatiale de 75 ° ;
— les filtres dits «ISD», calculés d'après les différences spectrales intéraurales dans une fenêtre spatiale de 30 ° ;
— les filtres «ISD» dans une fenêtre spatiale de 75 ° ;
— les filtres dits «Sennheiser», calculés d'après la méthode présentée par Merimaa avec un coefficient d'aplanissement $c = 0$, dans une fenêtre spatiale de 30 ° ;
— les filtres «Sennheiser» calculés avec un coefficient $c = 0.5$, dans une fenêtre spatiale de 30 ° ;
— les filtres «Sennheiser» calculés avec un coefficient $c = 0$, dans une fenêtre spatiale de 75 ° ;

Notons que les filtres issus du jeu de HRTF hybrides n'ont pu être évalués car les développements relatifs à cette approche n'étaient pas aboutis au moment des tests.

L'ensemble de ces filtres regroupe donc trois méthodes différentes détaillées dans le chapitre précédent, les méthodes dites «dirac», «ISD» et «Sennheiser». (L'utilisation de calcul d'HRTF n'a malheureusement pas pu être entreprise car les développements en lien avec cette approche n'étaient pas finis avant le début des tests.) Deux angles de fenêtrage spatial de 30 ° et 75 ° ont été utilisés pour chacune des trois méthodes mentionnées afin de juger d'une limite maximale éventuelle dans l'utilisation d'une fenêtre spatiale. Les filtres 1040, les HRTF médianes et les filtres «Sennheiser» pour un coefficient $c = 0.5$, ont été employés pour vérifier si les résultats sont en concordance avec l'hypothèse suivante : l'altération des contenus spectraux correspondants n'est pas suffisante pour se distinguer de l'externalisation et de la coloration des filtres 1040.

Un stimuli de 21 secondes par test fût employé lors de ces évaluations. La durée relativement courte de ces stimuli a permis de s'attarder sur les mêmes passages musicaux. Il s'agit de mixages effectués en orienté objet au CNSMDP par Jean-Christophe Messonnier. L'écriture musicale ainsi que le projet esthétique de mixage ont permis de choisir ces stimuli en cohérence avec les critères à évaluer. Chaque scène audio est définie par vingt-quatre objets sonores dont la répartition spatiale figure sur le tableau 3.1. La figure 3.1 matérialise quant à

elle la position de ces objets dans l'espace sur le plan azimutal. La zone frontale est constituée de 13 objets sonores et les 11 restants permettent de décrire l'acoustique environnante.

Canal	1	2	3	4	5	6	7	8	9	10	11	12
Azimut	0	180	7	-7	15	-15	22	-22	30	-30	37	-37
Élévation	0	0	0	0	0	0	0	0	0	0	0	0
Canal	13	14	15	16	17	18	19	20	21	22	23	24
Azimut	45	-45	70	-70	110	-110	60	-60	120	-120	150	-150
Élévation	0	0	0	0	0	0	45	45	45	45	0	0

TABLE 3.1 – Répartition spatiale des 24 canaux constituant les stimuli.

Les stimuli étaient joués en boucle et le temps imparti aux sujets était de 15 minutes maximum par test afin d'éviter d'introduire un biais dû à la fatigue. On leur demandait d'écouter préalablement l'ensemble des filtrages binauraux avant de répondre au questionnaire, afin de mieux appréhender l'échelle d'appréciation. Les sujets avaient la main sur la sélection de HRTF à écouter. Le changement de filtres pouvait s'effectuer lors de la lecture des contenus audio, pour éviter trop de manipulations de transport du son et ainsi favoriser une comparaison directe des filtres. Enfin, seuls les mouvements de tête de moins de 60 ° par rapport à la position d'origine était autorisés étant donné que les traitements réalisés se concentraient dans la zone frontale seulement pour la plupart des jeux de HRTF.

Dans la mesure où les développements sont destinés aux conditions d'utilisation professionnelles, seuls des ingénieurs du son et étudiants ingénieurs du son ont fait partie des sujets. Sachant que le temps imparti à disposition, le nombre relativement faible de sujets susceptibles de passer les tests ainsi que la multitude de variables à étudier ne permettraient pas d'identifier clairement une méthode à adopter dans l'immédiat, les résultats de ces tests perceptifs préliminaires permettront d'envisager des voies de développement futures.

La perception de l'externalisation

Le stimuli utilisé pour ce test est un mixage composé de quatre voix situées dans la zone frontale à 45 °, 30 °, 0 ° et -30 ° et d'éléments percussifs (claquements de doigts) perceptibles de manière symétrique à 60 °. Cet extrait musical interprété par *Les Dyvettes d'en Face* est une reprise de *Fever*, morceau écrit par Eddie Cooley et Otis Blackwell. Il a l'avantage

de présenter un nombre réduit d'éléments sonores répartis distinctement dans l'espace. Les claquements de doigts présents dans l'enregistrement fournissent des indices de localisation réguliers. Leur caractère impulsionnel permet de plus d'exciter fréquentiellement une grande partie des filtres HRTF. D'autre part, la voix semble être un bon support d'analyse de la localisation tant cette source sonore, très souvent vecteur d'information, est fréquemment analysée par l'audition pour en déterminer la provenance.

Ce test a permis de juger du sentiment d'externalisation selon les HRTF employées par comparaison aux filtres 1040, faisant ici figure de référence. Ceux-ci ont servi de base d'implémentation aux méthodes présentées au chapitre 2, en raison de la qualité avérée de leur externalisation qui justifie son usage répandu pour le traitement binaural. Les filtres soumis aux tests sont en effet issus de déclinaisons des indices spectraux contenus dans ce jeu 1040, à l'exception des HRTF médianes. Cette comparaison vise alors à juger si ces modifications sont critiques pour l'appréciation de l'espace.

Nous faisons l'hypothèse que l'usage d'une fenêtre spatiale plus large que la zone frontale vient altérer le sentiment d'externalisation, car les objets sonores destinés au rendu de l'acoustique d'une salle, aux ambiances ou à la reverbération artificielle seraient soumis à l'altération des indices spectraux de localisation. Ils seraient donc moins aptes à retranscrire une sensation d'espace. D'autre part, ce test a été effectué afin d'identifier une limite dans la modification du contenu spectral au-delà de laquelle la sensation d'externalisation n'est plus présente. Dans cette optique, une attention particulière a été portée au jugement des filtres «diracs» qui font totalement abstraction du contenu spectral présent dans le jeu 1040.

Cinq qualificatifs ont été proposés aux sujets ayant passé le test. Ces derniers devaient déterminer si, par comparaison à l'utilisation du jeu de HRTF 1040, l'externalisation était :

— meilleure ;
— similaire ;
— moins bonne ;
— beaucoup moins bonne ;
— ou si ils ne percevaient aucune externalisation.

A priori la première et la dernière mention ne semblent pas pertinentes. D'une part, l'externalisation ne pourrait vraisemblablement être améliorée du fait de la modification du contenu spectral plus ou moins importante qui joue sur les indices de localisation, d'autre

part, on peut s'attendre à ce que l'usage d'un *head-tracker* garantisse un taux d'externalisation minimum. Mentionner ces deux options est une occasion de vérifier ces hypothèses.

Le nombre relativement restreint de choix pour qualifier l'externalisation a semblé plus pertinent par rapport à l'utilisation d'une échelle comprenant une graduation de 1 à 10. En effet, il parait complexe de juger de manière très précise la sensation d'externalisation. De plus, l'ambition de ce test est de rendre compte d'un jugement qualitatif plutôt que quantitatif dans le but d'aiguiller de futurs développements.

La perception du timbre

Le second test a permis de juger de la qualité du timbre résultant des filtrages HRTF comparés au signal stéréophonique non binaural écouté au casque puisque celui-ci rassemble les objets sonores non filtrés et respecte ainsi leur timbre. Le but est de vérifier si la réduction de la coloration est significative et convaincante.

Le signal stéréophonique est issu d'une réduction automatique du signal multicanal en deux canaux. Ce processus s'est effectué en appliquant la loi de répartition en intensité suivante :

$$L(\theta_i) = \sqrt{\frac{\sin\theta_i + 1}{2}} \tag{3.1}$$

$$R(\theta_i) = \sqrt{1 - \frac{\sin\theta_i + 1}{2}} \tag{3.2}$$

$$S(\theta_i) = 10\log_{10}\left(10^{\frac{L_{dB}(\theta_i)}{10}} + 10^{\frac{R_{dB}(\theta_i)}{10}}\right) \tag{3.3}$$

Où $L(\theta_i)$ et $R(\theta_i)$ sont les coefficients de pondération des sources situées à l'azimut θ_i. Cette loi permet d'obtenir une sommation des puissances acoustiques du canal gauche et droit, $S(\theta_i)$, égale à 0 dB.

L'hypothèse à vérifier est celle d'une plus grande similitude entre les timbres de la réduction stéréophonique et les filtrages utilisant très peu d'information spectrale. D'autre part, il est intéressant de vérifier l'influence de la largeur du fenêtrage spatial pour le traitement sur la perception du timbre selon les différentes méthodes employées. *A priori*, plus cette fenêtre

est large, plus le nombre d'objets sonores épargnés par le filtrage des HRTF 1040 augmente. Ce qui devrait améliorer la concordance avec le signal stéréophonique.

Il a été demandé aux sujets de mesurer la différence de timbre par rapport au signal stéréophonique. Quatre mentions figurent dans le questionnaire afin de qualifier la coloration par rapport à cette référence :

— similaire ;
— peu coloré ;
— coloré ;
— très coloré.

Pour les raisons évoquées plus haut, l'utilisation d'une échelle de jugement comportant si peu de graduations permet de mettre en avant une tendance plutôt qu'une appréciation de la différence de timbre quantitative.

Le stimuli utilisé pour ce test est un mixage composé de deux saxophones, un trombone, un piano, un vibraphone et une batterie. Cet extrait musical interprété par l'ensemble de jazz *Rémi Fox*, fût enregistré au CNSMDP par Lucas Rémond et Jean-Christophe Messonnier puis mixé en objet par ce dernier. Il présente une grande variété de timbres différents, le mixage procure un sentiment de fusion entre les instruments et de «masse» musicale spectralement riche, tout en permettant une distinction des éléments qui le compose.

3.2 La mise en œuvre du protocole

Les tests perceptifs présentés dans ce chapitre ont utilisé le moteur binaural, *MyBino*, développé sous MATLAB au CMAP. Des pré-traitements sur les filtres HRTF sont nécessaires afin de les incorporer dans le moteur binaural. De ce fait, toutes les modifications du jeu 1040 présentées au chapitre 2 ont été appliquées aux HRTF 1040 pré-traitées. Une vérification par l'écoute de l'intégrité de ces filtres adaptés au moteur a eu lieu préalablement. Nous nous sommes assurés que leur capacité d'externalisation ainsi que leur rendu spectral étaient similaires à celles de la 1040 originales. De plus, les HRTF figurant dans ces tests ont toutes subi le même pré-traitement et les jugements demeuraient relatifs et non absolus dans la mesure où ils portaient sur la comparaison à une référence donnée.

L'interface utilisée pour les tests est présentée figure 3.1. On peut y distinguer les différents objets sonores situés dans l'espace, la position de la tête (matérialisée par une croix rouge) ainsi qu'un panneau de contrôle. Celui-ci a permis aux sujets de sélectionner les stimuli ainsi que les jeux de filtres HRTF à écouter. Ce changement de filtre pouvait s'effectuer pendant l'écoute. Le commutateur désigné par la mention «3D» permettait de rendre actif ou non le traitement binaural. Si il n'était pas enclenché, le downmix stéréophonique était alors diffusé. Un autre nommé «HT», avait pour conséquence la prise en compte des mouvements de la tête en traitant les données issues du *head-tracker* (HT). Il permettait par la même occasion de le calibrer : le moteur prenait alors comme point d'origine la position de la tête de l'auditeur au moment où il enclenchait le bouton. Les trois angles de rotation de la tête - à savoir le lacet, le tangage et le roulis (*yaw*, *pitch*, *roll*) - étaient pris en compte pour déplacer les sources sonores en fonction des mouvements de l'auditeur. Le *head-tracker* utilisé lors de ces tests s'interface rapidement et facilement avec logiciel MATLAB. Son intégration est détaillée sur le projet Github de Peter Bratz *.

La latence due à l'implémentation du moteur binaural sous MATLAB, s'élevait à 146 ms. Cette valeur très élevée procure une sensation d'instabilité de la scène sonore qui n'apparait plus comme fixe. Il a été demandé aux sujets de prendre en considération cette valeur et de bouger la tête lentement afin d'éviter d'amplifier la perception du phénomène. Les tests ne portant pas sur l'évaluation de la localisation des sources sonores, cette latence n'a pas été considérée comme un biais dans la perception de l'externalisation et du timbre.

3.3 L'analyse des résultats

Vingt sujets ont passé les tests perceptifs. Les moyennes et les médianes des résultats sont visibles sur les figures 3.3 et 3.2. Il a été choisi de représenter ces deux valeurs par HRTF afin de faire figurer les éventuelles dissymétries dans les résultats. L'utilisation d'écart-types n'étant pas pertinente au vue de l'amplitude des échelles de valeurs. Les échelles des différents qualificatifs des tests ont été converties en échelles numériques allant de -3 (pour les mentions «très coloré» et «beaucoup moins bonne») à 0 (pour «similaire») et 1 (pour «meilleure», dans le cas de la perception de l'externalisation).

Un test de Wilcoxon-Mann-Whitney [30,31] a été réalisé afin d'aider à l'interprétation des résultats. Ce test statistique non paramétrique permet de tester l'hypothèse selon laquelle

∗. https ://github.com/ptrbrtz/razor-9dof-ahrs

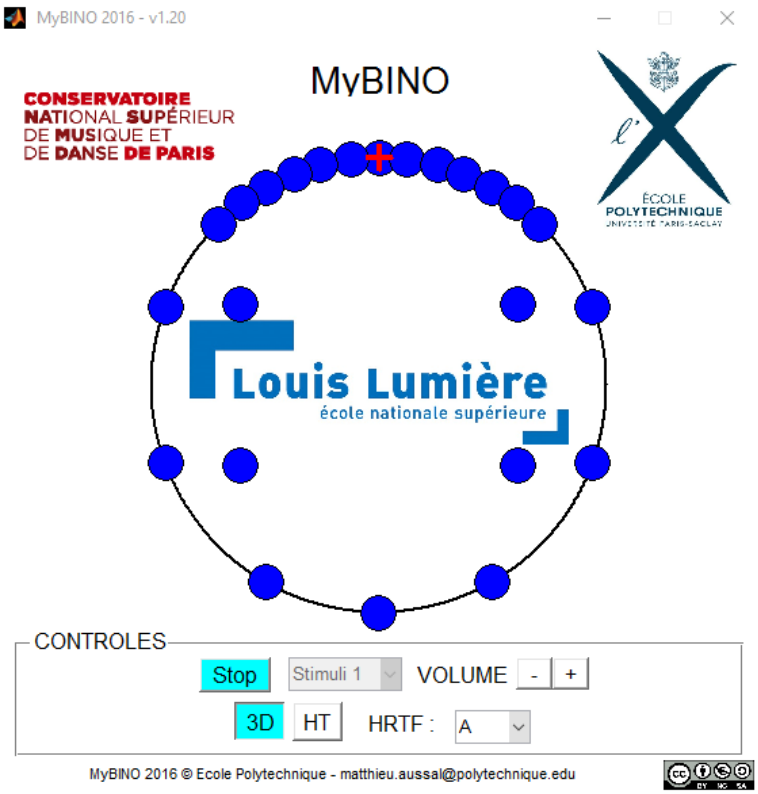


FIGURE 3.1 – Interface graphique ayant servi aux tests perceptifs.

la distribution des données est la même pour deux observations données. Il est adapté à l'usage de données qualitatives obtenues d'après une échelle dite ordinal ou de rangement, c'est-à-dire lorsqu'il existe une relation entre les variables à qualifier du type : plus petit que, inférieur à, plus facile que, *etc* ... Des comparaisons deux à deux ont donc été employées entre chacune des HRTF et par rapport aux références. Ceci a permis de déterminer quels filtres étaient significativement différents ou non, entre eux et vis-à-vis des références.

La perception de l'externalisation

Les neuf étiquettes présentes sur la droite de la figure 3.2 font référence aux différents filtres. Elles correspondent à la nomenclature suivante : *methode_angle du fenetrage spatial* - hormis le label «MED» faisant référence aux HRTF médianes calculées d'après la base Listen et le label «sen05_30» dont la mention «05» correspond au coefficient $c = 0.5$ du traitement «Sennheiser».

Un résultat sans équivoque apparait suite au test de Wilcoxon : les filtres «dirac_30»,

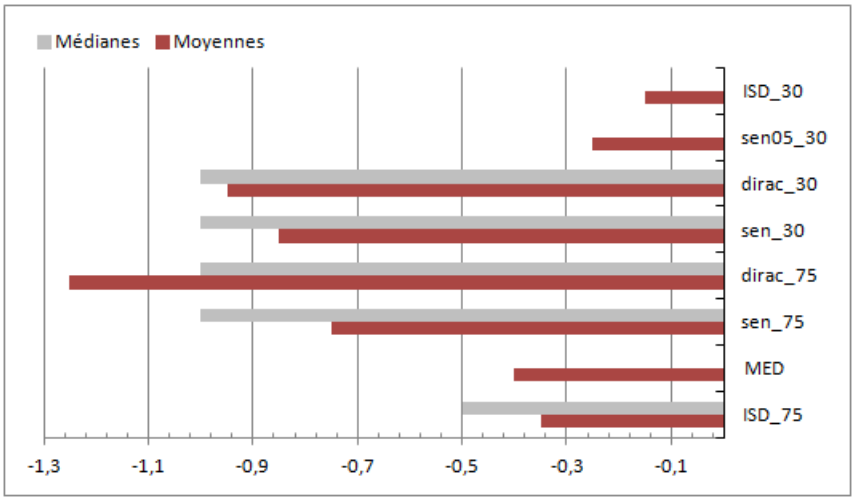


FIGURE 3.2 – Moyennes et médianes des notes obtenues par les différents filtres HRTF du point de vue de la perception de l'externalisation.

«dirac_75», «sen_30» et «sen_75» sont significativement différents des HRTF 1040 du point de vue de l'externalisation. Les mentions attribuées à ces filtres s'apparentent au qualificatif «moins bonne». Les traitements effectués avec ces méthodes et ces fenêtrages spatiaux semblent donc avoir altéré les indices de localisation contenus dans l'information spectrale. *A contrario*, les filtres «ISD_30», «ISD_75», «sen05_30» et «MED» ne sont pas significativement différents de la référence. Ce qui concorde avec l'hypothèse selon laquelle, les traitements relatifs aux deux dernières méthodes n'impactaient pas suffisamment les indices spectraux pour se distinguer de l'externalisation des filtres 1040. La méthode «ISD» semble donc conserver les indices de localisation nécessaires à l'externalisation en comparaison aux autres méthodes de fenêtrage spatial du chapitre 2.

Le test de Wilcoxon met également en évidence que les filtres «dirac_75» sont significativement différents de «ISD_75», «ISD_30», «MED» et «sen05_30» (les quatre jeux d'HRTF proche de 0). Une tendance similaire est observée pour les filtres «dirac_30». Un nombre plus conséquent de sujets pourrait venir confirmer ou infirmer cette tendance. La méthode «dirac» semble alors la moins apte à procurer un sentiment d'externalisation, ce qui était prévisible dans la mesure où aucun indice spectral ne figure dans ces filtres. Notons néanmoins que les moyennes et médianes des notes correspondantes à l'externalisation sont relativement éloignées de la plus faible valeur, -3, correspondant à la sensation de ne percevoir aucune externalisation. Le *head-tracking* est vraisemblablement responsable de ce taux d'externalisation minimum.

La perception du timbre

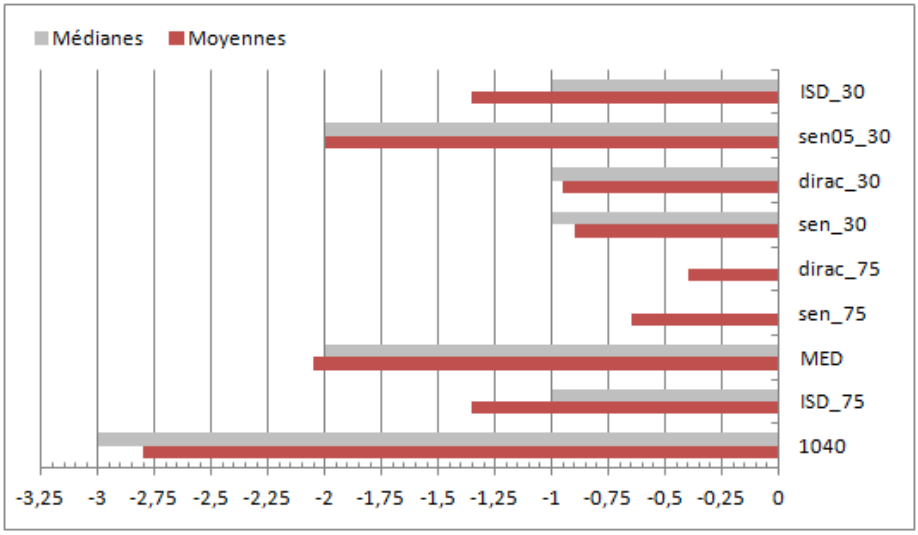


FIGURE 3.3 – Moyennes et médianes des notes obtenues par les différents filtres HRTF en rapport à la perception du timbre.

Les résultats mettent en avant une différence significative du timbre entre les signaux issus des filtrages binauraux et le signal stéréophonique pour tous les HRTF employées. Aucun des traitements utilisés n'a donc permis d'obtenir une similitude avec la référence. Même si les HRTF «dirac_75» et «sen_75» sont significativement différentes de 0, elles présentent les notes moyennes et médianes les plus proches de la mention «similaire». Sur une grande partie de la zone frontale, les HRTF «dirac» et la moyenne quadratique des filtres «sen_75» présentent en effet une réponse fréquentielle droite.

De même que dans la comparaison au signal stéréophonique, les filtres 1040 se distinguent de tous les autres HRTF au niveau du timbre. Tous les traitements effectués ont permis une modification significative du rendu spectral du jeu de filtres 1040. La figure 3.3 fait état des mauvais résultats obtenus par les filtres 1040 lors de ce test. Même si les filtres «sen05_30» et «MED» sont significativement différents des filtres 1040, les notes moyennes et médianes sont les plus proches de celles de ces derniers. Notons que pour l'externalisation et la perception du timbre, ces deux jeux de HRTF, «sen05_30» et «MED», sont non significativement différents entre eux.

Il est intéressant de constater que le test de Wilcoxon nous informe que les filtres HRTF «dirac_30», «ISD_30» et «sen_30» ne sont pas significativement différents entre eux. Alors que pour un fenêtrage spatial de 75 °, les filtres «dirac_75», «ISD_75» et «sen_75» le sont. L'usage d'une fenêtre relativement réduite ne permet vraisemblablement pas de discriminer les traitements effectués dans la zone frontale.

Notons également que pour la méthode «dirac», l'usage d'une fenêtre frontale de 30 ° et 75 ° a pour conséquence une appréciation significativement différente du timbre. *A contrario*, ceci ne semble pas avoir d'importance pour les méthodes «Sennheiser» et «ISD» car le test Wilcoxon nous informe qu'il n'y a aucune différence significative entre «ISD_30» et «ISD_75» d'une part, «sen_30» et «sen_75» d'autre part. Cette non différenciation est également constatée pour ces deux méthodes dans le test traitant de l'externalisation. Le fenêtrage inférieur à 75 ° ne semble pas avoir d'utilité pour ces filtres «Sennheiser» et «ISD». La nécessité d'un fenêtrage spatial apparait donc dépendante de la méthode employée.

3.4 Conclusions locales

Deux tests perceptifs ont été réalisés afin de juger de la qualité des timbres et de la capacité d'externalisation des jeux de HRTF modifiés dans la zone frontale. Les protocoles de test mis en œuvre ont tenté d'isoler chaque grandeur afin de les juger indépendamment bien que l'externalisation fasse également intervenir le contenu spectral du signal car la perception de la localisation est intimement liée au timbre.

Pour résumer les résultats des tests, les filtres offrant les plus mauvaises capacités d'externalisation, à savoir «dirac_30», «dirac_75» et les HRTF «Sennheiser», présentent également une meilleure fidélité vis-à-vis du timbre des sources traitées. Les HRTF réalisant le meilleur compromis pour ce test sont celles issues de la méthode «ISD». Nous pouvons constater que sa capacité d'externalisation n'est pas significativement différente de celle des HRTF 1040 et qu'elle présente une qualité de timbre proche du signal stéréophonique. Il reste à savoir, quel degré d'externalisation nous sommes prêt à sacrifier pour s'approcher au plus près des timbres d'origines des sources. Il se peut que l'externalisation minimale constatée dans le premier test, vraisemblablement due à l'usage du *head-tracker*, puisse constituer une externalisation suffisante. L'apprentissage étant une variable clé de l'externalisation, une écoute régulière employant ces HRTF permettrait de déterminer si elles fournissent une sensation d'espace convaincante et suffisante pour éviter toute perception intra-cranienne.

Il est intéressant de noter que le jugement de l'externalisation a été considéré le plus difficile à réaliser tant cette notion semble complexe à mesurer. Les différences perçues entre les signaux ont paru relativement faibles et déterminer la moins bonne ou meilleure qualité d'une externalisation en comparaison à une autre ne semble pas forcément pertinent. Le rendu parait parfois simplement différent, sans jugement de valeur à apporter.

Le jugement du timbre a quant à lui mené à des réflexions sur la référence que constitue le signal stéréophonique. Il est arrivé que ce dernier ne soit pas considéré comme restituant le plus fidèlement possible le timbre. En effet, par la capacité de démasquage du traitement binaural, certaines HRTF ont vraisemblablement permis de révéler un timbre plus en adéquation avec le timbre naturel, ou de référence, que reconnait le sujet. Un signal stéréophonique peut cependant être considéré comme une référence dans la mesure où il est le fruit d'un travail effectué par le mixeur qui donne à entendre un projet esthétique, une intention artistique. L'utilisation d'un *down-mix* automatique peut être remise en cause dans la mesure où l'interaction entre les sources peut intégrer des problèmes de masquage remettant en cause le projet esthétique original souhaité lors du mixage.

Un second test destiné à évaluer de manière plus quantitative les HRTF pourrait être entrepris afin d'identifier clairement une méthode permettant de réaliser le meilleur compromis entre l'externalisation et la qualité du timbre. Une échelle de 0 à 10 serait plus adaptée à l'analyse quantitative. En effet un test paramétrique tel que l'ANOVA, pourrait alors s'appliquer. Notons cependant que l'amplitude de cette échelle peut s'avérer non pertinente vis-à-vis du jugement de l'externalisation qui parait complexe à mesurer. D'autre part, en ce qui concerne la perception de l'espace, l'usage des filtres 1040 en tant que référence peut être remis en cause, dans la mesure où ils ne constituent pas nécessairement un idéal d'externalisation. Encore faut-il déterminer cet éventuel idéal si il existe. S'agit-il de l'écoute réelle de la diffusion des sources sonores ? Une sensation d'espace quelconque, sans exigence quant à la profondeur des sources sonores et alliée à un apprentissage, peut également s'avérer suffisante. Dans ce dernier cas, l'appréciation absolue de l'externalisation serait plus pertinente.

Dans le but d'effectuer des tests plus rigoureux une répétition de chaque évaluation peut être mise en place. Il faut cependant veiller à ce que la durée du test ne soit pas trop importante afin de ne pas introduire de biais en lien avec la fatigue. Le nombre de filtres à étudier peut alors être réduit suite à ce premier test. En effet, les méthodes relatives à l'utilisation de diracs numériques peuvent être écartées tant leur externalisation ne semble pas être convaincante. De plus, l'utilisation des HRTF médianes et «sen05_30» ne serait pas justifiée

car le test a confirmé que les modifications spectrales correspondantes ne se distinguent pas suffisamment de celle de la 1040. En revanche, il serait intéressant de trouver une limite au-delà de laquelle le fenêtrage spatial est significatif pour les méthodes «ISD» et «Sennheiser». De plus, pour cette dernière méthode, l'utilisation de coefficients c différents mais proches de 0 pourrait être pertinente pour améliorer la capacité d'externalisation tout en conservant une faible coloration spectrale. L'introduction de plusieurs filtres calculés numériquement pourrait également être intéressant. La flexibilité de cette méthode permettrait d'évaluer un grand nombre de facteurs différents potentiellement responsables de la coloration.

On peut également imaginer mélanger l'ordre de présentation des filtres à évaluer selon les sujets, afin de supprimer tout biais éventuel en lien avec l'effet de précédence. L'usage de sujets professionnels familiers de l'écoute binaurale pourrait également améliorer la fiabilité des résultats, tout comme la multiplication du nombre de stimuli par test.

Conclusion Générale

Le mixage orienté objet peut constituer une solution à l'évolution des habitudes d'écoute en lien avec la multiplication des systèmes de diffusion. Il permet de s'adapter aux avancées technologiques offrant une multitude de moyens de reproduction toujours plus performants et qui ne sont pas nécessairement standardisés. Par l'usage d'un format compatible flexible, l'intérêt de cette approche est de répondre aux choix d'écoute des auditeurs de manière cohérente et fidèle au mixage effectué par l'ingénieur du son. Et ce malgré la grande variabilité des conditions d'écoute et des systèmes de diffusion. La nécessité de compatibilité entre les dispositifs implique un monitoring sur différents systèmes de reproduction. La stéréophonie sur enceintes et le traitement binaural constituent deux approches différentes dans la constitution d'une scène sonore.

L'usage de ces deux modes de reproduction permet de balayer un large éventail de dispositifs présentant des caractéristiques différentes liées à l'effet de masquage et aux sources fantômes. La création de ces dernières est un point important du mixage orienté objet du fait de sa nécessité de s'adapter à des dispositifs spatialement peu définis. L'usage de prises de son adéquates, d'objets sonores décorrélés ou d'une description spatiale dense permet de prévenir d'éventuelles interférences négatives entre les objets.

Tout système de reproduction nécessite un apprentissage minimal afin d'exploiter au

mieux les possibilités offertes par le système. Il est important d'apprendre à utiliser le binaural interactif tant au niveau de la mise en espace des contenus que des conditions de restitution. Ceci implique d'identifier les prises de sons adéquates, de s'adapter à la restitution peu précise des images fantômes et de s'accommoder de l'absence éventuelle d'acoustique d'écoute. Étant donné la plasticité de l'audition, l'apprentissage est une alternative au problème complexe qu'est l'individualisation. Il permet notamment de se familiariser avec l'outil afin d'assimiler les distorsions de localisation et améliorer sa capacité d'externalisation.

Le contenu fréquentiel des sources est une composante essentielle à prendre en compte dans la réalisation d'un mixage. Comme nous l'avons exposé dans les chapitres précédents, le binaural présente cependant des colorations spectrales défavorables pour cette tâche en particulier mais également vis-à-vis de la compatibilité avec la stéréophonie au casque à laquelle il est souvent comparé. Nous considérons le binaural comme un moyen de reproduction sonore, auquel on demanderait d'assurer une fidélité vis-à-vis du timbre et de procurer une sensation d'espace. Malgré l'impact de cette coloration sur la perception de l'espace, il présente un désavantage sans justifier un réel atout du fait de la mauvaise capacité d'externalisation dans les zones frontale et arrière. Il a alors été entrepris de trouver un meilleur compromis entre la sensation d'espace et la fidélité du timbre.

Des méthodes de traitement des filtres HRTF ont alors été développées afin d'optimiser ce compromis. Des modifications spectrales plus ou moins importantes ont été effectuées sur les filtres de la zone frontale de l'espace, jugée plus significative. Afin d'évaluer de la pertinence des traitements réalisés et dans le but d'identifier des axes de développement vers lesquels s'orienter, des tests perceptifs en lien avec la perception de l'externalisation et du timbre ont été réalisés. Alors que la suppression des indices spectraux altère significativement la capacité d'externalisation, l'utilisation des différences spectrales relatives entre les filtres gauche et droite, ou ISD, semble fournir un bon compromis et ce pour une zone frontale large. L'usage de ces HRTF fournit une capacité d'externalisation non significativement différente de celle des filtres utilisés en guise de référence et une qualité de timbre proche de la stéréophonie. Le paramétrage de la méthode proposée par Sennheiser pourrait faire l'objet d'une étude plus conséquente afin d'améliorer la capacité d'externalisation tout en conservant une bonne qualité spectrale. Certaines méthodes ont pu être écartées et d'autres mériteraient de figurer dans des tests plus formels destinés à quantifier l'apport des traitements réalisés. Des améliorations relatives aux nombres de stimuli, de sujets et aux choix des références peuvent être apportées en ce sens.

De nombreuses voies de développements sont donc envisageables afin de permettre au binaural la possibilité de constituer un outil de travail fiable pour les professionnels et un moyen de reproduction de qualité pour les particuliers.

Références Bibliographiques

[1] E. Tech, "3364, audio definition model," *Geneva, January*, 2014.

[2] J.-C. Messonnier and A. Moraud, "Auditory distance perception : Criteria and listening room," in *Audio Engineering Society Convention 130*, Audio Engineering Society, 2011.

[3] G. Theile and G. Plenge, "Localization of lateral phantom sources," *Journal of the Audio Engineering Society*, vol. 25, no. 4, pp. 196–200, 1977.

[4] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, 1992.

[5] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[7] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2053–2064, 2002.

[8] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3248–3259, 1996.

[9] Y. Kahana, P. A. Nelson, M. Petyt, and S. Choi, "Boundary element simulation of HRTFs and sound fields produced by virtual acoustic imaging systems," in *Audio Engineering Society Convention 105*, Audio Engineering Society, 1998.

[10] Y. Kahana, P. A. Nelson, M. Petyt, and S. Choi, "Numerical modelling of the transfer functions of a dummy-head and of the external ear," in *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, Audio Engineering Society, 1999.

[11] J. Faure and G. PALLONE, "Evaluation de la synthese binaurale dynamique," tech. rep., Tech. Rep., France Telecom, 2005.

[12] D. Begault, E. Wenzel, A. Lee, and M. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *AES Convention108, Paper Number 5134*, 2000.

[13] G. Theile, *On the localisation in the superimposed soundfield*. PhD thesis, Technische Universität Berlin, 1980.

[14] A. Blum, B. F. Katz, and O. Warusfel, "Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training," in *Proc. CFA/DAGA*, vol. 4, 2004.

[15] G. S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, 1995.

[16] J.-C. Messonnier, J.-M. Lyzwa, D. Devallez, and C. de Boishéraud, "Object-based audio recording methods," *Journal of the Audio Engineering Society*, vol. 140, 2016.

[17] J.-C. Messonnier and A. Baskind, "Produire pour plusieurs types de systèmes de restitution : une approche de la production par objets.," *VDT Magazin*, 2016.

[18] R. Nicol, L. Gros, C. Colomes, E. Roncière, and J.-C. Messonnier, "Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation," *CFA / VISHNO*, 2016.

[19] V. Larcher, J.-M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Audio Engineering Society Convention 105*, Audio Engineering Society, 1998.

[20] A. van Opstal and T. v. Esch, "Estimating spectral cues underlying human sound localization," 2003.

[21] P. J. Bloom, "Creating source elevation illusions by spectral manipulation," *Journal of the Audio Engineering Society*, vol. 25, no. 9, pp. 560–565, 1977.

[22] B. Friedlander and B. Porat, "The modified yule-walker method of ARMA spectral estimation," *Aerospace and Electronic Systems, IEEE Transactions on*, no. 2, pp. 158–173, 1984.

[23] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[24] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.-l. Thomas, and F. Wu, "Time-reversed acoustics," *Reports on progress in Physics*, vol. 63, no. 12, p. 1933, 2000.

[25] M. Aussal, *Méthodes numériques pour la spatialisation sonore, de la simulation à la synthèse binaurale*. PhD thesis, École Polytechnique, 2014.

[26] F. Alouges and M. Aussal, "The sparse cardinal sine decomposition and its application for fast numerical convolution," *Numerical Algorithms*, vol. 70, no. 2, pp. 427–448, 2015.

[27] M. Morimoto, "The contribution of two ears to the perception of vertical angle in sagittal planes," *The Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1596–1603, 2001.

[28] P. Hofman and A. Van Opstal, "Binaural weighting of pinna cues in human sound localization," *Experimental brain research*, vol. 148, no. 4, pp. 458–470, 2003.

[29] J. Merimaa, "Modification of HRTF filters to reduce timbral effects in binaural synthesis," *Journal of the Audio Engineering Society.*, vol. AES 127th Convention, 2009.

[30] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[31] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[32] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 99–102, IEEE, 2001.

[33] L. Rayleigh, "XII. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[34] J. Blauert, *Spatial hearing : the psychophysics of human sound localization*. MIT press, 1997.

[35] M. B. Gardner, "Historical background of the Haas and/or precedence effect," *The Journal of the Acoustical Society of America*, vol. 43, no. 6, pp. 1243–1248, 1968.

[36] R. S. Woodworth, H. Schlosberg, J. W. Kling, and L. A. Riggs, *Woodworth & Schlosberg's Experimental psychology*. Holt, Rinehart and Winston, 1971.

[37] R. Nicol, *Représentation et perception des espaces auditifs virtuels*. PhD thesis, Université du Maine, 2010.

[38] V. Larcher, *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Paris VI, 2001.

[39] B. Rakerd, W. M. Hartmann, and T. L. McCaskey, "Identification and localization of sound sources in the median sagittal plane," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2812–2820, 1999.

[40] E. A. Macpherson, *Spectral cue processing in the auditory localization of sounds with wideband non-flat spectra*. University of Wisconsin–Madison, 1998.

[41] W. R. THURLOW and C. E. JACK, "Certain determinants of the" ventriloquism effect"," *Perceptual and motor skills*, vol. 36, no. 3c, pp. 1171–1184, 1973.

[42] E. Hendrickx, M. Paquier, V. Koehl, and J. Palacino, "Effet ventriloque pour des sources sonores variant simultanément en azimut et elévation," in *CFA 2016 (Congrès Français d'Acoustique)*, pp. 2313–2319, 2016.

[43] D. Devallez, "Auditory perspective : perception, rendering, and applications," *PhD thesis. University of Verona*, 2009.

[44] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller, "The interaural time difference in binaural synthesis," in *Audio Engineering Society Convention 108*, Audio Engineering Society, 2000.

[45] P. Minnaar, F. Christensen, H. Moller, S. K. Olesen, and J. Plogsties, "Audibility of all-pass components in binaural synthesis," in *Audio Engineering Society Convention 106*, Audio Engineering Society, 1999.

[46] A. Kulkarni, S. Isabelle, and H. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2821–2840, 1999.

[47] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[48] S. Busson, R. Nicol, and B. Katz, "Subjective investigations of the interaural time difference in the horizontal plane," in *118th Audio Engineering Society Convention*, 2005.

[49] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, A. I. Tew, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *Multimedia, IEEE Transactions on*, vol. 16, no. 1, pp. 37–46, 2014.

Annexes

ANNEXE A

La localisation sonore

Les différences interaurales de niveau et de temps

Ces différences sont liées à la morphologie de l'auditeur et à la distance interaurale, de 14.49 cm en moyenne dans l'étude menée par le laboratoire CIPIC [32]. Ces caractéristiques jouent sur le retard et l'intensité du son. La conjugaison des différences d'intensité, ou ILD, et des différences de temps, ou ITD, permet à l'oreille de localiser les sources situées dans l'espace [33]. La prépondérance de l'ITD ou de L'ILD en tant qu'indice de localisation dépend de la nature du son en lui-même. En effet, les études de Kuhn et de Blauert ont mis en évidence la prépondérance de l'ILD pour des sons dénués de fréquences en-deçà de 2 kHz environ [34].

A titre d'exemple, l'ILD permet de situer une source sonore car un son est localisé du côté de l'oreille l'ayant perçu avec le niveau le plus élevé. Par effet de précédence ou effet Haas, l'oreille localise également une source sonore du côté du premier front d'onde qui lui parvient au delà d'un certain ITD [35]. La figure A.1 illustre schématiquement la différence de marche d'une onde sonore entre les deux oreilles.

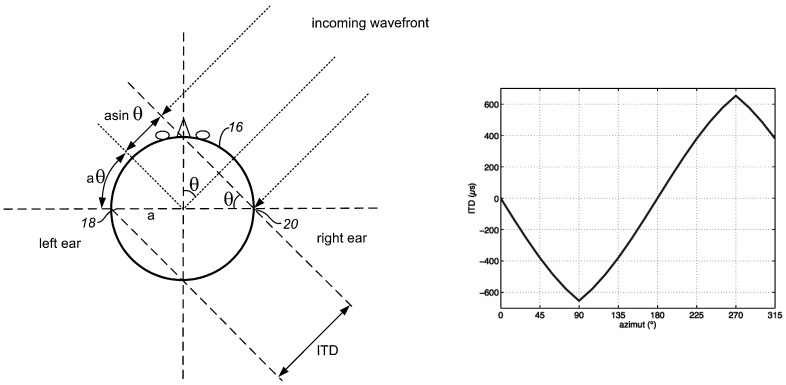
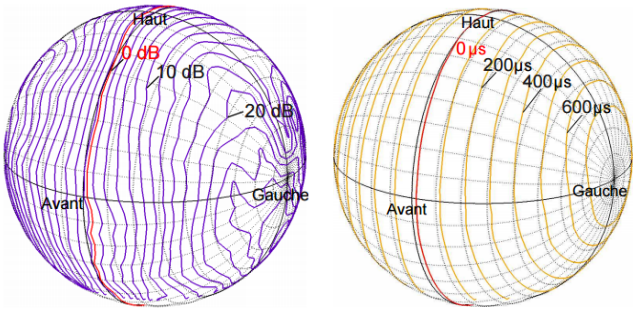


FIGURE A.1 – Modèle de tête représentant les décalages temporels entre les signaux parvenant aux oreilles [36].

En élévation, la localisation est moins précise : les différences interaurales de temps ou d'intensité sont les mêmes pour deux sons situés à un azimuth donné et des élévations différentes. De plus, ne considérer que l'ITD ou l'ILD pour caractériser la localisation auditive entraînerait l'apparition d'un cône de confusion, représenté sur la figure A.3. En effet, les différences se répètent dans l'espace par symétrie. La première conséquence est une ambiguïté entre les parties avant et arrière de l'espace sonore.



(a) (b)

FIGURE A.2 – Exemples de lignes iso-ILD (*a*) et iso-ITD (*b*) présentés dans [37], p. 122. Ces courbes illustrent les cônes de confusion relatifs aux différences interaurales de niveau et de temps de mêmes valeurs dans l'espace.

Les indices spectraux

Les différences de temps et d'intensité ne permettent pas à elles seules la localisation d'une source sonore dans l'espace. Des modifications spectrales existent et sont dues à des réflexions et diffusion dépendantes de caractéristiques morphologiques du sujet et de son positionnement par rapport à la source. Les travaux de Blauert montrent l'existence de "bandes directionnelles", des intervalles fréquentiels qui induisent une localisation dans une zone de l'espace déterminée [34]. Ces indices spectraux permettent de distinguer l'élévation d'une source en discriminant les incidences hautes/basses et lèvent les ambiguïtés avant/arrière.

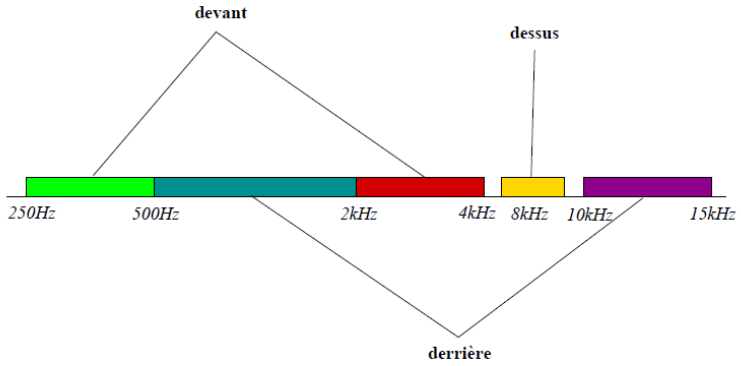


FIGURE A.3 – Intervalles fréquentiels identifiés par Blauert [34], d'après [38], p. 10

L'identification des indices spectraux fait l'objet de deux interprétations différentes :

— le système auditif analyse les caractéristiques locales des HRTF telles que les bosses et les creux [21] ;

— le système auditif analyse le spectre dans son intégralité et la localisation sonore s'effectuerait alors à la suite d'une comparaison entre le contenu fréquentiel perçu et une base de données interne constituée par apprentissage à l'aide de ses propres HRTF [13, 20] ;

Actuellement l'avancée des recherches ne permet pas de trancher quant à la validité de l'une ou l'autre option, les mécanismes de la perception vis-à-vis de la composante spectrale seule n'étant pas assez connus. Une autre question est de savoir si le système auditif extrait les informations spectrales issues des signaux droit et gauche de manière indépendante ou conjointement. Les études effectuées par Morimoto ou Hofman mettent en avant le rôle prédominant de l'oreille ipsolatérale, la plus proche de la source sonore, pour des azimuts supérieurs à 30 ° [27, 28]. Au-delà de cet angle les indices spectraux peuvent être considérés comme monoraux en dehors de la zone frontale. Dans cette zone cependant, il reste à savoir

si l'analyse s'effectue de manière absolue ou de manière relative. En d'autres termes, si ce sont les différences spectrales seules qui importent. Il s'avère que l'analyse des différences spectrales seules ne suffit pas à garantir une performance de la localisation en élévation [39, 40].

La perception de la distance

La distance est déduite de manière complexe à travers la perception de plusieurs paramètres :

— le rapport entre le champ direct et le champ réverbéré ;
— le rapport entre l'énergie précoce et l'énergie tardive de la réverbération ;
— le détimbrage du son direct ;
— le niveau sonore de la source.

L'appréciation de la distance est donc fonction du niveau d'une source sonore, de son spectre et de l'effet de salle. Quand la distance entre une source sonore et la tête d'un auditeur augmente, l'intensité et l'énergie dans les fréquences aigües diminuent. En champ réverbéré, le rapport entre le champ direct et le champ diffus diminue également. Notons qu'en binaural, le point de mesure des HRTF influe sur la perception de la distance. Hormis ce facteur et le niveau sonore absolu qui dépendent du moteur binaural, l'information de distance est intrinsèque aux signaux traités.

Les facteurs cognitifs

De nombreux facteurs liés à notre connaissance peuvent interférer avec notre perception auditive. Le contenu sémantique des sons peut notamment influencer la localisation des sources. L'écoute d'enregistrement d'oiseaux, d'avions ou d'hélicoptères peut par exemple faciliter une perception zénithale de leur incidence. Le contexte culturel de l'auditeur peut alors constituer une variable significative pour la localisation.

La vue peut également avoir une influence non négligeable sur la perception auditive. En effet, la localisation d'une source sonore peut être assimilée à celle d'une source visuelle avec des différences angulaires très élevées. Selon Jack et Thurlow un sujet peut être trompé par un écart angulaire de 20 ° dans le plan azimutal [41]. Cet effet est un phénomène d'aimantation spatiale appelé l'effet ventriloque. Une étude réalisée par Hendrickx a montré que cet

effet fonctionnait bien mieux en élévation qu'en azimuth [42]. Selon Devallez [43], des effets similaires peuvent être perçus avec la distance.

L'influence de la vision sur notre système auditif peut être une des causes de la perception intra-crânienne et du défaut d'externalisation dans la zone frontale. On peut imaginer que de très importantes différences de localisations sonores puissent apparaître entre deux écoutes ; avec et sans représentations visuelles des sources.

ANNEXE B

Les filtres HRTF à phase minimale

Une méthode commune pour modéliser un filtre HRTF consiste à calculer le filtre à phase minimale correspondant et à introduire un retard calculé en fonction de la position de la source. Le filtre à phase minimale possède le même contenu fréquentiel que celui de la HRTF d'après lequel il est calculé. La modélisation consiste seulement en une manipulation de la réponse en phase des filtres. Une HRTF étant un filtre causal et stable, il peut être décomposé en une composante à phase minimale et une composante passe-tout. Considérons un filtre HRTF donné $H(\omega)$.

$$H(\omega) = |H(\omega)|.e^{j\phi(\omega)}. \tag{B.1}$$

La phase peut être divisée en une composante à phase minimale et une composante de phase en excès.

$$H(\omega) = |H(\omega)|.e^{j\phi_{min}(\omega)}.e^{j\phi_{exces}(\omega)}. \tag{B.2}$$

Or, jusqu'à 1.5 kHz environ, on peut considérer que la phase en excès est linéaire [44]. La composante de phase en excès peut alors être décomposée en une composante linéaire et

une composante passe-tout.

$$H(\omega) = |H(\omega)|.e^{j\phi_{min}(\omega)}.e^{j\phi_{lin}(\omega)}.e^{j\phi_{passe-tout}(\omega)}. \tag{B.3}$$

Selon l'étude menée par Paulli Minnaar, la composante passe-tout peut être négligée sans perturber la perception spatiale [45]. Il a en effet été montré qu'elle est inaudible pour la plupart des directions d'incidence. On peut alors approcher de manière satisfaisante le filtre HRTF en lui substituant son homologue à phase minimale auquel on ajoute une phase linéaire, soit un retard pur.

$$H(\omega) = H_{min}(\omega).e^{j\phi_{lin}(\omega)}, \tag{B.4}$$

où $H_{min}(\omega) = |H(\omega)|.e^{j\phi_{min}(\omega)}$ correspondant au filtre à phase minimale associé à $H(\omega)$.

La figure B.1 illustre deux réponses impulsionnelles dont l'une est la version à phase minimale de l'autre. Les deux ont la même réponse en fréquence, mais des réponses en phase différentes. Afin d'introduire des différences de temps entre les HRTF d'une position donnée, il est alors nécessaire de calculer une phase linéaire fonction de l'angle d'incidence.

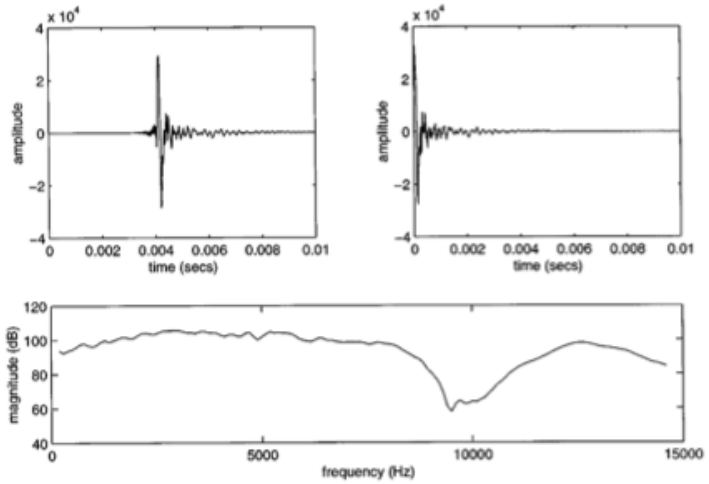


FIGURE B.1 – HRIR avec et sans la composante de phase en excès ainsi que la réponse en fréquence commune aux deux [46].

L'estimation du retard interaural

Deux approches sont possibles pour estimer le retard interaural : la modélisation géométrique ou l'extraction de l'information temporelle d'après les mesures. Dans le cadre de la première approche, la formule B.5 permet d'estimer l'ITD, qui déterminera le retard pur des filtres HRIR, en se basant sur un seul paramètre anthropométrique, le rayon de la tête et sur la position de la source [36] :

$$ITD = \frac{r}{c}.(\sin(\theta) + \theta).\cos(\psi) \tag{B.5}$$

avec r : le rayon de la tête, c : la vitesse du son, θ : l'azimuth et ψ : l'élevation.

Comme l'illustre la figure A.1, il s'agit du calcul de la différence de marche entre les deux oreilles situées sur une tête considérée sphérique et dans l'approximation des ondes planes. Le terme $\cos(\psi)$ est introduit en raison d'un meilleur accord avec les résultats empiriques obtenus en élévation [47].

Ainsi les filtres correspondant à l'oreille contralatérale (opposée à l'incidence du son) et à l'ipsolatérale sont respectivement retardés de :

$$\tau_{contra} = \tau_{offset} + \frac{ITD}{2} \tag{B.6}$$

et

$$\tau_{ipso} = \tau_{offset} - \frac{ITD}{2}. \tag{B.7}$$

D'autres méthodes en lien avec l'extraction de l'ITD d'après les mesures d'HRTF peuvent être utilisées. Nous pouvons les classer en trois catégories [48] :

— les méthodes d'inter-corrélation entre le filtre gauche et le filtre droit ou entre les HRIR et leurs représentations à minimum de phase respectives ;
— la détection d'un seuil de montée de la HRIR pour évaluer le temps d'arrivée de l'onde à l'oreille ;
— la détection de la pente de la phase en excès ou l'estimation de son retard de groupe.

ANNEXE C

Modélisation géométrique par maillage 3D

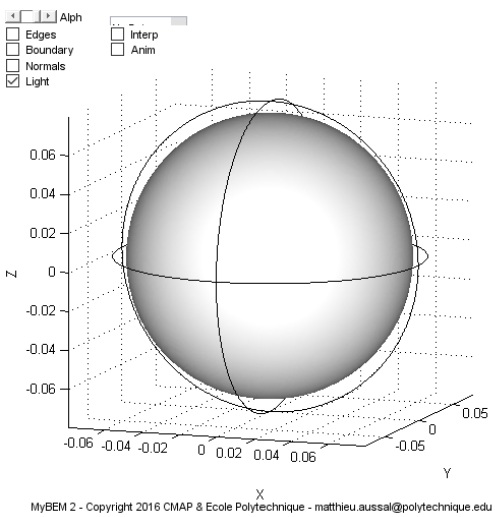


FIGURE C.1 – Maillage d'une sphère de rayon 14.49 cm

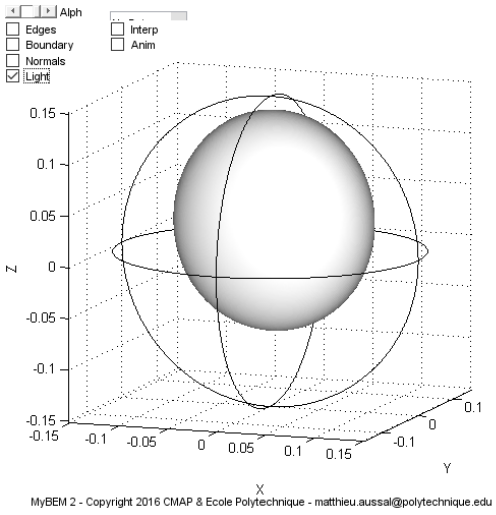


FIGURE C.2 – Maillage d'un ovoïde d'une largeur, hauteur et profondeur respectivement égales à 14.49 cm, 21.46 cm et 19.96 cm.

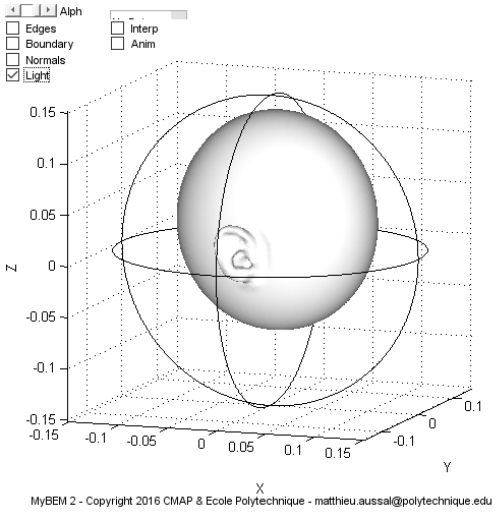


FIGURE C.3 – Maillage d'un ovoïde d'une largeur, hauteur et profondeur respectivement égales à 14.49 cm, 21.46 cm et 19.96 cm auquel est associée une numérisation de la paire d'oreille de la tête artificielle Neumann.

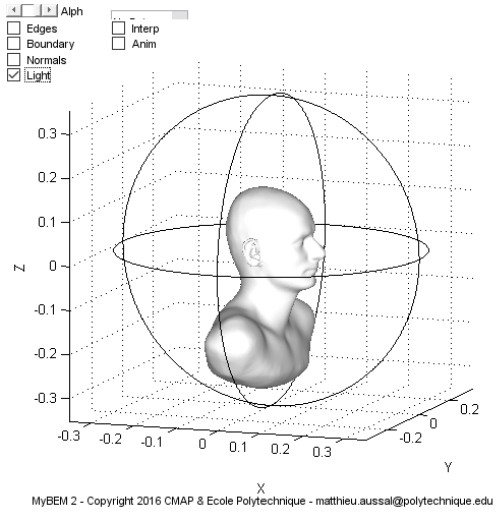


FIGURE C.4 – Maillage de la tête et du torse HTE03 issu de la base SYMARE [49].

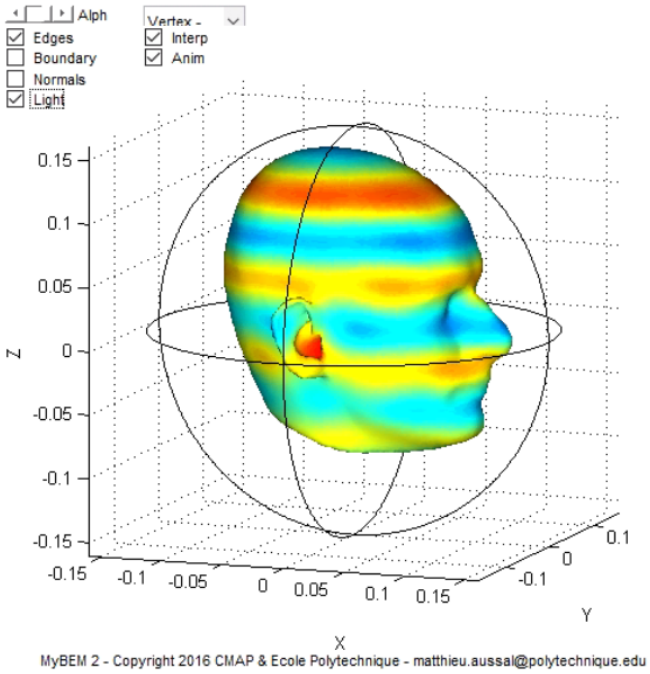


FIGURE C.5 – Calcul dans MyBEM de la propagation acoustique d'une onde sur le maillage tri-dimensionnel de la tête HE03, issu de la base SYMARE [49].

ANNEXE D

Questionnaire des tests perceptifs préliminaires

Optimisation d'un moteur binaural interactif pour le mixage orienté objet

François Salmon

***

L'objet de ce mémoire consiste à définir dans quelle mesure le traitement binaural peut être amélioré afin d'apporter aux professionnels un outil performant dans le cadre d'une production en mixage orienté objet.

Le moteur binaural interactif utilisé lors de ce test contient des filtres modifiés destinés à s'affranchir de la coloration spectrale. Pour considérer un monitoring binaural fiable, il est en effet nécessaire pour les professionnels de posséder une qualité de timbre non dégradée par le traitement du signal.

La pertinence des filtres sera évaluée au cours de deux tests perceptifs. Ces développements étant destinés aux conditions d'utilisation professionnelles, seuls des ingénieurs du son et étudiants ingénieurs du son font partie des sujets. Les tests présentés ici ne seront pas soumis à une étude statistique mais à une appréciation qualitative. En effet, le temps imparti dont nous disposons ne permet pas d'utiliser un nombre de sujets conséquent. De plus, la diversité des stimuli implique de nombreuses variables à étudier. Les résultats permettront d'envisager des voies de développement futures.

Le moteur binaural utilisé dans ce test est un prototype issus des recherches en son3D effectuées au Centre de Mathématiques Appliquées de l'école Polytechnique. Le panneau de contrôle permet aux sujets de sélectionner entre autres les stimuli ainsi que les jeux de filtres HRTF à écouter. Ce changement de filtre peut s'effectuer pendant l'écoute. Les fonctions de deux autres boutons méritent d'être précisées :

Ce commutateur permet de rendre actif ou non le traitement binaural. Si il n'est pas enclenché, un *down-mix* stéréophonique est alors diffusé.

Celui-ci permet la prise en compte des mouvements de la tête en traitant les données issues du *head-tracker* (HT). Il permet de plus calibrer le capteur : le moteur prendra comme point d'origine la position de la tête de l'auditeur au moment où il enclenchera le bouton. Veillez donc à garder votre tête fixe en regardant l'interface lorsque vous appuyez sur ce bouton.

La perception de l'externalisation

Ce test permet de comparer le sentiment d'externalisation selon les HRTF employées.

- **Durée : 5 à 10 min**
- **Stimuli : 1**
- **HRTF : A à I**

Qualifier l'externalisation des sources sonores selon les différentes HRTF comparées au jeu d'HRTF A en entourant la mention adéquate. Il est demandé d'écouter préalablement l'ensemble des filtrage binauraux avant de répondre au questionnaire. Seuls des mouvements de tête de moins de 60 ° par rapport à la position d'origine sont autorisés.

HRTF B :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF C :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF D :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF E :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF F :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF G :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF H :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure
HRTF I :	aucune -	beaucoup moins bonne -	moins bonne -	similaire -	meilleure

La perception du timbre

Ce test consiste à vérifier si la réduction de la coloration spectrale est significative.

- **Durée : 5 à 10 min**
- **Stimuli : 2**
- **HRTF : A à I**

Qualifier le timbre de la scène sonore selon les différentes HRTF comparées à la stéréophonie et entourer la mention adéquate. Il est demandé d'écouter préalablement l'ensemble des filtrage binauraux avant de répondre au questionnaire. Seuls des mouvements de tête de moins de 60 ° par rapport à la position d'origine sont autorisés.

HRTF A :	très coloré -	coloré -	peu coloré -	similaire
HRTF B :	très coloré -	coloré -	peu coloré -	similaire
HRTF C :	très coloré -	coloré -	peu coloré -	similaire
HRTF D :	très coloré -	coloré -	peu coloré -	similaire
HRTF E :	très coloré -	coloré -	peu coloré -	similaire
HRTF F :	très coloré -	coloré -	peu coloré -	similaire
HRTF G :	très coloré -	coloré -	peu coloré -	similaire
HRTF H :	très coloré -	coloré -	peu coloré -	similaire
HRTF I :	très coloré -	coloré -	peu coloré -	similaire

ANNEXE E

Représentations graphiques des HRTF

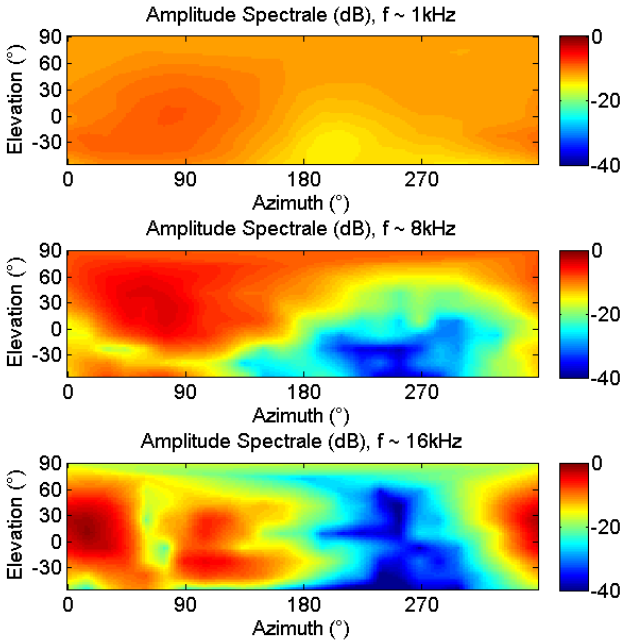
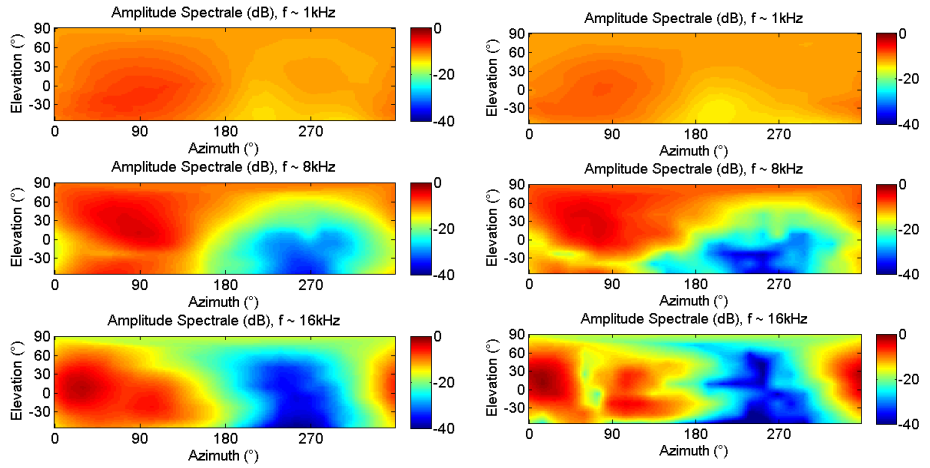
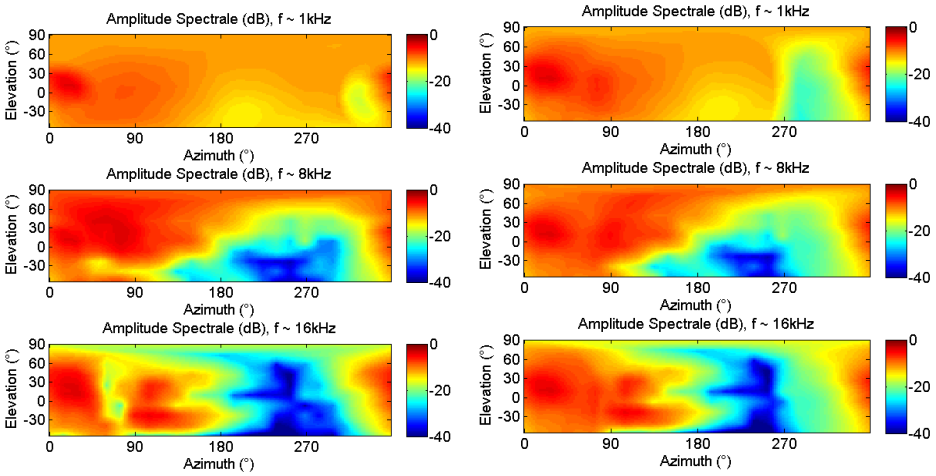


FIGURE E.1 – HRTF 1040 de la base Listen.



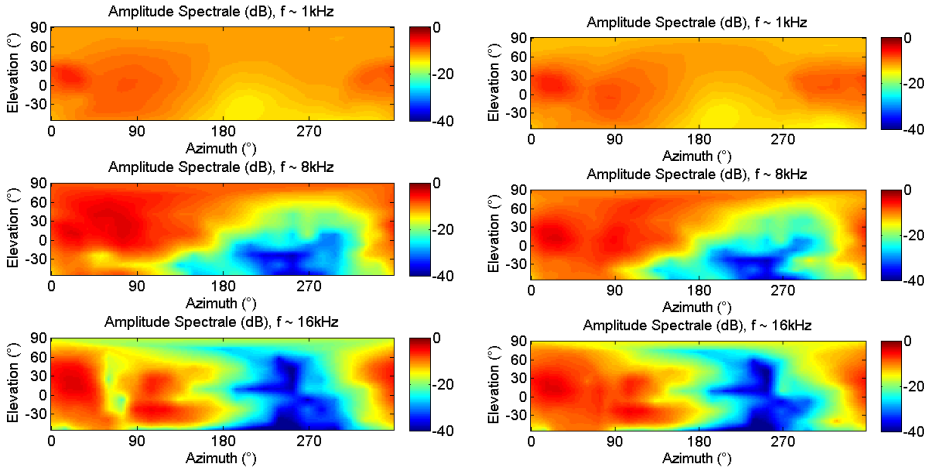
(a) (b)

FIGURE E.2 – HRTF médianes obtenues d'après la base Listen (a) et HRTF issues du traitement «sennheiser» sur une fenêtre de 30 ° pour un indice de traitement de 0.5 (b).



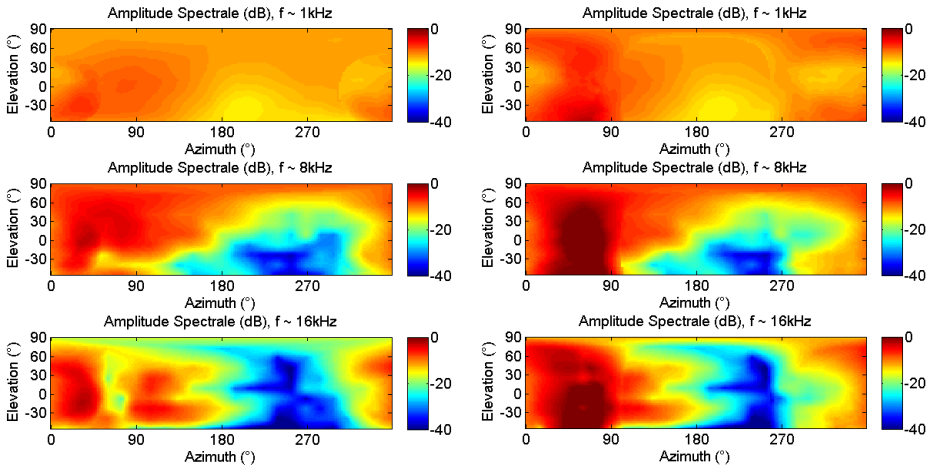
(a) (b)

FIGURE E.3 – HRTF issues d'un filtrage par dirac sur une fenêtre de 30 ° (a) et sur une fenêtre de 75 ° (b).



(a) (b)

FIGURE E.4 – HRTF issues du traitement «Sennheiser» sur une fenêtre de 30 ° (a) et sur une fenêtre de 75 ° (b).



(a) (b)

FIGURE E.5 – HRTF issues du traitement des ISD sur une fenêtre de 30 ° (a) et sur une fenêtre de 75 ° (b).

École nationale supérieure Louis-Lumière 2016