EXTERNALISATION DU SON AU CASQUE

Réduction de la perception intracrânienne des sources en synthèse binaurale par design spectral des HRTF

STAGE DE FIN D'ETUDES – RAPPORT FINAL

— Étudiant —

Arthur MINGASSON

Troisième année à l'École Centrale de Marseille

Master 2 Recherche au Laboratoire de Mécanique et d'Acoustique

Année scolaire 2014-2015

— Tuteur entreprise —

Rozenn NICOL – Orange Labs

2 av. Pierre Marzin, 22300 Lannion

— Rapporteurs —

Muriel ROCHE – École Centrale de Marseille

Sophie SAVEL – Laboratoire de Mécanique et d'Acoustique

Sommaire

Remerciements

Je tiens à remercier Rozenn Nicol pour m'avoir donné l'opportunité de réaliser ce stage et s'être montrée disponible durant ce stage malgré son implication dans de nombreux projets. Un grand merci pour ton soutien durant la rédaction de ce rapport.

Je remercie également les enseignants du LMA, et plus précisément Sabine Meunier et Sophie Savel pour leurs précieux conseils au long de ce stage.

Enfin, je remercie tous ceux dont j'ai partagé le quotidien à Orange durant ces six beaux mois en pays breton !

Conventions et vocabulaire

Conditions anéchoïques (ou de champ libre) : Environnement ne renvoyant aucun écho

ITD : Différence interaurale de temps (Interaural Time Difference)

ILD : Différence interaurale d'intensité (Interaural Level Difference)

HRTF : Fonction de transfert relative à la tête (Head-Related Transfer Function)

HRIR : Réponse impulsionnelle relative à la tête (Head-Related Impulse Response)

BRIR : Réponse impulsionnelle binaurale mesurée dans une chambre réverbérante (Binaural Room Impulse Response)

[θ, ϕ, r] : Coordonnées sphériques avec θ l'azimut de la source, ϕ l'élévation de la source et r la distance du sujet à la source (θ_{pv} et ϕ_{pv} sur la Figure 1). La coordonnée r sera parfois omise.

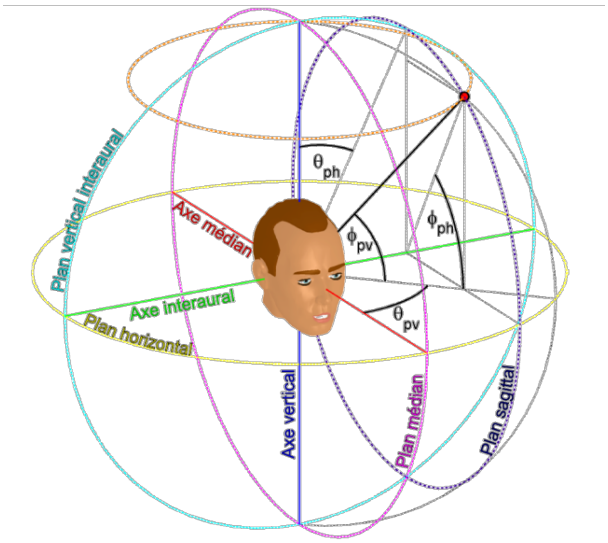

Figure 1. Système de coordonnées, d'après [20].

Introduction

Nos oreilles nous offrent la formidable capacité de pouvoir localiser des sons dans l'espace. Cette information spatiale est précieuse car elle nous permet de prendre conscience d'évènements ayant lieu hors de notre champ de vision. Le son est intégré à notre perception de l'espace. Par conséquent, pour réaliser un espace de réalité virtuelle immersif et réaliste, les sources sonores doivent être spatialisées. De nombreuses technologies de spatialisation du son existent, comme la restitution sur plusieurs haut-parleurs entourant l'auditeur.

Une autre approche, la technologie binaurale, consiste à diffuser au casque le son que l'auditeur aurait perçu dans l'environnement simulé. Pour spatialiser une source sonore, on applique au son d'origine toutes les transformations qu'il aurait subies lors de sa propagation de la source aux oreilles de l'auditeur.

Cette technologie a fait ses preuves depuis de nombreuses années. Son principe est simple et sa mise en œuvre ne requière qu'un casque et deux pistes audio. Elle peine néanmoins à se diffuser auprès du grand public car l'écoute binaurale au casque entraine parfois des perceptions erronées. Parmi ces artéfacts, les auditeurs constatent fréquemment que les sources sonores ne sont pas localisées à l'extérieur mais au contraire très proche de la tête voire à l'intérieur.

La perception de l'espace sonore se base sur une analyse fine des signaux qui arrivent à chaque oreille pour extraire des indices de spatialisation. Afin de résoudre le problème de perception intracrânienne, nous allons étudier la possibilité de renforcer certain de ces indices, les indices spectraux, afin de produire une externalisation des sources.

Ce stage de fin d'études s'inscrit dans le cadre du Master 2 Recherche au Laboratoire de Mécanique et d'Acoustique et de la troisième année à l'École Centrale Marseille et s'est déroulé à Orange Labs, sur le site de Lannion.

Le présent rapport expose dans un premier temps les fondements de la spatialisation sonore et de sa restitution binaurale. Il présente dans une seconde partie nos travaux de recherche et les outils développés pour étudier le défaut d'externalisation sous l'angle du design des indices spectraux.

Partie 1.
Contexte

1 Perception du son en 3D

Grâce à nos oreilles, nous sommes capables de localiser des sources sonores dans l'espace tridimensionnel qui nous entoure. La source sonore émet une onde de pression, laquelle atteint nos tympans avec un retard, une atténuation, et de multiples réflexions sur notre corps et éventuellement sur des objets qui nous entourent. Le signal mono émis par la source sonore est ainsi transformé en deux signaux légèrement différents que l'on appelle signal binaural. Notre système auditif extrait alors plusieurs indices de localisation de ce signal binaural et en déduit la position de la source sonore.

1.1 Indices interauraux

Si un auditeur est capable de localiser des sons, c'est principalement parce que ses deux oreilles reçoivent des signaux légèrement différents. La distance qui les sépare cause un retard de propagation ou ITD tandis que notre tête diffracte l'onde sonore et cause une chute d'amplitude ou ILD.

Nous sommes sensibles à l'ITD pour les fréquences inférieures à 2 kHz ; au-delà, le système auditif ne sait plus interpréter ce retard [**6**]. En revanche, les longueurs d'onde diffractées par notre tête correspondent à des fréquences supérieures à 1.5 kHz [**32**]. Ainsi l'ILD remplace l'ITD aux hautes fréquences. Ces deux indices sont donc complémentaires et nous permettent de latéraliser les sons, c'est-à-dire de distinguer la gauche et la droite, pour l'ensemble des fréquences audibles (de 20Hz à 20kHz).

Cependant, ITD et ILD ne permettent pas de distinguer un son provenant de l'avant d'un son venant de l'arrière. Cette ambiguïté se généralise sur un cône de confusion (cf. Figure 2) sur lequel ILD et ITD sont constantes. Pour lever cette ambiguïté, le système auditif doit alors extraire d'autres indices des signaux qu'il reçoit.

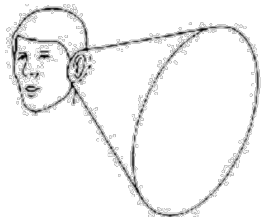

Figure 2. Cône de confusion sur lequel ITD et ILD sont constantes et ne suffisent pas pour déterminer la position de la source.

1.2 Indices spectraux monauraux

Lors de la propagation du son de la source vers l'auditeur, de nombreuses réflexions perturbent l'onde sonore avant qu'elle n'atteigne le tympan de chaque oreille. Ces réflexions ont lieu sur le pavillon de l'oreille, la tête et dans une moindre mesure sur le torse, seulement aux longueurs d'onde correspondant aux dimensions morphologiques de l'auditeur. Par exemple, une onde provenant de la gauche ($\theta = 90°$, $\phi = 0°$) occasionnera une réflexion sur l'épaule avant d'atteindre l'oreille, alors qu'une source frontale ($\theta = 0°$, $\phi = 0°$) occasionnera une réflexion sur le torse, avec un retard et une atténuation différents. D'un point de vue fréquentiel, ces perturbations de l'onde incidente se traduisent par des résonances et antirésonances qui colorent le spectre du signal source différemment selon la direction de la source. On parle d'indices monauraux car ils apparaissent indépendamment pour chaque oreille.

1.2.1 Head-Related Transfer Function (HRTF)

On appelle HRTF la fonction de transfert décrivant les changements de phase et de magnitude d'un son lorsqu'il voyage d'une source vers l'oreille d'un auditeur, dans un environnement anéchoïque. Une HRTF contient donc les indices spectraux monauraux pour une oreille. La connaissance de la paire de HRTF gauche et droite associée à une position permet aussi de déduire les indices interauraux (ITD et ILD) en comparant les phases et les amplitudes des HRTF. On a tracé Figure 3 le module spectral de HRTF pour deux directions.

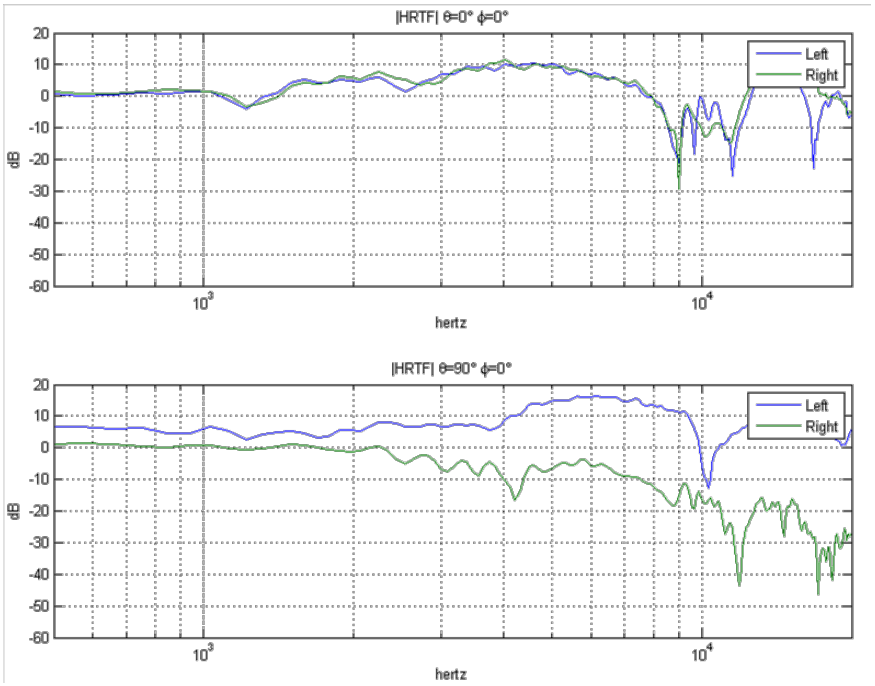

Figure 3. Paire de HRTF pour une source placée en face du sujet puis pour une source placée à gauche. Chaque courbe présente les indices monauraux pour chaque oreille, tandis que l'écart entre les courbes de chaque oreille constitue l'ILD. Alors qu'à azimut nul les courbes sont très semblables, on observe à 90° la diffraction par la tête avant que l'onde n'atteigne l'oreille droite. Sujet ArMi.

En champ libre, pour un auditeur fixe, les HRTF contiennent tous les indices disponibles pour localiser une source sonore dans une direction donnée. En revanche en condition d'écoute naturelle, le système auditif dispose de nombreux autres indices tout aussi importants.

1.2.2 Bandes de directivité

En faisant écouter des bandes de bruit étroites, Blauert montre que l'élévation perçue par l'auditeur est directement déterminée par la fréquence centrale de ces bandes, alors que la source réelle demeure fixe [**5**]. Middlebrooks confirme ces observations en diffusant des bandes d'⅙ d'octave de fréquences centrales variables, depuis des positions variables [**47**]. Il observe que tandis que l'azimut perçu correspond toujours à l'azimut réel de la source, l'élévation perçue est uniquement liée aux fréquences diffusées. L'auditeur perçoit la source dans la direction où cette fréquence correspond à un maximum d'amplitude dans la réponse directionnelle associée.[1] Middlebrooks nomme ce type d'indice « référent spatial ». Ces indices correspondent à un

[11]Maximum spatial à une fréquence donnée de la SFRS, diagramme de directivité correspondant aux HRTF évaluées pour toutes les directions à une fréquence donnée.

maximum spatial, à fréquence fixe. Ainsi dans une direction donnée, cette fréquence ne correspondra pas forcément à un maximum des HRTF tracées suivant la fréquence. C'est pourquoi le terme de pic dissimulé (ou *covert peak*) est aussi employé par Humanski et Butler pour désigner un maximum qui ne peut être repéré qu'après avoir écouté toutes les directions [**27**].

1.2.3 Indices manifestes

Il n'est pas évident de transposer ces résultats à l'écoute quotidienne des sons. Ils ont été obtenus à partir de bandes étroites tandis qu'au quotidien, nous sommes plutôt soumis à des sons large-bande.

Dans ce cas d'un spectre large à une position fixe, notre système auditif se baserait alors aussi sur la forme du filtrage réalisée dans cette direction. Ce sont les indices spectraux manifestes (ou *overt features*) dans le sens où l'écoute dans une seule direction fait apparaître tous ces indices. On leur oppose les indices dissimulés évoqués plus haut, lesquels requièrent une connaissance des filtrages dans toutes les directions.

Il n'y a à ce jour pas de consensus sur la prédominance d'un indice sur l'autre, et par conséquent, les indices manifestes et dissimulés doivent être pris en compte, selon [**65**], section 4.5.3.

1.2.4 Région fréquentielle des indices spectraux

Stevens et Newmann déterminent dans [**68**] l'étendue spectrale d'un signal requise pour éviter une confusion avant/arrière. Les auteurs montrent que tant que le spectre est limité à [0—2 kHz], le taux de confusion reste élevé et constant. À partir de 2 kHz, le taux de confusion baisse et devient faible et constant au-delà de 4 kHz. On considère couramment que les indices spectraux se situent aux fréquences supérieures à 3 ou 4 kHz. Algazi *et al.* démontrent l'existence d'indices spectraux en dessous de 3 kHz [**1**] mais il est probable que ces derniers n'entrent en jeu que lorsque le spectre est nul aux fréquences supérieures.

1.2.5 Interprétation des indices spectraux

Le système auditif se base sur les indices spectraux pour distinguer un son venant de l'arrière d'un son venant de l'avant et plus généralement pour déterminer l'élévation de la source. L'interprétation de ces indices par le système nerveux dépasse le cadre de ce stage, et l'on se contentera de citer Guillon, section 1.2.2 de [**20**]:

> « *Grossièrement, la localisation sur la base des indices spectraux s'apparenterait à un processus d'identification entre les filtrages subis par le signal source et les indices spectraux "stockés" en mémoire. Le système auditif utiliserait en quelque sorte des facultés de reconnaissance de formes.* »

1.3 Autres indices

On a présenté ci-dessus les principaux indices mis en jeu dans des conditions de laboratoire, c'est à dire pour un dont la tête est sujet fixe, dans un environnement anéchoïque, avec une unique source présentée à distance constante. Notre perception des sons aux quotidiens est évidemment bien plus complexe et il est rare qu'une seule source sonore soit présente.

1.3.1 Indices de distance

En champ libre, la décroissance du niveau sonore est uniquement liée à la distance séparant l'auditeur de la source. La décroissance est de 6 dB par doublement de distance. Le niveau constitue donc un bon indice de distance à condition que le niveau de la source soit invariant et que l'auditeur le connaisse [**46**].

Il est néanmoins rare que nous n'entendions que le son direct d'une source. Souvent, un effet de salle est présent, lequel se traduit par des réflexions sur les objets et plus globalement par un temps de réverbération. Dans ces conditions, la loi de décroissance n'est plus vérifiée mais on peut alors définir le ratio d'énergie entre le champ direct et le champ réverbéré. Ce ratio constitue généralement un bon indicateur de distance [**71**] : à proximité de la source, le champ direct est prépondérant tandis que loin de la source, le champ direct se fond dans le champ réverbéré.

1.3.2 Indices dynamiques

À l'écoute d'un son, l'auditeur a le réflexe d'orienter sa tête dans la direction de la source sonore. Cela lui permet d'utiliser sa vision et lui procure aussi la localisation auditive dans la région frontale qui est bien plus précise. De plus, les mouvements de la tête permettent une meilleure spatialisation, même pour des oscillations de faibles amplitudes.

On appelle indices dynamiques les informations que l'auditeur déduit du lien entre le mouvement de sa tête et la variation des signaux à ces tympans. Les indices dynamiques permettent d'améliorer grandement la résolution de l'élévation et font fortement diminuer les inversions avant/arrière (cf. section 1.3 de [**20**]).

2 Écoute binaurale

Nous entendons naturellement les sons des objets qui nous entourent par une écoute binaurale. Arrivent à nos tympans des signaux audio contenant des indices qui nous permettent de localiser les sources. Cette écoute immersive en 3D peut être simulée en diffusant à l'entrée du conduit auditif de chaque oreille le son qu'il aurait reçu s'il avait été réellement placé dans cet environnement. Cette partie présente comment obtenir ce signal binaural.

2.1 Enregistrements binauraux

On peut enregistrer un signal binaural en plaçant une paire de microphones dans les oreilles d'un individu, lequel se trouve dans un lieu quelconque, entouré de sources sonores réelles. L'individu peut aussi être remplacé par un mannequin sur lequel on pose des pavillons artificiels munis de microphones de mesure (cf. Figure 4).

Figure 4. Tête de mannequin Neumann KU 100.

Il ne reste alors plus qu'à diffuser les signaux enregistrés au niveau du tympan de chaque oreille, de façon préférentielle en utilisant un casque. Cette méthode présente des limites. Premièrement, elle exige que l'on puisse réellement placer le sujet équipé de micros ou le mannequin dans la scène sonore. Ensuite, on pourra difficilement modifier ces enregistrements binauraux ultérieurement. Impossible par exemple pour l'ingénieur du son de déplacer dans l'espace tel ou tel évènement sonore à partir du signal binaural. Plus généralement, on aimerait pouvoir créer des scènes sonores en plaçant les sources sonores où bon nous semble, à des fins artistiques ou de réalité virtuelle, ce que la synthèse binaurale permet.

2.2 Synthèse binaurale

La synthèse binaurale consiste à positionner virtuellement un son dans l'espace en lui appliquant le filtrage qu'il aurait subi lors de sa propagation de cette position vers chaque oreille. Il s'agit donc dans un premier lieu de déterminer la fonction de transfert de cette propagation.

2.2.1 Mesure des HRTF

2.2.1.1 Individualité

Comme l'illustre la Figure 5, la morphologie de nos pavillons d'oreille varie beaucoup d'un individu à l'autre. Par conséquent, les filtres HRTF ont des réponses très variables (Figure 6).

Cela signifie que deux individus reçoivent à leurs tympans respectifs des sons différents pour une source placée de manière identique. Les HRTF sont donc individuelles et leur mesure doit idéalement se faire pour chaque individu.

De nombreux auteurs ont montré que si l'utilisation de HRTF non-individuelles perturbe peu la perception de l'azimut, en revanche elle dégrade grandement la localisation en élévation, phénomène amplifié pour les positions médianes [**72**].

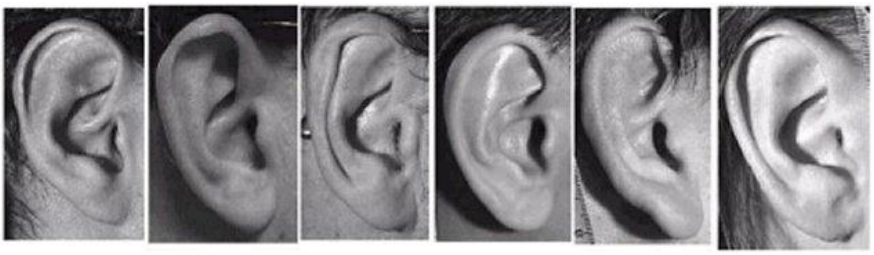

Figure 5. Illustration de la grande variabilité des morphologies des oreilles. D'après [14].

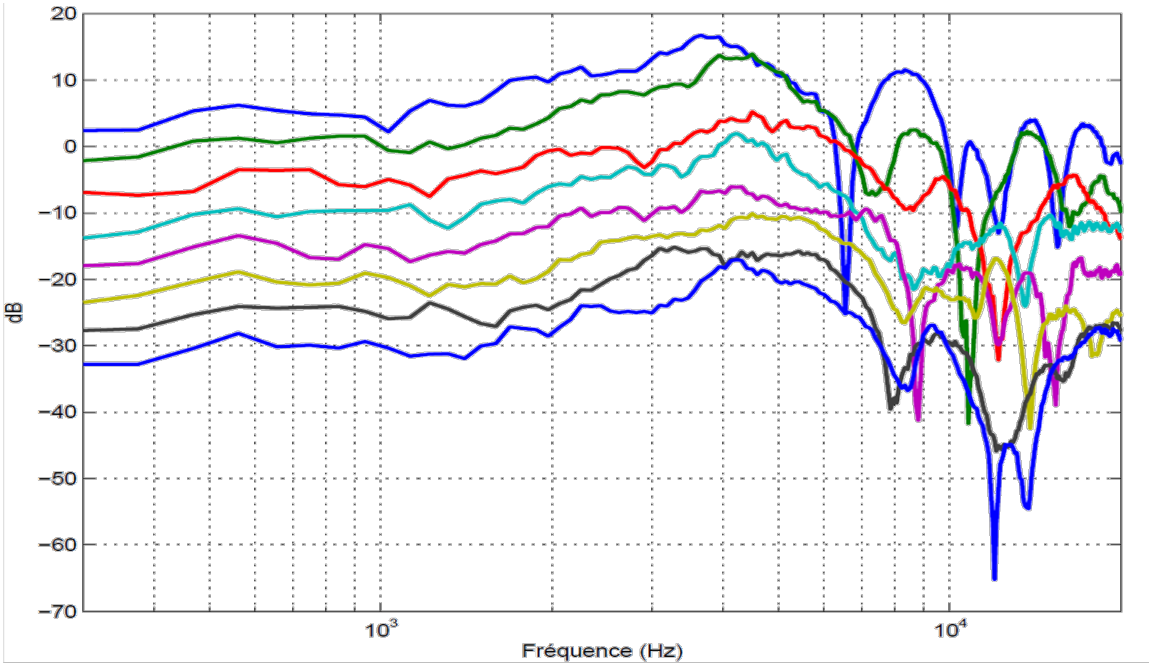

Figure 6. Illustration des variations des HRTF dans une même direction (θ = 0°, ϕ = -45°), pour 8 individus de la base Jean-Marie Pernaux, décalés de 5dB pour une meilleure lisibilité. D'après [51].

2.2.1.2 Principe de la mesure

La mesure des HRTF consiste à placer un microphone à l'entrée de chaque conduit auditif du sujet puis à déplacer une source sonore autour de ce dernier. On souhaite mesurer uniquement la contribution de la morphologie du sujet. Par conséquent, le sujet doit être placé dans un environnement anéchoïque.

En pratique, la mesure ne peut pas se faire dans toutes les directions et l'on doit se limiter à un échantillonnage suffisamment fin des directions de l'espace. Le nombre minimal de directions à

mesurer est de l'ordre du millier. On émet successivement dans chaque direction un signal et l'on enregistre la réponse pour chaque oreille. Selon la finesse du maillage et le type de signaux émis, la mesure peut ainsi durer plusieurs heures durant lesquelles le sujet ne doit pas bouger. À partir des enregistrements, on peut alors calculer les fonctions de transfert en déconvoluant la réponse dans l'oreille par la chaîne de mesure (signal source, haut-parleur, microphone).

Orange a mis en place un protocole réduisant la mesure à seulement 20 minutes. La méthode est présentée en Annexe 2.

2.2.1.3 Méthodes de mesure alternatives

Nous avons abordé ci-dessus la mesure des HRTF en chambre sourde. On peut néanmoins calculer les HRTF par modèle physique, par BEM à partir d'un maillage de la morphologie, ou encore par interpolation des HRTF aux positions voisines lorsque la position désirée n'a pas été mesurée. De nombreux travaux portent actuellement sur ces questions, mais la mesure en chambre sourde reste à ce jour la référence.

2.2.2 Synthèse binaurale

Une fois que l'on dispose des HRTF, on peut procéder au placement de sources dans l'espace par synthèse binaurale. Le principe est assez direct et consiste à filtrer le signal anéchoïque de la source x (cf. Figure 7) par la paire d'HRTF associées h_L et h_R, puis à diffuser ce signal binauralisé aux oreilles de l'auditeur. On crée alors l'illusion d'une source x' provenant de la direction pour laquelle les HRTF ont été mesurées.

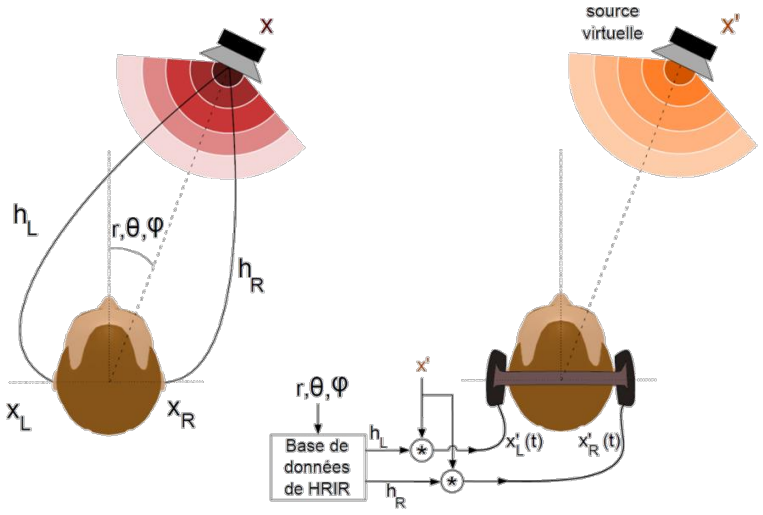

Figure 7. Principe de la synthèse binaurale. D'après [20].

La diffusion se fait idéalement au casque afin de contrôler aussi précisément que possible le signal diffusé à chaque oreille. Cependant, des méthodes de restitution de signaux binauralisés sur haut-parleurs ont aussi été mises au point comme le *Transaural™*. Pour s'assurer que chaque oreille reçoit uniquement le signal filtré par la HRTF correspondante, les signaux doivent alors être traités par des méthodes d'annulation des chemins croisés.

Afin que la scène sonore ne se décale pas lorsque la tête de l'auditeur tourne, des technologies de synthèse dynamique ont été développées. Un suivi de la position de la tête est réalisé et les filtres binauraux sont alors mis à jour en temps réel, en suivant les mouvements de la tête.

2.3 Artéfacts de perception

2.3.1 Inversions avant/arrière

Le problème des inversions avant/arrière est récurrent et ne concerne pas uniquement la synthèse binaurale au casque. Il est plus général et s'observe fréquemment lors des expériences de localisation de sources. On trouve dans la littérature des taux moyens de confusions assez variables. En champ libre, pour des sources réelles, Makous et Middlebrooks rapportent un taux de confusion moyen de 6 % [**41**] alors que Wenzel *et al.* distinguent deux populations d'auditeurs avec un premier groupe associé à un taux 6.5 % et un second avec un taux moyen de 32 % [**72**].

Pour des sources virtuelles spatialisées par synthèse binaurale, le taux de confusion augmente beaucoup chez les sujets qui présentent de faibles taux avec des sources réelles. Il double voire quadruple, selon l'utilisation de HRTF individuelles ou non. Au regard de plusieurs études, on peut s'accorder sur un taux moyen de 25% [**72**] [**74**].

De tels taux de confusion ne s'observent pas au quotidien. Ceci est principalement lié au fait que les conditions expérimentales offrent une perception de l'espace sonore limitée. Les résultats présentés ci-dessus ont été obtenus en environnement anéchoïque, sans mouvement de tête possible. De nombreux indices de spatialisation sont ainsi supprimés.

Il est difficile d'établir une hiérarchie parmi les indices pour déterminer lesquels doivent absolument être simulés. On voit par exemple que les deux comportements mis en évidence par Wenzel *et al.* traduisent des stratégies de localisation différentes selon les individus [**72**]. Le premier groupe base probablement la distinction avant/arrière sur des indices spectraux, dont la dégradation par l'utilisation de HRTF non-individuelles entraine une augmentation des inversions d'un facteur 4. Le second groupe se baserait sur les indices dynamiques, lesquels ne sont pas disponibles, que ce soit pour la source réelle ou pour la source virtuelle : le taux de confusion reste alors stable et élevé.

2.3.2 Effet ventriloque

La perception de l'espace qui nous entoure est multisensorielle. Elle est principalement régie par la vision tandis que l'audition et nos autres sens viennent en quelque sorte confirmer la position des objets sonores visibles ou prévenir de la présence d'objets hors de notre champ de vision. Ces différentes perceptions doivent être cohérentes sans quoi l'audition est en général écrasée par la vision. C'est l'effet ventriloque mis en évidence par Jack et Thurlow [**30**]. Ainsi, la perception des sources frontales serait dégradée par l'absence de stimulus visuel cohérent : « Si je ne vois rien devant moi, alors ce que j'entends provient d'une autre direction. »

2.3.3 Perception intracrânienne

La perception intracrânienne est un artéfact de perception qui se produit particulièrement lorsque l'on écoute des signaux issus de synthèse binaurale : les sources ne sont pas perçues à l'extérieur mais au contraire très proches de la tête, au niveau des oreilles voire à l'intérieur de la tête.

Les indices dynamiques influent sur la perception intracrânienne, surtout aux positions médianes. L'utilisation d'indices dynamiques n'étant pas envisagé dans le cadre de ce stage, le

lecteur pourra se référer à la section 3.3.1.8 de [**50**] pour une revue détaillée de l'influence des mouvements de la tête sur l'externalisation dans le plan médian.

Kim et Choi montrent que l'utilisation de HRTF génériques est une cause de mauvaise externalisation et remarquent qu'augmenter le taux de réverbération est plus bénéfique que de passer de HRTF génériques à des HRTF individuelles [**33**], constat partagé par Völk *et al.* [**70**]. Begault *et al.* confirment dans [**3**] que l'ajout de réverbération entraine une meilleure externalisation pour des sons de paroles spatialisés par synthèse binaurale. Entre les conditions anéchoïques et réverbérantes (T_{30} à 1.5s), les taux de bonne externalisation varient alors du simple au double (40% contre 79%). Sakamoto *et al.* ont étudié dans [**62**] l'influence du ratio entre champ réverbéré et champ direct sur l'externalisation. Ils montrent que l'augmentation du ratio acoustique améliore l'externalisation, mais que cet effet est moindre dans le plan médian.

Plus généralement, l'externalisation serait dégradée lorsque le sujet est soumis à des percepts incohérents ne permettant pas une localisation précise [**15**]. Ces incohérences peuvent provenir de l'absence de certains indices auditifs, de l'utilisation d'indices non-individuels ou même de l'incohérence entre vision et audition.

3 Conclusion

3.1 Un espace auditif fragile

Cette première partie s'est largement appuyée sur la littérature pour présenter l'audition en tant que sens de localisation. La perception de la localisation se fait grâce à l'analyse par le système auditif de nombreux indices contenus dans les deux signaux arrivant à nos oreilles.

Pour créer un d'espace auditif virtuel, on applique ces indices à un signal audio. Le signal binaural résultant est diffusé au casque, afin que l'auditeur ait l'illusion que le signal provient d'une certaine direction : c'est la synthèse binaurale.

A minima, on applique les différences interaurales (de phase et d'intensité) et les indices spectraux monauraux dus au filtrage du son par notre morphologie. Ces indices sont indispensables pour permettre une localisation en azimut et en élévation.

Se contenter de ces indices risque souvent d'aboutir à un espace sonore peu immersif, pour plusieurs raisons. La perception de la distance est très difficile en l'absence d'effet de salle. De plus, l'auditeur ne peut pas se déplacer dans la scène sonore : ses mouvements feront bouger toute la scène sonore avec lui.

Enfin, si les indices d'ITD et ILD sont peu dépendants de l'auditeur, les indices spectraux varient beaucoup d'un individu à l'autre. On peut néanmoins utiliser des HRTF génériques pour synthétiser des indices non-individuels mais cela se traduit souvent par une dégradation de l'espace virtuel auditif (erreurs de localisation, confusions avant/arrière, perception intracrânienne).

La synthèse binaurale d'un espace sonore est donc complexe, car sa perception est souvent fragilisée par l'absence de certains indices auditifs et par l'utilisation d'indices non-individuels.

3.2 Vers l'externalisation

La perception intracrânienne des sources est très problématique car elle supprime la sensation de volume et comprime l'espace sonore au niveau de notre tête.

L'externalisation est probablement l'attribut perceptif le plus fragile dans la synthèse binaurale. Nous avons vu que l'absence d'externalisation pouvait être résolue par l'ajout de réverbération ou par une synthèse dynamique suivant les mouvements de la tête. On aimerait cependant pouvoir créer la sensation d'externalisation, même dans le cas idéalisé où l'auditeur est fixe, dans une pièce non réverbérante, soumis à une source sonore immobile elle aussi. Dans de telles conditions, la solution se trouve donc au niveau des indices spectraux contenus dans les HRTF.

Partie 2.
Travail de recherche

4 Objectifs du stage

Des travaux antérieurs ont montré que certaines modifications des HRTF pouvaient rendre les indices spectraux plus saillants et ainsi faire diminuer les confusions avant/arrière. Dans le cadre de ce stage, nous souhaitons déterminer si en modifiant les indices spectraux contenus dans des HRTF génériques, nous pouvons supprimer le problème récurrent de perception intracrânienne.

Le déroulement du stage s'organise en deux temps. Une première phase, préliminaire, consiste déterminer quelles sont les conditions optimales pour juger l'externalisation :

1. Comparaison de la qualité d'externalisation selon l'utilisation de HRTF individuelles ou non-individuelles.
2. Choix des directions spatiales selon lesquelles étudier le défaut d'externalisation.
3. Définition d'un signal binaural non-externalisé de référence.

Puis la deuxième phase cherche à corriger le défaut d'externalisation:

1. Développement d'un outil de lissage et de sculptage de HRTF.
2. Exploration de l'influence des modifications spectrales sur la perception intracrânienne.

À terme, l'objectif de l'étude est d'identifier un ensemble de paramètres permettant de corriger le défaut d'externalisation. Ces paramètres devront être validés par un test d'écoute. Cette dernière étape n'a pas pu être réalisée dans le temps imparti du stage. Les résultats perceptifs que nous présentons dans cette Partie 2 sont des résultats préliminaires observés pour un individu.

Les sections 5, 6 et 7 présentent les prérequis à l'étude de l'externalisation (outils, stimulus et direction étudiée). La section 8 présente une méthode de lissage fréquentiel, indispensable pour notre étude sur l'amélioration de l'externalisation par design spectral section 9.

5 Outils expérimentaux

5.1 Choix du système de restitution

Nous utilisons une carte son USB et deux types de casque sont à notre disposition (cf. Figure 8).

Figure 8. Carte son USB *Scarlett 6i6* de Focusrite – Casque ouvert *DT 990 PRO 250 ohm* de Beyerdynamic – Casque fermé *HFI-580* d'Ultrasone.

Concernant le choix du casque, Møller *et al.* ont comparé différents casques disponibles sur le marché en 1995 [**49**]. Cette étude conduit les auteurs à recommander les casques proposant un couplage équivalent à des conditions de champ libre (*Free-air Equivalent Coupling*), c'est à dire en première approximation les casques de type ouvert. Ces derniers permettraient en effet à l'auditeur d'oublier qu'il porte un casque bien qu'aucun résultat perceptif ne vienne confirmer ce résultat.

La sensation de pression des coussins du casque sur les oreilles pourrait aussi dégrader notre perception. Inanaga *et al.* ont réalisé un casque n'exerçant aucune pression sur l'oreille. Ce dernier ne permet cependant pas de résoudre le défaut d'externalisation [**28**].

Pour ces raisons, nous avons privilégié le casque ouvert DT 990 PRO, car ce dernier ne procure pas la sensation d'isolement des oreilles de l'extérieur. Il est mécaniquement très bien conçu car ses coussinets exercent une pression assez faible sur les oreilles, ce qui le rend bien plus confortable que le casque fermé d'Ultrasone. Néanmoins, dans certains cas, pour une écoute plus affinée des détails, nous avons ponctuellement utilisé le casque Ultrasone.

5.2 Interface graphique de design spectral d'HRTF

Pour réaliser le design spectral des HRTF, nous avons développé une interface graphique présentée Figure 9. Elle permet de sculpter aussi bien à la souris qu'en ligne de commande les filtres à une position donnée, puis d'écouter un son spatialisé par ces filtres modifiés ou par les filtres originaux. La plupart des traitements audio effectués en arrière-plan seront présentés dans la section 8.2 du présent rapport.

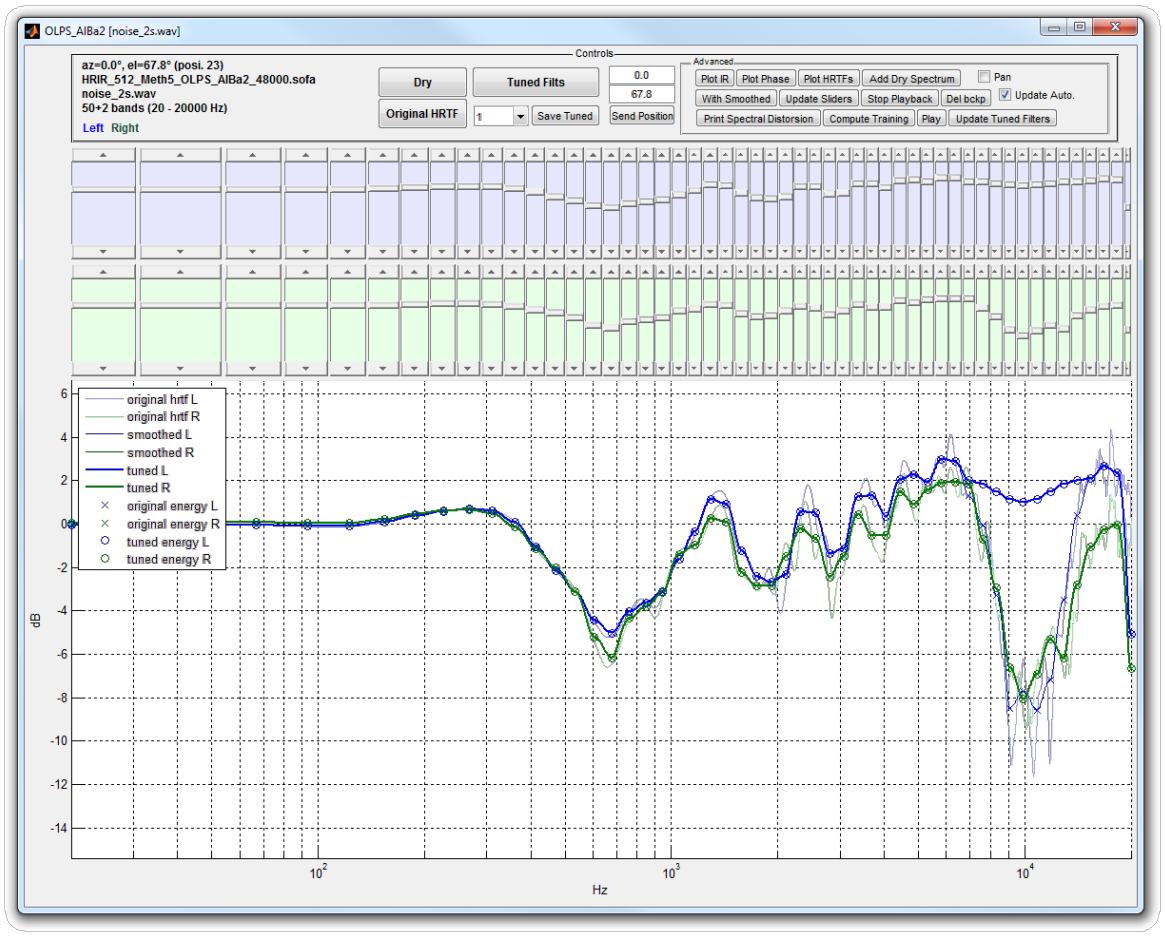

Figure 9. Capture d'écran de l'interface de design spectral des HRTF.

6 Choix d'une position

6.1 Idée

Nous ne pouvons pas d'emblée attaquer le défaut d'externalisation dans toutes les directions. Le profil des HRTF varie beaucoup d'une position angulaire à l'autre comme l'illustre la Figure 10. Ainsi, avant de développer des solutions de tuning spectral pour toutes les régions de l'espace, nous devons choisir une direction dans laquelle expérimenter les modifications spectrales des HRTF. Ce n'est que dans un second temps que l'on pourra chercher à généraliser à d'autres directions les résultats obtenus.

Nous cherchons à étudier le rôle des indices spectraux contenus dans le profil des HRTF. Ainsi, afin d'écarter l'influence des indices interauraux sur l'externalisation, nous choisissons une position dans le plan médian ($\theta = 0$) permettant d'annuler ILD et ITD. Les positions médianes présentent des HRTF gauche et droite très proches puisque nos oreilles sont a priori symétriques par rapport à ce plan. Ainsi, le tuning spectral peut être réalisé de manière symétrique.

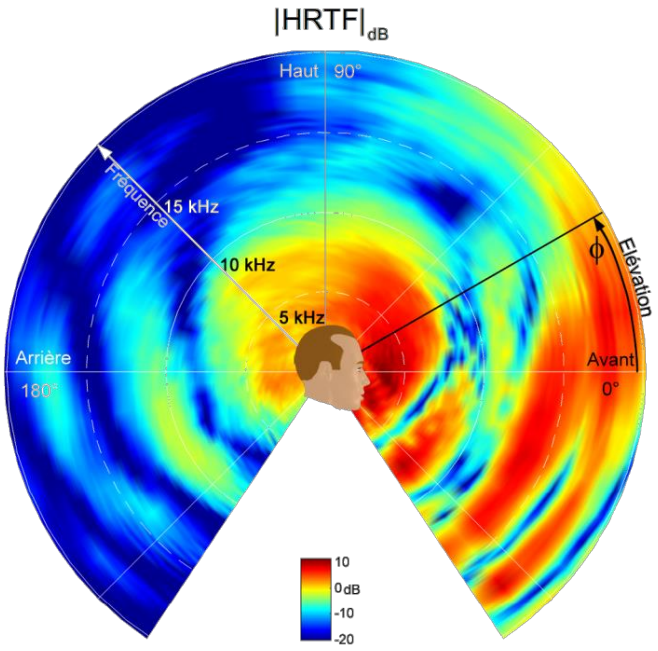

Figure 10. Représentation polaire du module des HRTF dans le plan médian (oreille droite, sujet n°5 de la base privée de HRTF d'Orange). D'après [20].

6.2 Problème du plan médian : revue bibliographique

La perception des sources dans le plan médian est souvent problématique lors de leur synthèse binaurale. Une revue relativement complète du problème est faite par A. H. Moore section 3.3.1 de [**50**], dont on reprendra ici l'essentiel.

6.2.1 Augmentation du risque de perception intracrânienne

À $\theta = 0°$, la localisation en champ libre se fait uniquement par les indices spectraux. Si les indices spectraux ont été mal mesurés ou bien ne correspondent pas aux HRTF individuelles de

l'auditeur, alors ce dernier n'a aucun autre indice disponible pour localiser. Dans ce cas, faute de pouvoir placer la source, elle sera perçue intracrânienne.

6.2.2 Pas de certitude sur la symétrisation

On suppose fréquemment que dans le plan médian, ITD et ILD sont nulles. Plusieurs études rapportent cependant de légères différences et montrent qu'elles sont en réalité perceptibles [**13**] [**73**]. Wightman et Kistler [**73**] ont mesuré l'asymétrie interaurale à [θ=0°, ϕ=0°] et [θ=180°, ϕ=0°] par bandes de tiers d'octave sur 10 sujets. Tandis qu'elle reste inférieure à 3 dB en dessous de 5 kHz, les auteurs observent un écart moyen supérieur à 5 dB atteignant pour certains sujets 20 dB au-delà de 5 kHz (cf. Figure 11).

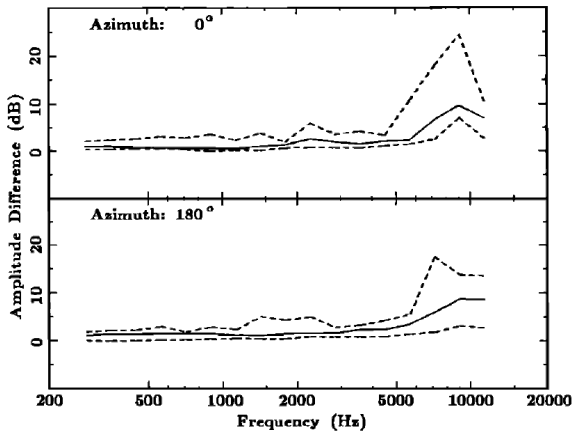

FIG. 9. Mean ⅓-oct band interaural asymmetry in the HRTF measurements. The two panels show the absolute value of the dB difference between the left and right HRTFs for a source at 0-deg azimuth and 0-deg elevation (upper panel), and for a source at 180-deg azimuth and 0-deg elevation (lower panel). The solid lines indicate the mean interaural asymmetry and the dashed lines indicate the upper and lower limits of the range of asymmetries.

Figure 11. D'après [73].

On oppose parfois à ces observations la question de la précision des mesures. Les écarts s'observent à des longueurs d'onde de l'ordre de quelques centimètres (34 mm à 10 kHz) et correspondent donc aux dimensions des pavillons. Deux questions pertinentes : La part d'erreur du protocole de mesure est-elle négligeable devant les écarts mesurés ? Ces écarts, qu'ils soient liés à la morphologie du sujet ou bien au protocole de mesure, sont-ils perceptivement influents ?

Iwaya et Suzuki déterminent dans [**29**] l'effet de légères modifications morphologiques sur les HRTF. On voit Figure 12 qu'un décalage du pavillon de seulement 5 mm décale le creux N1 de près de 1000 Hz tandis qu'une rotation du pavillon de 5° se traduit par des variations d'amplitude considérables.

Pour répondre à la seconde question, Brookes et Treble [**10**] ont montré que l'utilisation de HRTF génériques mesurées sur des mannequins aux pavillons légèrement asymétriques conduisait à une externalisation sensiblement meilleure que pour des pavillons symétriques (63 % des sujets).

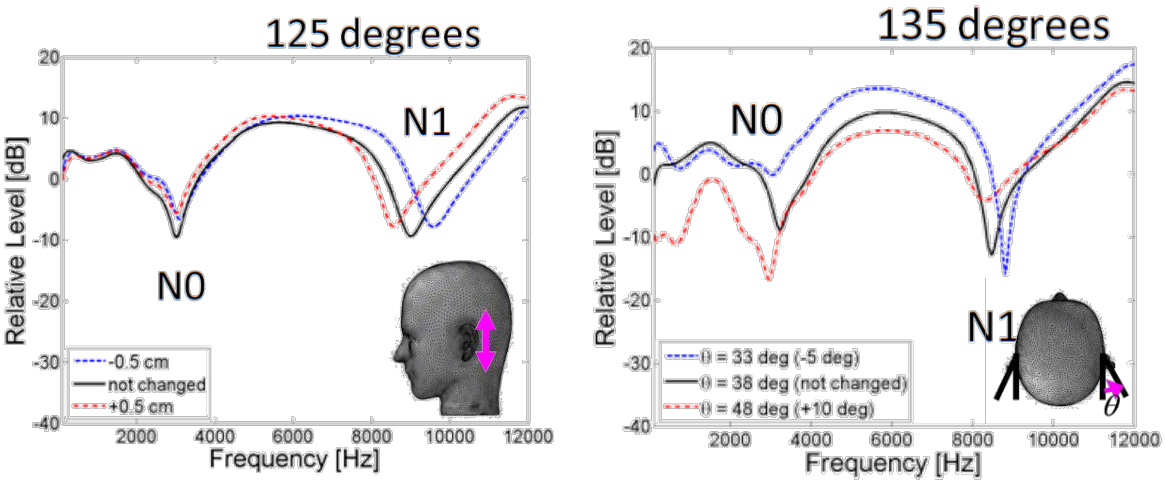

Figure 12. Décalage des HRTF en selon différents décalages morphologiques. D'après [29]

6.3 Démarche et résultats

Nous choisissons une source dans le plan médian pour les raisons expliquées précédemment et afin d'éviter l'effet ventriloque, nous choisissons une source hors du champ de vision[2].

Nous considérons la position [0°,70°] car, pour cette dernière, mes HRTF (sujet « ArMi » de la base BiLi) présentent une dynamique assez faible : entre -5 et +2 dB en dessous de 10 kHz et des variations entre -30 et 0 dB au-dessus de 10 kHz (cf. Figure 13).

Cette faible variation limite les variations de timbre lors du filtrage et permettra de se concentrer sur l'externalisation. En effet lors des premières écoutes, nous constatons que je me concentre sur la coloration spectrale plus que sur la localisation quand les différences de timbre sont importantes entre le stimulus et le stimulus spatialisé.

L'inconvénient est qu'avec ces faibles variations, il est parfois difficile de faire la part des choses entre les indices spectraux liés à notre morphologie et les artéfacts de mesure uniquement liés au protocole expérimental. En effet, on observe Figure 13 une oscillation régulière du spectre, rappelant un filtrage en peigne. Ce filtrage est en général causé par l'ajout d'une version retardé du signal à lui-même. Martens *et al.* expliquent dans [**43**] que ces oscillations proviennent des réflexions sur la source, lesquelles viennent s'ajouter au signal original et atteignent l'oreille avec un retard et une atténuation liés à la distance entre le sujet et la source. Étant liées au protocole de mesure davantage qu'à notre morphologie, nous concluons que l'on peut supprimer ces oscillations par un lissage par bandes.

[2] Le champ visuel se limite verticalement aux élévations $-70° < \phi < 60°$ [**37**].

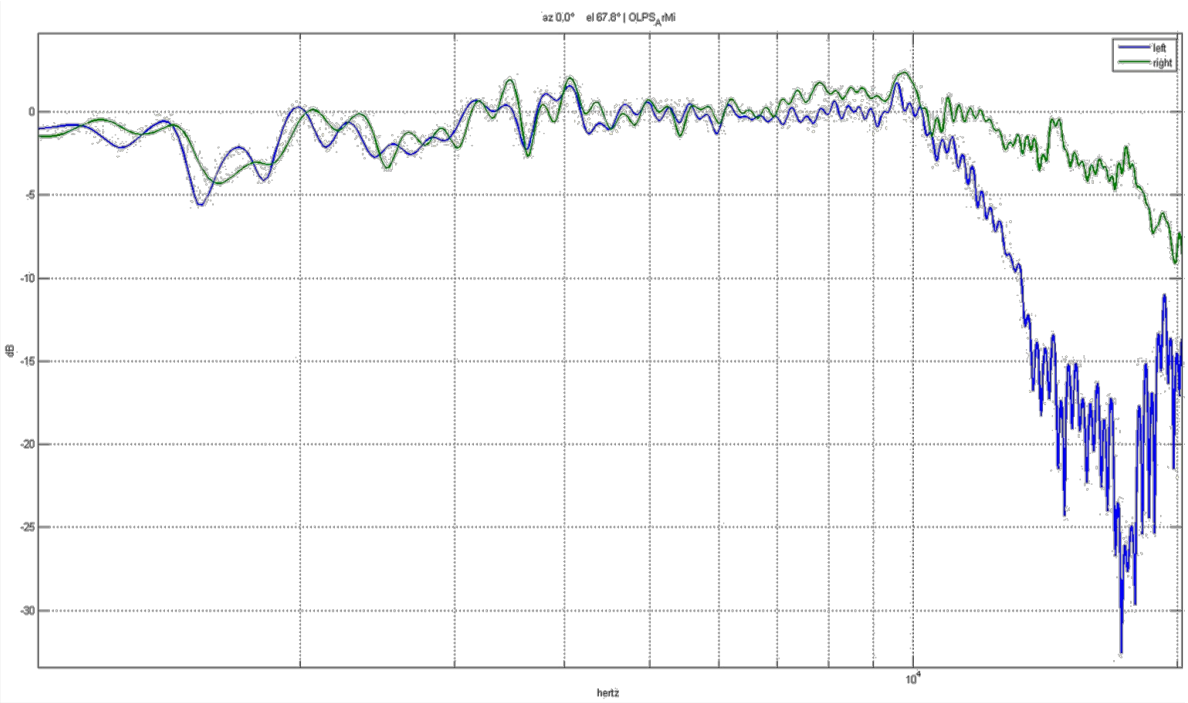

Figure 13. Amplitude des HRTF à la position choisie (θ = 0°,ϕ = 70°). Sujet ArMi.

7 Choix du stimulus

Le choix du signal anéchoïque utilisé pour synthétiser des signaux binauraux n'a pas été évident et a évolué au cours du stage.

7.1 Premiers essais

Lors des premières écoutes de sources binauralisées, que ce soit avec des HRTF individuelles ou non, l'externalisation des sources nous apparait très compliquée. Des signaux impulsifs sont utilisés en premier lieu, tels que des claquements de doigts, mais les transitoires semblent trop courts pour que l'on puisse déterminer avec constance leur perception intracrânienne ou externalisée. En nous inspirant de Begault [**3**], nous utilisons alors des enregistrements de voix. Ces derniers sont toujours perçus dans la tête pour les positions médianes et sur les oreilles pour les positions latérales, ce qui peut s'expliquer par un spectre pauvre au-delà de 4 ou 5 kHz. Enfin, les mêmes problèmes se retrouvent pour des contenus plus riches comme de la musique, d'autant plus qu'il n'est pas évident de trouver des enregistrements anéchoïques de morceaux de musique.

7.2 Bruit blanc

Le bruit blanc apparaît comme une bonne solution, car son spectre contient toutes les fréquences audibles et peut durer indéfiniment [**53**]. Le point faible d'un tel son est qu'il n'est

absolument pas écologique, c'est-à-dire qu'il ne correspond pas à ce qu'un humain peut entendre dans sa vie de tous les jours. Ainsi, les premières écoutes de bruits blancs spatialisés conduisent presque toujours à une localisation dans la tête, quand bien même on utilise des HRTF individualisées.

7.3 Apprentissage du signal

L'externalisation d'un bruit blanc spatialisé est probablement compliquée car l'écoute de signaux de bruits blancs est peu familière. On peut supposer que ce type de son n'a pas été mémorisé et qu'il parait alors peu vraisemblable à notre cerveau qu'une source réelle puisse diffuser ce type de contenu. Une phase d'apprentissage de ce signal a donc été mise en place.

Le dispositif de mesure des HRTF en chambre sourde permet de contrôler simultanément une trentaine de haut-parleurs répartis régulièrement en élévation (cf. Annexe 2). Nous utilisons MAX/MSP pour développer un patch[3] permettant de diffuser successivement du bruit blanc[4] dans chaque direction. L'ordre dans lequel les directions s'enchaînent est connu du sujet, choix inspiré par le dispositif d'apprentissage accéléré de Blum *et al.* [**7**]. L'apprentissage consiste simplement à se placer au centre du cercle de haut-parleurs et à écouter le bruit blanc se déplacer, en gardant la tête fixe.

Cet exercice permet de se familiariser avec la perception d'un bruit blanc spatialisé. Il est aussi l'occasion de chercher ce qui procure la sensation d'externalisation. On constate que toutes les fréquences du spectre ne présentent pas la même qualité d'externalisation (cf. 9.2.2). On remarque aussi que l'on peut assez facilement se persuader que la source sonore se trouve au niveau de nos oreilles voire à l'intérieur de notre tête, ce qui démontre la fragilité du caractère externalisé d'une source, quand bien même on écoute une source réelle.

7.4 Durée des stimuli

Nous fixons arbitrairement une durée de 2 secondes. Ce choix est en partie motivé par [**23**] où Hassager *et al.* utilisent des stimuli de bruit de 4 s. Les auteurs expliquent que cette durée permet d'avoir une réponse stable.

7.5 Conclusion

Le bruit blanc est probablement le meilleur stimulus pour évaluer la synthèse binaurale au casque. Ceci est vrai à condition que le sujet connaisse bien ce signal lorsqu'il est issu d'une source réelle. On utilisera un signal de bruit blanc de 2 secondes.

[3] MAX/MSP est un langage et un environnement de programmation visuelle. Initialement dédié à l'audio, il permet de programmer simultanément des traitements audio et une interface utilisateur. On appelle « patch » un tel programme.

[4] Pour obtenir un bruit au spectre parfaitement plat, il aurait fallu prendre en compte les courbes de réponse des haut-parleurs, lesquelles ne sont pas plates (creux de -10 dB autour d'1 kHz).

8 Lissage fréquentiel des HRTF

Afin de pouvoir modifier le spectre d'une HRTF, on se propose de filtrer cette dernière par un banc de filtres passe-bande pour ensuite pouvoir la recréer en conservant uniquement le niveau moyen bande par bande. Il importe cependant que la différence entre le filtre lissé et la HRTF originale soit imperceptible.

8.1 Bibliographie

8.1.1 Physiologie du système auditif : la tonotopie

Lorsqu'un son atteint notre oreille, il se propage jusqu'à notre tympan. La vibration du tympan est mécaniquement transmise à une membrane via plusieurs éléments que nous ne détaillerons pas ici. Cette membrane dite basilaire peut être vue comme un long tissu cellulaire mis en vibration par le son. Tout le long de la membrane basilaire sont disposées des cellules vibratiles, dont la mise en mouvement par une onde déclenche la libération de neurotransmetteurs vers le système nerveux.

Le point clef est que les propriétés mécaniques de la membrane ne sont pas constantes. Ainsi, un son de fréquence élevée mettra en vibration uniquement le début de la membrane tandis qu'une fréquence grave excitera l'autre extrémité de la membrane. Des neurotransmetteurs différents seront ainsi mis en jeu selon la fréquence de l'onde sonore. Du type d'influx nerveux reçu, le cerveau déduit alors quelles fréquences étaient contenues dans le son, selon des schémas qui dépassent largement nos connaissances.

8.1.2 Notion de bande critique

Même excitée par une fréquence pure, le mouvement de la membrane basilaire ne peut se limiter à la vibration d'un point. Une zone restreinte autour du point correspondant à la fréquence pure est mise en vibration, comme l'illustre la Figure 14. À cette région de la membrane correspond un intervalle de fréquence. Le système auditif analyse donc les sons selon des bandes de fréquence correspondant à l'étalement des vibrations le long de la membrane basilaire. Harvey Fletcher propose en 1940 la notion de bandes critiques pour décrire cette analyse fréquentielle des sons par bande de fréquences [**17**].

À l'écoute simultanée de deux fréquences éloignées, par exemple 440 Hz et 4000 Hz, ces dernières sont perçues séparément. En revanche, lorsque ces deux fréquences se trouvent dans la même bande critique, par exemple 1000 Hz et 1010 Hz, le système auditif est incapable de résoudre ces deux fréquences, c'est-à-dire de les percevoir séparément. Les deux composantes fusionnent et l'on entend un son modulé en amplitude[5]. Par exemple, une fréquence de 1 kHz fusionnera avec une fréquence *f* si et seulement si cette dernière se trouve dans la bande critique centrée sur 1 kHz, soit donc $f \epsilon$ [936 Hz, 1068 Hz]. Il faut garder à l'esprit que ces effets se produisent uniquement si les fréquences apparaissent simultanément. Pour des fréquences diffusées successivement, l'auditeur est capable de distinguer des différences bien plus faibles, de l'ordre de ±2 Hz à 1 kHz.

[5] Les deux fréquences interfèrent. On entend alors un son pur dont la fréquence est la moyenne des deux fréquences initiales, modulé en amplitude par un battement dont la fréquence est la différence des deux fréquences initiales. Mathématiquement, $\sin(\omega_A * t) + \sin(\omega_B * t) = 2 * \cos\left(\frac{\omega_A - \omega_B}{2} * t\right) * \sin\left(\frac{\omega_A + \omega_B}{2} * t\right)$.

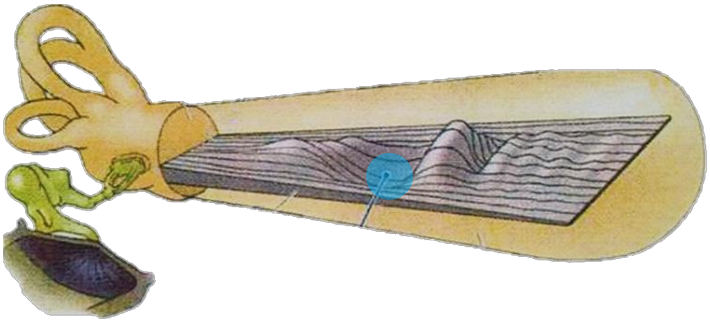

Figure 14. Membrane basilaire (en gris). La vibration au niveau du point bleuté occasionne des vibrations de plus faible amplitude dans les zones voisines. D'après [64].

8.1.3 Largeur de bande rectangulaire équivalente (ERB)

On parle couramment de filtre auditif pour désigner le filtre passe-bande correspondant à une bande critique. Patterson propose en 1976 un protocole pour estimer les propriétés des filtres auditifs [**56**]. Il propose ensuite d'approximer les filtres auditifs par des passe-bandes rectangulaires et mesurer leur largeur de bande sur des sujets. On appelle cette largeur *Equivalent Rectangular Bandwidth* pour « largeur de bande rectangulaire équivalente ».

Glasberg et Moore proposent dans [**19**] de modéliser la largeur ERB à 3 dB par la relation ci-dessous, avec f_c fréquence centrale du filtre en Hertz:

$$ERB(f_C) = 24.7 * (0.00437 * f_C + 1)$$

8.1.4 Échelle du taux de bande rectangulaire équivalente (ERBS)

Glasberg et Moore définissent une échelle fréquentielle correspondant à notre perception [**19**]. Elle indique le nombre de bandes critiques en dessous d'une fréquence donnée. On l'appelle ERBS pour « *Equivalent Rectangular Bandwidth rate Scale* », c'est-à-dire une échelle du taux de bandes critiques. Par exemple, si ERBS(f_2)=ERBS(f_1)+2, alors exactement deux bandes critiques séparent f_1 de f_2. L'expression mathématique de ce taux ERBS se déduit de l'expression de l'ERB[6]:

$$ERBS(f) = 21.4 * \log_{10}(1 + 0.00437 * f), \text{ avec } f \text{ en Hertz}$$

Cette notion de nombre de bandes ERB en dessous d'une fréquence donnée est peu intuitive, et il faut davantage voir l'échelle ERBS comme un moyen de travailler sur des bandes fréquentielles perceptivement équivalentes.

8.1.5 Banc de filtres auditifs

Au risque de nous répéter, bien que l'onde sonore arrivant à notre oreille contienne une infinité de fréquences, notre système auditif est incapable de toutes les discriminer et procède davantage en regroupant les fréquences voisines par paquets. Ces paquets résultent du filtrage du signal par les filtres auditifs. Excité simultanément par plusieurs fréquences, le système auditif les analyse ensemble si elles sont dans la même bande critique. Ceci se généralise à l'ensemble du spectre audible sur lequel on trouve alors une succession de bandes critiques, régulièrement espacées sur l'échelle ERBS avec un pas de 1 ERBS [**19**].

[6] La fonction ERBS se calcule par intégration de la fonction réciproque d'ERB et on fixe ERBS(0)=0.

Pour se rapprocher autant que possible du fonctionnement de la membrane basilaire, les bandes critiques ne doivent pas être fixes, mais doivent suivre le contenu spectral du signal reçu [**56**]. Schématiquement, à l'écoute de deux sons purs successifs à 1000 Hz puis 1012 Hz, la bande critique doit se décaler légèrement de 12 Hz pour être toujours centrée sur le maximum de vibration. Il est cependant fréquent de simplifier le modèle en considérant les filtres fixes.

8.1.6 Lissage du spectre d'amplitude des HRTF

Il est possible de réduire ou augmenter la résolution du banc de filtres auditifs en utilisant un pas inférieur ou supérieur à 1 ERBS. Les fréquences de coupure ne correspondent alors plus à celles de nos bandes critiques, mais chacune des bandes a néanmoins une largeur perceptivement équivalente.

Hassager *et al.* estiment dans [**23**] comment les détails du spectre d'amplitude des HRTF jouent sur l'externalisation de la source. Sept sujets ont participé au test, leurs BRIR ont été mesurées en chambre réverbérante avec le même casque sur les oreilles que celui ensuite utilisé pour le test s'écoute. Les HRIR sont obtenues en supprimant des BRIR le filtrage du casque et les réflexions dans la pièce. Les HRTF lissées sont calculées en imposant dans chaque sous-bande l'énergie moyenne calculée dans la même sous-bande que dans les HRTF originales. Plusieurs bancs de filtres sont utilisés en faisant varier la largeur de bande sur l'échelle ERBS. Les auteurs constatent que pour des bandes inférieures à 1 ERB l'externalisation n'est pas affectée tandis que pour un lissage avec des bandes plus larges, la perception se dégrade comme l'illustre la Figure 15.

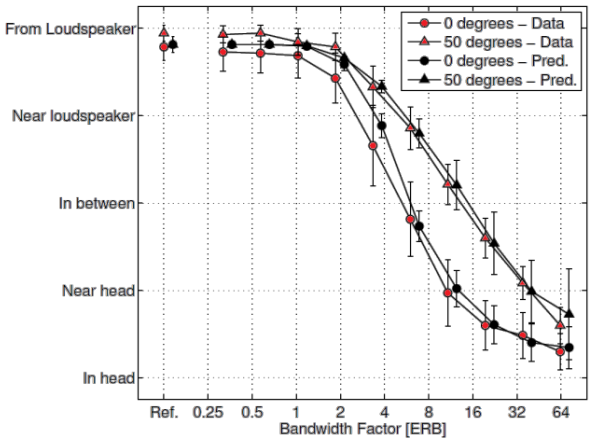

Figure 2. The mean of the seven listeners perceived sound source location (open symbols) as a function of the bandwidth factor and the corresponding model predictions (filled symbols). The model predictions have been shifted slightly to the right for a better visual interpretation.

Figure 15. Dégradation de l'externalisation selon le lissage fréquentiel (symboles rouges). D'après [23].

Une autre façon d'évaluer l'influence du lissage en conservant une approche basée sur les bandes critiques est proposée par Breebaart et Kohlrausch dans [**9**]. Les auteurs choisissent comme paramètre de lissage l'ordre utilisé pour modéliser les filtres cochléaires. La bande passante à 3 dB d'un filtre correspond toujours à 1 ERB mais sa raideur diminue avec l'ordre, ce qui se traduit globalement par une amplification du lissage. Les auteurs soumettent aux sujets le même signal, filtré par deux lissages différents de HRTF et ces derniers doivent dire s'ils ont entendu deux fois la même chose ou non. Les auteurs concluent que le lissage est perceptible

seulement quand l'ordre des filtres est divisé par trois ou plus d'un lissage à l'autre. La méthode de lissage est intéressante, car la largeur des bandes critiques à 3dB est conservée quel que soit l'ordre des filtres. En revanche, seulement trois sujets ont passé le test. Notons en outre que les auteurs utilisent des HRTF génériques.

8.1.7 Lissage du spectre de phase des HRTF : ITD et phase minimale

Tandis que l'amplitude des HRTF doit être approximée sur des bandes fréquentielles au moins aussi étroites que les bandes critiques, la phase peut être modélisée plus globalement.

Kulkarni et Colburn furent les premiers à s'interroger sur le rôle des détails spectraux des HRTF [**34**]. Ils ont montré que la phase des HRTF peut être réduite à la somme d'un retard pur (c'est-à-dire une phase linéaire) et de la phase minimale du spectre d'amplitude. Cette modification de la phase n'est pas perceptible, en accord avec [**23**] et [**9**].

8.2 Démarche

De manière classique, nous modélisons notre système auditif comme un banc de filtres analysant le spectre audible. Les fréquences centrales (ou « de coupure ») des filtres sont réparties régulièrement sur l'échelle ERBS.

Nous faisons cependant le choix de mailler plus finement les fréquences en imposant aux filtres des largeurs d'une moitié de bande ERB. Ce choix est motivé par les résultats de [**23**] : on observe un gain d'externalisation pour un lissage à 0.6 bande ERB par rapport à 1 bande ERB dans la direction frontale (cf. Figure 15).

Ce qui est présenté ci-dessous est réalisé dans le domaine fréquentiel des ERBS. On retourne dans le domaine des Hertz seulement à la fin, via la réciproque de la fonction ERBS :

$$ERBS^{-1}(n_{ERB}) = (10^{n_{ERB}/21.4} - 1)/0.00437$$

8.2.1 Fréquences de coupures

Le nombre de bandes équivalentes a été déterminé entre f_{min}=20Hz et f_{max}=20kHz d'après [**19**] : ERBS(f_{max})–ERBS(f_{min}) ≈ 41. On choisit donc N=82 bandes pour avoir des bandes de largeurs sensiblement égales à 0.5 bande ERB. On répartit ainsi régulièrement N points sur l'intervalle [ERBS(f_{min})—ERBS(f_{max})] (cf. Figure 16.).

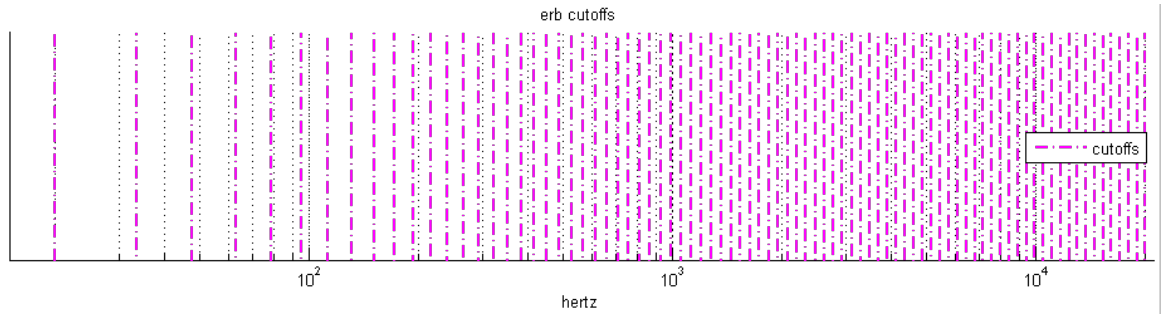

Figure 16. Fréquences centrales des 82 filtres ERB.

8.2.2 Filtres

Il s'agit à présent de générer des filtres passe-bande centrés sur ces N fréquences de coupures. Nous choisissons d'utiliser des filtres dont la réponse fréquentielle a la forme d'un cosinus et dont la phase est nulle. Comme l'explique McDermott dans [**44**], de tels filtres s'adaptent bien à la création d'un banc de filtres ERB. En effet, pour créer le i-ème filtre (avec 1<i< N), on paramètre le cosinus afin qu'il vaille 1 à la fréquence de coupure f_i et 0 aux deux fréquences de coupures voisines f_{i-1} et f_{i+1}. Par définition du cosinus, il vaudra $\sqrt{2}/2$ aux milieux[7] de f_{i-1} et f_i et de f_i et f_{i+1}. Cela permet d'avoir une bande passante à -3dB qui s'arrête exactement au milieu des deux fréquences de coupures. Cela permet aussi que le banc de filtres soit transparent lors de la reconstruction du signal. En effet, les filtres adjacents sont des cosinus déphasés de $\pi/2$: énergétiquement, leur somme vaut donc 1. On trouvera l'illustration de trois filtres adjacents Figure 17.

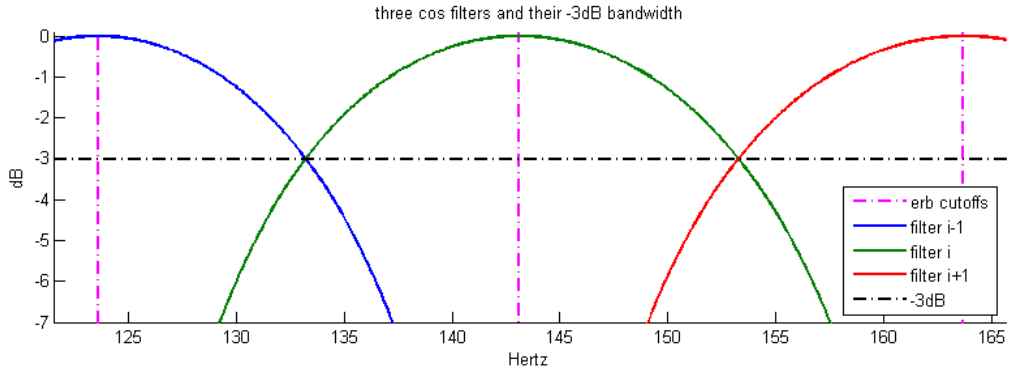

Figure 17. Trois filtres adjacents dans le domaine fréquentiel des Hertz. On a tracé en magenta les trois fréquences centrales et en noir une atténuation de -3dB ou de $\sqrt{2}/2$ en amplitude.

8.2.3 Passe-haut et passe-bas

Afin de décrire l'intégralité du spectre, c'est-à-dire de 0Hz jusqu'à la fréquence de Nyquist, on ajoute un passe-bas et un passe-haut aux extrémités du banc de filtres. Leur intérêt est limité dans le cas où $[f_{min}—f_{max}]$ décrit le spectre audible. En revanche, on verra ultérieurement que ces deux filtres permettent de lisser par bande un domaine fréquentiel plus restreint sans introduire de modifications majeures du timbre. On trouvera Figure 18 le banc de filtres complet.

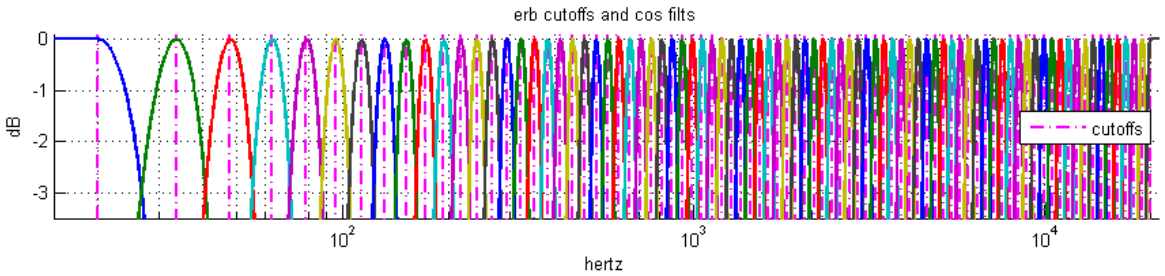

Figure 18. Banc de 82 filtres de largeur perceptivement équivalente, ainsi que 2 filtres passe-bas et passe-haut aux extrémités du spectre.

[7] Les « milieux » sont à prendre au sens de l'échelle ERBS, c'est-à-dire en réalité à $[ERBS(f_{i-1})+ERBS(f_i)]/2$ et $[ERBS(f_i)+ERBS(f_{i+1})]/2$.

8.2.4 Reconstruction des HRTF lissées

Le banc de filtres présenté ci-dessus permet d'analyser les HRTF. Cette section présente une méthode pour reconstruire une version lissée de ces dernières.

Quelques conventions :

- On désigne le i-ème filtre passe bande BP_i, et, par commodité, le passe bas BP_0 et le passe-haut BP_{N+1}.[8]
- On note le filtrage avec de A par B comme $[A * B](f) = A(f) * B(f)$
- On définit l'énergie spectrale d'un spectre S comme $\mathbb{E}(S) = \sum_{f \geq 0} |S(f)|^2$

8.2.4.1 Spectre d'amplitude

On analyse la HRTF par le banc de filtres. On calcule l'énergie dans chacune des N+2 bandes à laquelle on retranche l'énergie du filtre associé pour obtenir l'énergie spectrale $E_i$de la HRTF dans cette i-ème bande ERB :

$$E_i(\text{HRTF}) = \frac{\mathbb{E}(\text{HRTF} * BP_i)}{\mathbb{E}(BP_i)}$$

Nous choisissons de diviser par $\mathbb{E}(BP_i)$, car bien que chaque filtre soit équivalent sur l'échelle ERBS, il n'a pas la même largeur en Hertz. Cela revient à calculer l'énergie moyenne dans chaque sous-bande à un facteur constant près.

Après cette phase d'analyse, on procède à la synthèse du spectre d'amplitude. On applique aux filtres ERB les énergies calculées E_i puis on somme les sous-bandes pour générer le spectre lissé SMOO :

$$|\text{SMOO}|(f) = \sqrt{\sum_{0 \leq i \leq N+1} E_i * \left(BP_i(f)\right)^2}$$

8.2.4.2 Spectre de phase

Pour synthétiser la phase des HRTF lissées, on procède en deux étapes. On calcule tout d'abord la phase minimale $\psi_{\min}$ à partir du spectre d'amplitude, selon la formule tirée de [**54**] :

$$\psi_{\min}\{\text{SMOO}\}(f) = Imag\left(\text{Hilbert}(-\log(|\text{SMOO}|(f))\right)$$

Dans un second temps, on estime le retard interaural (ITD) sur la paire de HRTF originales, entre le filtre gauche et droite. Un retard pur T<0 correspond spectralement à l'ajout d'une phase négative linéaire $\psi_T(f) = T * 2\pi * f$. Nous cherchons donc à estimer les retards T_{Left} et T_{Right} de la paire de HRTF, leur différence constituant l'ITD. D'un point de vue psychoacoustique, l'estimation de l'ITD devrait se limiter aux basses fréquences [20—2000Hz]. D'un point de vue signal, la linéarité de la phase s'observe néanmoins sur un intervalle plus large.

On calcule la dérivée de la phase $Angle(\text{HRTF})$ pour estimer dans quelle région la phase est linéaire. On observe à la position [0°,70°] étudiée une pente non linéaire sur [0—60Hz] puis des instabilités au-delà de 10Khz. On a observé un comportement sensiblement identique pour

[8] BP pour « Band Pass », bien que BP_0 et BP_{N+1}soient en réalité des passe-bas et passe-haut.

d'autres positions. Nous choisissons donc d'estimer les retards sur la bande [100—10000Hz] où nous jugeons la phase suffisamment linéaire.

Il nous a été par la suite conseillé de retrancher la phase minimale calculée sur les HRTF originales avant d'estimer la phase $\left(Angle(\mathrm{HRTF}) - \psi_{\min}\{\mathrm{HRTF}\}(f)\right)$. On constate Figure 20 que la phase est alors beaucoup plus régulière. Nous choisissons donc d'estimer le retard T sur cette phase en conservant l'intervalle [100—10000Hz] :

$$\mathrm{T}(\mathrm{HRTF}) = \mathrm{Moyenne}_{100 \leq f \leq 10000} \left\{ \frac{Angle(\mathrm{HRTF})(f) - \psi_{\min}\{\mathrm{HRTF}\}(f)}{2\pi * f} \right\}$$

Finalement, la phase qu'on appliquera à chaque HRTF lissée SMOO est donc la somme du retard estimée sur la HRTF originale et de la phase minimale de SMOO:

$$Angle(\mathrm{SMOO})(f) \triangleq \psi_{\min}\{\mathrm{SMOO}\}(f) + 2\pi * \mathrm{T}(\mathrm{HRTF}) * f$$

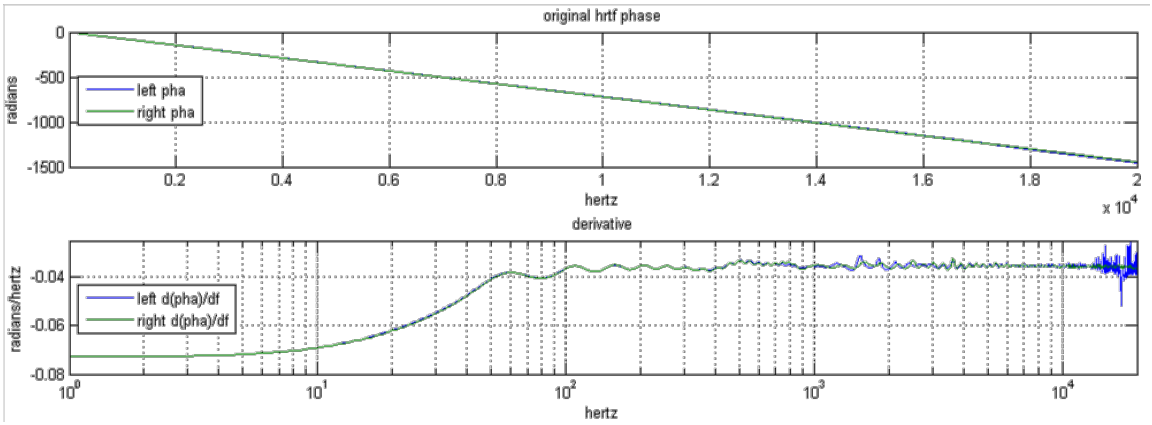

Figure 19. Phase (en haut) et sa dérivée (en bas). HRTF égalisées en champ diffus, position [0°,70°]. Sujet ArMi.

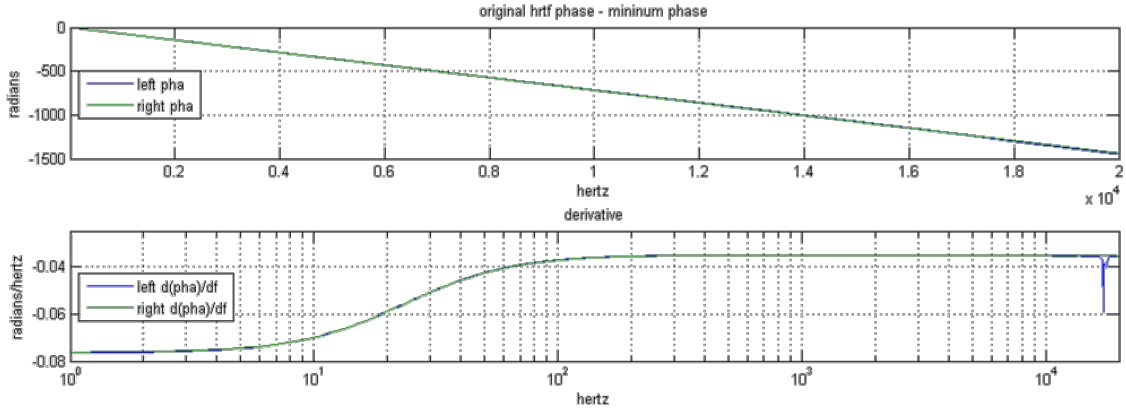

Figure 20. Phase à laquelle on a retranché la phase minimale du spectre d'amplitude.

8.3 Résultats

Nous avons à notre disposition un banc de filtres capable d'analyser tout le spectre d'une paire de HRTF donnée et pouvons reconstruire des filtres lissés conservant l'énergie par sous-bandes, ainsi que les indices interauraux de la paire originale. La section suivante abordera la modification des énergies par sous-bandes tandis que nous nous contentons ici de valider ce modèle et de préciser ces paramètres (nombre de sous bandes et intervalle de fréquence).

8.3.1 Choix du nombre de sous-bandes

Nous cherchons à valider ce lissage en comparant différents grains de lissage sur la paire de HRTF à [0°,70°]. Pour valider un lissage, nous avons comparé les HRTF lissées aux HRTF originales. De plus, il nous semble judicieux de juger à la fois, mais séparément les différences de timbre et les différences de localisation (largeur de la source, externalisation, direction).

Pour un lissage avec 20 bandes, soit donc un taux de 2 ERBS par bande, on entend une source diffuse, vers l'avant, perçue sur les oreilles. On n'entend pas, à proprement parler, une fusion des deux signaux binauraux en une source ponctuelle, mais plutôt deux sources excitant chacune une oreille. En se concentrant sur le timbre, on entend clairement que le son est plus « brut », c'est-à-dire plus proche de la référence de bruit blanc non filtré. Pour N=30, le timbre est plus agréable même si toujours éloigné des originaux. À partir de 40 bandes, la différence n'est plus nette. Au-delà de 60, aucune différence n'est perceptible.

8.3.2 Exclusion des basses fréquences

On peut se demander s'il est indispensable de lisser les HRTF par bandes ERB sur tout le spectre audible. La question porte principalement sur les basses fréquences. Pour toutes les fréquences inférieures à une certaine f_{min}, on pourrait imposer sur la HRTF lissée une amplitude moyenne calculée sur la HRTF originale sur [0—f_{min}].

Comme expliqué section 1.2.4, les indices spectraux influents sur la perception en 3D se trouveraient à des fréquences au moins supérieures à 2kHz. On constate néanmoins que les HRTF varient déjà pour des fréquences plus basses. Par exemple, à la position choisie, on observe un creux de -5dB de 400Hz à 900Hz (cf. Figure 13). En prenant f_{min} = 1kHz, cette atténuation sera étalée sur [0—1kHz] dans la version lissée et cette modification sera audible. Afin de déterminer pour quelle fréquence f_{min} aucun changement n'est audible entre les HRTF originales et les lissées, nous écoutons aléatoirement un bruit blanc de 2s filtré par l'une des deux. Nous constatons qu'en deçà de f_{min} = 200 Hz, leur identification relève de la chance.

8.4 Conclusion

La conception de ce modèle d'analyse-synthèse de HRTF s'est largement appuyée sur la littérature psychoacoustique ainsi que sur quelques références de traitement du signal. Pour ajuster les paramètres du modèle, nous avons cherché les valeurs garantissant un traitement perceptivement transparent, afin que l'auditeur ne puisse différencier HRTF originales et lissées. Ce résultat étant obtenu, nous pouvons à présent aborder le tuning des bandes d'énergies pour améliorer l'externalisation.

9 Design spectral

Nous abordons à présent l'objet de ce stage : le design spectral des HRTF pour améliorer l'externalisation des sources lors de leur restitution en binaural.

Le terme d'égalisation est souvent utilisé pour désigner une normalisation des HRTF dans toutes les directions par un même filtre (égalisation champ diffus, champ libre). Ces modifications des HRTF visent à corriger ou minimiser les détimbrages introduits par les filtres HRTF.

Nous employons le terme tuning spectral pour désigner une modification des HRTF bien plus localisée. Il s'agit d'ajuster finement l'amplitude des HRTF, pour une direction donnée et pour certaines fréquences, dans l'espoir de renforcer la perception de sources externalisées.

Figure 21. Le terme anglais de tuning rassemble les notions de réglage, d'ajustement, mais aussi l'idée d'exagérer certaines propriétés, de manière analogue au « car tuning » qui consiste à transformer une voiture grand public en une voiture de course.

9.1 Motivations

9.1.1 Inversions et détimbrage

L'idée de modifier le spectre des HRTF pour améliorer la qualité d'un espace auditif virtuel est déjà assez répandue. La difficulté réside dans le fait que les modifications spectrales sont susceptibles d'affecter la perception de deux attributs : la localisation (c'est ce que l'on cherche) et le timbre (c'est ce que l'on souhaite éviter). La question du détimbrage est un problème inhérent aux des technologies binaurales et constitue d'ailleurs l'un des principaux freins à leur diffusion. En effet, à l'écoute de contenus binauraux obtenus à partir de HRTF non-individuelles, l'auditeur n'identifie qu'approximativement les indices spectraux, car ces derniers diffèrent plus ou moins des siens. En simplifiant, l'écart entre ces indices non-individuels et les indices individuels de l'auditeur peut être interprété comme un détimbrage de la source.

Le problème se pose ainsi particulièrement aux positions médianes (θ=0° ou θ=180°) où les HRTF sont sensiblement symétriques. Les oreilles gauche et droite reçoivent un signal binaural de spectre $h_L(f)*S(f)$ et $h_R(f)*S(f)$ (en notant h_L et h_R la paire de HRTF non-individuelles et S le spectre de la source). En principe l'auditeur analyse correctement le signal binaural comme la source S placée à une certaine position. Cependant, les HRTF ne correspondent pas à ses HRTF individuelles avec lesquelles il a l'habitude de percevoir les sources au quotidien. De plus, on

peut supposer $h_L = h_R = h$, ce qui est acceptable dans le plan médian. Par conséquent, le filtrage par h peut être mal interprété et le signal binaural perçu par l'auditeur comme une source S'(f)= h(f)*S(f) à une position indéterminée.

Une première illustration est fournie par les inversions avant/arrière. Zhang *et al.* [**78**] ont montré que l'on pouvait obtenir des taux d'inversion plus faibles en modifiant les HRTF afin d'exagérer les différences spectrales entre les HRTF de directions symétriques avant/arrière.

Silzle [**66**] s'est penché sur un protocole visant à trouver un équilibre entre spatialisation et détimbrage. Le choix des HRTF, des différents lissages et des égalisations est confié à un expert en son 3D. Ce choix est ensuite validé par test d'écoutes avec d'autres sujets.

On a le sentiment que ces deux approches vont dans des directions opposées. Tandis que la résolution des inversions se fait en accentuant les indices spectraux, Silzle résout le problème de détimbrage en lissant les HRTF afin de minimiser les indices spectraux tout en permettant la localisation [**66**].

9.1.2 Apprentissage des HRTF

L'apprentissage du signal présenté précédemment (cf. 7.3) se traduit par une meilleure capacité à percevoir des sources virtuellement spatialisées et a été très bénéfique pour ensuite juger l'externalisation. Il semble néanmoins que ces gains sont aussi imputables à un apprentissage des HRTF. Nous avons en effet constaté que des synthèses binaurales réalisées avec des HRTF individuelles et non-individuelles présentaient une qualité d'externalisation meilleure au fil du temps.

Hofman *et al.* proposent de simuler une écoute par des HRTF non individualisées en obstruant certaines cavités des pavillons des oreilles [**26**]. Ils montrent qu'au bout de quelques semaines d'adaptation, les sujets localisent les sources avec autant de précision qu'avec leurs propres filtres. La plasticité du système auditif offre de nombreuses possibilités pour pallier le besoin de HRTF individualisées. Le lecteur pourra se référer à la section 4.2.7 de [**20**] pour quelques applications concrètes de ces effets.

Ces résultats sont en soit très positifs, car ils suggèrent que l'on puisse améliorer d'externalisation par l'apprentissage de HRTF génériques. Toutefois, dans le cadre de ce stage, nous voulons isoler les effets du design de HRTF et un apprentissage introduirait un biais.

On veillera donc à limiter les effets d'apprentissage en utilisant des HRTF jamais écoutées auparavant pour mener nos explorations.

9.2 Démarche

Dans le cadre de ce stage, nous ne chercherons pas à résoudre le défaut de détimbrage et nous inspirerons davantage des travaux visant à renforcer les indices spectraux. Il faut néanmoins garder à l'esprit que la modification du spectre des HRTF peut aussi bien être interprétée comme de nouveaux indices spectraux que comme une modification du spectre de la source, surtout si la modification altère de manière identique les HRTF gauche et droite.

La plupart des comparaisons se feront sur le filtrage d'un bruit blanc de 2s (cf. 7). On travaillera toujours sur les HRTF mesurées à la position [0°,70°] (cf. 6). Le lissage des HRTF devra être

perceptivement indétectable, on choisit pour cela au moins 80 bandes ERB de lissage sur le spectre audible (cf. 8.3.1).

Le défaut d'externalisation est principalement rencontré lors de l'écoute avec des HRTF non-individuelles mais peut aussi apparaître avec ses propres HRTF. C'est pourquoi il nous semble judicieux d'explorer le tuning de filtres non-individuels aussi bien que le filtrage de ses propres HRTF.

9.2.1 Tuning de HRTF non-individuelles : Comparaison Clément-ArMi

Cette partie présente nos recherches sur le tuning de HRTF non-individuelles. Nous sélectionnons un jeu de HRTF mesurées avec le même système de mesure que les nôtres, en l'occurrence celui d'Orange Labs à Lannion. Comme de nombreux sujets y ont été mesurés, notre choix parmi ces derniers s'est basé sur les critères suivants :

- HRTF jamais utilisées auparavant : Avant de disposer de mes HRTF individuelles mesurées, j'ai dû me contenter d'autres HRTF. Les nombreuses écoutes avec ces filtres se sont certainement traduites par un apprentissage des filtres (cf. 9.1.2). Ainsi, les HRTF déjà écoutées sont exclues.
- HRTF permettant une bonne localisation : Nous choisissons un jeu de HRTF pour lequel les positions perçues sont cohérentes. En effet, avec des HRTF non individuelles, les confusions avant/arrière sont fréquentes et la perception de l'élévation est parfois floue.
- HRTF produisant une perception intracrânienne : Nous souhaitons créer un effet d'externalisation et choisissons donc des HRTF ne le permettant pas par défaut. Notons que toutes les HRTF remplissent cette condition, à l'exception de celles déjà écoutées à plusieurs reprises.

Nous optons pour les HRTF mesurées de Clément Cerles, lesquelles permettent une bonne localisation malgré une perception intracrânienne. Nous choisissons les filtres associés à l'azimut de -6°, car mieux centrés (cf. Annexe 5). On trouvera Figure 22 mes HRTF ArMi et celles de Clément.

Nous nous assurons que le lissage ne fait pas apparaître d' artéfacts puis débutons notre exploration sonore.

On compare les HRTF (versions lissées), en les écoutant successivement et l'on constate que la première différence qui nous frappe est la variation du timbre. Nous cherchons donc à ajuster ce timbre en recopiant les bandes en dessous de 1 kHz des filtres ArMi sur les filtres de Clément. La principale différence provient cependant des hautes fréquences. Nous amplifions donc les HRTF de Clément pour les fréquences au-delà de 10 kHz jusqu'à ce que l'écoute des filtres d'ArMi et de Clément procure la sensation d'un timbre similaire.

Nous avons aussi envisagé d'imposer les énergies des bandes d'ArMi sur [10k—20kHz] sur les bandes de Clément. Cela donne des résultats semblables donc nous nous en tenons la première solution, car notre objectif est d'ajuster les indices spectraux de HRTF non-individuelles et non pas de recopier ceux mesurés individuellement.

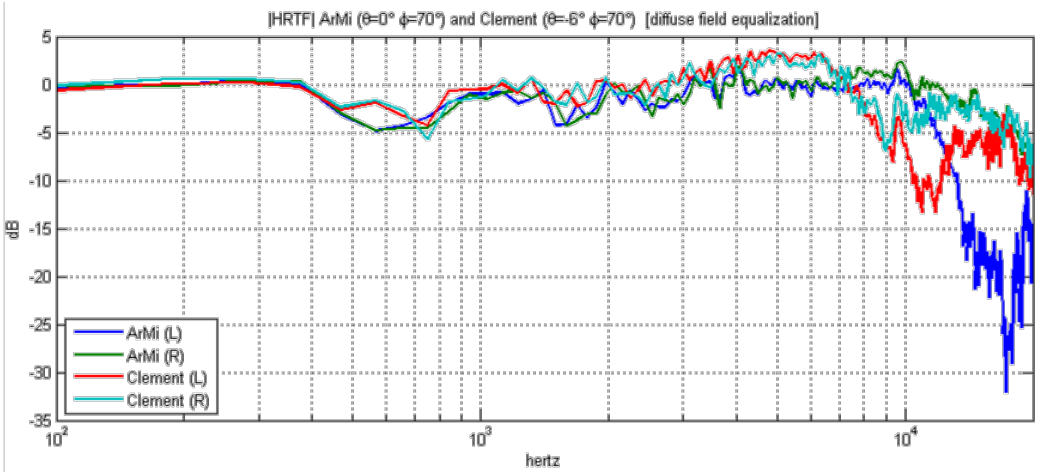

Figure 22. HRTF pour les sujets ArMi à la position [0°,70°] (bleu et vert) et Clément (rouge et turquoise).

9.2.2 Région fréquentielle

Les modifications des HRTF de Clément que nous avons présentées ci-dessus n'ont pas amené une meilleure externalisation. En revanche, elles ont mis en évidence une perception de la direction et de l'externalisation différente selon les bandes de fréquences. Nous décidons donc d'approfondir cette question, en travaillant directement sur mes propres HRTF, toujours à la position [0°,70°].

Nous avons observé qu'à l'écoute d'une source spatialisée à cette position, la source est externalisée sur une bande assez étroite tandis que le signal présent en dehors de cette bande perturbe l'externalisation de la source. Nous percevons la source bien externalisée, à la bonne position, sur une bande centrée sur 4 kHz, de bande passante d'approximativement 2 kHz. L'observation des HRTF montre une amplification aux fréquences proches de 4 kHz dans de nombreuses directions (cf. Annexe 3). Cette fréquence correspond au premier mode de résonance du pavillon. On peut alors supposer qu'une amplification par ce mode se traduirait perceptivement comme un marqueur du passage de l'onde sonore par le pavillon et donc un marqueur de l'externalisation de la source. La plupart des sons contenant encore de l'énergie à ces fréquences moyennes, cet indice serait toujours révélé.

Pour déterminer la fréquence et largeur de la bande bien externalisée, nous développons un patch en MAX/MSP permettant de générer des bandes de bruit en contrôlant en temps réel leur fréquence centrale, leur largeur et leur raideur. En écoutant alternativement le son spatialisé et la bande de bruit, nous faisons progressivement correspondre la bande de bruit avec la bande bien externalisée en jouant sur sa fréquence centrale et sa largeur.

En amplifiant ou atténuant uniformément (i.e. avec le même gain) les bandes d'énergie au-delà de 10 kHz, nous constatons que cela ajoute ou supprime un bruit parasite mais n'influe nullement sur la source bien externalisée. Ce son parasite est perçu autour des oreilles.

9.2.3 Trajectoire d'apprentissage

Afin de préparer l'auditeur à l'écoute de HRTF non-individuelles, nous mettons en place une phase d'apprentissage consistant à faire tourner la source avant d'atteindre la position que l'on souhaite modifier. Le but est double : habituer le sujet à ces HRTF et s'assurer que la position perçue est la bonne. Ainsi, avant l'écoute de la position étudiée [0°,70°], nous diffusons

successivement des positions d'azimut $\theta = 0°$ et d'élévation décroissante : $\phi = \{180°,170°,160°,\ldots,80°,70°\}$.

Nous pouvons remarquer que ces trajectoires permettent de mieux percevoir de manière moins ambigüe la position, et de pouvoir alors mieux se concentrer sur l'externalisation de la source, et le timbre. Cette trajectoire prend du temps avec le stimulus par défaut (bruit blanc de 2s) et nous optons donc pour réaliser la trajectoire d'apprentissage avec un extrait plus court (bruit blanc de 200ms).

Nous avons aussi envisagé de créer un effet d'oscillations des HRTF autour d'une position. Le pas de mesure des HRTF est malheureusement trop grand (6° en élévation) et il faudrait alors envisager l'interpolation d'HRTF, ce que la durée du stage ne permet pas.

9.2.4 Niveau sonore du stimulus

La question du niveau des signaux spatialisés est rarement abordée et les auteurs se contentent souvent de choisir un niveau confortable pour l'utilisateur. Macpherson et Middlebrooks dans [**39**] puis par Vliegen et Van Opstal dans [**69**] montrent que notre capacité de localisation varie avec l'intensité du stimulus. Pour un niveau à 45 dB au-dessus du seuil de détection (similaire aux 73 dB SPL dans [**69**]), les capacités de localisation sont dégradées. Seraient en cause d'après les auteurs, des phénomènes de compression dans l'oreille faisant interférer les indices. Schönstein propose une revue plus détaillée de ces questions partie 4.3 de [**65**] et évoque aussi l'activation du réflexe stapédien au-delà de 70 dB SPL.

Dans notre cas, la plupart des écoutes ont été faites à un niveau sonore relativement fort, de l'ordre de 70 à 80 dB SPL au niveau des oreilles. C'est assez tardivement que nous considérons cet aspect. Nous constatons qu'un niveau élevé fatigue beaucoup plus rapidement l'auditeur, mais mettons en évidence d'autres effets sur la spatialisation et l'externalisation.

9.2.4.1 Déplacement latéral

Nous remarquons qu'à la position étudiée [0°,70°], le son se déplace clairement sur la droite lorsque l'on augmente le niveau. Ceci peut s'expliquer par l'asymétrie des HRTF en haute fréquence et par les courbes d'isosonie.

Les lignes isosoniques tracées Figure 23 montrent que l'intensité perçue des sons varie avec la fréquence de ces derniers. À niveaux égaux, certaines fréquences sont perçues plus fortes que d'autres. Les courbes d'isosonie ont tendance à s'aplatir pour des niveaux plus forts. Ainsi, en augmentant le niveau global d'écoute, la contribution des hautes fréquences est plus importante.

Or, nous constatons sur mes HRTF que la réponse fréquentielle de l'oreille gauche est plus faible que la droite, au-delà de 10kHz (cf. Figure 13). Ainsi, la dysmétrie droite-gauche est plus perceptible à un niveau élevé et l'on déplace donc la source du côté le plus fort.

9.2.4.2 Internalisation

Nous constatons dans un second temps que la qualité de l'externalisation est dégradée par l'augmentation du niveau sonore.

On module en amplitude du bruit blanc (Figure 24) et l'on diffuse ce même signal à chaque oreille. Tandis qu'à faible niveau, le son peut être perçu comme externalisé, la brusque augmentation du niveau renvoie toujours le son à l'intérieur de la tête. Ces observations

semblent être partagées par d'autres sujets et ces résultats mériteraient d'être validés avec plus de rigueur.

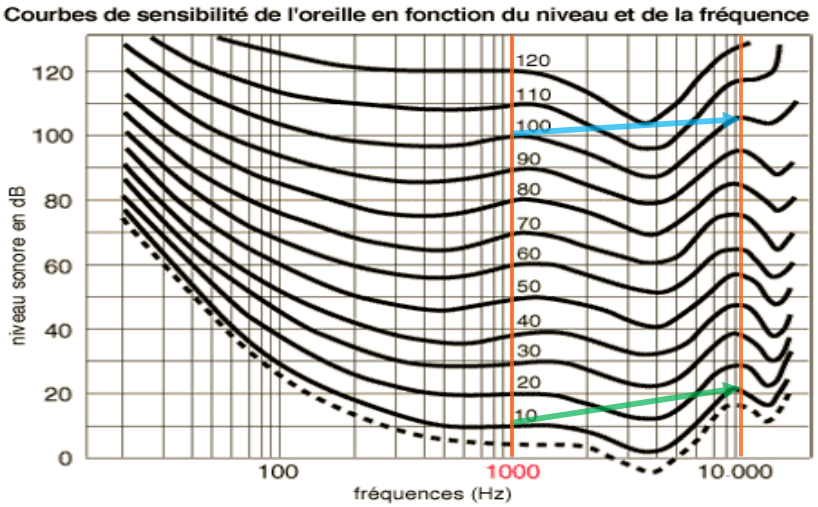

Figure 23. Courbes isosoniques : pour un signal à 10 dB SPL le niveau sonore varie de près de +10 dB entre 4 kHz et 10 kHz (en vert), pour un signal à 100 dB SPL, il varie de seulement +5 dB entre les mêmes fréquences (en bleu).

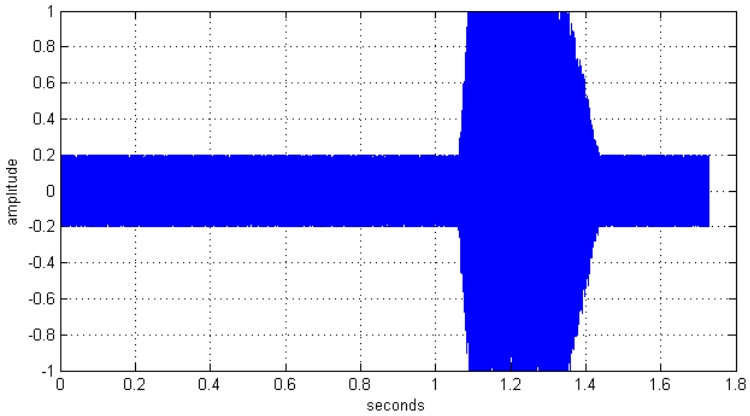

Figure 24. Bruit blanc modulé en amplitude (Salve générée avec les paramètres suivants : attaque +14 dB pendant 30 ms, maintien pendant 300 ms, décroissance de -100 dB pendant 100 ms.)

9.3 Conclusion

Les résultats suggèrent que l'externalisation est un attribut qui ne s'applique pas pour toutes les fréquences. Sa perception serait au contraire limitée à une région fréquentielle, en dehors de laquelle, le son serait perçu intracrânien. Nous rendons aussi compte de l'importance du niveau sonore des signaux diffusés. La perception humaine de l'intensité est loin d'être linéaire et des effets indésirables dégradant l'externalisation se produisent à des niveaux élevés. Enfin, le jugement de l'externalisation exige de nombreux prérequis : l'apprentissage nécessaire du stimulus par écoute de sources réelles, un « verrouillage » de la direction étudiée par la synthèse de trajectoires et enfin la minimisation de l'apprentissage de HRTF génériques par l'utilisation de HRTF jamais écoutées auparavant.

Conclusion générale

L'objectif de ce stage était de déterminer dans quelle mesure le design spectral de HRTF peut résoudre le problème de perception intracrânienne des sources en synthèse binaurale.

Une importante partie de ce rapport a présenté les fondements de la spatialisation sonore et de la synthèse binaurale, en insistant sur la notion d'indice spectral. Cette revue bibliographique souligne particulièrement la fragilité de l'externalisation lors de la synthèse d'espaces auditifs. Tandis que l'ajout de réverbération ou la synthèse dynamique s'adaptant au mouvement de la tête résolvent couramment la localisation intracrânienne, la contribution des indices spectraux est beaucoup plus ambigüe et ne fait pas l'unanimité.

Le travail de recherche s'est ensuite axé sur le design spectral des HRTF. L'idée de sculpter un spectre d'amplitude afin d'externaliser la perception des sources semble assez simple, mais appelle en réalité de nombreux développements techniques ainsi que des recherches bibliographiques complémentaires. S'il consomme beaucoup de temps, ce travail préliminaire a été en revanche très bénéfique. Le choix de la position sur laquelle travailler, le choix du stimulus à externaliser ou encore la validation d'une méthode de lissage spectral sont autant d'étapes qui ont nourri la phase d'exploration sur le design spectral. Ainsi, la majorité de nos travaux sur le *tuning* des HRTF furent imaginés lors de cette phase préliminaire. D'autre part, c'est ce travail qui aura été le plus formateur à titre personnel, m'apportant des compétences en traitement signal audio nouvelles ainsi qu'une meilleure compréhension de notre audition.

Ce rapport se termine par la présentation du travail de design spectral à proprement parler. Cette dernière étape nous a permis d'identifier des comportements différents selon les fréquences ainsi qu'un certain nombre de paramètres qui influencent la qualité d'externalisation. Ce stage était vraisemblablement ambitieux et nous n'avons pas été en mesure de valider ces résultats par des tests d'écoute sur d'autres sujets. Il laisse néanmoins une base bibliographique solide, plusieurs résultats intermédiaires sur la perception intracrânienne et livre des outils de design spectral de HRTF fonctionnels.

Bibliographie

[1] Algazi, Avendano, and Duda, "Elevation localization and headrelated transfer function analysis at low frequencies.," *Journal of the Acoustical Society of America*, vol. 3, no. 109, pp. 1110–1122, 2001.

[2] Association Belge d'Orthoptie. Le champ visuel. [Online]. http://www.orthoptie.be/fr/hoe-werkt-het-oog/gezichtsveld/

[3] D. R. Begault, E. M. Wenzel, and M. R & Anderson, "Direct comparison of the impact on head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *108th Convention of the Audio Engineering Society*, Paris, 2000.

[4] BiLi. [Online]. http://www.bili-project.org/le-projet/

[5] Jens Blauert, "Sound localization in median plane," *Acustica*, vol. 4, no. 22, pp. 205-213, 1969.

[6] Jens Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localisation*.: MIT Press, 1974.

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

[8] Boyd, "Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers," *Journal of the Acoustical Society of America*, no. 131, 2012.

[9] J. Breebaart and A. Kohlrausch, "The perceptual (ir) relevance of HRTF magnitude and phase spectra," in *Audio Engineering Society Convention*, Amsterdam, 2001.

[10] T. Brookes and C. Treble, "The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binuaral recordings," in *118th Convention of the Audio Engineering Society*, 2005.

[11] Burge and Burger, "Ear biometrics," in *Biometrics, Personal Identification in Networked Society*., 1996.

[12] Sylvain Busson, *Individualisation d'indices acoustiques pour la synthèse*., 2006.

[13] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America*, no. 104, 1998.

[14] Ramani Duraiswami, "Creating Virtual Spatial Audio Via Scientific Computing and Computer Vision," in *Acoustical Society of America 140th Meeting*, 2000.

[15] N. I Durlach et al., "On the externalization of auditory images," *Presence: Teleoperators and*

Virtual Environments, vol. 1, no. 2, pp. 251-257, 1992.

[16] Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique.," in *108th Convention of the Audio Engineering Society*, Paris, 2000.

[17] Harvey Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[18] Gardner, "Problem of localization in the median plane : effect of pinnae cavity occlusion," *Journal of the Acoustical Society of America*, 1973.

[19] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, 1990.

[20] Pierre Guillon, *Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF.*, 2009.

[21] Gupta, "Improved localization of virtual sound by spectral modification of hrtfs to simulate protruding pinnae," in *6th World Multiconference on Systemics, Cybernetics and Informatics*, 2002.

[22] Hartmann and Wittenberg, "On the externalization of sound images," *Journal of the Acoustical Society of America*, no. 99, 1996.

[23] H. G. Hassager, T. Dau, and F Gran, "The effect of spectral details on the externalization of sounds," in *7th Forum Acusticum*, Krakow, 2014.

[24] Hawley, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer.," *The Journal of the Acoustical Society of America*, no. 115, 2004.

[25] Hofman, "Spectro-temporal factors in two-dimensional human sound localization," *Journal of the Acoustical Society of America*, 1998.

[26] Hofman, Van Riswick, and Van Opstal, "Relearning sound localization with new ears," *Nature neuroscience*, 1998.

[27] R. A. Humanski and R. A. Butler, "The contribution of the near and far ear toward localization of sound in the sagittal plane," *Journal of the Acoustical Society of America*, vol. 6, no. 83, pp. 2300–2310, 1986.

[28] K. Inanaga, Y. Yamada, and H. Koizumi, "Headphone system with out-of-head localization applying dynamic HRTF (Head-Related Transfer Function)," in *Audio Engineering Society Convention 98*, Paris, 1995.

[29] Yukio Iwaya and Yôiti Suzuki, "Numerical Analysis of Effects of Pinna's Shape/Position on Characteristics of Head-Related Transfer Functions," in *Proceedings of Forum Acousticum 2008*, Paris, 2008.

[30] Jack and Thurlow, "Effects of degree of visual association and angle of displacement on the "ventriloquism" effect," *Perceptual and Motor Skills*, 1973.

[31] Françoise Jauzein and Anne Woehrlé. (2011) Interprétation des anomalies du champ visuel. [Online]. http://acces.ens-lyon.fr/acces/ressources/neurosciences/vision/comprendre/cas_anomalies_vision/tech_explo_vision/interpretationAnomalie

[32] Jot, Larcher, and Warusfel, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," in *Audio Engineering Society Convention 98*, 1995.

[33] Kim and Choi, "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach," *Journal of the Acoustical Society of America*, 2005.

[34] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747-749, 1998.

[35] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *Applications of Signal Processing to Audio and Acoustics, IEEE ASSP Workshop on*, 1995, pp. 84-87.

[36] Langendijk, "Contribution of spectral cues to human sound localization," *Journal of the Acoustical Society of America*, 2002.

[37] Le champ visuel. [Online]. http://www.orthoptie.be/fr/hoe-werkt-het-oog/gezichtsveld/

[38] Loomis, "Personal Guidance System for the Visually Impaired," in *Proc. of the first annual ACM conference on Assistive technologies.*, 1994, p. Proc. of the first.

[39] E. A. Macpherson and J. C Middlebrooks, "Localization of brief sounds: Effects of level and background noise.," *Journal of the Acoustical Society of America*, vol. 4, no. 108, pp. 1834–1849, 2000.

[40] Majdak, Balazs, and Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions.," *Journal of the Audio Engineering Society*, no. 55, pp. 623-637, 2007.

[41] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188-2200, 1990.

[42] Martens, "Uses and misuses of psychophysical methods in the evaluation of spatial sound reproduction," in *110th Convention of the Audio Engineering Society*, 2001.

[43] William L. Martens, Densil Cabrera, and Ken Stewart, "Close-range variation in binaural responses to orally-radiated sources," in *Proceedings of ACOUSTICS 2011, Gold Coast, Australia*, 2011.

[44] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926-940, 2011.

[45] McMullen, "Subjective Selection of Head-Related Transfer Functions (HRTFs) based on Spectral Coloration and Interaural Time Differences (ITD) Cues," in *133rd Convention of the Audio Engineering Society*, 2012.

[46] D. Mershon and L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409-415, 1975.

[47] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *Journal of the Acoustical Society of America*, no. 92, pp. 2607–2624, 1992.

[48] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividuallized external-ear transfer functions in frequency," *Journal of the Acoustical Society of America*, no. 106, 1999.

[49] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *Journal of the Audio Engineering Society*, vol. 43, no. 4, pp. 203-217, 1995.

[50] Alastair Howe Moore, *Towards the perception of externalised auditory images using binaural technology*., 2009.

[51] Rozenn Nicol, *Représentation et perception des espaces auditifs virtuels*., 2010.

[52] Nicol et al., "Le son 3D dans les futurs services de télécommunication," *Acoustique et Techniques*, no. 71, 2013.

[53] R. Nicol et al., "Reference tools and methods to assess the QoE of binaural sound," Rapport interne.

[54] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*.: Prentice Hall, 1975.

[55] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, *Discrete-time signal processing*.: Prentice Hall, 1989.

[56] Roy D. Patterson, "Auditory filter shapes derived with noise stimuli," *The Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 640-654, 1976.

[57] Jean-Marie Pernaux, *Spatialisation du son par les techniques binaurales : Application aux services de télécommunications*., 2003.

[58] Radio France. NouvOson. [Online]. http://nouvoson.radiofrance.fr/

[59] Lord Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, 1907.

[60] Roginska, "User selected HRTFs: Reduced complexity and improved perception.," in

Proceedings of the Undersea Human Systems Integration Symposium, 2010.

[61] Pascal Rueff. 3D-Radio. [Online]. http://www.binaural.fr/

[62] Sakamoto, "On “out-of-head localization” in headphone listening," *Journal of the Audio Engineering Society*, no. 24, 1976.

[63] N. Sakamoto, T. Gotoh, and Y. & Kimura, "On “out-of-head localization” in headphone listening," *Journal of the Audio Engineering Society*, vol. 24, no. 9, pp. 710-716, 1976.

[64] Sophie Savel, "Introduction à la psychophysique," Laboratoire de Mécanique et d'Acoustique, Marseille, 2015.

[65] David Schönstein, *Individualisation of spectral cues for applications in virtual auditory space: study of inter-subject differences in head-related transfer functions using perceptual judgements from listening tests*.: Université Pierre et Marie Curie-Paris VI, 2012.

[66] Silzle, "Selection and tuning of HRTFs," in *112th Audio Engineering Society Convention*, 2002.

[67] SNOF. L'acuité visuelle. [Online]. http://www.snof.org/encyclopedie/acuite-visuelle

[68] S. S. Stevens and E. B. Newman, "The localization of actual sources of sound," *The American Journal of Psychology*, pp. 297-306, 1936.

[69] J. Vliegen and A. J. Van Opstal, "The influence of duration and level on human sound localization," *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1705-1713, 2004.

[70] F. Völk, F. Heinemann, and H. Fastl, "Externalization in binaural synthesis : effects of recording environment and measurement procedure," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3935, 2008.

[71] Georg Von Békésy, *Experiments in hearing*. New York: McGraw-Hill, 1960.

[72] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, no. 94, 1993.

[73] F. L. Wightman and D. J. Kistler, "Headphone simulation of freefield listening. I: Stimulus synthesis," *Journal of the Acoustical Society of America*, no. 85, 1989.

[74] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II:Pyschophysical validation," *Journal of the Acoustical of America*, no. 85, 1989.

[75] Wikimedia Commons. File:Band-pass filter.svg. [Online]. https://commons.wikimedia.org/wiki/File:Band-pass_filter.svg

[76] Wikipédia. Chirp. [Online]. http://fr.wikipedia.org/wiki/Chirp

[77] Wikipédia. Pavillon de l'oreille (anatomie humaine). [Online]. http://fr.wikipedia.org/wiki/Pavillon_de_l%27oreille_%28anatomie_humaine%29

[78] Zhang, Tan, and Er, "Three-dimensional sound synthesis based on Head-Related Transfer Functions," *Journal of the Audio Engineering Society*, no. 46, pp. 836-844, 1998.

[79] Xiao-li Zhong and Bo-sun Xie, "Head-Related Transfer Functions and Virtual Auditory Display," in *Soundscape Semiotics - Localization and Categorization*., 2014.

Annexes

1 Démarche ingénieur de recherche

Ce stage constitue un travail de recherche. Cependant, les travaux réalisé à Orange Labs sur le son 3D aboutissent aussi à la commercialisation par Orange de deux types de produits :

1. Les communications immersives, qu'elles soient dédiées au grand public ou aux solutions professionnelles.
2. La qualité d'expérience avec l'enrichissement des contenus avec le son immersif pour la TV d'Orange, les services de vidéo à la demande ou encore le streaming de musique via son partenariat avec Deezer.

Ce rapport a montré que le travail d'un ingénieur de recherche en son 3D fait appel à des compétences en traitement du signal, en psychoacoustique mais aussi en programmation ou encore en manipulation de systèmes plus ou moins complexes (système de mesure de HRTF par exemple).

2 Protocole de mesure à Orange Labs

Orange a mis en place un protocole de mesure réduisant le temps de mesure à seulement 20 minutes. Le sujet est placé sur une table tournante. Autour de lui, 26 haut-parleurs sont régulièrement espacés en élévations ($-56° \leq \phi \leq +85°$) tandis que la table tournante fait varier l'azimut sur 360°. Un pas angulaire de 6° en azimut et élévation a été choisi et la distance à la source est maintenue constante (r = 2 m).

Ce sont donc les fonctions de transfert associées à 26*360/6 = 1560 positions angulaires qu'il faut identifier. On émet alors un sweep successivement dans chaque direction. On parle de sweep pour désigner un signal pseudo-périodique dont la fréquence instantanée croit avec le temps. Dans notre cas, la fréquence suit une rampe exponentielle de 100Hz à 24000Hz pour une durée de 7s.

Le gain de temps réside dans l'émission quasi simultanée des sweeps via les 26 haut-parleurs (déphasage de 100ms entre deux sweeps successifs). Les micros recueillent donc le son provenant de plusieurs élévations simultanément.

La déconvolution des enregistrements par chaque sweep correctement déphasé permet de conserver uniquement la contribution de la direction associée à ce déphasage, et d'obtenir la réponse impulsionnelle associée à chaque position.

Cette méthode de mesure simultanée de réponses impulsionnelles par sweeps exponentiels fut initialement proposée par [**16**]. On pourra se référer à [**40**] pour une explication claire et détaillée.

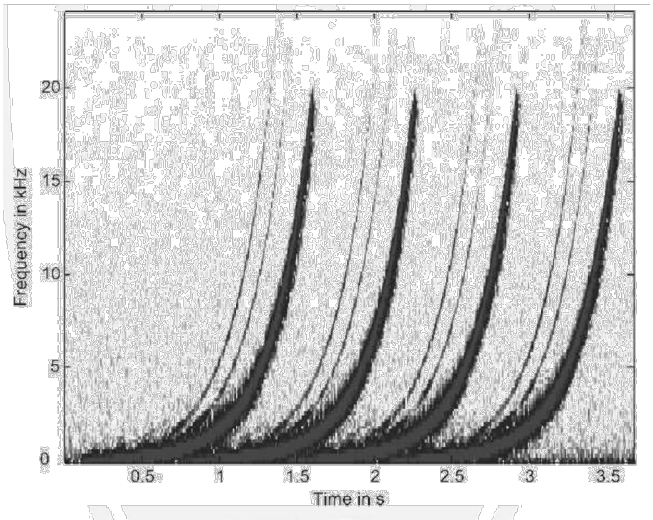

Fig. 7. Response-signal spectrogram as an example of four overlapped sweeps.

Figure 25. D'après [40].

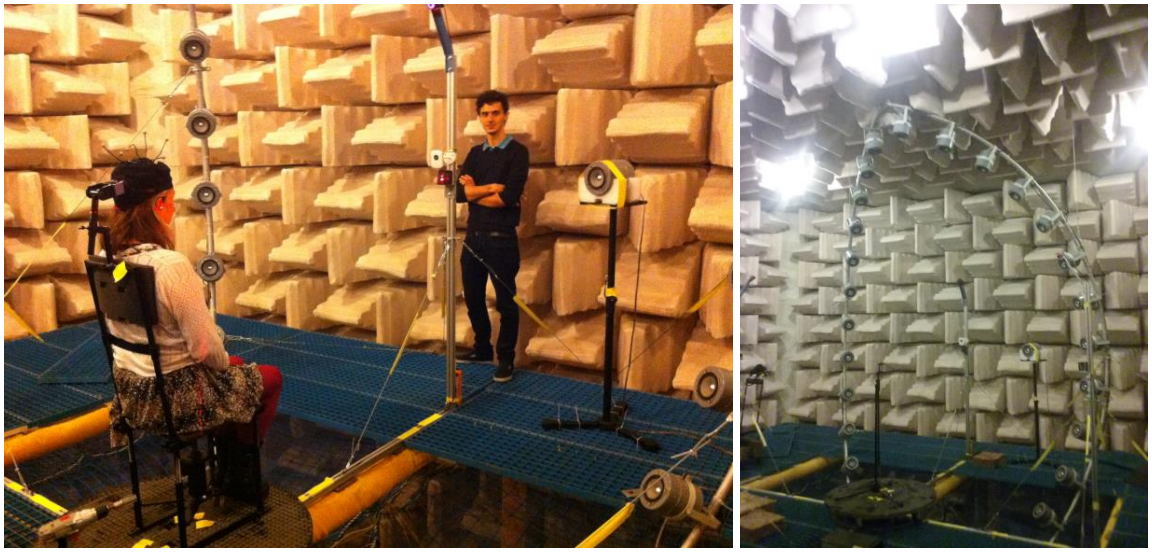

Figure 26. Dispositif de mesure de HRTF en chambre anéchoïque à Orange.

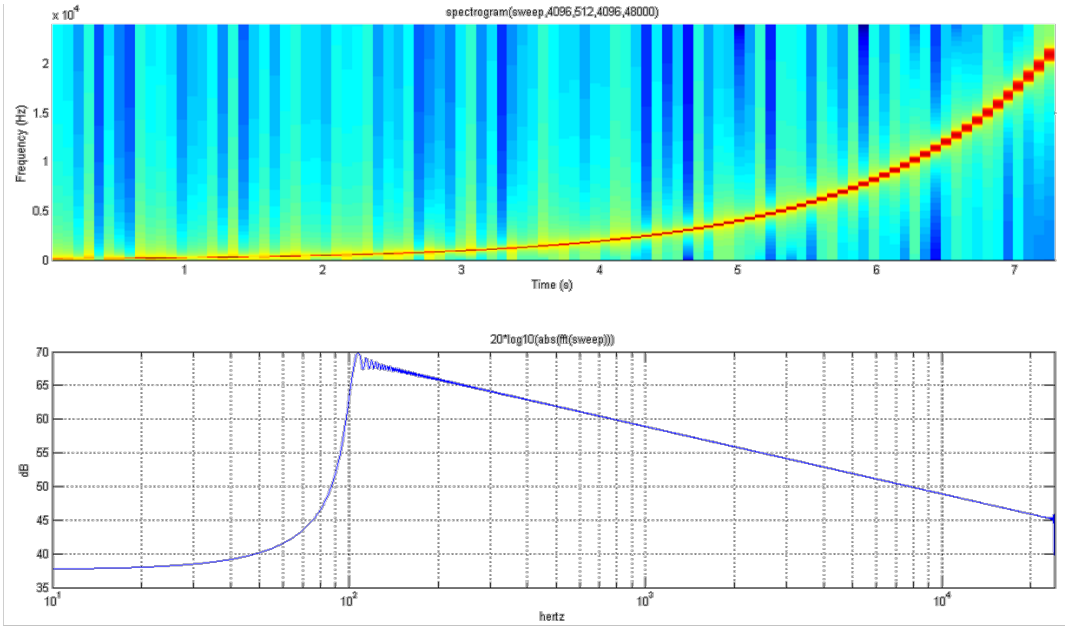

Figure 27.

3 Région fréquentielle autour de 4 kHz

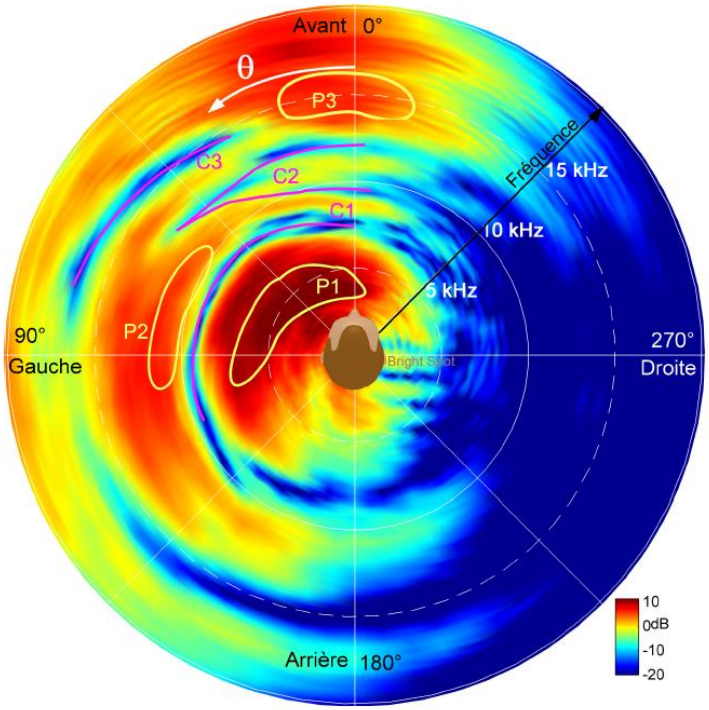

Figure 3.7 – Illustration des creux et pics caractéristiques observés dans le plan horizontal : représentation polaire du module des HRTF d'un sujet de la base privée d'Orange Labs (sujet n°5, oreille gauche, mesurées à l'entrée du conduit auditif, conduit bloqué), en fonction de l'azimut dans le système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

Figure 28. D'après [20].

4 Patch bruit filtré

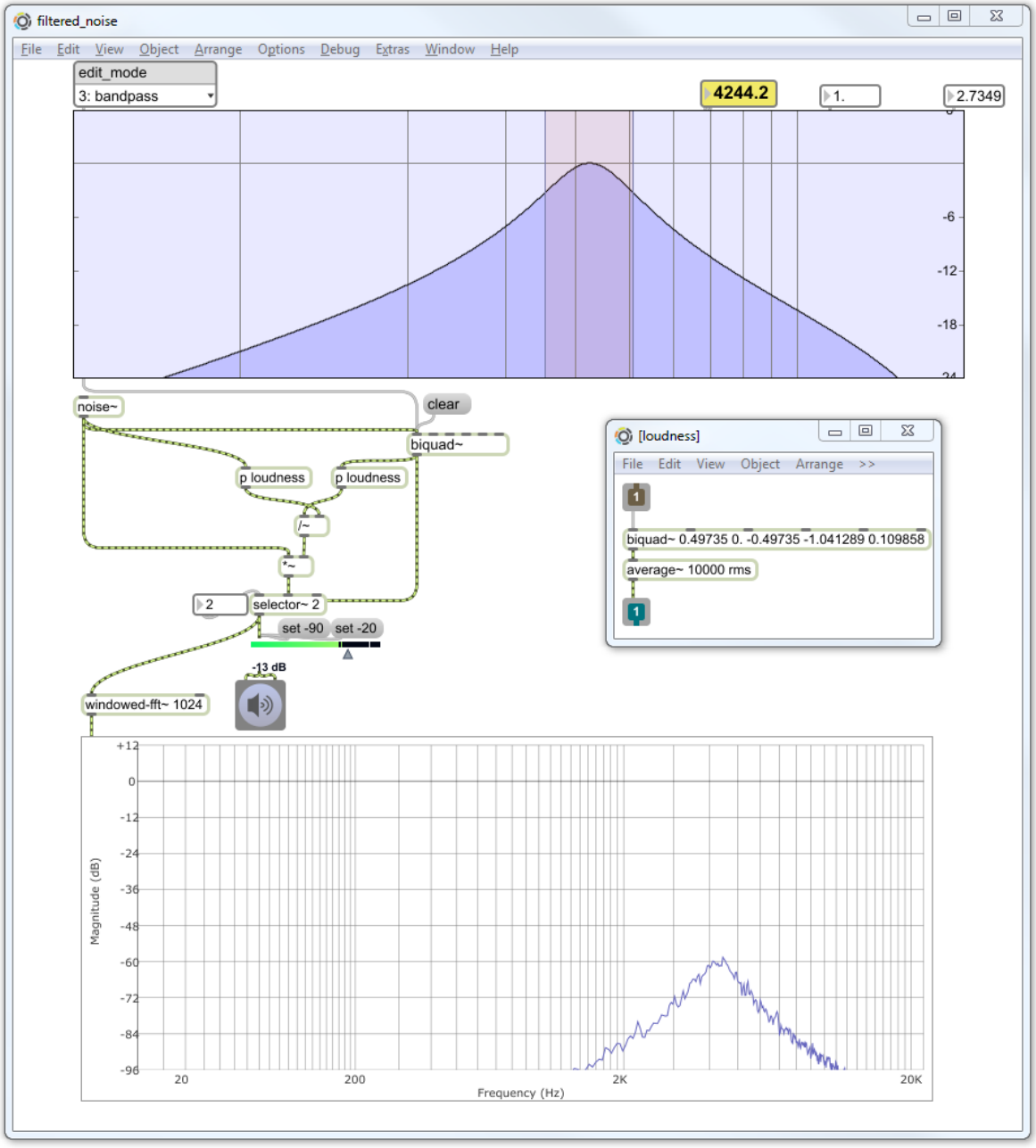

Figure 29. Capture du patch

5 Décalage des HRTF en azimut

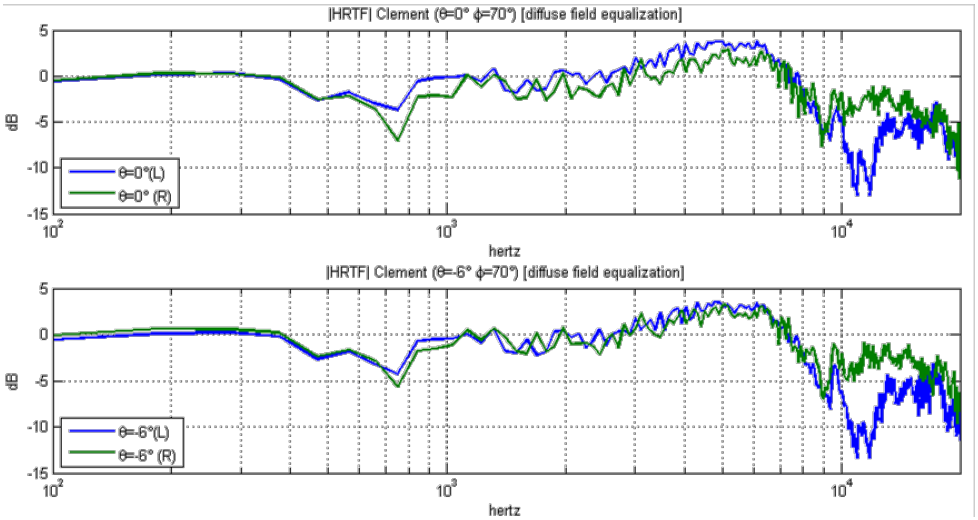

Figure 30. Les HRTF associées à la position [0°,70°] sont perçues décalées à droite tandis que celles associées à [-6°,70°] voire [-12°,70°] sont perçues bien dans l'axe médian. Nous avons constaté que l'ITD était bien nulle. En revanche, les dysmétries droite/gauche des amplitudes pourraient expliquer une ILD non nulle et donc un décalage latéral.