Reproduction transaurale adaptative temps réel appliquée à la réalité virtuelle

Rapport de stage

PARCOURS MASTER 2
ATiAM
Parcours multi-mentions du Master Sciences et Technologies
Université Pierre et Marie Curie - Paris 6
en collaboration avec TELECOM ParisTech et l'Ircam

Auteur :
Van Phan Quang

Encadrants :
Olivier Warusfel
Thibaut Carpentier

Remerciements

Je tiens à remercier mon maître de stage Olivier Warusfel, Thibaut Carpentier pour son aide précieuse et Markus Noistering pour leur encadrement au sein de l'équipe Espaces Acoustiques et Cognitifs de l'IRCAM. Je remercie l'IRCAM pour m'avoir accueilli durant mon stage. Cette année dans le master ATIAM fut très enrichissante. Je souhaite donc remercier Carlos Agon, Fleur Gire et l'UPMC pour l'organisation du master ATIAM.

Table des matières

Introduction

Dans le cadre du master ATIAM, j'ai effectué un stage de recherche de 5 mois au sein de l'IRCAM dans l'équipe Espaces Acoustiques et Cognitifs. Le stage a commencé le 1er mars 2011. L'équipe Espaces Acoustiques et Cognitifs est composée d'Olivier Warusfel, chef d'équipe et chercheur, de Markus Noisternig, chercheur, et Thibaut Carpentier, ingénieur de recherche. Les disciplines scientifiques concernées sont principalement le traitement de signal pour l'élaboration de techniques de reproduction de champs sonores et l'informatique appliquée à la conception d'interface de contrôle de la spatialisation.

Les techniques de spatialisation sonore sont utilisées dans différents domaines applicatifs : production musicale, télécommunications, jeux et pour les systèmes de réalité virtuelle. Pour ce domaine, elles permettent de renforcer l'immersion au sein d'un environnement virtuel. De nombreuses techniques de spatialisation sonore ont été développées ces dernières décennies. Cependant, le contexte de la réalité virtuelle suppose que l'auditeur soit capable d'évoluer dans la scène au moins sur une zone de quelques mètres carrés. A priori, seules les techniques basées sur une approche physique, comme la WFS ou l'Ambisonic, permettent de reproduire un champ sonore sur une zone élargie et donc de remplir cette contrainte. Cependant elles nécessitent un large réseau de haut-parleurs et sont donc inexploitables dans un environnement de réalité virtuelle car les contraintes pratiques de placement des enceintes sont trop importantes (non masquage des écrans). La technique binaurale permet une reproduction à l'aide d'un casque et offre une solution simple à condition toutefois d'être associée à un système de suivi de la position de l'auditeur de sorte à adapter en temps réel la scène sonore en fonction de ses mouvements. Le port d'un casque peut être cependant vécu comme une entrave à l'immersion. Afin d'obtenir des conditions d'écoutes plus naturelles on peut recourir à la technique dite transaurale, dérivée de la technique binaurale et permettant de restituer un champ sonore aux oreilles d'un auditeur à l'aide de deux enceintes seulement. Cependant, la reproduction du champ sonore n'est alors valable qu'en un unique point de l'espace et pour une orientation fixe de l'auditeur. Dès que l'auditeur change de position, la cohérence spatiale de la restitution n'est plus garantie. L'objectif de l'étude est donc d'étendre la reproduction transaurale pour le contexte d'un auditeur naviguant dans une zone étendue.

La technique dite transaurale se base sur la technique binaurale. C'est pourquoi, il convient de rappeler les principes de base du binaural. Dans la partie 1, nous verrons comment les indices interauraux et monauraux sont primordiaux pour comprendre le fonctionnement de la localisation sonore chez un homme. Et nous verrons comment la synthèse binaurale au casque permet de créer cette impression de spatialisation sonore, à l'aide d'un filtrage bicanal par des filtres HRTF spécifiques.

Une fois la technique binaurale comprise, nous verrons comment on peut s'affranchir de l'utilisation du casque grâce à la technique transaurale. Cette technique a pour effet d'annuler l'onde acoustique transmise de chaque haut-parleur vers l'oreille opposée ; on parle d'annulation de la diaphonie transaurale (CrossTalk Cancellation). La synthèse binaurale combinée à un décodage transaural permet donc de restituer des scènes sonores spatialisées. Dans la partie 2, nous verrons les principes et les effets non désirés liés à cette technique, tels que les effets de coloration spectrale liés à l'inversion du déterminant, et nous décrirons les principales architectures de traitement de la synthèse transaurale.

Ensuite dans la partie 3, les aspects dynamiques liés au filtrage adaptatif seront détaillés. Le filtrage adaptatif doit pouvoir se faire en temps réel. Pour pouvoir répondre à ces nouvelles contraintes, il nous faudra mettre en place une chaîne de traitement capable de s'adapter à toutes les positions d'un auditeur à l'intérieur d'un domaine d'excursion. Pour cela, de nombreux paramètres doivent être pris en compte pour réaliser un système de reproduction transaurale adaptive en temps réel. Le choix du type de filtre, la mise en place d'un fondu enchaîné entre deux crosstalk cancellers, des pré-traitements sur les filtres transauraux, la définition de secteurs angulaires entre autres sont autant de paramètres à régler. Leurs effets seront étudiés à l'aide d'une

simulation MATLAB.

Puis finalement, dans la partie 4 nous décrirons également l'implémentation choisie pour réaliser le système.

Conventions

Pour représenter les phénomènes de localisation auditive on se réfère à un système de coordonnées sphériques, dont les dimensions sont l'azimut, le site et la distance. Dans la littérature anglo-saxonne le site est désigné par "elevation". Bien que l'utilisation du terme "élévation" en français soit donc un anglicisme, force est de constater que celle-ci s'est considérablement répandue et il nous a semblé plus commode de l'utiliser car sa compréhension est plus intuitive que le terme de site. Si un sujet est placé au centre de l'espace ainsi défini, une incidence de 0° d'azimut et 0° d'élévation correspond à une incidence frontale au sujet dans le plan horizontal de l'équateur. Les incidences comprises entre 0° et 180° d'azimut sont situées dans le demi-espace à gauche du sujet, celles comprises entre 180° et 360° dans le demi-espace à sa droite. Les incidences comprises entre 0° et 90° d'élévation sont situées dans l'hémisphère supérieur, celles comprises entre 0° et -90° dans l'hémisphère inférieur. Il pourra être fait allusion à l'espace angulaire défini par les azimuts -30° et 30°, il s'agira alors des azimuts compris entre 330° et 360°, et entre 0° et 30°.

HRTF	:	Head Related Transfer Function
HRIR	:	Head Related Impulse Response
H_{XX}	:	Désigne indifféremment l'un des filtres transauraux H_{LL}, H_{RL}, H_{LR} et H_{RR}
FIR	:	Réponse impulsionnelle finie (Finite Impulse Response)
IIR	:	Réponse impulsionnelle infinie (Infinite Impulse Response)
LS	:	Least Squares (moindres carrés)
FFT	:	Fast Fourier Transform (Transformée de Fourier Rapide)
BF	:	Basses fréquences, ici considérées de 20 Hz à 200 Hz
MF	:	Fréquences médium, ici considérées de 200 Hz à 2 kHz
HF	:	Fréquences aiguës, ici considérées de 2 kHz à 20 kHz
CTC	:	Annulation des trajets croisés (Crosstalk Cancellation)
ITD	:	Différences interaurales de temps (Interaural Time Differences)
ILD	:	Différences interaurales d'intensité (Interaural Level Differences)
span	:	écartement angulaire entre les haut-parleurs

1 Synthèse binaurale

La synthèse binaurale est une technique de spatialisation sonore destinée à l'écoute sur casque audio. Elle permet de simuler le procédé naturel humain de localisation en s'appuyant sur un principe d'échantillonnage. Des mesures de réponses impulsionnelles sont réalisées sur des sujets ou sur des têtes artificielles pour un ensemble d'incidences et sont consignées dans des bases de données. Ces réponses impulsionnelles, dénommées HRIR (*Head Related Impulse Response*), sont spécifiques à chaque personne et chaque incidence. Elles contiennent de manière exhaustive les informations de localisation, monaurales et interaurales, naturellement disponibles pour la localisation d'une source sonore. A partir de ces mesures, il est éventuellement possible de modéliser ces HRIR sous forme de filtres RII de sorte à réduire le temps de calcul du filtrage binaural. Ces optimisations nous seront utiles pour réaliser notre filtrage adaptatif en temps réel.

1.1 Les indices interauraux

Il est convenu que l'on voit en trois dimensions grâce à nos deux yeux. Et donc que si l'on entend en trois dimensions, ce doit être grâce à nos deux oreilles. En effet, les indices interauraux sont issus des différences entre les signaux parvenant aux deux oreilles de l'auditeur. Rayleigh en 1907 dans sa théorie duplex de localisation ([Ray07]) a mis en évidence l'importance de ces différences dans la latéralisation de sources sonores. Selon lui, les deux indices prépondérants sont :

- les différences interaurales de temps (ITD) représentant les décalages temporels entre les signaux parvenant aux oreilles
- les différences interaurales de niveau (ILD) représentant les différences d'amplitude entre les spectres des signaux des oreilles

L'ITD est due à la différence de parcours de l'onde sonore entre les deux oreilles. Selon Blauert [Bla74], cette différence interaurale de temps n'est prépondérante que jusqu'à 2 kHz. Kuhn dans son étude [G. 77], montre que l'ITD est minimum entre 1.4 et 1.6 kHz.
Pour un signal avec des fréquences au-delà de $2kHz$ l'ILD devient prépondérant par rapport à l'ITD. En effet, l'ILD qui est due au masquage de la tête de l'onde sonore est plus importante en hautes-fréquences (supérieures à 2 kHz). Ceci est dû aux dimensions moyenne de la tête comparativement aux longueurs d'onde. De sorte à établir une valeur moyenne de l'ILD, Jot dans [JO95] en propose une modélisation :

$$ILD = 10log_{10}\frac{\int_{f_l}^{f_u}|X_L(f)|^2df}{\int_{f_l}^{f_u}|X_R(f)|^2df} \quad (1)$$

avec X_L et X_R respectivement le spectre du signal gauche et droit. L'intervalle d'intégration proposé est $[f_L; f_U]$ avec $f_L = 1kHz$ et $f_U = 5kHz$, domaine pour lequel l'ILD est un indice prépondérant. Finalement ces deux indices sont complémentaires. Cependant ces indices peuvent être ambigus. Par exemple, une même valeur d'ITD peut être observée selon un ensemble d'incidences différentes. Ce lieu est appelé cône de confusion (cf figure 1). Dans le cadre d'une écoute limitée au plan horizontal, cette ambiguité se traduit par une confusion avant-arrière, qui peut naturellement s'avérer déstabilisante pour l'auditeur. Cependant, cette confusion peut être résolue par des mouvements de la tête, ce qui souligne l'importance d'associer un système adaptatif asservissant la reproduction de la scène sonore à un système de suivi de l'orientation de la tête de l'auditeur. Finalement, ces deux indices permettent de représenter principalement les phénomènes mis en jeu pour la perception de la latéralisation des sources. D'autres phénomènes sont à prendre en considération pour la perception de localisation en élévation (cf 1.2).

1.2 Les indices monoraux

Les réflexions sur le torse, les épaules, la tête et l'oreille externe transforment le signal d'une source sonore entre son point d'origine et l'oreille de l'auditeur. Cette transformation est spécifique à chaque incidence

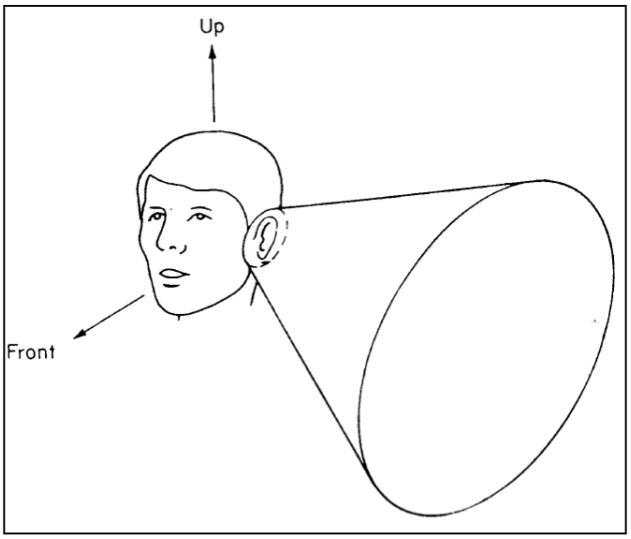


Fig. 1 – Les indices ITD et ILD sont ambigus. Pour toutes les courses sur le cône dessiné, on constate le même ITD. Concernant l'ITD, étant donné la symétrie des oreilles situées de part et d'autre de la tête. Pour un ITD donné, l'origine possible de cette ITD peut être représenté par la surface d'un cône, ce cône est appelé cône de confusion. Cette figure est tiré de [Moo04].

et chaque morphologie. Ces réflexions modifient le spectre du signal sonore. Plusieurs recherches ont décrit ces phénomènes. Parmi celles-ci, Blauert dans [Bla74] à l'aide d'expériences psychoacoustiques a pu mettre en évidence ce mécanisme de codage fréquentiel de l'oreille pour la localisation de sources sonores. Pour ses expériences, il utilisa des bruits blancs à bandes étroites, présentés en écoute distique, c'est à dire avec un même signal diffusés dans les deux oreilles. Il remarqua par exemple, qu'un signal riche en énergie autour de $8kHz$ sera perçu comme provenant d'une incidence élevée. Ces transformations fréquentielles dues à diverses réflexions sur le corps permettraient donc à l'auditeur de localiser des sources sonores en trois dimensions.
Le mécanisme de localisation sonore de l'humain est un mécanisme complexe. Il met en jeu comme nous l'avons vu des indices interauraux et monauraux. Afin de reproduire ce mécanisme de manière fidèle, l'utilisation de HRTF (Head Related Transfer Function) est donc une technique efficace puisqu'elle se base sur un échantillonnage de la fonction de directivité de la tête. Elle est assez répandue dans les systèmes de spatialisation sonore interactive comme le domaine de la réalité virtuelle.

1.3 Filtres HRTF

1.3.1 Mesure

Les filtres HRTF que nous utiliserons lors de notre étude sont issus de plusieurs campagnes de mesures :
- mesure de la tête artificielle KEMAR par Bill Gardner et Keith Martin au MIT([GM94])
- LISTEN [LIS03]
- CROSSMOD [CRO08]
- CIPIC [ADTA01]

La procédure usuelle de mesure d'HRTF consiste à enregistrer un signal large bande en chambre anéchoïque à l'aide d'un microphone miniature placé à l'entrée du canal auriculaire. Le recours à de tels microphones se justifie par la nécessité d'avoir un système de mesure le plus transparent possible, qui limite les effets de réflexion et de résonance non désirés. L'utilisation et l'avantage de ces microphones pour ce genre de mesure sont présentés dans [Car96]. L'emploi d'une chambre anéchoïque permet de s'affranchir de l'empreinte acoustique d'une salle dans la mesure réalisée.
En ce qui concerne le placement des microphones, les travaux de Møller [Møl92] montrent que des mesures

d'HRTF convenables pour la restitution binaurale d'espaces acoustiques virtuels peuvent s'obtenir en plaçant le microphone miniature à l'entrée du canal auriculaire que l'on aurait préalablement obstrué par un bouchon d'oreille. L'idée est ainsi de mesurer la contribution du torse, du pavillon et de la tête en s'affranchissant des réflexions causées par le canal auriculaire, qui sont considérées comme indépendantes de l'incidence de l'onde acoustique.
Les HRTF disponibles sont au départ consignées sous forme de réponses impulsionnelles (*Head Related Impulse Response*) échantillonnées à 44.1 kHz de longueur 512 échantillons. La méthode utilisée pour ces mesures est celle de [Far00].
La restitution d'espaces auditifs virtuels nécessite qu'un événement sonore puisse être généré à n'importe quel endroit de l'espace. Les HRTF ne pouvant se mesurer que de manière ponctuelle, il est alors nécessaire d'interpoler les mesures afin d'obtenir des fonctions de transfert pour chaque position de l'espace. Différentes méthodes d'interpolation sont présentées dans [K. 99], [T. 08] et [FK08]. Ce procédé induit une erreur, proportionnelle à l'écart de distance entre deux mesures voisines. Ainsi plus le nombre de mesures est grand, plus l'erreur est minimisée. Il y a donc un compromis à faire entre le temps requis pour réaliser l'ensemble des mesures et l'erreur induite par interpolation. On estime en fait qu'un espacement de 6° dans le plan horizontal est suffisant pour pouvoir reconstruire sans erreur audible les directions intermédiaires [LB00], ce qui est le cas des bases de données CIPIC (1250 mesures par sujet), KEMAR et CROSSMOD (650 mesures par sujet).

1.3.2 Egalisation

L'égalisation est une étape essentielle pour éliminer les contributions de la chaîne de mesure (principalement amplificateur et microphone). Plusieurs techniques ont été élaborées :

- **Egalisation par rapport au champ libre** : l'égalisation se fait par rapport à une incidence particulière, la plupart du temps celle correspondant à l'incidence frontale.
- **Egalisation par rapport au champ diffus** : l'égalisation se fait par rapport à la moyenne énergétique des HRTF mesurées sur toutes les incidences. Cette technique a l'avantage de ne favoriser aucune incidence particulière et n'altère pas trop la couleur du signal. Gardner dans [Gar97] a opté pour cette technique ;
- **Egalisation découplée** : cette méthode présentée par Blauert ([Bla74]) et reprise par Larcher dans [Lar01] permet de séparer la contribution due au système de mesure. Elle est basé sur un champ de référence, qui peut être soit un champ libre, soit le champ diffus. Le principe est de "pré-déconvoluer" le signal par la fonction de transfert d'un casque en particulier afin de s'affranchir de l'effet du casque. Le signal résultant est alors indépendant du système de restitution et peut être utilisé pour une restitution à l'aide d'enceintes, ce qui est le cas pour un système transaural. Dans la suite on utilisera, l'égalisation au champ diffus.

1.3.3 Symétrisation

L'opération de symétrisation consiste à rendre identiques les HRTF de l'oreille gauche et de l'oreille droite, à la seule différence que leurs valeurs sont opposées symétriquement par rapport au plan vertical défini par l'origine de l'espace et le vecteur d'azimut 0°.
Cette opération s'avère utile lorsque l'on vise une écoute binaurale non-individualisée. En effet, les dissymétries qui peuvent exister dans les HRTF d'un sujet ou d'un mannequin sont dues à des caractéristiques très propres à sa physionomie. La symétrisation permet de corriger ces dissymétries, qui n'ont pas lieu d'être si la synthèse transaurale est par exemple destinée à plusieurs auditeurs. D'un point de vue mathématique, elle s'opère en moyennant d'une part le spectre de magnitude des HRTF gauche et droite respectivement mesurées pour deux incidences symétriques et, d'autre part, en moyennant les retard interauraux relevés pour ces mêmes directions.

1.3.4 Filtre à réponse impulsionnelle finie(FIR)

La mesure des HRTF est obtenue à partir des HRIR (*Head Related Impulse Response*) mesurées. Soit H une HRTF correspondant à une incidence donnée, sa transformée en Z s'exprime de la manière suivante :

$$H(z) = \sum_{i=0}^{N-1} h_i z^{-i} \tag{2}$$

avec N le nombre d'échantillons des HRIR, soit 512 dans notre cas, et h_i avec $i \in [0..N-1]$ les échantillons de la réponse impulsionnelle mesurée. Le filtrage binaural peut se faire à partir de HRTF sous forme de HRIR. Cependant le temps de calcul pour ce genre de filtrage par convolution est important. C'est pourquoi on peut utiliser ces HRTF sous une autre forme plus efficace d'un point de vue temps de calcul.

1.3.5 Filtre à réponse impulsionnelle infinie(RII)

Il est possible aussi de modéliser les filtres HRTF en filtre RII permettant d'améliorer le temps de calcul, cette méthode est décrite par Larcher dans [Lar01]. Oppenheim et Schafer ([OSB99]) ont montré que tout système déterminé, tel que H dont on connaît la réponse en fréquence et en phase peut se décomposer en deux composantes. L'une à phase minimale, que nous appellerons H_{min} et l'autre passe-tout à excès de phase, H_{exc}. Cette décomposition est décrite à l'équation (3).

$$H(z) = H_{min}(z)H_{exc}(z) \tag{3}$$

avec

$$H_{min}(z) = H(z)e^{j\phi_{min}}$$
$$H_{exc}(z) = e^{-j\tau}$$

ϕ_{min} est la phase minimale, proche de 0 pour toutes les fréquences, et τ le retard monaural. Ce retard est arbitraire. Il convient de le fixer de manière à ce que la différence entre les deux oreilles correspondent au retard interaural estimé. Le filtre à phase minimal peut se modéliser sous la forme d'un filtre RII. Sa transformée en Z s'exprime alors de la manière suivante

$$H_{min}(z) = \frac{B(z)}{A(z)}$$
$$A(z) = \sum_{i=0}^{M} a_i z^{-i}, B(z) = \sum_{i=0}^{L} b_i z^{-i} \tag{4}$$

Les ordres M et L de chacun des polynômes A et B sont paramétrables.
Larcher dans [Lar01] a comparé plusieurs méthodes de modélisation et d'implémantation de filtres RII, elle s'appuie sur [J. 99b] et [J. 99a]. Finalement la méthode retenue est celle de Steiglitz et l'implémentation choisie est une structure en cellule d'ordre 2 mis en cascade. La stabilité du filtre modélisé est garantie en contrôlant la stabilité de chacune des cellules d'ordre 2 indépendamment. Sur la figure 2 on peut comparer le filtre HRTF original et sa modélisation en cellule d'ordre 2.

On peut remarquer quelques effets de cette modélisation :

- erreur en basses fréquences en-dessous de 1 kHz de l'ordre de quelques dB
- erreur en hautes fréquences à partir de 16 kHz

Malgré ces effets la modélisation en cascade de cellule d'ordre 2 reste très bonne. En effet, l'erreur de modélisation en basses fréquences n'est pas primordiale car sur cette plage de fréquence l'ITD est l'indice de localisation prépondérant. Or cette erreur n'agit que sur l'ILD. De plus, l'erreur en hautes fréquences n'apparaît qu'au dessus de 16 kHz, à partir de ces fréquences l'oreille humaine est beaucoup moins sensible. La modélisation sur la plage utile 2 kHz à 16 kHz est bonne et le gain d'un point de vue temps de calcul est très intéressant.

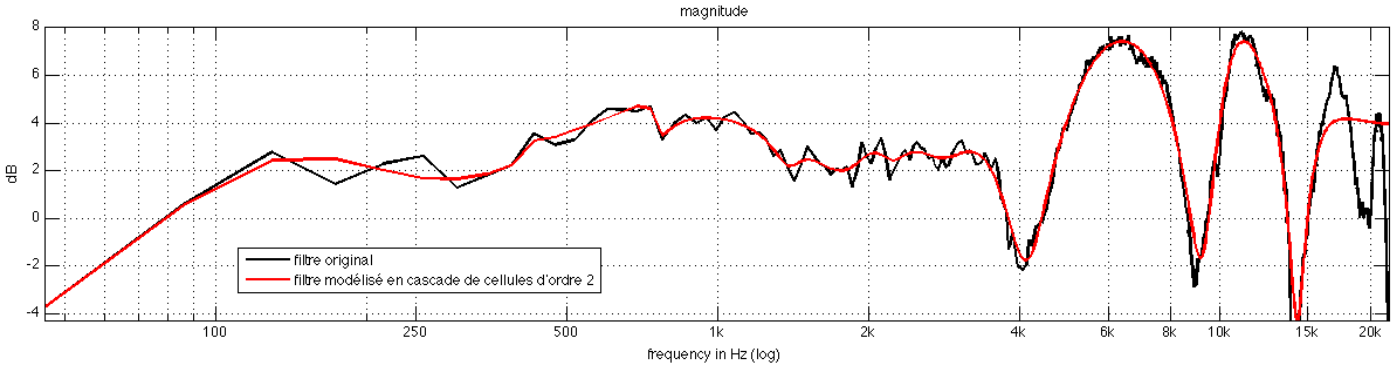


Fig. 2 – Filtre HRTF original et sa modélisation en cellules d'ordre 2 en cascade d'ordre 24.

1.4 Implantation bicanale de la synthèse binaurale

Pour réaliser une synthèse binaurale, le signal d'entrée monophonique est retardé, par un retard monaural puis filtré par le filtre à phase minimale modélisé en cellules d'ordre 2 en cascade. On parle d'implantation bicanale car le signal monophonique x est réparti sur deux filtrages spécifiques à chacune des oreilles pour obtenir le signal du canal gauche y_L et celui du canal droit y_R.
La forme générale d'une synthèse est la suivante :

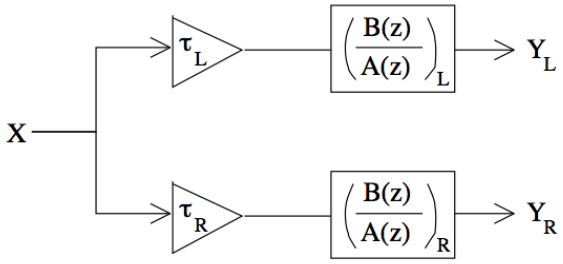


Fig. 3 – Schéma général de la synthèse binaurale bicanale. Figure tiré de [Lar01].

2 Synthèse transaurale

La partie précédente présentait une méthode de spatialisation sonore destinée à l'écoute sur casque audio. La synthèse transaurale permet d'adapter cette technique binaurale à la restitution sur deux enceintes. L'intérêt est de préserver la fidélité de restitution spatiale offerte par la technique binaurale tout en s'affranchissant du port d'un casque ce qui peut favoriser l'impression d'immersion du sujet. Cependant la diffusion des signaux binauraux sur deux enceintes ne permet plus de préserver la nécessaire confidentialité des signaux gauche et droit vers les oreilles respectives. La technique transaurale requiert un traitement particulier permettant de restaurer cette confidentialité par l'insertion d'un décodeur visant à annuler les trajets croisés c'est à dire les informations délivrées par le haut-parleur droit vers l'oreille gauche et réciproquement. Plus exactement, il s'agit d'inverser la matrice de fonctions de transfert bi-canal entre les deux canaux de diffusion et les deux oreilles. Dans cette partie, nous nous attacherons à décrire le principe de ce décodeur, dénommé crosstalk canceller. Nous étudierons deux types d'inversion du déterminant de cette matrice de transfert : l'inversion par transformée de Fourier et l'inversion par la méthode des moindres carrés. Différentes architectures de traitement existent dans la littérature. Certaines d'entre-elles exploitent la symétrie usuelle d'un système de diffusion stéréophonique pour un auditeur centré. Cependant dans notre cas, l'auditeur sera mobile, par conséquent nous restons dans le cadre général des systèmes transauraux asymétriques. Nous décrirons l'architecture general assymetric feedforward proposée par Gardner dans [Gar97] permettant de réaliser le crosstalk canceller à l'aide de filtres FIR. Nous adapterons ensuite cette architecture au type de filtre RII que nous avons choisi pour la suite.

2.1 Principe du crosstalk canceller

Le principe du crosstalk canceller est de reproduire le signal binaural aux oreilles de l'auditeur à l'aide d'enceintes. Pour cela, on effectue dans un premier temps une synthèse binaurale du signal d'entrée. Sans crosstalk canceller on ne peut pas directement diffuser un signal binaural sur des enceintes. L'effet produit ne serait pas le même que celui au casque. Des contributions parasites apparaîtraient, car une partie du signal destinée à l'oreille gauche arriverait à l'oreille droite et une partie du signal destiné à l'oreille droite arriverait à l'oreille gauche. Le but du crosstalk canceller est de filtrer et mélanger le signal binaural afin d'éliminer ces contributions parasites aux oreilles de l'auditeur. Le principe de ce traitement est décrit dans la suite de cette partie. Le formalisme utilisé dans cette partie est celui de Gardner [Gar97]. Le signal binaural est obtenu par convolution du signal d'entrée à l'aide d'une paire de HRTF :

$$\mathbf{x} = \mathbf{h}s \tag{5}$$

$$\mathbf{x} = \begin{bmatrix} x_L \\ x_R \end{bmatrix}, \mathbf{h} = \begin{bmatrix} H_L \\ H_R \end{bmatrix}$$

avec s le signal d'entrée monophonique, $\mathbf{x}$ le vecteur colonne des signaux binauraux et $\mathbf{h}$ l vecteur colonne des filtres HRTF. Le signal binaural est ensuite filtré pour produire les signaux des haut-parleurs.

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{6}$$

$$\mathbf{y} = \begin{bmatrix} y_L \\ y_R \end{bmatrix}, \mathbf{C} = \begin{bmatrix} C_{11}\ C_{12} \\ C_{21}\ C_{22} \end{bmatrix}$$

$\mathbf{y}$ est le vecteur de signaux des haut-parleurs. $\mathbf{C}$ est une matrice de filtres d'annulation de chemins croisés. $\mathbf{x}$ et $\mathbf{y}$ sont l'entrée et la sortie de notre système. Ensuite le signal $\mathbf{y}$ est transmis par les haut-parleurs jusqu'aux oreilles de l'auditeur. En posant $\mathbf{e}$ le vecteur signal présent à l'entrée des canaux auditifs de l'auditeur et $\mathbf{H}$ la matrice des fonctions de transfert entre les haut-parleurs et les oreilles de l'auditeur, la propagation dans l'air peut s'exprimer de la manière suivante :

$$\mathbf{e} = \mathbf{H}\mathbf{y} \tag{7}$$

$$\mathbf{e} = \begin{bmatrix} e_L \\ e_R \end{bmatrix}, \mathbf{H} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix}$$

Le signal **e** peut être vu comme un signal mesuré par des microphones insérés à l'entrée des conduits auriculaires de l'auditeur comme ceux utilisés pour les mesures de HRTF. Le but du crosstalk canceller est donc de choisir **C** de telle sorte que les vecteurs **e** et **x** soit identiques, dans le cas idéal. On en tire donc que :

$$\mathbf{e} = \mathbf{H}\ \mathbf{C}\ \mathbf{x} \tag{8}$$

D'où la condition, dans le cas d'un crosstalk canceller idéal :

$$\mathbf{C} = \mathbf{H}^{-1} \tag{9}$$

$$\begin{gathered} \mathbf{H}^{-1} = \begin{bmatrix} H_{RR} & -H_{RL} \\ -H_{LR} & H_{LL} \end{bmatrix} \frac{1}{D} \\ D = H_{LL}.H_{RR} - H_{LR}.H_{RL} \end{gathered} \tag{10}$$

D est le déterminant de la *head transfer matrix*. Pour notre système ce déterminant est calculé à partir des FIR des HRTF. Puis il est modélisé en filtre de cellules d'ordre 2 en cascade, par la même méthode que celle des filtres binauraux (voir 1.3.5).

2.2 Inversion du déterminant

L'inversion du déterminant est un point critique du décodeur transaural, car l'inversion d'un filtre peut donner un filtre instable, si les méthodes d'inversion sont trop imprécises. Ceci implique donc de recourir à des méthodes permettant de trouver des approximations du filtre inverse avec un minimum d'erreur. L'inversion d'un filtre revient à résoudre un système d'équations. Comme tout système d'équations il est possible d'étudier son conditionnement pour évaluer les risques liés à l'inversion. Dans un premier temps nous définirons donc le conditionnement de notre problème d'inversion. Ensuite, nous présenterons deux méthodes d'inversion du déterminant différentes.

2.2.1 Conditionnement

Il existe un critère permettant d'évaluer si un système est facilement inversible ou non, il s'agit du conditionnement. On dit d'un système linéaire $b = Ax$ qu'il est bien conditionné si son conditionnement est faible. A l'inverse, on dira de ce système qu'il est mal conditionné si son conditionnement est élevé. Le conditionnement peut également s'interpréter comme une mesure de la robustesse d'un système face à une erreur lui étant introduite en entrée. Si ce système est bien conditionné, une faible erreur en b résultera en une faible erreur en x. En revanche, si le système est mal conditionné, une faible erreur en b conduira à une forte erreur en x. Ainsi, le conditionnement est un indice mathématique permettant d'évaluer la précision du résultat en sortie du système testé.

Le nombre de conditionnement est défini par le ratio entre la norme de l'erreur relative en Δx et la norme de l'erreur relative en Δb.

$$\frac{\|A^{-1}\Delta b\|/\|A^{-1}b\|}{\|\Delta b\|/\|b\|} \tag{11}$$

On peut simplifier cette expression par

$$\kappa(A) = \|A^{-1}\| \cdot \|A\| \tag{12}$$

Le choix de la norme reste ouvert. La norme la plus communément utilisée est la norme L_2, qui mène alors à un conditionnement égal au ratio entre la plus grande et la plus petite valeur singulière.

$$\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \tag{13}$$

Plusieurs études se sont penchées sur le conditionnement de systèmes visant à opérer une annulation des trajets croisés [NR05] [MRK99] [NHE92] [RM06]. Elles montrent que le conditionnement dépend à la fois de la fréquence et de l'angle entre les hauts parleurs (ou span) [**?**]. Elles font également référence au phénomène des "ringing frequencies" : elles correspondent aux fréquences mal reconstruites par le processus d'inversion, et sont susceptibles d'être nettement audibles dans la restitution des signaux transauraux. Kirkeby et Nelson montrent que la configuration optimale en regard du critère d'inversion s'obtient avec un écart angulaire de 10° entre les haut-parleurs (disposés frontalement et de manière symétrique par rapport à l'auditeur). Ils déposent un brevet pour ce système, qu'ils nomment "stereo dipole" [KN98]. Dans [TN00], Takeuchi et Nelson proposent un système idéal constitué d'une distribution continue de haut-parleurs autour de l'auditeur, permettant de minimiser le conditionnement du système à chaque fréquence. Une solution pratique suggérée par les auteurs consiste en un système 3 voies constitué de 3 paires de hauts parleurs à +/- 90° pour les fréquences graves, +/- 16° pour les fréquences médiums, et +/- 3,1° pour les fréquences aigües. Le conditionnement du déterminant de la matrice de transfert est ainsi optimisé pour 3 bandes de fréquences et résulte en une minimisation de l'erreur due à l'opération d'inversion.

2.2.2 Inversion par transformée de Fourier

Gardner dans [Gar97] utilise cette méthode pour inverser le déterminant. Comme son nom l'indique cette inversion se fait dans le domaine fréquentiel. Tout d'abord la transformée de Fourier du déterminant est calculée par FFT. Ensuite, on calcule directement l'inverse de cette réponse en fréquence. Cependant afin de garantir une certaine stabilité de cet inverse, Gardner limite l'amplitude maximale de la réponse en fréquence avant de calculer l'inverse. Il prévient ainsi des effets de résonance, en particulier pour les hautes fréquences.
Pour que l'inversion du déterminant soit complète, il faut également inverser sa réponse en phase. Cette phase est décomposée en deux phases, la phase minimale et l'excès de phase ([OSB99]). La phase minimale peut être négligée car quasi nulle. Il suffit donc de prendre l'opposé de l'excès de phase pour obtenir celle du déterminant inverse.
Cette méthode proposée par Gardner a l'inconvénient d'altérer la réponse en fréquence du déterminant. Les effets sont difficilement prévisibles. Une autre méthode existe l'inversion par la méthode des moindre carrés.

2.2.3 Inversion par la méthode des moindres carrés

Cette méthode est généralement utilisé pour résoudre des systèmes surdéterminés. Soit un filtre h et son inverse g, on cherche g tel que,

$$h * g = \delta \tag{14}$$

où δ représente le symbole de kronecker, $\delta[n] = 0$ lorsque $n \neq 0$ sinon $\delta[n] = 1$. Le filtre g est solution lorsqu'il minimise l'erreur au sens de la norme $\mathcal{L}_2$, en comparant la solution trouvé à une solution idéale.

2.3 Implémentation du crosstalk canceller

2.3.1 Architecture general feedforward

La structure générale du crosstalk canceller est celle de Gardner ([Gar97]) *general asymmetric feedforward*. Cette structure est présentée dans la figure 4 et correspond à celle que nous avons adoptée pour la suite. Il existe d'autres structures symétriques permettant de factoriser certains filtrages. Ces structures, certes plus

économes en terme de coût de calcul, doivent être utilisées dans des conditions strictes où l'auditeur doit rester centré entre les deux haut-parleurs en gardant sa tenté orientée vers le milieu des deux haut-parleurs. Ces conditions d'utilisantes sont bien trop restrictives dans le cadre d'un système de réalité virtuelle. D'où notre choix vers cette structure *general asymetric feedforward* qui est la plus générale. Pour plus de précision sur le cadre d'utilisation de cette structure voir 3.

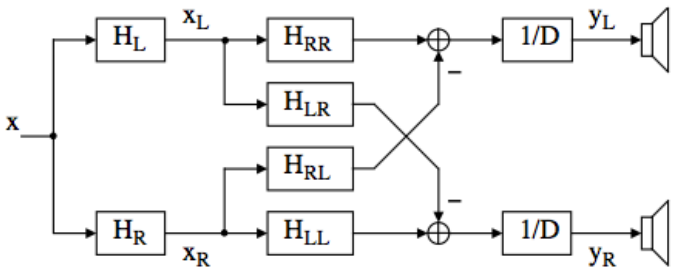


Fig. 4 – Schéma général de la structure*general asymmetric feedforward* de Gardner pour une seule source binaurale. Figure tiré de [Gar97].

2.3.2 Architecture general feedforward avec des filtres RII

Nous avons opté pour une modélisation des filtres transauraux à l'aide de filtres RII sous forme de cellules d'ordre 2 en cascade. Sur la figure 5 on peut voir la structure du crosstalk canceller pour des filtres RII. Cette architecture reprend celle de Gardner à la différence près qu'il faut intégrer des retards τ_A, τ_B, τ_C, τ_D dans chaque branche du crosstalk canceller. Il est nécessaire de bien ajuster ces retards pour reproduire l'ITD aux oreilles de l'auditeur. Pour cela on peut montrer que la condition 15 est suffisante.

$$\tau_D - \tau_B - (\tau_C - \tau_A) = ITD \tag{15}$$

Avec $ITD = \tau_L - \tau_R$. τ_L et τ_R sont définis comme étant respectivement le retard monaural de l'oreille gauche et celui de l'oreille droite. Pour cela il suffit que, $\tau_D = \frac{\tau_L}{2}$, $\tau_B = -\frac{\tau_L}{2}$, $\tau_C = \frac{\tau_R}{2}$, $\tau_A = -\frac{\tau_R}{2}$

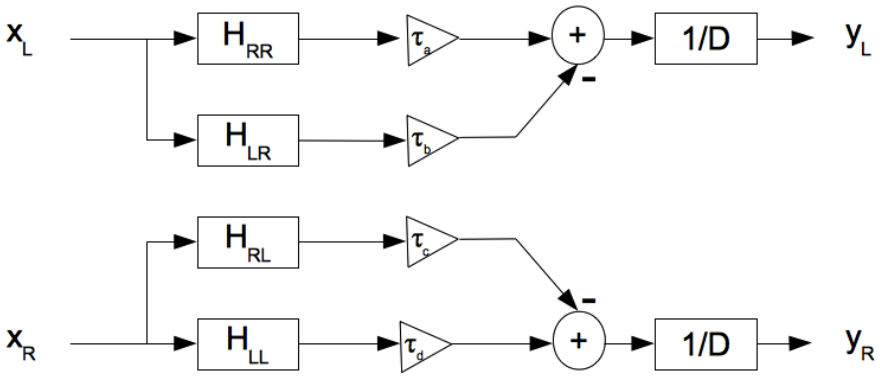


Fig. 5 – Schéma général de la structure *general asymmetric feedforward pour des filtres RII* de Gardner pour une seule source binaurale avec des filtres RII modélisés par un cascade de cellules d'ordre 2.

Finalement nous avons vu dans cette partie un système de spatialisation sonore permettant de s'affranchir de l'utilisation du casque. Dans la suite nous reprenons l'architecture general feedforward avec des filtres RII

que nous venons de présenter pour l'adapter à un système transaural adaptatif en temps réel.

3 Système transaural adaptatif

Nous venons de voir le fonctionnement d'un système transaural statique permettant de spatialiser des sources sonores sur un couple de haut-parleurs. Ce système n'est cependant valable que pour une seule position et orientation précise de l'auditeur. Si l'auditeur se déplace ou modifie l'orientation de sa tête, le traitement précédent n'est plus valable. Les relations entre les signaux gauche et droit ne préservent plus les indices interauraux contenus dans le couple de signaux binauraux et l'image spatiale est corrompue. L'idée du système transaural adaptatif est de relever la position de l'auditeur et l'orientation de sa tête en temps réel afin d'adapter le traitement transaural à sa position courante. Ainsi l'auditeur pourra utiliser librement les mouvements de tête pour confirmer une localisation sonore. Cette liberté supplémentaire a pour effet de renforcer l'impression de spatialisation sonore et de présence des sources qui composent la scène sonore puisque la position des sources dans l'espace semblera indépendante des mouvements de l'auditeur. Dans la suite de cette partie, nous allons étudier les aspects liés à cette fonctionnalité supplémentaire. Nous verrons tout d'abord les aspects dynamiques, tels que la résolution spatiale, la réalisation d'une base de donnée de filtres HRTF et de filtres transauraux, l'interpolation de filtres. Ensuite nous décrirons l'architecture et le fonctionnement du système transaural adaptatif qui a été développé durant le stage.

3.1 Principe général et cadre d'étude

La synthèse transaurale que nous avons décrite précédemment n'est valable qu'en un point unique de l'espace et pour une orientation fixe de l'auditeur. Le but de notre système transaural adpatatif en temps réel est de permettre à l'auditeur une totale liberté de mouvement au sein d'une scène virtuelle. Cette liberté a un coût et plusieurs aspects rentrent alors en considération. Nous balayerons ici les nouveaux aspects à prendre en compte.

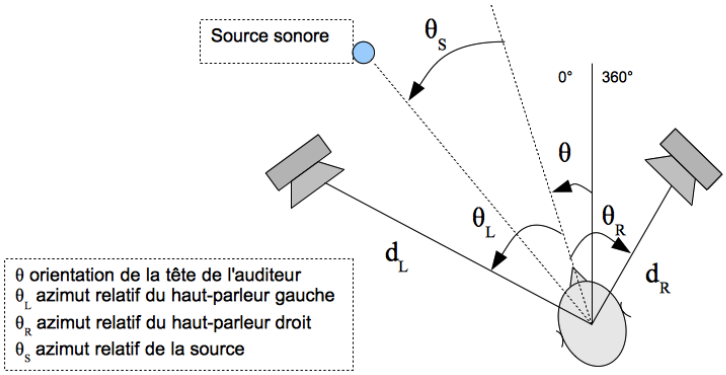


Fig. 6 – Figure d'un auditeur dans une configuration asymétrique d'un système transaural.

La figure 3.1 montre une configuration asymétrique de synthèse transaurale où l'auditeur est décentré. La position relative de la source ainsi les filtres H_L et H_R de l'équation (5) doivent être actualisés à chaque mouvement de l'auditeur. On peut noter que ce problème est équivalent à celui d'une source en mouvement par rapport à un auditeur fixe, le mécanisme de l'adaptation du filtrage est déjà traité dans le cas du binaural, seul le calcul de la position relative évolue.

Le coût de cette liberté de mouvement ne s'arrête pas là. Les positions relatives des haut-parleurs changent également. Ceci affecte le choix des filtres transauraux et donc la *head transfer matrix* $\mathbf{H}$ de l'équation7. En effet, ces filtres dépendent des azimuts relatifs θ_L et θ_R des haut-parleurs gauche et droit (voir figure 3.1). De plus, puisque la matrice $\mathbf{H}$ évolue avec les mouvements de l'auditeur, son inverse $\mathbf{H}^{-1}$ doit évoluer avec évidemment ; si l'on veut que la synthèse transaurale soit valable. Le coût de calcul lié à une inversion de filtre par transformée de Fourier ou la méthode des moindres carrés (2.2) rend cette inversion lourde à réaliser en

temps réel. C'est pourquoi, nous avons opté pour la réalisation d'une base données de filtres transauraux. Cette base de données de filtres nous économise ce coût de calcul. La description de cette base sera vue plus loin dans 3.3.1.

Il convient de préciser que pour notre étude nous nous concentrons sur le cas de sources se situant en champ lointain, car les HRTF de l'IRCAM ont été mesurées en champ lointain. De plus, nous ne traitons pas la compensation de la directivité des enceintes faute de temps. Seule une compensation en distance est réalisée par l'ajout d'un retard et d'un gain de compensation.

Un système transaural composé de deux enceintes n'est performant que pour un espace restreint, compris entre les deux enceintes. Afin d'élargir l'effet de notre système transaural à 360°, nous avons choisi d'utiliser quatre enceintes encerclant l'espace d'excursion de l'auditeur. Ce choix est principalement dû à un compromis entre la qualité de la restitution, l'encombrement et le coût de calcul. Nous disposons de quatre haut-parleurs pour réaliser un système transaural adaptatif. Le choix de la paire de haut-parleurs pour réaliser la synthèse transaural est un point crucial dans la conception de notre système. Il faut distinguer deux cas, le cas d'une source sonore unique et le cas d'une scène complexe avec plusieurs sources sonores. Dans le premier cas, le choix de la paire de haut-parleurs peut être guidé par la position de la source, en choisissant la paire entourant la source sonore, dans ce cas le décodeur transaural ajouterait un peu plus de réalisme à la restitution de cette source sonore. Mais dans le cas de plusieurs sources sonores, qui est le cas le plus général, cette technique nous obligerait à exécuter un décodeur binaural et un décodeur transaural par source sonore. Mais pour limiter la charge de calcul, nous adopterons une autre stratégie pour le choix de la paire de haut-parleurs. Ce choix sera guidé par l'orientation de la tête de l'auditeur. La paire de haut-parleurs active devra tout le temps être de part et d'autre du plan médian de l'auditeur.

Ce critère de sélection n'est pas suffisant, car nous avons encore le choix entre deux spans différents 90° et 180°. Il nous faut donc définir un critère objectif pour choisir entre les deux. Quel span doit-on choisir selon l'orientation de la tête ? Peut-on utiliser les deux ? Si oui cela a-t-il un effet négatif sur la spatialisation ou sur la qualité de la restitution du son ? Pour répondre à ces questions, nous verrons dans la partie qui suit une caractérisation objective concernant ces différentes interrogations.

3.2 Caractérisation objective

L'élaboration d'un système transaural adaptatif demande une étude approfondie des nombreux paramètres disponibles. Cette étude est nécessaire pour choisir une implémentation répondant aux exigences du système. Cette partie rend compte des choix que nous avons retenus. Nous présenterons dans un premier temps le cadre de l'étude basée sur une simulation Matlab. Ensuite nous verrons en quoi le critère de conditionnement peut fournir un critère objectif dans le choix de secteurs angulaires. Ces secteurs angulaires provoquent des changements brusques de coloration du son restitué. Nous verrons deux techniques permettant d'atténuer ces changements de colorations, le fondu enchaîné et la compression de la dynamique du déterminant.

3.2.1 Environnement de simulation

Dans cette partie nous nous attacherons à étudier les effets de chaque traitement que nous utilisons. Pour cela, une simulation à l'aide de Matlab a été développée en reprenant les travaux de l'équipe. Dans [Cor11] une chaîne de traitement Matlab simulant un système de synthèse transaural a été développé. Ce travail a servi de base pour nos simulations. Cette simulation se découpe en plusieurs parties :

Définition des paramètres Les paramètres disponibles sont nombreux :

- positions des haut-parleurs
- type de filtre utilisé FIR ou IIR
- taille des filtres FIR
- ordre des filtes IIR

- définition des secteurs angulaires
- taille de la zone de fondu enchaîné entre deux secteurs
- type d'inversion à utiliser pour inverser le déterminant : inversion par transformée de Fourier discrète ou inversion par méthode des moindres carrés
- traitement sur le déterminant : seuillage, compression de la dynamique, warping en fréquence

Préparation des filtres transauraux Les jeux de filtres transauraux sont pré-traités en amont de la chaîne de traitement.

Etage binaural Le signal d'entrée est un Dirac. Ce signal est filtré par une paire de filtres HRTF correspondant à la position de la source virtuelle.

Etage transaural Les signaux binauraux du canal gauche et droit sont ensuite traités dans la chaîne de traitement transaural.

Canal acoustique La propagation des signaux des enceintes au haut-parleurs est alors simulé par un filtrage HRTF en fonction de la position des haut-parleurs.

Evaluation des résultats Les signaux obtenus sont comparés à nos références. On évalue les différences spectrales ainsi que l'ITD reproduite.

La situation reproduite par notre simulation Matlab est la suivante : Un auditeur se trouve au centre des haut-parleurs et la source sonore se trouve à l'incidence 0°. Nous faisons varier l'orientation de la tête, θ de l'auditeur de 0°à 359°. La figure 7 montre la situation simulée. Le signal d'entrée étant un dirac, nous sommes censés retrouvés les HRTF de l'auditeur. Afin de visualiser, les effets de chaque paramètre, nous simulons un

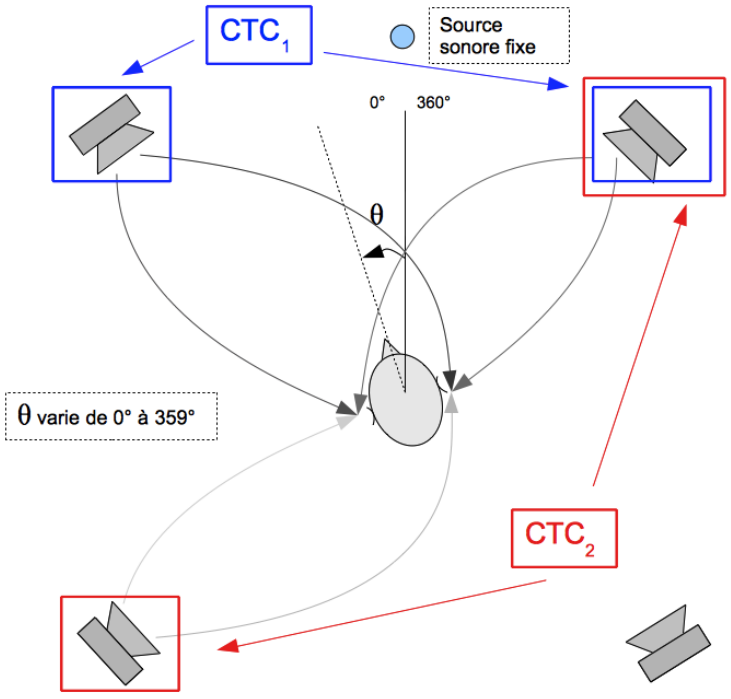
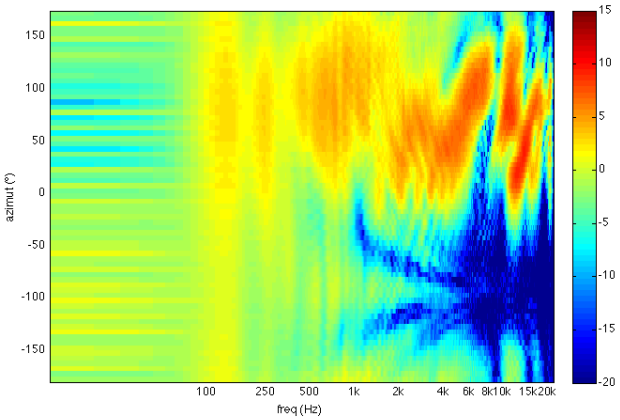


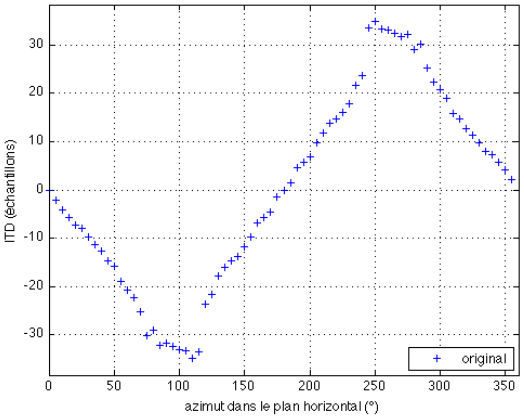
Fig. 7 – Situation reproduite par la simulation Matlab d'une source fixe avec un auditeur au centre tournant sa tête de 0° à 359°. Dans le cas où il y a deux décodeurs transauraux qui fonctionnent en même temps.

auditeur se plaçant au centre de notre dispositif. Nous simulons une rotation de la tête à 360° par pas de 5° tout en maintenant sa position centrée, en calculant pour chaque position le signal arrivant à ses oreilles. Le signal d'entrée de notre simulation est directement la réponse impulsionnelle d'un auditeur mesurée à l'IRCAM.

La référence que nous essayons d'atteindre est donc directement la HRTF de l'auditeur considéré et l'ITD estimé sur les mesures des HRTF originales. La figure 8 montre les données de références que nous essaierons d'atteindre.



(a) Amplitudes des filtres HRTF



(b) ITD en nombre d'échantillons évalué pour le sujet 1087 pour tous les azimuts de 0° à 359°

Fig. 8 – Amplitude des HRTF mesurées pour le sujet 1087 et ITD évalué pour le sujet 1087 à partir des mesures des HRTF originales.

Dans la suite, les erreurs de reconstruction présentées seront calculées par la valeur absolue de la différences des HRTF originales et reconstruites :

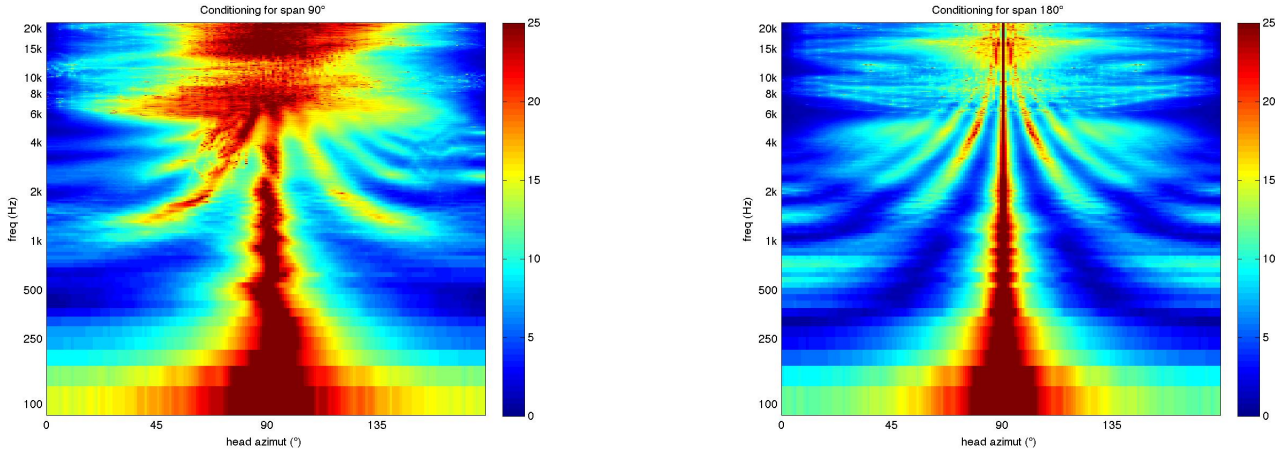
$$\varepsilon(f,\theta) = |20\log_{10}\frac{HRTF_{orig}(f,\theta)}{HRTF_{reconst}(f,\theta)}| \tag{16}$$

Cette erreur dépend également de l'élévation mais dans le cadre de notre étude nous nous limitons au cas où l'élévation est nulle.

3.2.2 Critère de conditionnement

Le conditionnement de l'inversion du déterminant agit directement sur la coloration du son du décodeur transaural. Nous savons également, que le conditionnement dépend à la fois de la fréquence et du span des haut-parleurs. Un conditionnement proche de 0 dB nous indique que l'inversion se fera sans trop de problème, c'est à dire que la restitution transaurale sera d'autant plus fidèle. Au contraire un conditionnement très élevé sur une plage de fréquences et un span donnés, nous indique que l'inversion risque fortement de détériorer la reconstruction de ces fréquences lors de la synthèse transaurale. Dans cette partie, nous comparerons le conditionnement pour les deux spans présents dans notre configuration, c'est à dire le span de 90° et celui de 180°. Le but de cette comparaison est de déterminer de manière objective le choix des haut-parleurs en fonction de la position de l'auditeur. Les données présentées dans cette partie ont été obtenues en faisant la moyenne des conditionnements pour 22 sujets différents.

Sur la figure 9, est représenté le conditionnement en fonction de l'azimut de la tête de l'auditeur d'une part pour un span de 90°, figure 9(a) et d'autre part pour un span de 180° 9(b). Sur ces figures, on peut remarquer que les problèmes de conditionnement ne se situent pas aux mêmes fréquences selon le span. Ces figures nous renseignent sur les effets de chacun de ces spans indépendamment l'un de l'autre. Ceci n'est pas suffisant pour déterminer une limite angulaire pour le passage d'un span à un autre. De plus, le conditionnement n'a aucun effet sur l'ITD reproduit. C'est pourquoi afin d'affiner notre analyse du conditionnement, nous nous concentrons sur la plage de fréquence comprise entre 4 kHz et 16 kHz.



(a) span de 90°

(b) span de 180°

Fig. 9 – Conditionnement moyen de la *head transfer matrix* sur 22 sujets différents. Les positions simulées correspondent à une tête au centre tournant de 0° à 180°. 0° correspond à la position ou la tête se trouve exactement entre les deux enceintes.

Les figures 10(a) représentent une moyenne du conditionnement sur la plage de fréquences comprises entre 4kHz et 8kHz, pour le span de 90° et pour le span de 180°. Ces deux courbes sont disposées de manière à superposer les deux spans actifs en même temps. La courbe rouge représente le conditionnement pour un span de 180° pour les deux haut-parleurs se situant à 135° et 315° d'azimut. La courbe bleue représente le conditionnement en dB pour un span de 90° pour les haut-praleurs se situant à 45° et 315° d'azimut. Sur ces deux figures, on retrouve le résultat intuitif qui est que le système transaural ne fonctionne pas lorsque l'azimut de la tête se retrouve en face d'un haut-parleur ou lorsque les deux haut-parleurs sont du même côté par rapport au plan médian de la tête.

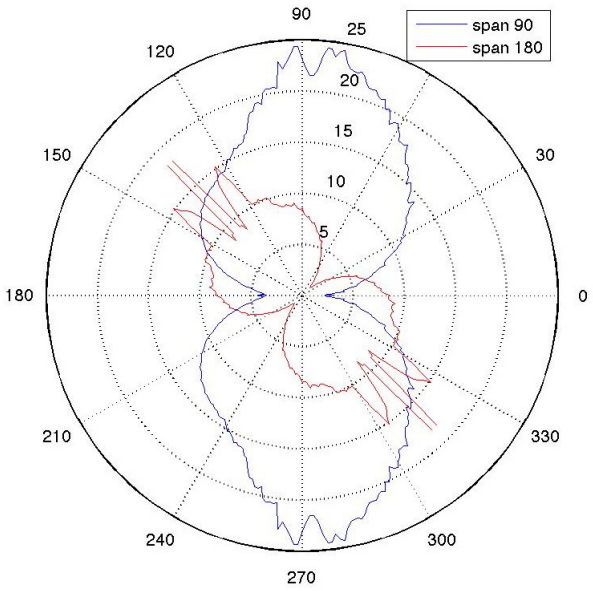
On observe que les courbes de conditionnement se coupent pour un azimut d'environ 22.5°, cet azimut est approximativement le même sur les deux plages de fréquences représentées. Si l'on se fie à ces courbes, on en déduit que :

- pour un azimut de la tête compris entre 0° et 22.5°, le conditionnement du span 90° est meilleur
- pour un azimut de la tête compris entre 22.5° et 45°, le conditionnement du span 180° est meilleur

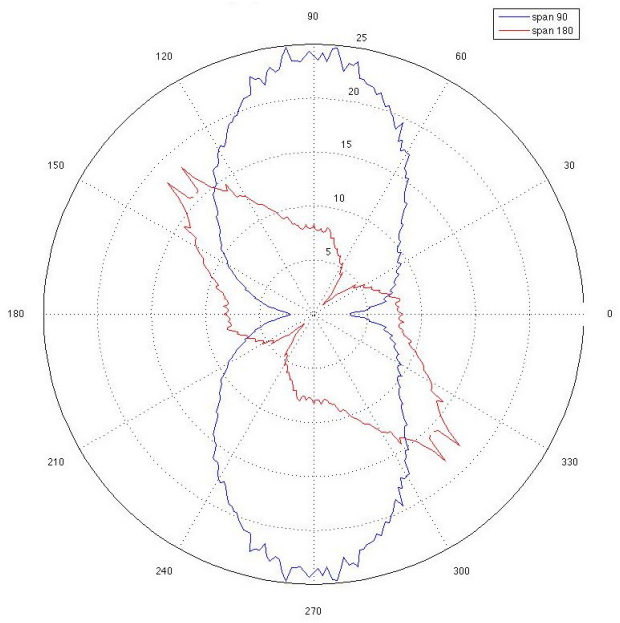
Il n'est pas nécessaire d'analyser la courbe autre part que pour les azimuts compris entre 0° et 45° car la configuration en carré de notre dispositif de haut-parleur nous permet de déduire par symétrie les azimuts limites de passage d'un span à un autre. Selon le critère de conditionnement, on en déduit donc les limites entre chaque span. Le choix des haut-parleurs en fonction de l'azimut θ de la tête est :

- pour $\theta \in [0°, 22.5°] \bigcup]337.5°, 360°]$, la paire de haut-parleurs respectivement à 45° et 315° est activée
- pour $\theta \in]22.5°, 67.5°]$, la paire de haut-parleurs respectivement à 135° et 315° est activée
- pour $\theta \in]67.5°, 112.5°]$, la paire de haut-parleurs respectivement à 135° et 45° est activée
- pour $\theta \in]112.5°, 157.5°]$, la paire de haut-parleurs respectivement à 225° et 45° est activée
- pour $\theta \in]157.5°, 202.5°]$, la paire de haut-parleurs respectivement à 225° et 135° est activée
- pour $\theta \in]202.5°, 247.5°]$, la paire de haut-parleurs respectivement à 315° et 135° est activée
- pour $\theta \in]247.5°, 292.5°]$, la paire de haut-parleurs respectivement à 315° et 225° est activée
- pour $\theta \in]292.5°, 337.5°]$, la paire de haut-parleurs respectivement à 45° et 225° est activée

Dans la suite nous utiliserons ces secteurs angulaires pour effectuer nos simulations.



(a) bande de fréquences de 4000 à 8000 Hz



(b) bande de fréquences 8000 à 16000 Hz

Fig. 10 – Conditionnement en dB obtenu à partir de la moyenne sur 22 sujets pour les bandes de fréquence de 4000 à 8000 Hz et 8000 à 16000 Hz. La courbe rouge correspond au span de 180° correspondant

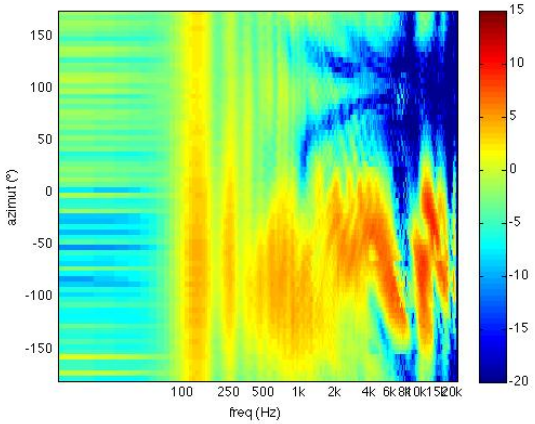
3.2.3 Effet du fondu enchaîné

Les secteurs angulaires définis précédemment nous assurent d'obtenir une reproduction transaurale satisfaisante pour chaque secteur. Cependant les effets de distorsion fréquentielle aux extrémités de chaque secteur sont spécifiques à chaque span. Ces distorsions créent une coloration du son dans la synthèse transaurale. Cette coloration est différente pour chacun des spans. De plus, ces effets de coloration seront d'autant plus perceptibles qu'ils vont dépendre de l'orientation de la tête du sujet. Il faut donc éviter qu'un petit mouvement de la tête dans la zone de transition des spans provoque un changement brusque de coloration. Lors de l'écoute nous avons pu constaté ces changements brusques de coloration. Afin de lisser ces effets, nous avons mis en place un mécanisme de fondu enchaîné pour passer d'un secteur à un autre. Initialement, les secteurs angulaires avaient une largeur de 45°. Dans cette configuration, il n'y avait aucun recouvrement entre les secteurs et donc aucun fondu enchaîné. Après plusieurs essais nous avons décidé d'élargir les secteurs de 45° pour qu'il y ait un recouvrement de 100% entre les zones avec naturellement une pondération.

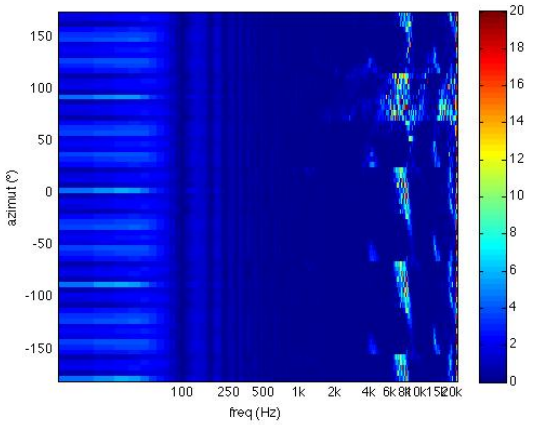
La figure 11 illustre l'effet de ce fondu enchaîné. Les figures 11(a) et 11(d) montrent les HRTF de l'oreille droite reconstruites après simulation Matlab. Il est difficile de noter une différence notable si l'on s'en tenait à ces figures. Par contre sur la figure 11(b) représentant l'erreur de reconstruction entre l'HRTF de l'oreille droite originale et celle reconstruite sans fondu enchaîné, on remarque une distorsion fréquentielle aux alentours des 8 kHz. Ces distorsions provoquent des changements brusques de coloration. Ces ruptures marquent le passage entre deux secteurs et traduisent le changement de coloration que nous avions remarqué lors de l'écoute.

Cette figure doit être comparée à la figure 11(e), représentant la même erreur de reconstruction mais cette fois-ci avec une zone de fondu enchaîné de 45° entre chaque secteur. On remarque sur cette figure que les brusques changements fréquentiels ont été lissés, mais il reste encore une partie de ces distorsions. L'écoute de cette configuration nous a confirmés cette tendance. Les changements de coloration n'ont pas totalement disparu, mais on ne distingue plus de limite flagrante au passage entre deux secteurs lors de l'écoute.

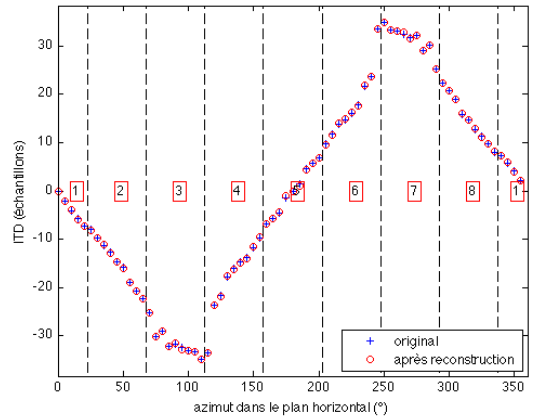
Les figures 11(c) et 11(f) montre que l'ITD reproduite est quasiment parfaite dans chacun des deux cas. L'impression de spatialisation est donc conservée quelque soit la configuration choisie.



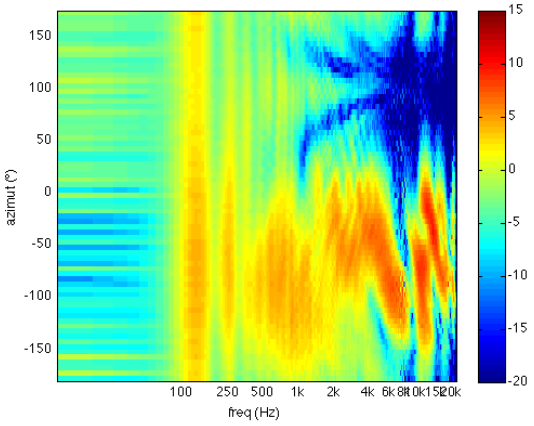
(a) HRTF de l'oreille droite reconstruite pour le sujet 1087 à partir de filtres FIR à 1024 coefficients sans zones de fondu enchaîné



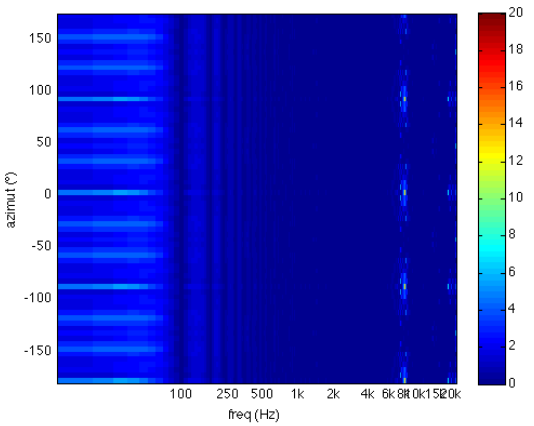
(b) Erreur de reconstruction



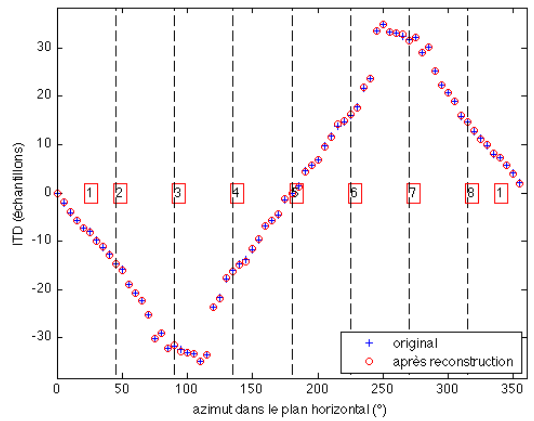
(c) Comparaison entre l'ITD évaluée sur les HRTF originales et l'ITD évaluée sur les HRTF reconstruites sans zone de fondu enchaîné



(d) HRTF de l'oreille droite reconstruite pour le sujet 1087 à partir de filtres FIR à 1024 coefficients avec des fondus enchaînés à toutes les positions



(e) Erreur de reconstruction

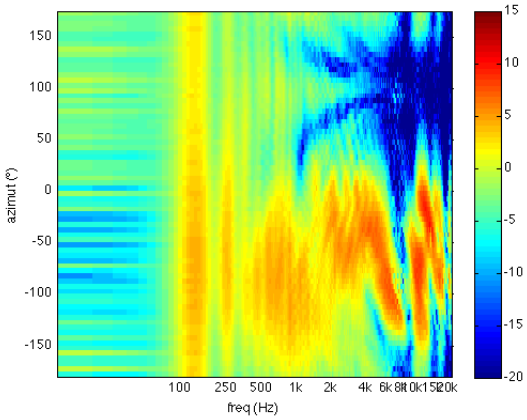


(f) Comparaison entre l'ITD évaluée sur les HRTF originales et l'ITD évaluée sur les HRTF reconstruites avec une zone de fondu enchainé de 45°

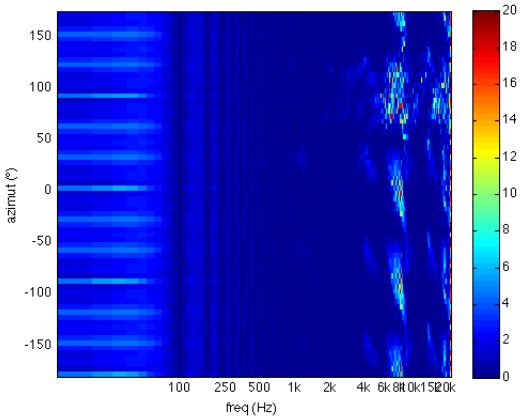
Fig. 11 – Comparaison d'amplitudes des HRTF de l'oreille droite du sujet 1087 reconstruites. Les figures de gauches ont été calculées sans aucun fondu enchaîné. Les figures de droite ont été calculées avec une zone de fondu enchaîné de 45°.

3.2.4 Effet de la compression de la dynamique du déterminant

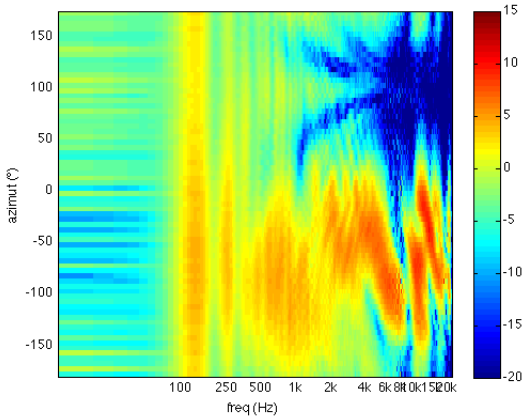
Cornuau dans [Cor11] a mis au point un traitement limitant la dynamique du déterminant dans le cas transaural statique. Cette compression n'est cependant pas linéaire, elle est appliquée de manière privilégiée sur les zones de "zéros" du spectre de ce déterminant, puisque ce sont elles qui sont à l'origine du mauvais conditionnement et des effets de coloration. La dynamique du déterminant est divisé par trois dans ces zones pour atténuer l'effet de coloration du système transaural. Nous avons adapté sa technique pour notre système. En comparant l'erreur de reconstruction en utilisant cette méthode, à celle présente à la figure 11(b) on constate un léger lissage de l'erreur de reconstruction. Mais par rapport à l'erreur présentée à la figure 11(e), l'effet est beaucoup moins important. Dans la suite, nous nous placerons donc dans la configuration de secteurs avec une zone de fondu enchaîné de 45°. Dans la suite, les nous n'utiliserons pas cette compression dynamique.



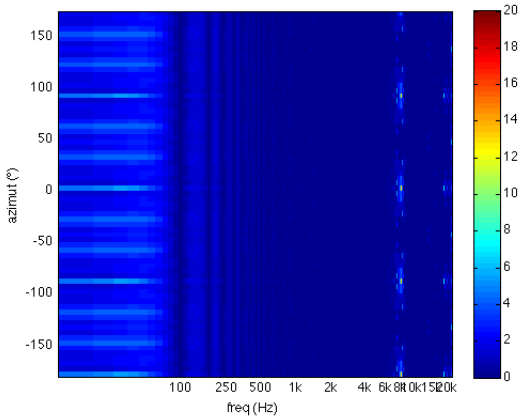
(a) HRTF de l'oreille droite reconstruite pour le sujet 1087 à partir de filtres FIR à 1024 coefficients avec compression de la dynamique



(b) Erreur de reconstruction



(c) HRTF de l'oreille droite reconstruite pour le sujet 1087 à partir de filtres FIR à 1024 coefficients sans compression de la dynamique



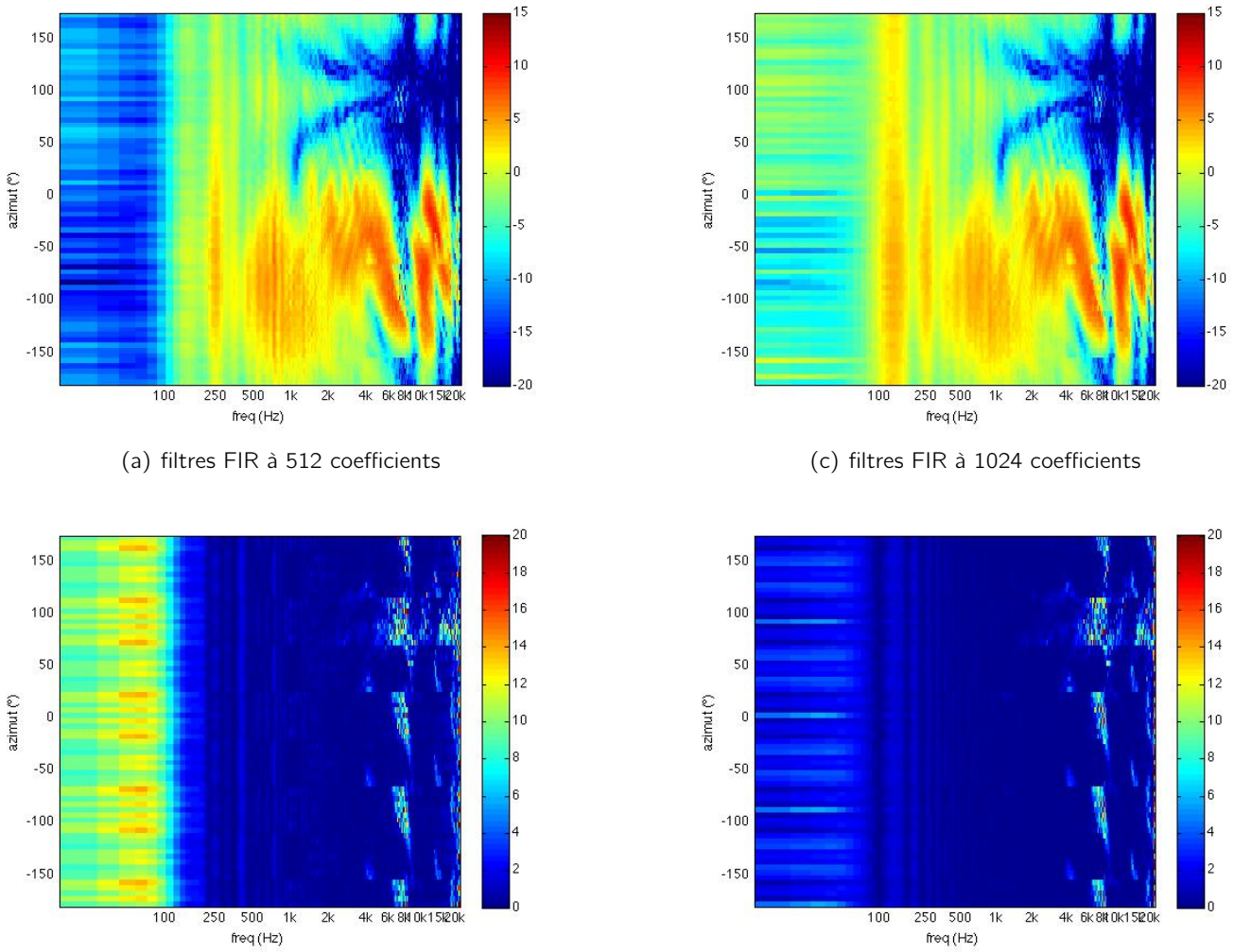
(d) Erreur de reconstruction

Fig. 12 – Effet de la compression de la dynamique du déterminant. La dynamique en haute fréquence a été divisé par 3 ainsi, les variations en hautes fréquences sont atténués par ce traitement.

3.2.5 Effet de la longueur de filtres FIR

La longueur d'un filtre FIR joue sur la précision de fréquentielle (**??**). Sur la figure 13, on peut voir l'effet de la longueur des filtres FIR. L'effet est particulièrement important pour les basses fréquences en dessous de 500

Hz. La résolution fréquentielle est trop basse pour suivre précisément les variations en des basses fréquences, d'où l'erreur relativement importante en basse fréquence , de l'ordre de $10dB$.



(a) filtres FIR à 512 coefficients

(c) filtres FIR à 1024 coefficients

(b) Erreur de reconstruction

(d) Erreur de reconstruction

Fig. 13 – Amplitude des HRTF reconstruites avec deux longueur de filtres FIR différentes.

3.3 Aspects dynamiques

La possibilité pour un l'auditeur de pouvoir se déplacer dans un domaine d'excursion délimité implique que les sources soient spatialisées de manière cohérente avec les mouvements de l'auditeur. Ce déplacement cohérent renforcera l'effet de spatialisation pour l'auditeur. La solution proposée devra répondre à de nouvelles exigences que nous allons présenter.

3.3.1 Résolution spatiale et base de données de filtres

A chaque incidence correspond une HRTF donnée. Malheureusement, comme il n'est pas possible de changer de HRTF de manière continue, on se doit de définir un incrément entre chaque incidence pour changer de HRTF. Cet incrément définit la résolution spatiale. Toute la difficulté réside dans le choix de l'incrément pour lequel le changement de filtres HRTF ne produise pas de clics audibles. Des études menées par Perrot et Saberi

dans [PS90], ainsi que Blauert dans [Bla74] ont montré qu'un incrément supérieur à 5° provoquait des sauts de localisation. Perrot et Saberi arrivent à des résultats de l'ordre de 1° dans le plan horizontal, 0.97° selon leurs mesures. En ce qui concerne l'incrément sur le plan vertical, pour une source placé à 90° d'incidence, il est de 3.65°. Quant à Lentz [Len07], les résultats de ses expériences lui donnent un incrément maximum de 1°.

Il paraît donc raisonnable de créer une base de filtres balisant le domaine d'excursion tous les 1°. Malheureusement, le protocole de mesure de HRTF étant long et fastidieux, les mesures disponibles de HRTF ont été réalisées pour un incrément de 5°. C'est pourquoi, les mesures de HRTF sont ensuite interpolées géométriquement à l'aide des quatres plus proches voisins pour affiner la résolution spatiale jusqu'à 1°.

Une base de données de filtres de crosstalk est calculée pour toutes les configurations possibles des haut-parleurs. C'est à dire que pour une résolution angulaire de 1°, le nombre de filtres à calculer est 360*360*6 = 775440. En effet, il y a 6 filtres transauraux (H_{LL}, H_{RR}, H_{LR}, H_{RL}, D, 1/D). Chaque filtre a une taille de 49 * 4= 196 octects. Ce qui donne une taille de base de données d'environ 145 Mo pour une élévation donnée. La base de filtres binauraux était déjà élaborée lors de notre étude. Cette base de filtres n'est calculée que pour une seule élévation, de plus nous ne prenons en compte que la dépendance angulaire. Les dépendances radiales sont calculées à la volée sous forme de retards et gains.

3.3.2 Effet de la résolution angulaire

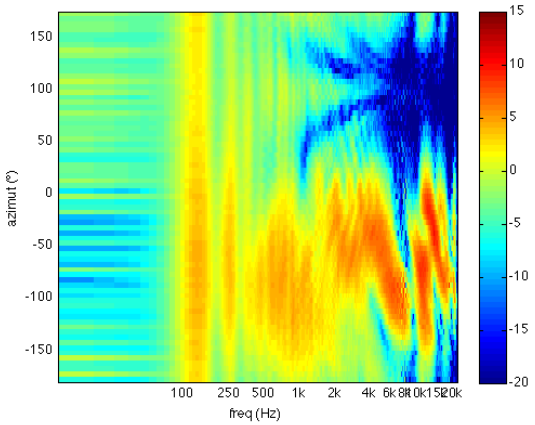
Lors de la première écoute de notre système, nous nous sommes rendus compte qu'on pouvait entendre un changement de coloration pour chaque changement de position de la tête. Ceci était dû à la base de filtres transauraux que nous utilisions. Car dans un premier temps nous avions testé notre système avec une base de filtres transauraux balisant l'espace d'écoute tous les 5°. La figure 14 illustre notre propos. Cette figure présente l'amplitude des HRTF de l'oreille droite pour le sujet 1087 de la base de données des HRTF de l'IRCAM. Ces amplitudes de HRTF ont été obtenues après simulation Matlab. Nous nous sommes placés dans une configuration angulaire à quatre haut-parleurs avec des secteurs angulaires se recouvrant. Sur ces figures, on peut observer que le passage d'une incidence à une autre est plus lisse dans le cas d'une résolution angulaire à 1°. Ce lissage de l'image se traduit à l'écoute par une disparition des changements brusques de coloration lors de l'écoute.

3.3.3 Interpolation et charge de calcul

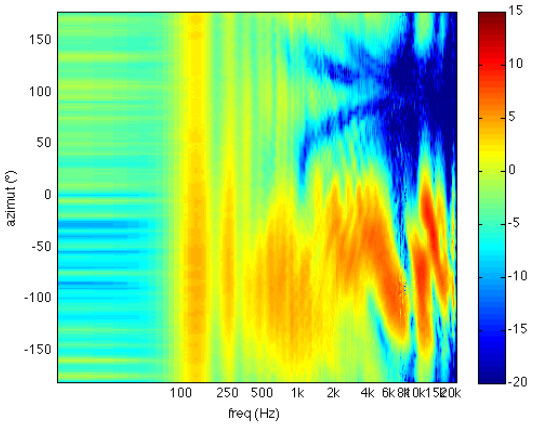
Comme énoncé ci-dessus le changement de filtres entre deux positions successives peut créer des clics audibles. Outre la nécessité de définir une résolution suffisamment précise, il faut également gérer le mécanisme permettant de charger un nouveau jeu de filtres. Un filtre a une mémoire proportionnelle à son ordre. Si l'on changeait brutalement les coefficients des filtres à chaque rafraîchissement, cela provoquerait des clics audibles. Jusqu'à présent dans le moteur de synthèse binaurale du Spat, le passage entre deux filtres se faisait par un fondu enchaîné de deux synthèses binaurales en parallèle entre deux positions successives. Dans le cas de la synthèse binaurale cela revenait à réaliser quatre filtrages en parallèle, deux filtres pour chacune des positions. Dans le cas de la synthèse transaurale, pour une seule position il faut faire huit filtrages en parallèle :

- deux filtres binauraux : H_L et H_R
- 5 filtres transauraux : H_{RR}, H_{LL}, H_{RL}, H_{LR} et deux filtres inverse du déterminant

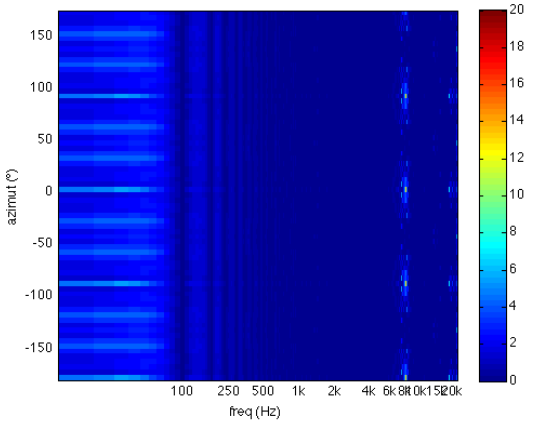
Si l'on voulait utiliser la même méthode que pour la synthèse binaurale, cela reviendrait à exécuter seize filtres en parallèle. Le coût de calcul serait bien trop important.
Pour éviter d'exécuter deux synthèses transaurales en parallèle, un procédé de fondu enchaîné directement sur les coefficients des filtres a été élaboré pour les filtres FIR par T.Carpentier durant le stage. Ainsi on limite à huit le nombre de filtrage à exécuter. Dans ce procédé il est possible de régler le temps de commutation entre chaque position. Ce temps a été règlé à 5 ms.



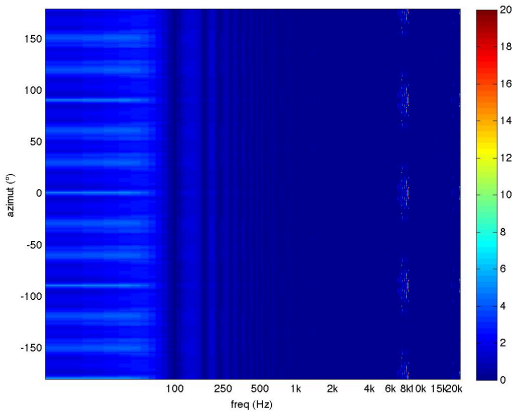
(a) Amplitudes de HRTF du sujet 1087 de la base de l'IRCAM modélisés par des filtres FIR à 1024 coefficients avec une résolution angulaire de 5°



(c) Amplitudes de HRTF du sujet 1087 de la base de l'IRCAM modélisés par des filtres FIR à 1024 coefficients avec une résolution angulaire de 1°



(b) Erreur de reconstruction



(d) Erreur de reconstruction

Fig. 14 – Figures d'amplitude des HRTF avec deux résolutions angulaires avec une zone de fondu enchaîné étendu. Effet de la résolution angulaire.

3.3.4 Temps de rafraîchissement

Le temps de rafraîchissement dépend de la cadence du système de tracking. Le système de tracking utilisé ici est un système *Optitrack* dont la cadence est réglable entre 60 Hz et 200 Hz. Soit F_R, la cadence de rafraîchissement choisie. Cette cadence fixe la vitesse maximale de rotation de la tête à F_R deg./s. Au-dessus de cette vitesse, il n'est pas possible de garantir un incrément de 1° à chaque rafraîchissement de la position. Les caractéristiques des mouvements de la tête n'ont pas étudiés ici. L'optimisation de ce paramètre serait intéressant à étudier mais faute de temps cela n'a pas été le cas.

3.3.5 Charge de calcul

La charge de calcul est dépendante de tous ces paramètres énoncés auparavant et bien sûr des moyens de calcul disponibles. Dans notre cas, la solution proposée avec une cadence de rafraîchissement de 60 Hz a pu être testée de manière satisfaisante sur un Mac Book Pro avec un processeur Intel Core 2 Duo de 2.5 *GHz* avec 2 *Go* de mémoire vive. La charge CPU lors des tests ne dépassait pas les 10%.

4 Architecture du programme

4.1 Environnement de développement

Le programme développé durant cette étude a été basé sur les bibliothèques du *Spat* de l'IRCAM. Le *Spat* est produit par l'équipe Espaces Acoustiques et Cognitifs de l'IRCAM. permettant de contrôler la spatialisation en temps réel de sources sonores. Il permet également de contrôler l'effet de salle. Il est entièrement développé en C++ et est compatible avec le logiciel *Max/MSP*. Le logiciel *Max/MSP* nous a permis de développer une interface de contrôle simplifiée pour nos tests. Mais toute la suite du développement de notre système a été réalisée en C++. De plus, il nous a permis de connecter le système de capture de mouvement *Optitrack* à notre logiciel. Un client VRPN permet de connecter le système de capture *Optitrack* par liaison Ethernet au logiciel *Max/MSP*. Les données utiles fournies par ce système de tracking sont la position et l'orientation de la tête de l'auditeur.

4.2 Architecture logicielle

Le *Spat* intègre un moteur de synthèse binaurale et transaurale statique. Nous nous sommes donc appuyés sur ce code existant pour développer notre moteur de synthèse transaurale adaptative en temps réel. Le moteur de synthèse binaurale a été complètement réutilisé pour fournir les signaux binauraux x_L et x_R d'entrées de notre système. Par contre, le moteur de synthèse transaurale a dû complètement être ré-implémenté car il ne gérait pas le changement de filtres transauraux en temps réel. La gestion de cette base de données consiste à pouvoir retrouver le bon jeu de filtres en fonction de la position et de l'orientation de l'auditeur et à le charger dans le crosstalk canceller.
Le système de capture fournit directement la position et l'orientation de la tête de l'auditeur. En fonction de cette orientation, le système détermine les secteurs actifs concernés par le crosstalk canceller. Ce fonctionnement est plus détaillé dans 4.3.
A partir des données du système de capture, les positions et orientations relatives des haut-parleurs sont calculées. Ces positions et orientations sont utilisées pour rechercher le jeu de filtres transauraux à activer et pour calculer les gains et retards à appliquer en fonction de la distance de l'auditeur à chaque haut-parleur.
Le crosstalk canceller reçoit les données du moteur de synthèse binaurale du *Spat*. La chaîne de traitement de crosstalk cancellation est décrite dans 4.4.
Une fois le traitement de crosstalk cancellation terminé, les gains et les retards sont appliqués aux signaux de chaque haut-parleur y_L et y_R. Sur la figure 15, l'architecture générale de notre système est présentée.
La base de données de filtres transauraux, contient tous les jeux de filtres H_{RR}, H_{LL}, H_{LR}, H_{RL} et l'inverse du déterminant possibles pour toutes les incidences. Cette base de données est pré-calculée à l'aide d'un programme Matlab.

4.3 Secteurs angulaires

Dans notre système transaural adaptatif nous avons choisi d'utiliser 4 haut-parleurs pour que le traitement transaural soit convenable pour toutes les positions possibles à l'intérieur du cercle de haut-parleurs. Lentz dans lentz2006 pour le système de réalité virtuelle CAVE, développé à l'université d'Aix-La-Chapelle en Allemange, utilise 4 haut-parleurs pour couvrir tout l'espace sonore. La différence avec Lentz est que nos haut-parleurs sont situés au niveau de la tête de l'auditeur. Lentz était obligé de placer ses haut-parleurs au-dessus de l'auditeur car les écrans tout autour de l'auditeur le lui imposait.
Le bon fonctionnement synthèse transaurale repose sur le conditionnement de l'inversion du déterminant. Cette inversion est particulièrement mal conditionnée lorsque l'auditeur se retrouve face à un haut-parleur. Le partage en secteurs angulaires a été détaillé dans 3.2.2.
Sur la figure 16, l'auditeur y est représenté au milieu des haut-parleurs. Avec un nombre de 4 haut-parleurs, l'espace est divisé en 8 secteurs. Sur chacun de ces secteurs est attribuée une paire de haut-parleurs. En fonction

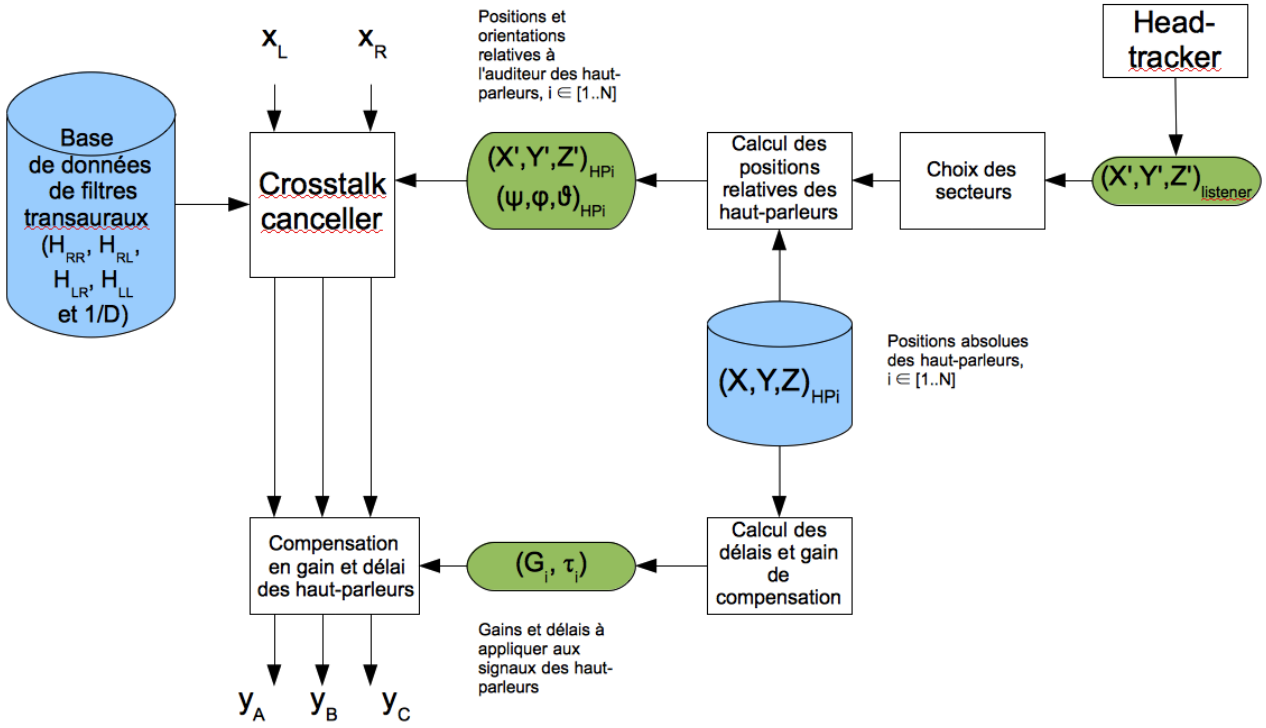


Fig. 15 – Diagramme général présentant la structure du système transaural adaptatif.

de l'orientation de la tête, le système active une paire de haut-parleurs bien spécifique. Pour l'instant notre système ne prend en compte que l'orientation de la tête de l'auditeur, c'est à dire son angle de rotation autour de l'axe vertical.

Chaque secteur $S_i, i \in [1..8]$ utilise une paire de haut-parleurs spécifique et est défini par deux orientations θ^i_{min} et θ^i_{max}, excepté pour le secteur 1 qui doit être défini en deux intervalles. La définition des secteurs est la suivante, les angles sont ici exprimés en degrés :

- S_1=$[0°, 45°] \bigcup]315°, 360°]$, utilise la paire de haut-parleurs $\{3, 4\}$
- S_2=$]0°, 90°]$, utilise la paire de haut-parleurs $\{2, 4\}$
- S_3=$]45°, 135°]$, utilise la paire de haut-parleurs $\{2, 3\}$
- S_4=$]90°, 180°]$, utilise la paire de haut-parleurs $\{1, 3\}$
- S_5=$]135°, 225°]$, utilise la paire de haut-parleurs $\{1, 2\}$
- S_6=$]180°, 270°]$, utilise la paire de haut-parleurs $\{4, 2\}$
- S_7=$]225°, 315°]$, utilise la paire de haut-parleurs $\{4, 1\}$
- S_8=$]270°, 360°]$, utilise la paire de haut-parleurs $\{3, 1\}$

L'ordre des haut-parleurs est important, car il désigne en premier le haut-parleur à gauche de l'auditeur et en second le haut-parleur à sa droite. Ainsi la paire $\{1, 3\}$ est différente de la paire $\{3, 1\}$.

Cette configuration n'est valable que pour quatre haut-parleurs. Il est tout à fait possible de créer un système avec six ou huit haut-parleurs. Dans ce cas là les secteurs doivent être redéfinis en fonction, en suivant le même principe.

4.4 Fondu enchaîné entre deux crosstalk canceller

Dans le paragraphe précédent nous avons divisé l'espace en 8 secteurs angulaires. Afin de réduire des effets de clics audibles à chaque changement de secteur pour rendre le passage le moins audible possible, nous avons convenu d'étendre les limites de secteurs pour ainsi appliquer un fondu enchaîné d'un secteur à un autre. Dans

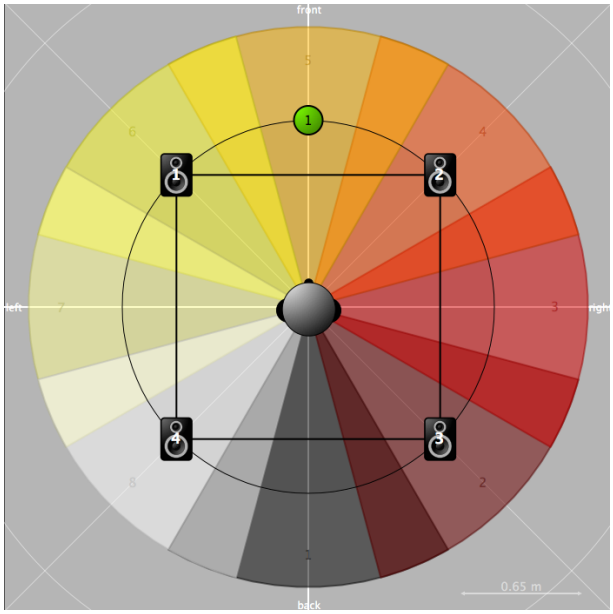


Fig. 16 – Configuration des haut-parleurs et des secteurs angulaires. Cette figure est tirée du patch Max/MSP développé avec T. Carpentier pour l'étude, fonctionnant avec le *Spat* de l'IRCAM.

cette configuration, nous forçons notre système à faire fonctionner deux crosstalk cancellers en parallèle. Si l'on considère que le fondu enchaîné commence à θ_0 et fini à θ_1. Par exemple, pour le fondu enchaîné entre S_1 et S_2, $\theta_0 = 0°$ et $\theta_1 = 45°$. Pour réaliser le fondu enchaîné entre deux secteurs, on définit un coefficient d'amplification α dépendant de l'orientation, θ de la tête,

$$\alpha(\theta) = \frac{\theta - \theta_0}{\theta_1 - \theta_0} \quad \text{pour } \theta_0 \leq \theta \leq \theta_1 \tag{17}$$

Le diagramme 17 représente la chaîne de traitement complète du système transaural adaptatif. Tous les filtres transauraux sont modélisés sous la forme de filtres en cascade de cellules d'ordre 2 (cf **??**).
Dans cette partie, nous avons décrit le système de synthèse transaurale adaptative et les aspects liés à l'adaptation en temps réel. Nous avons pu constater que cette adaptation est utile pour renforcer l'impression de spatialisation sonore. Mais elle apporte également de nouveaux problèmes liés principalement à la diminution des artefacts audibles et à la charge de calcul. Pour répondre à ces problématiques, l'architecture de notre crosstalk canceller a été basée sur l'architecture *general asymmetric feedforward* de Gardner adaptée pour des filtres IIR. Cette chaîne a été dupliquée pour pouvoir mélanger deux traitements de crosstalk canceller. De plus, un coefficient de pondération entre ces deux chaînes est calculé pour mélanger les sorties de ces deux chaînes de traitement. De cette manière, on réalise un fondu enchaîné entre deux crosstalk canceller permettant de diminuer les artefacts audibles lors d'un passage entre deux secteurs angulaires. Toujours dans le soucis de lisser les transitions entre chaque position, la résolution spatiale a été fixée à 1°. Pour limiter la charge de calcul, un mécanisme de fondu enchaîné sur les coefficients de filtres a été mis en place et la base de filtres transauraux a été pré-calculée pour éviter de faire l'inversion de filtre en temps réel. Tous ces choix d'implémentation ont fait l'objet d'une étude objective dans la partie 3.2.

4.5 Description du matériel

Système de haut-parleurs Pour la restitution transaurale, nous avons utilisé 4 haut-parleurs KEF HTS 2001. Ces haut-parleurs font partie d'un dôme de haut-parleurs de l'IRCAM. Nous n'utiliserons que 4 enceintes disposées en carré à 1.80 m du sol.

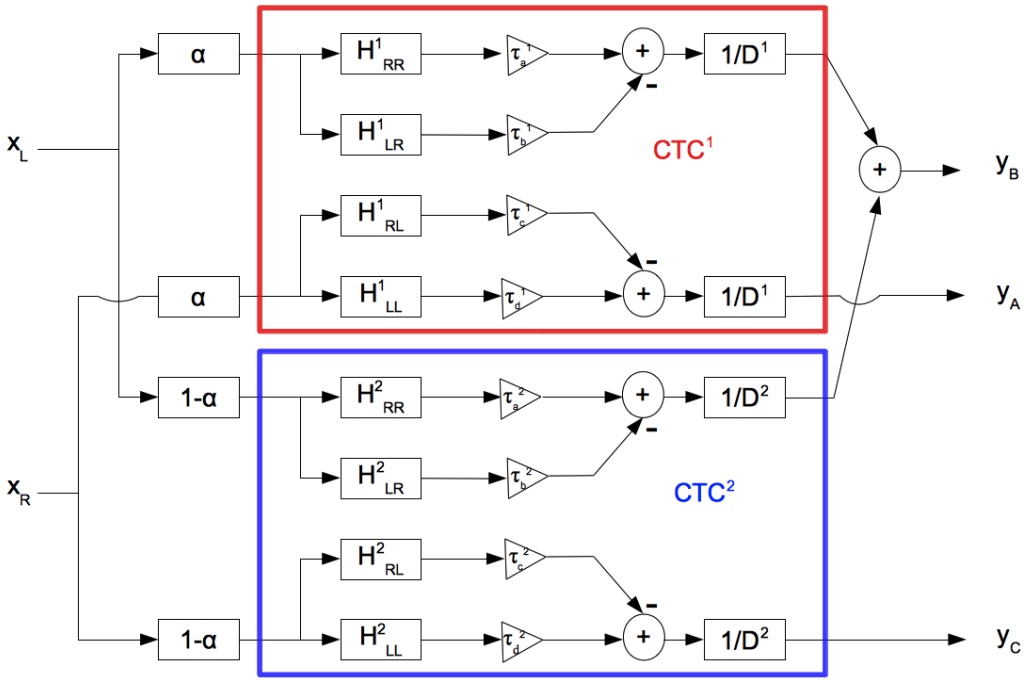


Fig. 17 – Schéma d'un système à deux crosstalk canceller en parallèle.

(a) Enceinte KEF HTS 2001 utilisée pour la restitution transaurale

(b) Caméra Optitrack utilisée pour la capture de mouvement de l'auditeur.

Fig. 18 – Enceintes et caméras utilisées pour le système transaural adapatatif.

Système de capture de mouvement Le système de tracking utilisé est un système Optitrack composé de 6 caméras infrarouges pouvant capturer la position et l'orientation de marqueurs réfléchissants. Toutes les caméras sont synchronisées par un hub usb. Le taux de rafraichissement des caméras peut atteindre 100 (FPS), donc la latence minimum entre deux mesures est de 10 ms. La précision de ces caméras est inférieure au millimètre.

Conclusion

La spatialisation sonore est un domaine dans lequel il existe de nombreuses techniques de différents types. Dans le cadre d'un système de réalité virtuelle les contraintes sont nombreuses. Le but de ce travail était de réaliser un système de spatialisation sonore nécessitant peu de moyen matériel, car le masquage écran nous y oblige. Pour cela la technique, dite transaurale, n'utilise que deux enceintes. Cette technique est basée sur la synthèse binaurale, qui permet de restituer un champ sonore aux oreilles de l'auditeur à l'aide d'un casque audio. Cette technique se base sur un filtrage à l'aide de filtres HRTF spécifiques à chaque morphologie et chaque incidence. Grâce à ces filtres, il est possible de reproduire les indices perceptifs, que sont les indices interauraux et les indices monoraux, permettant à un auditeur de localiser une source sonore. La synthèse transaurale cherche à reproduire le même champ sonore aux oreilles de l'auditeur que celui calculé par une synthèse transaurale. Pour cela, on utilise un filtrage inverse permettant d'annuler les contributions de chaque haut-parleur sur l'oreille opposée, on parle alors d'annulation de chemins croisés. C'est le principe du crosstalk canceller dont Gardner fourni une implémentation *general asymmetric feedforward*.

L'idée de cette étude est d'adapter ce filtrage en temps réel en fonction de la position de l'auditeur pour accentuer l'immersion dans une scène virtuelle. L'auditeur serait alors libre de se déplacer dans un espace d'excursion. Pour réaliser cela, nous avons repris le principe de la synthèse transaurale et l'avons adapté avec un système de capture de mouvement Optitrack. La réalisation de ce système a mis en exergue les défauts de la synthèse transaurale, c'est à dire une coloration supplémentaire du son dépendante de la position de l'auditeur et de la source. Pour atténuer ces effets indésirables, nous avons mis en place un mécanisme de fondu enchaîné entre deux crosstalk cancellers pour lisser cette coloration sur toutes les positions. Ces problèmes de coloration semblent inhérents au conditionnement du déterminant qui doit être inversé pour réaliser l'annulation de chemins croisés. Il semble difficile de les atténuer encore plus.

En revanche, lors de notre étude nous avons été confrontés à des problème de charge de calcul dus au type de filtre qu'on utilisait et à cette inversion lourde à faire en temps réel. C'est pourquoi nous avons réalisé un système de restitution transaurale basée sur des filtres IIR modélisés en cascade de cellules d'ordre 2. Grâce à cette modélisation le temps de calcul a été fortement diminué. Finalement, malgré des défauts de coloration résiduels, le système réalisé fonctionne globalement.

Bien entendu il reste des pistes d'investigation pour améliorer ce système. Tout d'abord il faudrait en priorité mettre en place une technique permettant de compenser la directivité des enceintes. Ensuite, dans notre étude nous nous sommes basés sur quatre haut-parleurs. Il faudrait étendre notre démarche pour un nombre quelconque de haut-parleurs. Le critère de conditionnement entre 4 kHz et 16 kHz semble adapté pour le choix des paires de haut-parleurs à utiliser en fonction de la position de l'auditeur. Il serait également intéressant d'étendre ce système pour des configuration de haut-parleurs à trois dimensions. Par exemple en disposant les hauts parleurs dans les quatre coins d'un cube.

En conclusion, cette étude a permis de montrer la faisabilité et les caractéristiques essentielles à un système de reproduction transaurale adaptative en temps réel. Maintenant il reste de nombreuses perspectives intéressantes à poursuivre.

Table des figures

Références

[ADTA01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pages 99–102, October 2001.

[Beg93] D.R. Begault. 3-d sound for virtual reality and multimedia. *Cambridge ,MA : Academic Press Professional*, 1993.

[Bla74] J. Blauert. *The Psychophysics of Human Sound Localization*. MIT Press, 1974.

[Car96] S. Carlile. *Virtual Auditory Space : Generation and Applications*. University of Sidney, 1996.

[Cor11] C. Cornuau. Etude et optimisation de la synthèse transaurale à deux canaux. Master's thesis, Conservatoire de Paris (CNSMDP) and IRCAM, 2011.

[CRO08] Crossmod project, 2008.

[D. 89] D. H. Cooper, and J. L. Bauck. Prospects for transaural recording. *J. Audio Eng. Soc.*, 37(1/2) :3–19, 1989.

[E.M93] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman. Localization using nonin- dividualized head-related transfer functions. *J. Acoust. Soc.*, 94 :111–123, 1993.

[Far00] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *J. Audio Eng. Soc. (Abstracts)*, 48 :350, April 2000.

[FK08] F. Keyrouz and K. Diepold. A new hrtf interpolation approach for fast synthesis of dynamic environmental interaction. *Journal of the Audio Engineering Society*, 56(1/2) :28–35, 2008.

[G. 77] G. F.Kuhn. Model for the interaural time differences in the azimuthal plane. *Acoustical Society of America, J. Acoust. Soc. Am.*, 62, July 1977.

[Gar97] W. Gardner. *3-D Audio Using Loudspeakers sdfsd*. PhD thesis, Masachusetts Institue of Technology, 1997.

[GM94] W.G. Gardner and K. Martin. Hrtf measurements on a kemar dummy-head microphone. *Technical Report 280, MIT Media Lab Perceptual Computing,*, 1994.

[J. 99a] J. Huopaniemi ; I. J. O. Smith,. Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, 3 1999.

[J. 99b] J. Huopaniemi, N. Zacharova and M. Karjalainen. Objective and subjective evaluation of head-related transfer function filter design. *J. Audio Eng. Soc*, 47(4) :218–239, 1999.

[JD91] J. Middlebrooks and D. Green. Sound localization by human listeners. *Ann Rev Psychol*, 42 :135–159, 1991.

[JO95] V. Larcher J.M. Jot and O. Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. *Audio Engineering Society*, February 1995.

[K. 99] K. Hartung, J. Braasch and S. J.Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, 3 1999.

[KN98] Ole Kirkeby and P.A. Nelson. The "stereo dipole" - a virtual source imaging system using two closely spaced loudspeakers. *Journal of the Audio Engineering Society*, 1998.

[Lac10] Y. P. Lacouture. *A Systematic Study of Binaural Reproduction Systems Through Loudspeakers : A Multiple Stereo-Dipole Approach*. PhD thesis, Aalborg University, Denmark, 2010.

[Lar01] V. Larcher. *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Université de Paris VI, 2001.

[Len06] T. Lentz. Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. In *117th Convention of the Audio Engineering Society*, 2006.

[Len07] T. Lentz. *Binaural Technology for Virtual Reality*. PhD thesis, Logos Verlag Berlin, 2007.

[LIS03] Listen database, May 2003.

[LS02] T. Lentz and O. Schmitz. Realisation of an adaptive cross-talk cancellation system for a moving listener. In *Proceedings of the 21st Audio Engineering Society Conference, St. Petersburg, Russia*, 2002.

[M. 95] S. Basu M. A. Casey, W. G. Gardner. Vision steered beam-forming and transaural rendering for the artificial life interactive video environment, (alive). In *99th Convention of the Audio Engineering Society*, 1995.

[MB93] R.M. Sachs M.D. Burkhard. Kemar the knowles electronics manikin for acoustic research. *Industrial Research Products, Inc., Elk Village, Illinois*, Report No. 20032-1, November 1993.

[Moo04] B. C. J. Moore. *An introduction to the psychology of hearing*. Elsevier Academic Press, London, UK, 2004.

[MRK99] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis. Non-minimum phase inverse filter methods for immersive audio rendering. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3077–3080 vol.6, 1999.

[Møl92] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3/4) :171–218, 1992.

[NHE92] P.A. Nelson, H. Hamada, and S.J. Elliott. Adaptive inverse filters for stereophonic sound reproduction. *Signal Processing, IEEE Transactions on*, 40(7) :1621–1632, 1992.

[NR05] P.A. Nelson and J.F.W. Rose. Errors in two-point sound reproduction. *Journal of the Acoustical Society of America*, 1(118) :193–204, 2005.

[OSB99] A. V. Oppenheim, W. Ronald Schafer, and J. R. Buck. *Discrete-time signal processing (2nd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[PS90] D. R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, 87(4) :1728–1731, 1990.

[Ray07] L. Rayleigh. On our perception of sound direction. *Philosophical magazine*, XIII :214–232, 1907.

[RM06] Harsha.I.K Rao and V. John Mathews. Inverse filter design using minimax approximation techniques for 3-d audio. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V, 2006.

[Rum01] F. Rumsey. Spatial audio. *Focal Press*, 2001.

[T. 08] T. Ajdler, C. Faller, L. Sbaiz and M. Vetterli. Sound field analysis along a circle and its applications to hrtf interpolation, 2008.

[TN00] Takashi Takeuchi and Philip A. Nelson. Optimal source distribution for binaural synthesis over loudspeakers. *J. Acoust. Soc. Am.*, 2000.

[Van01] G. Vandernoot. *Caractérisation et optimisation de la restitution haute-fidélité en véhicule*. PhD thesis, Université Paris 6 (Pierre et Marie Curie), 2001.

[Woo38] JR. S. Woodwortht. *Experimental Psychology*. New York : Holt, 1938.