N° d'ordre : 3869

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Joan, Mouba Ndjila**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

MANIPULATIONS SPATIALES DE SONS SPECTRAUX

Soutenue le : 09 novembre 2009

Après avis des rapporteurs :

Gianpaolo Evangelista	Professeur
Gaël Richard	Professeur

Devant la commission d'examen composée de :

Myriam Desainte-Cathérine	Professeur	Directrice de thèse
Sylvain Marchand	Maître de conférences HDR	Co-directeur de thèse
Gianpaolo Evangelista	Professeur	Rapporteur
Gaël Richard	Professeur	Rapporteur
Robert Strandh	Professeur	Président du jury
Udo Zölzer	Professeur	Examinateur

– 2009 –

Remerciements

Ce n'est pas à l'homme, quand il marche, à diriger ses pas.

Ce travail n'aurait pas été mature sans le soutien et la direction des mes directeurs de thèse : Professeur Myriam Desainte-Cathérine et maître de conférences HDR Sylvain Marchand. Ils m'ont orienté par leur expérience et leur expertise. Avec eux, j'ai véritablement fait les classes d'une école doctorale. Me limiter à leurs qualités scientifique serait incomplet, en effet ils ont largement contribué à ma stabilité, mes conditions de vie, mon financement par leur engagement et leurs conseils. Je vous remercie infiniment pour la bonté, l'amitié et la confiance.

Je tiens à remercier les membres du jury : les professeurs Gianpaolo Evangelista et Gaël Richard pour leur disponibilité, leur expertise, leur rigueur et leur sincérité dans l'analyse de cette thèse. Je veux remercier le professeur Udo Zölzer pour sa disponibilité, et de m'avoir reçu à l'université Helmut Schmidt afin de présenter des résultats de mes travaux. Je remercie le professeur Robert Strandh, président du jury. Merci pour votre présence dans mon jury.

Je veux remercier le LaBRI et le SCRIME pour l'opportunité offerte pour faire cette thèse dans un domaine passionnant. Merci à l'acousmaticien Jean-Michel Rivet, et à l'ingénieur recherche Boris Masencal. Merci pour vos contributions sur le plan artistique et technique.

Je veux remercier ma famille pour son soutien inébranlable, vous avez été pour moi une source de motivation, le carburant qui m'a exhorter à aller jusqu'au bout avec beaucoup d'enthousiasme. Je remercie particulièrement les parents Céline et Aloïse MOUBA, non seulement pour la vie, mais pour les merveilleux parents que vous êtes, car vous m'avez appris à aller à regarder loin devant et d'aller au-delà de ce que je m'imagine. Je vous aime pour votre amour et votre humilité. Merci à tous mes frères et soeurs. Merci Idriss pour ta proximité.

Je veux remercier le Centre d'Évangélisation Esprit et Vie de Talence, vous avez été pour moi une famille, je vous porte dans mon coeur comme un sceau. Merci particulièrement au couple Nina et Sosthène Mabouadi et à l'équipe Multimédia de CEEV (Igor, Micaël ...).

Je veux remercier mes amis et collègues au LaBRI (Rodrigue, Omer, Youssou, Alexander, ...), l'équipe Image & Son, le secrétariat, l'accueil, l'équipe système, l'AFODIB, les bibliothèques, l'école doctorale et la direction du LaBRI. Vous avez de près été des facteurs travaillant à l'achèvement et à la soutenance de cette thèse.

Le meilleur pour la fin, je remercie ma très chère fiancée et future épouse Liliane Nzoda, le moteur derrière cette thèse. Tu es magnifique et merveilleuse. Au delà du labeur continu au Laboratoire, le vrai bonheur, je l'ai trouvé avec toi par ton amour divin.

Certainement, citer tous ceux qui ont de près ou de loin jouer un rôle dans l'aboutissement de cette thèse donnerait lieu à une seconde thèse. Veuillez, m'excuser, mais sachez que vous avez une place dans mon coeur.

Table des matières

Table des figures

Introduction

Nous évoluons chaque jour dans un environnement où divers sons parviennent à nos oreilles. A partir des signaux perçus par les deux oreilles (signal binaural), le système auditif humain est capable de détecter, de localiser, de séparer ou de se concentrer sur une source précise. Cette faculté impressionnante et innée implique plusieurs traitements complexes. En effet, des processus physiques, physiologiques et psycho-acoustiques se combinent pour permettre à tout être humain de s'orienter, et ainsi de se préserver de divers dangers potentiels.

Le défi scientifique de reproduire techniquement les prouesses naturelles du système auditif humain reste ouvert, et continue à susciter autant d'engouement chez les psycho-acousticiens, et pas moins chez les musiciens. L'ensemble des processus de perception du système auditif humain sont regroupés sous le terme Analyse de Scène Sonore ou *Autidory Scene Analysis* (ASA). Les systèmes numériques implantant des algorithmes dédiés à l'ASA sont appelés Analyse de Scènes Auditives Computationnelles ou *Computational Auditory Scene Analysis* (CASA).

Dans le cadre de nos recherches, nous nous positionnons dans le plan horizontal, c'est-à-dire le plan qui se situe à la même hauteur du sol que nos oreilles. Nous nous intéressons particulièrement aux méthodes numériques pour la spatialisation de son, plus précisement, nous souhaitons imposer une position dans l'espace à une source sonore. La diffusion du son spatialisé s'effectue sur les deux écouteurs d'un casque audio ou à partir d'un nombre de haut-parleurs quelconque dans le cas d'une salle de diffusion.

Pour spatialiser, dans le passé, les musiciens positionnaient l'instrument directement à la position spatiale et à la distance désirées ; aujourd'hui la technique la plus simple consiste à filtrer le signal monophonique par les fonctions de transfert relatives à la tête (HRTF), ces filtres caractérisent les trajets acoustiques entre la source et chaque oreille.

Pour le calibrage d'un tel système, la mesure de HRTF des trajets pour chaque position cible et pour chaque personne est de règle. Ainsi, notre premier objectif est de s'affranchir de la dépendance des HRTF à l'aide d'un modèle paramétrique approprié, qui s'adapte facilement à chaque position et à toute personne, et d'en déduire une méthode de spatialisation efficace. Il s'agit précisément de modéliser les indices acoustiques pertinents pour la perception de l'espace à savoir la différence en amplitude et la différence en temps d'arrivée des signaux perçus par nos deux oreilles.

Dans le cas d'une diffusion multi-canal, les acousmaticiens utilisent des tables de mixage et disposent les haut-parleurs par rapport à la structure de la salle. Le second objectif de nos travaux est d'étendre la diffusion binaurale à une diffusion multi-canal intuitive, de sorte que ce système s'adapte à différentes configurations. Nous voulons également changer perceptivement la localisation d'une source dans l'espace (re-spatialisation). Ainsi, un instrument perçu à gauche pourrait être virtuellement re-positionné au centre ou à droite. Cette thèse s'inscrit ainsi dans un contexte d'écoute active, c'est-à-dire proposer la possibilité de changer les paramètres spatiaux d'un son pendant l'écoute. Toutefois cet objectif, ludique pour certains, pose des problèmes théoriques et techniques.

En effet, pour changer la position d'une source, il est primordial de la localiser. Notre troisième objectif est de proposer une méthode de localisation efficace adaptée au modèle paramétrique dans le cas binaural. Une exploration de techniques de localisation dans le cas multi-diffusion sera également réalisée.

La manipulation individuelle de chaque source suppose leur préalable séparation. La séparation de sources induit en général des phénomènes de distorsion et d'interférence. Notre quatrième objectif est de proposer des méthodes de séparation de sources à partir du signal binaural ; nous prendrons garde à ce que les méthodes proposées minimisent les effets indésirables dans les signaux reconstitués, et surtout que l'ensemble du système n'impose pratiquement aucune intervention humaine.

La distance est également un indice de spatialisation dans le plan horizontal qui ajouterait un autre degré de liberté à la manipulation spatiale du son. La maîtrise de la distance permettrait également de s'adapter aux dimensions de la salle. Notre cinquième objectif est de consigner la distance dans nos techniques de localisation et de spatialisation.

Ce manuscrit est structuré selon le plan suivant.

Chapitre 1 : Éléments de traitement du signal

Ce chapitre présente quelques éléments théoriques et pratiques du traitement du signal sonore. Nous nous limitons aux notions nécessaires pour la compréhension des techniques et algorithmes proposés dans ce travail. Tout d'abord, nous introduisons la notion de signal et mentionnons quelques caractéristiques. Ensuite, nous expliquons le principe de la transformation d'un signal temporel dans le domaine temps-fréquence et vice versa. Le domaine temps-fréquence est le domaine privilégié d'exécution de nos méthodes. Enfin, nous introduisons les principes fondamentaux de la manipulation spectrale du son, à savoir la manipulation de l'amplitude, de la phase, ainsi que leurs relations à la manipulation spatiale.

Chapitre 2 : Modélisation du son spatial

Les différences en amplitude (ILD) et en temps d'arrivée (ITD) constituent les indices acoustiques primordiaux dans la localisation binaurale. En effet, l'onde d'une source positionnée vers la gauche atteindra l'oreille gauche avant l'oreille droite, de la même manière l'amplitude à l'oreille gauche sera plus élévée qu'à l'oreille droite. Ces deux phénomènes sont causés par l'effet d'ombre de la tête. Aussi, la *Duplex Theory* de Lord Rayleigh (1907) stipule que l'ITD est crucial aux basses fréquences alors que l'ILD joue un rôle important pour la perception spatiale dans les hautes fréquences. Ces deux indices sont reliés à l'angle d'azimut

dans le plan horizontal. Afin de mieux utiliser ces indices dans des algorithmes, il est primordial de les modéliser. Nous recherchons un modèle réaliste et moins complexe que ceux proposés dans la littérature. Dans ce chapitre, nous proposons une amélioration du modèle paramétrique d'ITD basée sur l'hypothèse d'une tête idéalement sphérique, précedemment proposé dans [Kuh77]. Le modèle prend en paramètre l'angle d'azimut et le rayon de la tête de l'auditeur. Nous montrons qu'il existe un facteur d'échelle à chaque fréquence. Nous nous proposons de compléter le modèle qui n'était pas défini sur la bande fréquentielle de $[1-3]$ kHz. Enfin, nous évaluons le modèle sur une base de 45 auditeurs.

Chapitre 3 : Spatialisation de source

Une étude plus avancée de nos modèles paramétriques d'ILD et d'ITD confirme la symétrie de la tête dans les phénomènes de perception. Ce chapitre traite premièrement de la répartition de l'énergie d'un signal monophonique pour créer un son binaural perçu à une position d'azimut cible. Cette répartition entre les deux canaux est convenablement ajustée en s'appuyant sur les valeurs d'ITD et d'ILD. Cette technique paramétrique très simple et efficace permet de contrôler simultanément l'ILD et l'ITD, et d'apporter un gain en qualité.

Ce chapitre traite ensuite de la spatialisation sur un ensemble de haut-parleurs. La technique binaurale a été étendue à une spatialisation transaurale (sur deux enceintes). Cette technique considère les canaux statiques entre les haut-parleurs et les oreilles de l'auditeur. En approchant les signaux binauraux synthétisés, une méthode d'adaptation matricielle permet de déduire les coefficients nécessaires pour chaque canal. La spatialisation contrôlée entre deux haut-parleurs est reproductible sur les haut-parleurs autour de l'auditeur en utilisant à chaque fois la paire de haut-parleurs encadrant la localisation cible.

Ce chapitre se conclut par des commentaires de tests d'écoute sommaires et des tests objectifs qui ont permi d'évaluer leurs performances.

Chapitre 4 : Localisation binaurale mono-source

Ce chapitre est dédié à la localisation spatiale binaurale, c'est-à-dire à partir des signaux perçus par les deux oreilles. Ces signaux peuvent être issus d'une diffusion par casque d'écoute ou d'une diffusion transaurale. De nombreuses méthodes de détection et de localisation de sources sont limitées par l'ambiguïté de l'ITD au-delà de 2000 Hz. Nous nous basons sur la méthode d'évaluation conjointe de l'ILD et d'ITD proposée par Viste pour déduire un azimut robuste dans les hautes fréquences. Dans ce chapitre, nous adaptons cet algorithme à notre modèle, et nous en proposons une implantation efficace. Les sources sont localisables à partir d'un histogramme spatial construit. Des simulations montrent les performances de localisation en présence de plusieurs sources, et dans des environnements bruités et réverbérés.

Nous mettons en comparaison la méthode paramétrique et des méthodes classiques basées sur l'intercorrélation inter-canal, aussi bien dans des conditions synthétiques que dans des espaces réverbérés.

Chapitre 5 : Séparation binaurale de sources

Dans les compositions musicales, plusieurs sons sont mixés. L'isolement de chaque source est une condition pour leur manipulation individuelle. Ce chapitre donne un aperçu de plusieurs approches de séparation de sources existantes, tout en précisant leur utilité pour nos signaux d'intérêt et nos applications cibles. N'ayant à disposition que deux mélanges en présence d'une multitude de sources, la plupart de ces méthodes n'aboutissent pas à des résultats convaincants. Dans ce chapitre, nous proposons deux approches de filtrage spatial de source dans le domaine temps-fréquence. Ces méthodes sont fondées sur des masques temps-fréquence, la première fait usage d'un masque binaire et la deuxième d'un masque basé sur la probabilité *a posteriori* de chaque source. A cet effet, les sources sont considérées disjointes dans le domaine temps-fréquence. Toutefois, des points d'interférence fréquentielle de sources sont inévitables, nous évaluons les deux approches dans leur capacité à gérer et résoudre les situations d'interférences.

Des tests d'écoute et des mesures de qualités objectives pour différents type de mélange cloturent le chapitre.

Annexe A : Le logiciel RetroSpat

Dans cet annexe, nous exposons le logiciel d'informatique musicale initié à la suite de nos travaux. Le logiciel RetroSpat est actuellement structuré en deux modules principaux. Un module de localisation qui permet de détecter et de localiser automatiquement la configuration de haut-parleurs dans une salle. Un module de spatialisation qui prend en entrée des signaux monophoniques et les spatialise tout en permettant un contrôle spatial de chaque source.

RetroSpat est destiné à devenir une véritable plate-forme pour des acousmaticiens, afin de faciliter les étapes de calibrage, aussi de devenir une alternative à la table de mixage.

Chapitre 1

Éléments de traitement du signal

Ce chapitre présente les bases en traitement du signal nécessaires pour la compréhension des algorithmes proposés dans nos travaux de recherches. Dans la section 1.1, nous présentons le processus d'acquisition d'un signal numérique à partir d'un signal temporel. La section 1.2 décrit le passage du domaine temporel au domaine spectral, et dans la section 1.3, la transformation dans le domaine temps-fréquence est expliquée. Le domaine temps fréquence est le plan d'exécution principal des méthodes de manipulations spatiales (section 1.4) proposées tout au long de ce document.

1.1 Son numérique

1.1.1 Signal temporel

Lorsqu'une source sonore vibre, elle émet des ondes acoustiques par transmission d'énergie entre les molécules du milieu ambiant. L'amplitude de l'onde passe d'une valeur à une autre sans discontinuité. La représentation des amplitudes de l'onde en fonction du temps donne une courbe qui représente un *signal temporel continu* (figure 1). Les variations de pression acoustique transmises peuvent être enregistrées sur des supports physiques capables de prendre des valeurs continues, tel que le disque vinyle. On parle d'enregistrement analogique. Le signal analogique est continu. La diffusion de signaux analogiques sonores nécessite également des systèmes dit analogiques capable de décoder l'information analogique, à l'exemple de l'électrophone.

1.1.2 Numérisation du signal

Afin d'être utilisable et traitable sur un équipement numérique, le signal analogique doit être numérisé, c'est-à-dire converti en un signal discret en temps et discret en amplitude. Ces deux étapes de numérisation sont respectivement l'échantillonnage (*sampling*) pour le temps et la quantification (*quantization*) pour l'amplitude.

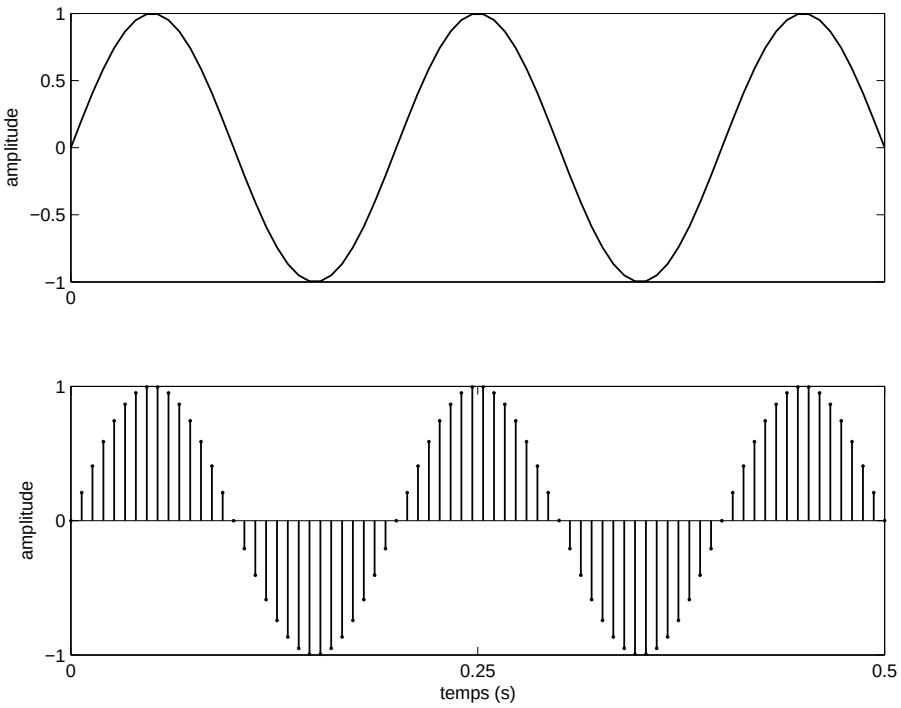


FIG. 1: *Signal continu (haut) et signal à temps discret (bas).*

Échantillonnage du signal continu

L'étape d'échantillonnage consiste à prélever l'amplitude du signal continu à différents instants. Le signal ainsi formé est dit à *temps discret* (ou *discrete-time signal*). En général, le signal sera considéré pour N instants à temps t_n ($n = 0, \cdots, N-1$) équidistants de T_s secondes. C'est l'*échantillonnage uniforme* où T_s représente la *période d'échantillonnage* (ou *sampling period*), voir figure 2. L'inverse de cette dernière est la *fréquence d'échantillonnage* (ou *sampling rate*) notée F_s et donnée par :

$$F_s = \frac{1}{T_s}. \tag{1}$$

La fréquence d'échantillonnage représente le nombre d'échantillons considérés par seconde, elle est exprimée en Hertz (Hz). Pour une fréquence d'échantillonnage de 44100 Hz, soit 44.1 kilo Hertz (kHz), 44100 échantillons du signal seront considérés par seconde à raison d'un échantillon chaque $1/44100 \simeq 22.67$ microsecondes. Au-delà de l'intervalle d'échantillonnage $T = N \cdot T_s$ le signal n'est pas connu.

Le signal temps-discret $x(n)$ est alors une suite de nombres donnés par :

$$x(t_n) = x(n \cdot T_s) = x(n), \qquad \text{avec} \quad n = 0, 1, \cdots, N, \tag{2}$$

où $x(n)$ est le n-ième échantillon de x.

Il est alors crucial que le signal temps-discret permette la reconstruction du signal analogique correspondant.

Le *théorème de Shannon-Nyquist* énonce qu'un signal analogique ne peut être reconstitué que s'il est limité en fréquence, et que la fréquence d'échantillonnage est au moins égale au double de la plus haute fréquence contenue dans le signal (fréquence maximale ou F_{max}), soit :

$$F_s > 2 \cdot F_{max}. \tag{3}$$

Des études psycho-acoustiques montrent que le système auditif humain peut percevoir des sons entre environ 20 Hz et 20000 Hz. Alors d'après l'équation 3, le signal doit être échantillonné à au moins 40 kHz afin de couvrir toute la bande fréquentielle audible. Dans l'industrie, la valeur normalisée généralement considérée est de 44.1 kHz, utilisée par le Disque Compact (CD). C'est cette fréquence d'échantillonnage qui sera d'usage dans l'ensemble de ce document.

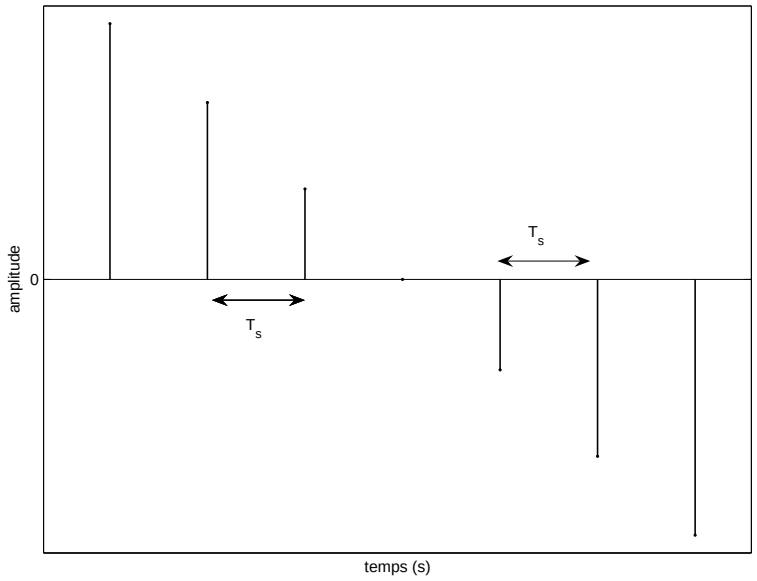


Fig. 2: *Période d'échantillonnage de la sinusoïde de la figure 1.*

Quantification du signal continu

L'étape de quantification consiste également à réduire les données, cette fois sur l'axe des amplitudes du signal. Le procédé consiste à approximer un signal analogique à valeurs dans un intervalle continu par un signal à valeurs dans un intervalle discret de taille raisonnable. A cet effet, un nombre w de bits est utilisé pour le codage des valeurs des amplitudes. Soit un intervalle d'entiers signés de $[-2^{w-1}, 2^{w-1} - 1]$. On obtient un pas de quantification $Q = 2^{-(w-1)}$ qui est différent de l'infinitésimalement petit (cas idéal). L'écart entre le signal réel et le signal quantifié est appelé *bruit de quantification*. L'erreur de quantification diminue d'environ 6 dB par bit. Dans le cas d'un CD codé sur 16 bits, on obtient $-16 * 6 = -96$ dB. Cette erreur reste encore audible par l'oreille humaine, mais tolérable. Une erreur de

quantification de -120 dB est nécessaire pour rendre l'erreur inaudible, un très bon système devrait donc compter au moins $120/6 = 20$ bits, Ce qui est le cas pour les simulations de cette thèse qui s'exécutent dans l'environnement Matlab version 7, qui utilise une arithmétique 64 bits.

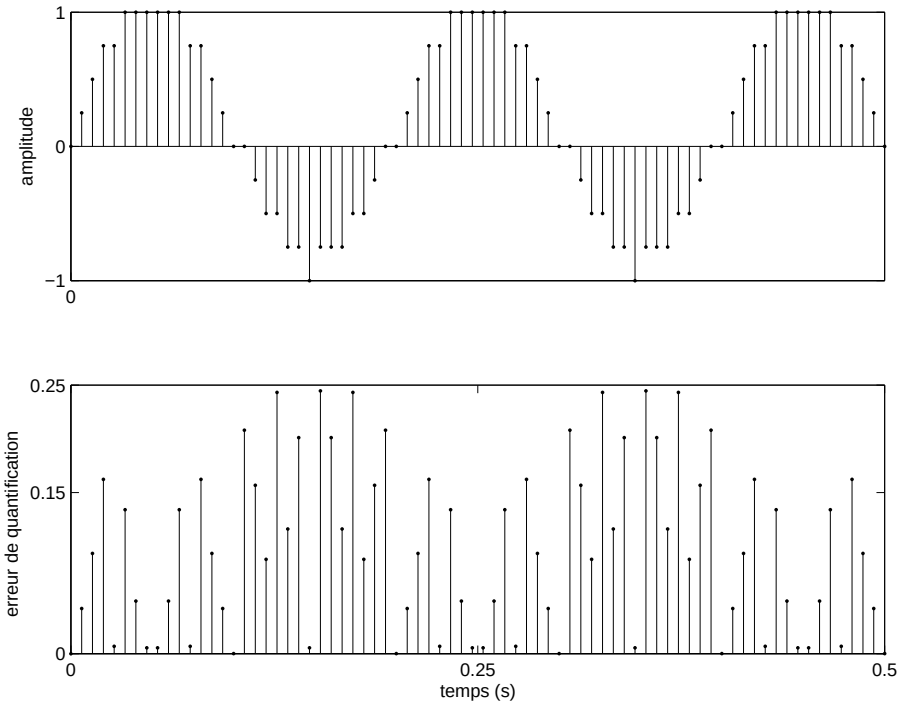


FIG. 3: *Signal quantifié (haut) et erreur de quantification (bas) à Q=3 bits .*

1.2 Représentation spectrale

1.2.1 Transformée de Fourier

La transformée de Fourier corrèle un signal quelconque avec une base de sinusoïdes à diverses fréquences. Chaque sinusoïde est caractérisée par son *amplitude*, sa *fréquence* (généralement exprimée en Hertz – Hz) et sa *phase* (généralement exprimée en radians – rads). Ainsi, on peut observer la distribution de l'énergie du signal en fonction de la fréquence. La transformée de Fourier est une fonction continue de la variable de pulsation ω (généralement exprimée en radians par seconde) ; elle n'est donc pas adaptée pour le traitement numérique, à partir d'un ordinateur.

Dans cette thèse, nous nous concentrons sur le traitement de signaux temps-discret finis.

La transformée de Fourier adaptée pour un système numérique est appelée *transformée de Fourier discrète* ou *Discrete Fourier Transform (DFT)*.

La transformée de Fourier discrète dite à N points d'un signal à longueur finie $x(n)$ est donnée par :

$$X(\omega_k) = \sum_{n=0}^{N-1} x(n) e^{-j\omega_k n} \tag{4}$$

avec

$$\omega_k = 2\pi \frac{k}{N} \qquad k = 0, \cdots, N-1, \tag{5}$$

où j est l'unité complexe ($j^2 = -1$) et k est l'indice fréquentiel. La pulsation ω_k est comprise entre $[0, 2\pi[$.

La fréquence discrète, en Hz, associée à chaque composante de la DFT $X(2\pi\frac{k}{N})$ est $f_k = 2\pi k/(NT_S)$.

La DFT peut alors s'énoncer sous la forme d'une fonction d'un entier avec :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi k}{N} n}. \tag{6}$$

$X(k)$ est dénommé *spectre du signal*. Les valeurs du spectre sont des valeurs complexes. On en déduit le spectre d'amplitude $|X(k)|$ (voir figure 4) qui sont les normes et le spectre de phase qui sont les angles en radians pour chaque fréquence.

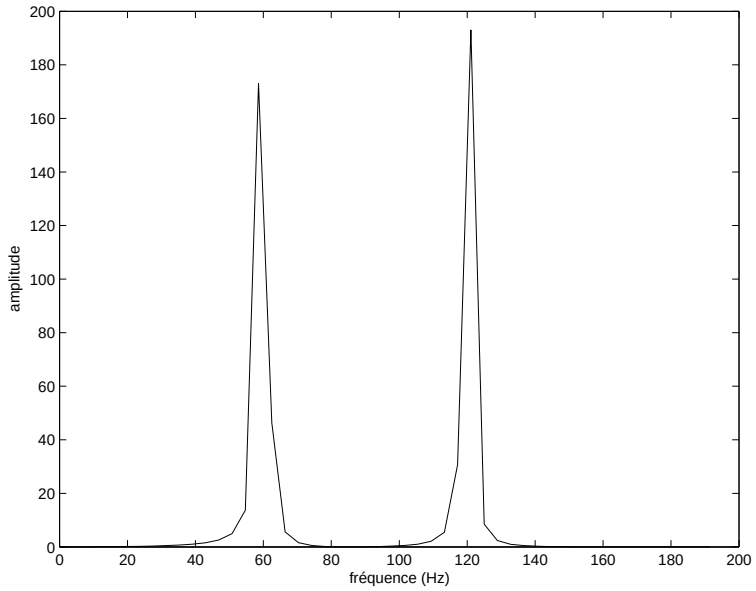


Fig. 4: *Spectre d'amplitude pour un mélange de deux sinusoïdes pures à* 60 *Hz et* 120 *Hz.*

1.2.2 Transformée de Fourier inverse

La transformée de Fourier discrète est une transformation inversible, en ce sens qu'à partir de la DFT, la version temporelle du signal peut-être reconstituée parfaitement à l'aide de la transformée discrète inverse (ou *Inverse Discrete Fourier Transform*–IDFT) selon :

opération	temporel	fréquentiel
	$x(n)$	$X(\omega)$
	$y(n)$	$Y(\omega)$
délai	$x(n-n_0)$	$\exp(-j\omega n_0)X(\omega)$
modulation	$\exp(-j\omega_0 n)x(n)$	$X(\omega+\omega_0)$
convolution	$x(n)\otimes y(n)$	$X(\omega)\cdot Y(\omega)$
multiplication	$x(n)\cdot y(n)$	$X(\omega)\otimes Y(\omega)$
intercorrélation	$x(n)\star y(n)$	$X(\omega)\cdot Y^*(\omega)$

TAB. 1: *Propriétés de la transformée de Fourier.*

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi}{N}kn}. \tag{7}$$

La DFT et la IDFT peuvent respectivement être calculées efficacement à l'aide de la transformée de Fourier rapide *Fast Fourier Transform* (FFT), et de la transformée inverse rapide ou *Inverse Fast Fourier Transform* (IFFT), avec une complexité en $\mathcal{O}(N\log(N))$.

1.2.3 Propriétés de la transformée de Fourier

La transformée de Fourier présente plusieurs propriétés qui sont très utiles dans des systèmes complexes (voir Table 1). Ces propriétés permettent également de rendre des programmes plus efficaces. Intéressons-nous à quelques propriétés données pour deux signaux finis à temps discret $x(n)$ et $y(n)$, de spectres respectifs $X(k)$ et $Y(k)$.

Symétrie hermitienne

Le spectre d'un signal possède une symétrie hermitienne. Pour N impair on a :

$$X(k) = X^*(N-k). \tag{8}$$

Ainsi le traitement de la moitié du spectre permet de réduire significativement le temps de calcul.

Théorème de linéarité

La transformée de Fourier de la somme de deux signaux pondérés $x(n)$ et $y(n)$, n'est autre que la somme pondérée de leurs transformées de Fourier individuelles.

$$a\cdot x(n) + b\cdot y(n) \quad \leftrightarrow \quad a\cdot X(k) + b\cdot Y(k), \tag{9}$$

où a, b sont des réels représentant les coefficients de pondération.

Délais temporel

Une translation du signal dans le domaine temporel entraîne une modulation dans le domaine fréquentiel.

$$x(n-\delta) \quad \leftrightarrow \quad X(k)e^{-j\frac{2\pi}{N}k\delta}, \tag{10}$$

où δ est un entier représentant le délai en échantillons.

Modulation fréquentielle

Une modulation dans le domaine temporel entraîne un décalage le domaine fréquentiel.

$$x(n)e^{-j\frac{2\pi}{N}k_0 n} \quad \leftrightarrow \quad X(k+k_0), \tag{11}$$

où la fréquence de modulation en Hz est donnée par $f_0 = \frac{k_0}{N}F_s$.

1.2.4 Convolution et corrélation

Convolution

La convolution discrète entre $x(n)$ et $y(n)$ est définie suivant :

$$(x \otimes w)[n] = \sum_{m=0}^{N-1} x(m) \cdot w(n-m). \tag{12}$$

Le théorème de convolution énonce que la convolution de deux signaux dans le domaine temporel correspond à une multiplication de leurs spectres dans le domaine fréquentiel.

$$(x \otimes y)(n) \quad \leftrightarrow \quad X(k) \cdot Y(k). \tag{13}$$

Inversement, le produit de deux signaux dans le domaine temporel, correspond à la convolution de leurs spectres.

$$x(n) \cdot y(n) \quad \leftrightarrow \quad X(k) \otimes Y(k), \tag{14}$$

avec

$$(X \otimes Y)(k) = \frac{1}{N}\sum_{m=0}^{N-1} X(m) \cdot Y(k-m). \tag{15}$$

La valeur de $X(k)$ à une fréquence particulière $k = k_0$ n'est autre que la somme de toutes les contributions à chaque k pondérée par la fenêtre centrée en k_0 et mesurée en k.

Corrélation

Par analogie à la convolution, la corrélation discrète est définie par :

$$(x \star y)(n) = \sum_{m=0}^{N-1} x(m) \cdot y^*(n+m). \tag{16}$$

Le théorème de corrélation est défini par :

$$(x \star y)(n) \quad \leftrightarrow \quad X(k) \cdot Y^*(k). \tag{17}$$

La relation d'auto-corrélation est définie par :

$$(x \star x)(n) \quad \leftrightarrow \quad |X(k)|^2, \tag{18}$$

$|X(k)|^2$ n'est autre que la puissance spectrale du signal (le carré de l'amplitude).

1.3 Transformée de Fourier à court terme

1.3.1 Définition et principes

La transformée de Fourier donne une idée en moyenne du contenu fréquentiel d'un signal à durée limitée, elle est adaptée aux signaux dits *stationnaires*, dont les caractéristiques fréquentielles ne changent pas avec le temps. Dans nos applications, les signaux d'intérêt sont des sons, la parole, les voix, la musique. Ces derniers sont dans la catégorie des signaux dits *non-stationnaires*, et ont la particularité d'avoir un contenu fréquentiel qui dépend du temps. En effet lorsqu'on parle, les syllabes alternent dans le temps. De tels signaux sont dits *non-stationnaires*. Pour une meilleure analyse de signaux non-stationnaires, la transformée de Fourier à court terme ou *Short-Time Fourier Transform* (STFT) est la mieux adaptée. La STFT consiste à projeter dans le domaine fréquentiel (par FFT) des blocs obtenus en appliquant une fenêtre glissante au signal. On obtient ainsi une représentation à deux dimensions, à savoir le temps et la fréquence. La transformée de Fourier discrète à court terme est exprimée mathématiquement par :

$$X(m,k) = \sum_{n=0}^{N-1} w(n+mH)x(n)e^{-j2\pi\frac{kn}{N}}. \tag{19}$$

Ici m est l'index temporel de chaque bloc, k est l'index fréquentiel, $w(n)$ est une fonction de pondération ou fonction de fenêtrage, N représente la longueur d'un bloc et H est le décalage de la fenêtre en échantillons entre deux index temporels. Le spectre à court-terme d'un signal peut être représenté par un spectrogramme (voir figure 5). Le spectrogramme est un diagramme qui associe à chaque instant d'un signal, son spectre de fréquence.

Appliquer un fenêtrage au signal permet aussi de remplir la condition de temps limité nécessaire pour la DFT, et permet aussi de s'accommoder à l'analyse fréquentielle de long

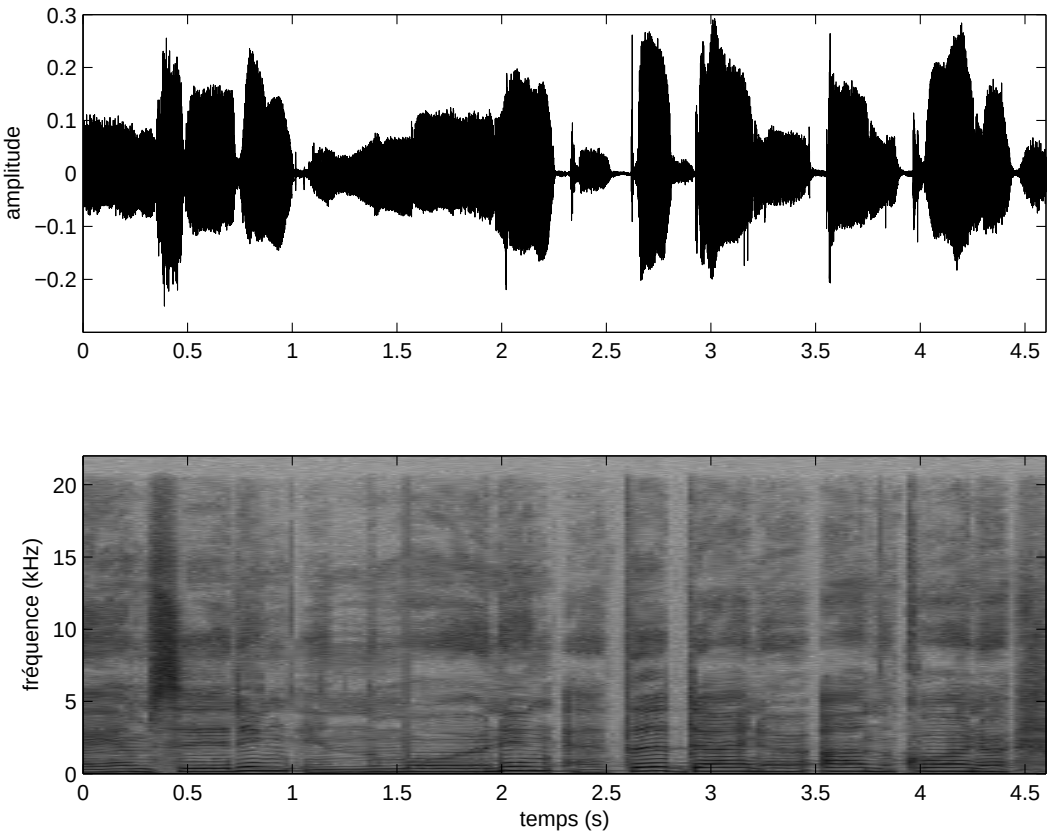


Fig. 5: *Signal temporel (haut) et spectrogramme (bas). Fenêtre de Hann, N=2048, τ=1024.*

signaux, comme dans des applications temps-réel. Toutefois, le fenêtrage du signal à des effets sur le spectre réel du signal. Dans la section suivante, nous justifions le choix de notre fenêtre de pondération parmi plusieurs.

1.3.2 Fenêtres d'analyse

De nombreuses fenêtres d'analyse existent, le choix de la fenêtre est crucial et dépend de l'application. Pour des discussions plus exhaustives sur les fenêtres de pondération consulter les références [Har78], [OSB99]. Nous analysons les quatre candidats les plus utilisés, à savoir les fenêtres rectangulaire, de Hamming, de Hann et de Blackman. Les versions temporelles et les spectres des fenêtres sont représentées dans les figures (6) et (7).

Les équations respectives des fenêtres centrées dans le domaine temporel et fréquentiel sont les suivantes :

Fenêtre	Ratio lobe princ.-lobe sec.(dB)	Dyn. d'atténuation (dB/Oct.)	Largeur du lobe princ.
Rectangulaire	-13	-6	$2\mathcal{B}$
Hamming	-43	-6	$4\mathcal{B}$
Hann	-32	-18	$4\mathcal{B}$
Blackman	-58	-18	$6\mathcal{B}$

TAB. 2: *Fenêtres d'analyse et quelques propriétés.*

$$w_{\text{Rect}}(n) = 1.0 \tag{20}$$

$$w_{\text{Hann}}(n) = \frac{1}{2}\left[1.0 - \cos\left[\frac{2n}{N}\pi\right]\right] \tag{21}$$

$$w_{\text{Hamming}}(n) = 0.54 - 0.46\cos\left(\frac{2n}{N}\pi\right) \tag{22}$$

$$w_{\text{Blackman}}(n) = 0.42 - 0.50\cos\left[\frac{2\pi}{N}n\right] + 0.08\cos\left[\frac{2\pi}{N}2n\right] \tag{23}$$

avec $n = 0, \cdots N-1$.

$$W_{\text{Rect}}(\omega) = \exp\Big(-j\frac{N-1}{2}\omega\Big)\frac{\sin\left[\frac{N}{2}\omega\right]}{\sin\left[\frac{1}{2}\omega\right]} \tag{24}$$

$$W_{\text{Hann}}(\omega) = 0.5D(\omega) - 0.25\Big[D\Big(\omega - \frac{2\pi}{N}\Big) + D\Big(\omega + \frac{2\pi}{N}\Big)\Big] \tag{25}$$

$$W_{\text{Hamming}}(\omega) = 0.54D(\omega) - 0.23\Big[D\Big(\omega - \frac{2\pi}{N}\Big) + D\Big(\omega + \frac{2\pi}{N}\Big)\Big] \tag{26}$$

$$\begin{aligned} W_{\text{Blackman}}(\omega) &= 0.21\Big[D\Big(\omega\Big)\Big] - 0.25\Big[D\Big(\omega - \frac{2\pi}{N}\Big) + D\Big(\omega + \frac{2\pi}{N}\Big)\Big] \\ &+ 0.04\Big[D\Big(\omega - \frac{4\pi}{N}\Big) + D\Big(\omega + \frac{4\pi}{N}\Big)\Big] \end{aligned} \tag{27}$$

où

$$D(\omega) = \exp\Big(+j\frac{\omega}{2}\Big)\left[\frac{\sin\Big(\frac{N}{2}\omega\Big)}{\sin(\frac{1}{2}\omega\Big)}\right]. \tag{28}$$

La résolution fréquentielle d'une transformée de Fourier à N-points est donnée par $\mathcal{B} = \frac{Fs}{N}$. De nombreux paramètres sont d'une importance dans le choix de la fenêtre, notamment, la hauteur relative du lobe principal par rapport au premier lobe secondaire, la largeur du lobe principal, l'évolution de l'atténuation de l'énergie des lobes secondaires. Ces paramètres sont résumés pour les quatre fenêtres dans la table 2.

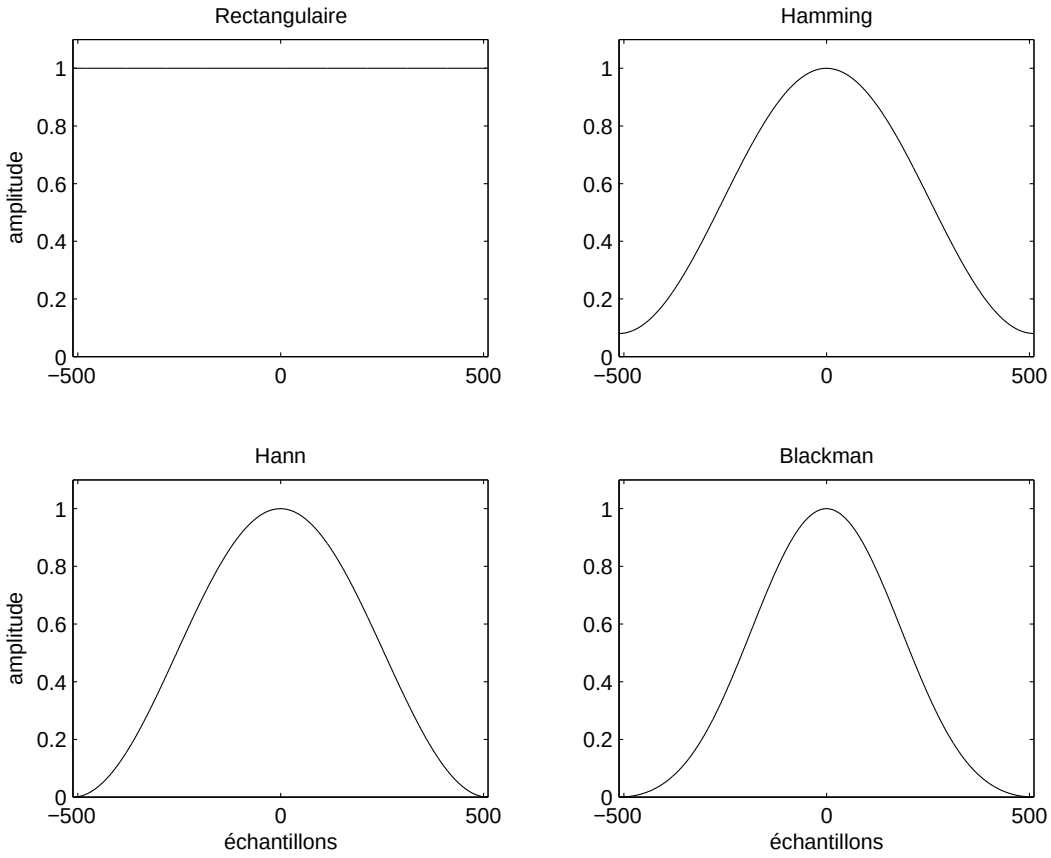


Fig. 6: *Versions temporelles des fenêtres d'analyse : fenêtre rectangulaire, Hamming, Hanning et Blackman.*

Choix de la fenêtre de Hann

Une condition importante est que le lobe principal domine significativement le premier lobe secondaire, c'est le cas de la fenêtre de Hann (-32 dB), et non de la fenêtre rectangulaire (-13 dB). Toutefois, l'atténuation des lobes secondaires accroît la largeur du lobe principal, qui doit être le plus fin possible. La fenêtre de Hann est un bon compromis pour nos applications Avec une largeur de $4\mathcal{B}$. La fenêtre de Hamming est également un bon candidat à ce niveau. Mais sa dynamique d'atténuation des lobes secondaires est beaucoup moins significative (-6 dB) que celle de la fenêtre de Hann (-18 dB). Ainsi, avec une longueur suffisamment grande, la fenêtre de Hann est utilisée dans le reste du document.

1.4 Manipulations dans le plan temps-fréquence

La transformée de Fourier à court terme fournit un espace à deux dimensions propice aux transformations musicales (spatialisation, filtrage, ...). La manipulation se déroule en trois étapes fondamentales :

– Analyse : création des blocs et transformation dans le domaine fréquentiel par FFT,

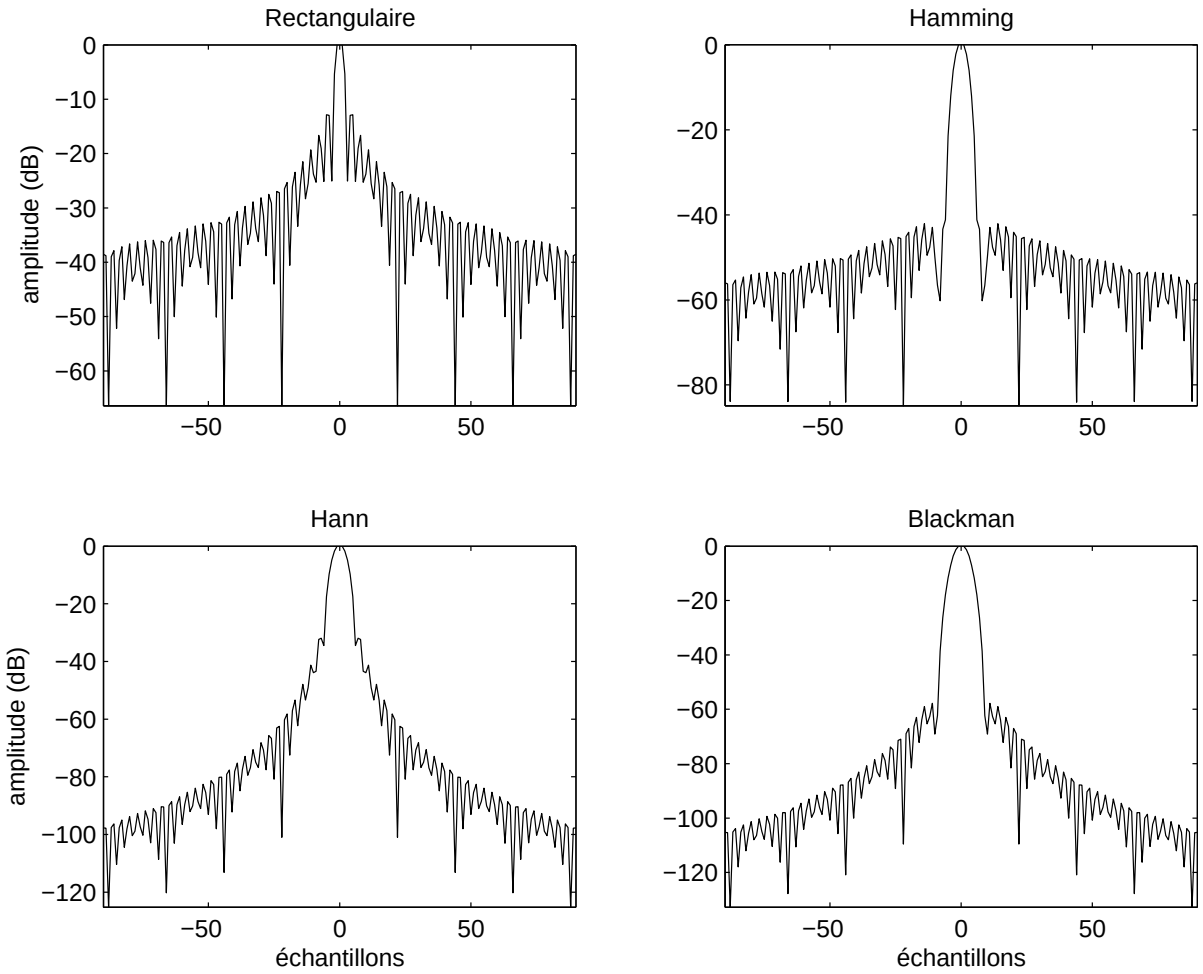


FIG. 7: *Spectres d'amplitude des fenêtres d'analyse : fenêtre rectangulaire, Hamming, Hanning et Blackman.*

- Modification : modification des paramètres spectraux (amplitude, phase, fréquence) par diverses opérations mathématiques,
- Synthèse : reconstruction de signaux temporels à partir des spectres modifiés par IFFT. Selon que la translation de la fenêtre d'analyse glissante est inférieure ou égale à la taille d'un bloc, les signaux temporels sont superposés et additionnés pour former le signal de sortie par *overlap-add*. Afin qu'une reconstruction parfaite soit possible, la somme des superpositions de la fenêtre glissante doit être égale à l'unité sur l'ensemble du signal traité.

Chaque spectre à court-terme $X(t,f)$ est filtré par une *fonction complexe* $G(t,f)$ avec :

$$\hat{X}(m,k) = G(t,f)X(m,k). \tag{29}$$

G peut s'écrire sous forme de Euler avec

$$G(m,k) = |G(m,k)|e^{j\phi_G(m,k)}. \tag{30}$$

Les spectres d'amplitude $|G(m,k)|$ et de phase $\phi_G(m,k)$ peuvent être modifiés indépendamment. La phase d'origine n'est pas modifiée si le spectre de phase de la fonction de manipulation est nul. L'amplitude d'origine n'est pas modifiée si le spectre d'amplitude de la fonction de manipulation est unitaire. En somme la manipulation spectrale d'une source consiste en un filtrage de cette source dans le domaine spectral. Il s'agit alors de déterminer le spectre du filtre approprié pour réaliser un effet donné (spatial, étirement, amplification, etc).

1.4.1 Manipulation spectrale de l'amplitude

L'amplification/affaiblissement correspond à une multiplication de l'amplitude du spectrale d'origine par l'amplitude du spectre de la fonction de manipulation, à chaque point temps-fréquence :

$$|\hat{X}(m,k)| = |G(m,k)| \cdot |X(m,k)|. \tag{31}$$

La manipulation d'amplitude est la fonction principale d'un égaliseur (equalizer), système qui permet de diminuer ou d'augmenter le volume sonore de certaines fréquences d'un son.

1.4.2 Manipulation spectrale de la phase

La transposition de la phase correspond à une addition/soustraction de la phase de la fonction de manipulation à la phase du spectre d'origine, à chaque point temps-fréquence.

$$\arg\{\hat{X}(m,k)\} = \arg\{\phi(m,k)\} + \arg\{\phi_G(m,k)\} + 2\pi p(m,k). \tag{32}$$

L'entier p précise que la phase n'est connue que modulo 2π, dû au caractère périodique de l'exponentiel. Dans certaines applications, il est crucial de dérouler correctement la phase pour retrouver la phase d'origine.

La manipulation de phase est utilisée pour des effets tels que le chorus, le flanger, le phaseur qui consistent généralement en un décalage de phase de différentes fréquences et de différentes valeurs d'une partie du spectre d'origine.

Chapitre 2

Modélisation du son spatial

Nous évoluons chaque jour dans un environnement sonore complexe, avec plusieurs sources sonores, chacune avec ses attributs (localisation, intensité, etc). Notre interaction avec le milieu dépend aussi de notre capacité à localiser et à caractériser chaque source présente, et à déchiffrer leurs interdépendances (fusion, groupement, masquage). En effet, le degré de diffusion d'une source est une empreinte du milieu environnant. Dans les milieux ouverts, on expérimente des réflexions de longue durée et les sources peuvent être très éloignées ; alors que dans les milieux fermés, comme des salles de concert, les sons subissent des réflexions de courte durée. Ainsi, l'environnement influence l'ambiance. Dans ce chapitre, nous étudierons la propagation du son et ses effets sur la localisation dans les espaces ouverts et dans les espaces fermés (section 2.1). Ensuite, un aperçu de la perception du son par le système auditif humain est présenté (section 2.2). Nous mettrons un accent sur les différents indices acoustiques liés à la perception spatiale (angle, distance), voir figure 9. Enfin, en se basant sur des modèles proposés dans la littérature par Kuhn et Viste, nous proposons un modèle paramétrique de la tête humaine par des indices acoustiques binauraux (section 2.3).

2.1 Propagation dans l'espace

Le son est dû à l'activité vibratoire d'une source à l'exemple de la membrane d'une haut-parleur. Il se propage par transmission d'énergie entre les particules du milieu ambiant. Dans nos conditions usuelles, l'air est le principal véhicule des sons que nous émettons. Le son se propage à une vitesse qui dépend des conditions physiques et du milieu de propagation. Sous les conditions atmosphériques standard, la vitesse moyenne du son dans l'air est d'environ 335 $m.s^{-1}$. Pendant leur voyage, les ondes sonores sont assujetties à divers phénomènes (réflexions, diffraction). L'intensité de ces phénomènes dépend de l'espace de propagation. Dans cette section, nous étudions la propagation en espace libre (section 2.1.1) et la propagation en espace confiné (section 2.1.2) ; dans la section 2.1.3, nous verrons les influences de la réverbération et de la distance sur la perception spatiale.

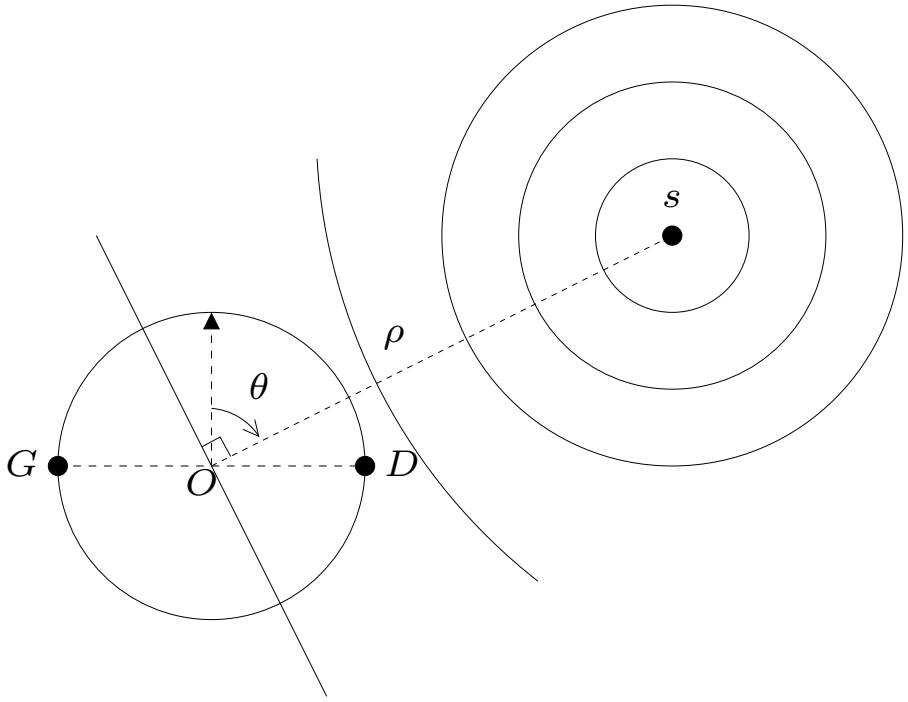


FIG. 8: *Une source s positionnée dans le plan horizontal à l'azimut θ, à la distance ρ propageant des ondes acoustiques planes vers la tête.*

2.1.1 Propagation en espace libre

L'espace libre ou champ libre désigne un environnement sans obstacle, sans réflexion. Artificiellement, un tel environnement est reproduit dans une large salle ayant des murs et le plafond constitués de matériel totalement absorbant aux ondes sonores, de manière à réduire au minimum toute émanation d'écho. Une telle chambre est appelée *chambre anéchoïque* ou *chambre sourde*. Les sons écoutés dans de telle salle sont souvent d'une qualité supérieure ; et ils sont relativement facile à localiser, car aucune réflexion ne vient troubler les indices acoustiques de localisation, principalement les différences entre les signaux perçus par les deux oreilles.

2.1.2 Propagation en espace confiné

L'espace confiné contient plusieurs obstacles qui causent des réflexions. Par exemple, une salle de concert. Dans un tel environnement, les murs, le sol, le plafond et même les sièges absorbent partiellement l'énergie du son émis, et en réfléchissent une partie dans l'environnement (voir figure 10). Ainsi, la localisation dans un milieu confiné est moins aisée, car des réflexions provenant de différentes directions se superposent au niveau des oreilles. Toutefois, le système auditif humain arrive toujours à des performances impressionnantes dans des conditions très complexes [FM04].

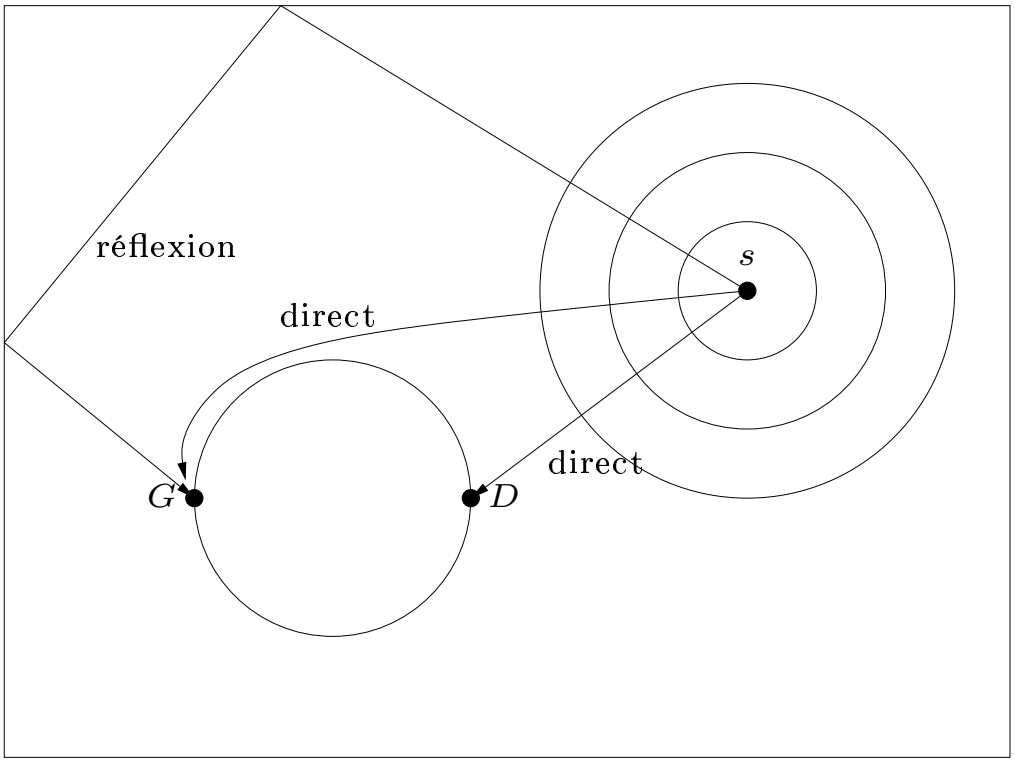


Fig. 9: *Une source s positionnée dans le plan horizontal à l'azimut θ, à la distance ρ propageant des ondes acoustiques planes vers la tête.*

2.1.3 Réverbération et distance

En espace libre

L'énergie du son est distribuée sur une surface de plus en plus grande au fur et à mesure que la source s'éloigne (voir figure 11). En conséquence, l'intensité du son décroît inversement proportionnelle au carré de la distance à la source. La *loi du carré inverse* (ou *Inverse Square Law*) précise que la pression sonore est réduite de moitié lorsque la distance double, elle diminue d'environ 6dB. L'évolution de la pression sonore peut donc permettre d'estimer la distance.

Toutefois, la localisation par la distance basée sur l'amplitude n'assure pas une estimation absolue de la distance, car en réalité, les hautes fréquences sont plus absorbées que les basses fréquences lorsque la source s'éloigne. Bass *et al.* [BSZ+95] proposent un facteur d'atténuation fréquentiel pour des conditions de température, d'humidité, et de pression atmosphérique. Les auteurs montrent que l'atténuation fréquentielle est grossièrement inversement proportionnelle au carré de la fréquence f^2. Similairement, la norme ISO 9613-1-93 aboutit au même résultat avec des formules analytiques différentes [Int93]. Intuitivement, l'estimation de la distance peut être biaisée par le type de son et le contenu fréquentiel du son.

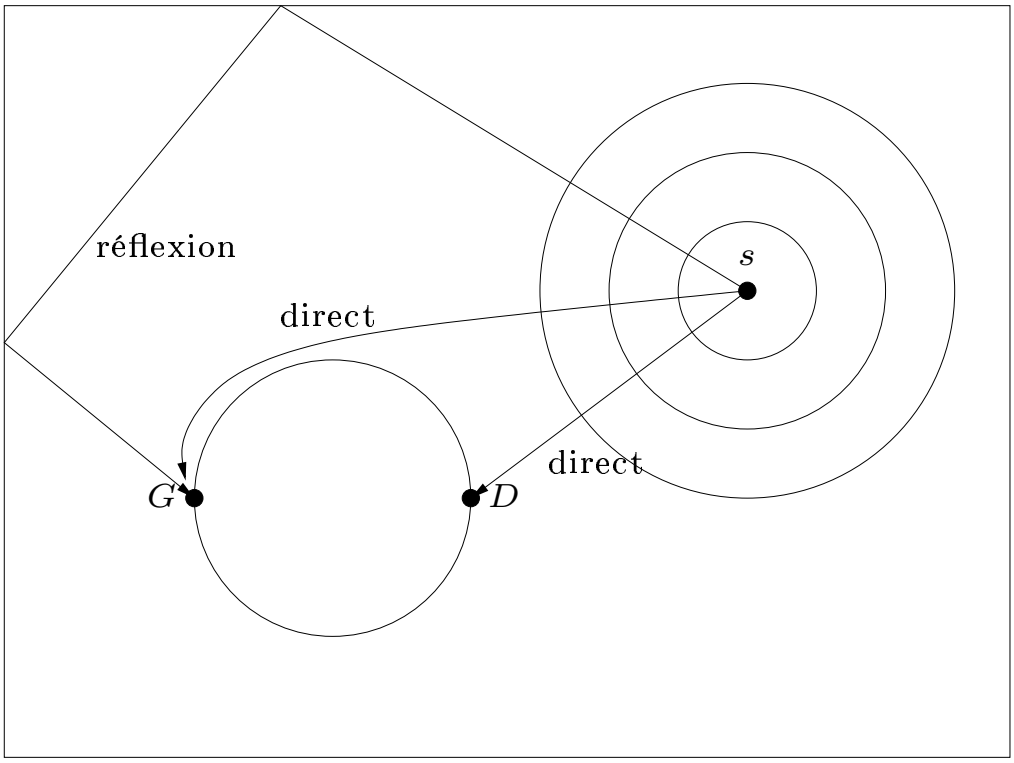


FIG. 10: *Son direct et premières réflexions.*

En espace confiné

L'énergie du son ne décroît pas très rapidement avec l'éloignement de la source. En effet, les réflexions contribuent à l'édification d'un son diffus qui maintient l'amplitude du son à un niveau assez élevé. La figure 12 montre la structure générale de la réponse impulsionnelle d'un espace confiné : le son direct d'amplitude élevée est suivi de réflexions précoces moins significatives (entre 10 ms et 80 ms) et de réflexions tardives qui forment un champ diffus ou *réverbération* (au-delà de 80 ms). La réverbération est semblable à un phénomène d'écho intensifié, caractérisée généralement par le *temps de réverbération*. Le temps de réverbération est le temps que prend l'amplitude du son pour décroître de 60 dB au-dessous de l'amplitude originelle du son émanant de la source. Contrairement à la position angulaire, la localisation par la distance dans un espace confiné semble plus aisée. En effet, nous verrons que les différents temps d'arrivée des réflexions et le rapport du son direct à la réverbération sont des informations fortement liées à la distance (section 2.2.3).

2.1.4 Modèle d'atténuation spectrale

La loi en carré inverse (figure 11) stipule que l'intensité d'un son est divisée par deux lorsque la distance à la source double [BS94]. Ce phénomène n'a aucune répercussion particulière sur le timbre du son. Mais en réalité, l'air absorbe sélectivement les différentes fréquences avec la distance, les basses fréquences étant moins absorbées que les hautes fréquences. Simuler la distance reviendrait à modifier scrupuleusement le spectre d'amplitude du son, donc à modifier

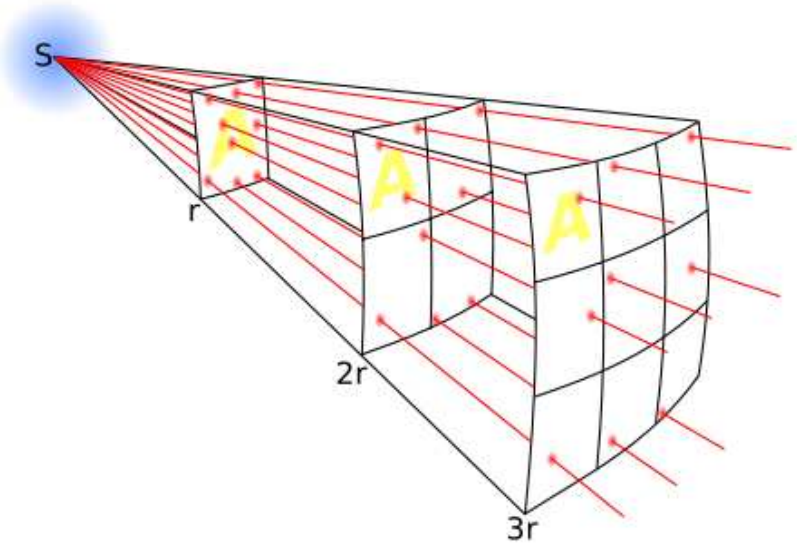


Fig. 11: *Ce schéma montre le principe de la loi en carré inverse. Les lignes représentent le flux émanant de la source. Le nombre total de lignes de flux dépend de l'intensité de la source et est constant avec l'accroissement de la distance. Une densité plus importante de lignes de flux (lignes par unité de surface) est la traduction d'un champ plus intense. La densité de flux est inversement proportionnelle au carré de la distance à la source car l'aire d'un secteur de disque s'accroit avec le carré de son rayon. L'intensité du champ est donc inversement proportionnelle au carré de la distance à la source.* `Référence : http://fr.wikipedia.org/Loi_en_carr%C3%A9_inverse`.

quantitativement son contenu fréquentiel. Plus précisément, la norme ISO 9613-1 [Int93] met en exergue l'atténuation fréquentielle que subirait grossièrement un son en tenant compte des conditions de température ambiante, d'humidité et de pression. Un facteur d'atténuation est calculable analytiquement. A la distance ρ, les amplitudes spectrales $X(f)$ seraient atténuées d'un facteur de $D(f,\rho)$ décibels, avec :

$$D(f,\rho) = \rho \cdot a(f), \tag{33}$$

où $a(f)$ est l'atténuation dépendante de la fréquence, qui impacte ainsi la brillance du son. De ce fait, nous recherchons une relation analytique entre la distance et la centroïde spectrale.

Formellement, l'absorption totale en décibels par mètre $a(f)$ est régie par la formule quelque peu complexe suivante :

$$\begin{aligned}\frac{a(f)}{P} \approx \; & 8.68 \cdot F^2 \Big\{ 1.84 \cdot 10^{-11} \left(\tfrac{T}{T_0}\right)^{\frac{1}{2}} P_0 + \left(\tfrac{T}{T_0}\right)^{-\frac{5}{2}} \\ & \Big[0.01275 \cdot e^{-2239.1/T} / [F_{r,O} + (F^2/F_{r,O})] \\ & + 0.1068 \cdot e^{-3352/T} / [F_{r,N} + (F^2/F_{r,N})] \Big] \Big\}\end{aligned} \tag{34}$$

où $F = f/P$, $F_{r,O} = f_{r,O}/P$, $F_{r,N} = f_{r,N}/P$ sont les fréquences normalisées par la pression atmosphérique P, et P_0 est la pression atmosphérique de référence (1 atm), f est la fréquence

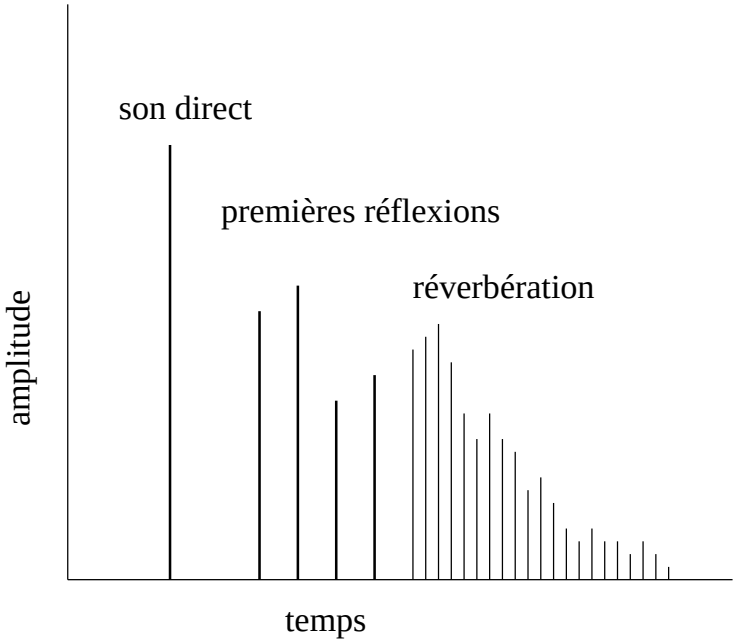


FIG. 12: *Réponse impulsionnelle dans un espace confiné. Le son direct est suivi de réflexions précoces, et de la réverbération qui décroît selon une allure exponentielle.*

en Hz, T est la température en Kelvin (K), T_0 est la température de référence (293.15K), $f_{r,O}$ est la relaxation fréquentielle des molécules d'oxygène, et $f_{r,N}$ est la relaxation fréquentielle des molécules de nitrogène. Pour plus de détails sur ces quantités, Bass [BSZ+95] constitue une excellente référence.

2.2 Perception spatiale binaurale

Dans cette section, nous donnons un aperçu de la perception spatiale du son en rapport avec la localisation spatiale et la projection spatiale du son. Des études exhaustives sur les phénomènes psychoacoustiques sont exposées par Blauert [Bla97]. Nos recherches se focalisent sur la perception des personnes à deux oreilles fonctionnelles, aussi appelée perception binaurale. La perception spatiale est liée à de nombreux indices acoustiques fournis par les signaux perçus au niveau des oreilles. Chaque indice acoustique est prépondérant selon la situation acoustique. Les indices binauraux permettent la localisation dans le plan horizontal (section 2.2.1) ; les indices spectraux sont nécessaires à la perception de l'élévation ; les indices acoustiques de la distance sont abordés dans la section 2.2.3 ; lorsque ces indices sont ambigus ou détériorés, des indices dynamiques tel les mouvements de la tête soutiennent la localisation (section 2.2.4), ainsi que l'effet de précédence (ou effet de Haas) qui favorise la localisation dans les milieux réverbérés (section 2.2.5).

Head-Related Transfer Function - HRTF

Les indices acoustiques sont des clefs physiques nécessaires au système auditif afin de localiser les sources. Pour les quantifier, il est primordial de caractériser les filtres de transfert

entre le point d'émission du son et les tympans (gauche, droite) de l'auditeur, à une position donnée. Le son propagé subit divers effets pendant son voyage, et précisément des phénomènes de diffraction, de diffusion et de réflexion sur le corps de l'auditeur (pavillons, torse, épaules). Tous ces effets sont modélisés par un filtre unique appelée *Head-Related Impulse Response* (HRIR) et dont la réponse fréquentielle est appelée *Head-Related Transfer Function* (HRTF). La HRTF dépend du rayon de la tête (r), de l'angle d'incidence (θ), de la fréquence (f), de la morphologie de l'auditeur, on la note $\mathrm{H}(r, \theta, f)$.

Les HRTF mesurées d'une source aux tympans des oreilles capturent tous les indices acoustiques relatifs à la localisation. Ainsi, un son binaural à une position (ρ, θ), où ρ est la distance de la source au centre de la tête et θ est l'angle d'azimut (voir figure 9), peut être simulé par convolution du signal sonore avec les HRTF gauche et droite (H_G, H_D).

Les HRTF sont généralement mesurées dans des milieux anéchoïques. Nous utilisons des HRTFs humain et de mannequin afin d'identifier et de modéliser les caractéristiques physiques liées à la localisation et à la spatialisation. Dans les sections suivantes, nous allons étudier des indices acoustiques liés à la localisation binaurale.

2.2.1 Indices acoustiques binauraux

De nombreuses recherches psycho-acoustiques montrent que la localisation et la spatialisation binaurale du son sont basées sur les différences des sons perçus entre les oreilles. Deux processus interviennent principalement dans la localisation binaurale et la spatialisation binaurale du son, il s'agit de la *différence interaurale en temps d'arrivée* ou *Interaural Time Difference* (ITD), et de la *différence interaurale en amplitude* ou *Interaural Level Difference* (ILD) [Ray07]. Ces deux indices acoustiques joueront un rôle principal dans nos recherches.

Différence interaurale en temps d'arrivée

La tête forme un obstacle pour le son incident. Un son d'une source positionnée vers la droite atteindra l'oreille droite (ipsilatérale) avant l'oreille gauche (contralatérale). En effet, l'onde acoustique devrait contourner la tête avant d'atteindre l'oreille gauche.

La différence en temps d'arrivée est liée à l'*angle d'incidence* ou *Direction Of Arrival* (DOA). Ainsi l'ITD est nulle pour un angle d'azimut zéro, et atteint son maximum pour un DOA de $\pm 90^\circ$, soit environ 0.7 ms pour une tête humaine typique [Bla97], c'est le *délai binaural*. Apparemment le système auditif humain serait capable de différencier la localisation à un ordre de 1°.

En considérant un signal sinusoïdal, la différence en temps peut s'exprimer en terme de différence de phase. On parle de *Interaural Phase Difference* (IPD).

Différence interaurale en amplitude

La tête forme un obstacle pour le son incident. Un son d'une source positionnée vers la droite sera plus élevé au tympan droit qu'au tympan gauche, ce qui conduit à une différence d'amplitude ou ILD, exprimée généralement en décibels (dB). L'ILD est liée à l'angle d'incidence. Ainsi, l'ILD sera nulle pour un angle d'azimut zéro, et atteint son maximum pour un

DOA de $\pm 90^{\circ}$. Apparemment le système auditif humain serait capable d'identifier une différence d'amplitude de l'ordre de 1 dB [Har]. La tête est une barrière particulière pour les hautes fréquences (au-delà de 1500 Hz), mais pas pour les basses fréquences. En effet, pour les basses fréquences, la longueur d'onde est plus large que la tête de l'auditeur et l'onde est diffractée autour de la tête, et atteint ainsi l'oreille contralatérale avec une atténuation moindre, alors que plus la fréquence augmente plus l'*effet d'ombre de la tête* devient importante, et l'amplitude à l'oreille contralatérale ne fait que décroître. L'ILD est primordiale pour les fréquences au-delà de 1500 Hz, et ne souffrent pas d'ambiguïté fréquentielle comme l'ITD. Ainsi l'ITD et l'ILD semblent être des indices acoustiques binauraux complémentaires sur l'ensemble de la bande de fréquence audible. L'ILD et l'ITD vont ainsi concentrer notre attention dans les section suivantes.

2.2.2 Indices acoustiques monauraux

Les indices monauraux ne sont pas considérés dans nos recherches, car ils sont liés au signal perçu par une seule oreille. Toutefois nous les introduisons brièvement du fait de leur implication dans la localisation naturelle. Les indices monauraux proviennent des réflexions sur le torse et sur les pavillons des oreilles. La forme des pavillons des oreilles et de ses cavités de résonance agissent comme des filtres acoustiques. Les réflexions constructives et destructives dans les oreilles entraînent des pics ou des trous dans le spectre, dont les positions fréquentielles sont liées à la localisation. Le premier trou spectral connu généralement sous le nom de *pinna notch* est reconnu comme étant très indicatif de l'élévation [Gar97]. Les indices spectraux sont très variable d'un sujet à un autre à cause de leur forte corrélation à la morphologie des oreilles et à la géométrie très fine des pavillons.

2.2.3 Indices acoustiques de distance

Plusieurs indices acoustiques influencent l'estimation de la distance par le système auditif humain. L'intensité [BS94] et le rapport du son direct à la réverbération [BH99] (D/R) sont reconnus comme les indices acoustiques les plus significatifs. L'intensité est liée au spectre et les indices acoustiques binauraux (ILD, ITD) sont significatifs pour les distances inférieures à 1m [BR99]. Ces différents indices sont relativement significatifs selon les conditions acoustiques du lieu.

Dans les environnements anéchoïques, les indices d'intensité sont d'une importance primordiale. Les indices acoustiques d'intensité proviennent du fait que la même source positionnée proche de l'auditeur est perçue avec un fort volume, alors que son volume décroît avec l'éloignement. Selon la "loi en carré inverse", le rapport d'intensité (I_1 et $I2$) d'une source à la distance $r_1 = r$ et $r_2 = 2r$, est donné par $I_1/I_2 = r_2^2/r_1^2$. De ce fait, lorsque la distance double, l'intensité du son à l'oreille de l'auditeur décroît d'à peu près 6dB. Aussi, les basses fréquences sont moins absorbées que les hautes fréquences. Ainsi, les sons lointains comportent relativement plus de basses fréquences. Le système auditif humain fait usage de ces indices acoustiques pour estimer la distance. Toutefois, une familiarité avec la source et ses caractéristiques (notamment spectrales) influencent l'estimation.

Dans les environnements réverbérés, la distribution du son dépend des caractéristiques de la réverbération de la salle. On considère que le champ sonore au-delà de la distance de

réverbération est diffuse, et théoriquement indépendant de la distance, d'où l'importance du rapport D/R comme un indice important de la distance dans un milieu réverbéré [Bla97], [Beg92]. Dans cette thèse, les méthodes d'estimation de la distance que nous proposerons se baseront particulièrement sur les indices d'intensité et le contenu fréquentiel (spectre).

2.2.4 Indices acoustiques dynamiques

Dans des situations d'écoute où les indices acoustiques ne fournissent pas suffisamment d'information au système auditif dans un but de localisation ; le sujet humain a tendance à user de mouvements de la tête pour résoudre les ambiguïtés perceptives, notamment la confusion *avant/arrière* [Wal40]. Ces ambiguïtés sont introduites dans des *cônes de confusion*. Un cône de confusion est un ensemble de points de l'espace pour lesquels la distance entre les deux oreilles est identique [BSC00]. A titre illustratif, considérons une tête parfaitement symétrique, une source sonore à l'azimut θ vers la droite dans le plan horizontal peut introduire des indices acoustiques binauraux similaires à ceux d'une source sonore à la position $180° - \theta$ (voir figure 13). En pivotant la tête vers la droite, si les indices binauraux sont minimisés, l'auditeur décide que la source est dans le plan avant, sinon, si ces derniers sont maximisés la source est dans le plan arrière. Ainsi, les mouvements de la tête favorisent la localisation dans l'écoute naturelle. Cependant la prise en charge des indices dynamiques dans des systèmes techniques n'est pas une tâche aisée. Certains systèmes assez coûteux (casques audio) sont munis de "traqueurs de mouvements de la tête". Leur fiabilité est souvent limitée par la complexité des trajectoires et la rapidité des mouvements de l'auditeur [TR67], [WRTR67].

2.2.5 L'effet de précédence

Un autre phénomène affecte la localisation dans les environnements réverbérés, c'est *l'effet de précédence*, ou la loi de la première onde. L'effet de précédence est un mécanisme d'inhibition utilisé par le système auditif humain pour affiner la localisation. Lorsque des sons concurrents (son direct, réflexions) se superposent au niveau des oreilles de l'auditeur, ce dernier localise le son comme provenant de la direction du son arrivant le premier aux oreilles, même si les réflexions semblent plus intenses que le son direct [Har97]. Toutefois, l'effet de précédence n'élimine pas l'effet des réflexions ; ces derniers donnent l'effet de salle, d'enveloppement. Dans le cas des signaux de parole, il a été prouvé que la suppression maximale se produit dans un délai de 10 à 20 ms. La compréhension est affectée pour des délais de réflexions au-delà de 50 ms.

CIPIC - Base de HRIR

Dans le cadre de nos travaux, nous avons employé la base de HRIR du CIPIC [ADTA01]. La base CIPIC est publique et contient des mesures de HRIR pour 45 sujets (43 humains, 1 mannequin avec deux moulages de pavillons différents). La base CIPIC contient des mesures haute résolution à 44.1 kHz, de 200 échantillons de long, pour 25 azimuts, 50 élévations, soit 1250 positions (voir figure 15). Les mesures sont réalisées dans une chambre anéchoïque, avec la technique à conduit auditif fermé (blocked-meatus) [Ma95]. Le sujet est assis au centre d'un arc de 5 haut-parleurs, qui diffusent tour à tour un signal large bande (code de Golay) (voir figure 14). L'arc est pivoté de sorte à positionner les haut-parleurs aux azimuts qu'on

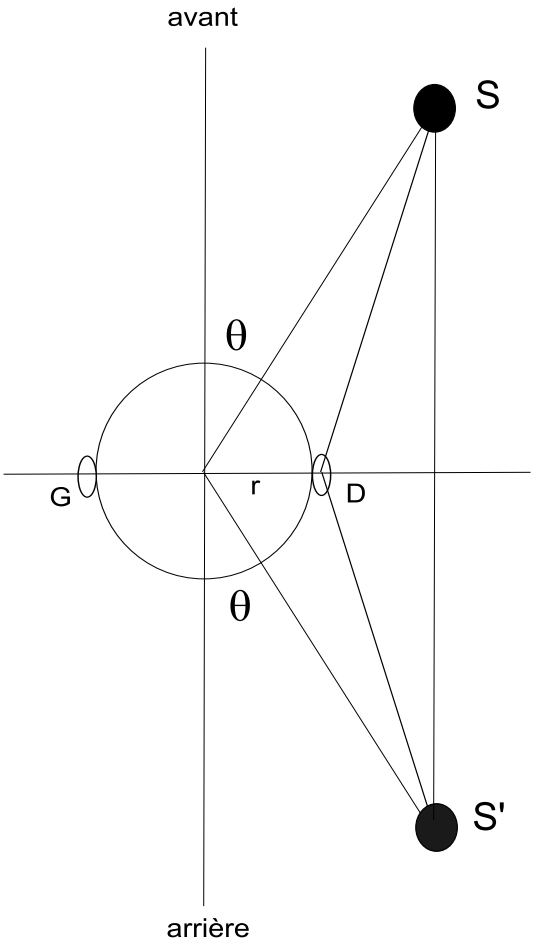


FIG. 13: *Pour une tête symétrique, une source S et son image S' introduisent des indices acoustiques similaires.*

souhaite mesurés. Les HRIR mesurées sont compensées afin d'éliminer les imperfections des haut-parleurs et microphones, et de manière à assurer un bon comportement de la fonction de transfert au-dessous de 400 Hz [Mol92]. La base contient également des mesures morphologiques de chaque sujets (rayon de la tête, dimensions des pavillons, dimensions du torse, largeur des épaules).

Les HRTF sont donc spécifiques à chaque sujet et représentent des fonctions complexes de la fréquence et de l'angle d'incidence. Il s'avère ainsi indispensable de les modéliser et de les simplifier afin qu'elles soient exploitables à des fins informatiques (avec un espace mémoire limité).

2.3 Modélisation des indices acoustiques

Avant de plonger dans la modélisation des indices acoustiques, quelques simplifications fondamentales sont à prendre en compte. On considère une source sonore ponctuelle et omnidirectionnelle dans le plan horizontal, à la position en coordonnées polaires (ρ, θ). En effet, nous considérons les situations où les auditeurs et les musiciens sont dans le même plan, comme

Fig. 14: *Système de mesure de HRIR d'un sujet humain pour la base de HRIR CIPIC.* `Référence : http://interface.cipic.ucdavis.edu/CIL_html/CIL_research.htm`.

c'est souvent le cas dans la vie quotidienne. Ainsi, nous considérons que l'élévation est petite, et peut ainsi être négligée.

La source s parviendra aux oreilles gauche (G) et droite (D) par différents trajets représentés par des lignes droite et courbe (figure 18). En effet, nous considérons les situations pour lesquelles les sources sonores sont éloignées de quelques mètres de l'auditeur (> 2 ms) ou *far-field*, au point où les ondes acoustiques parvenant aux oreilles peuvent être considérées comme des ondes planes. Cela implique que la distance à la source est d'une importance minime pour les indices binauraux (ITD et ILD).

2.3.1 Modèles d'ILD et d'ITD de Harald Viste

Modèle de l'ILD

Basé sur une étude de la base de HRIR CIPIC, Viste observe que l'ILD est sinusoïdale en azimut θ. Pour les azimuts dans l'intervalle $-90^\circ \leq \theta \leq 90^\circ$, une expansion de Fourier de premier ordre de l'ILD [DM97] conduit au modèle sinusoïdal d'ILD suivant :

$$\mathrm{ILD}(\theta, f) = \alpha(f) \sin(\theta), \tag{35}$$

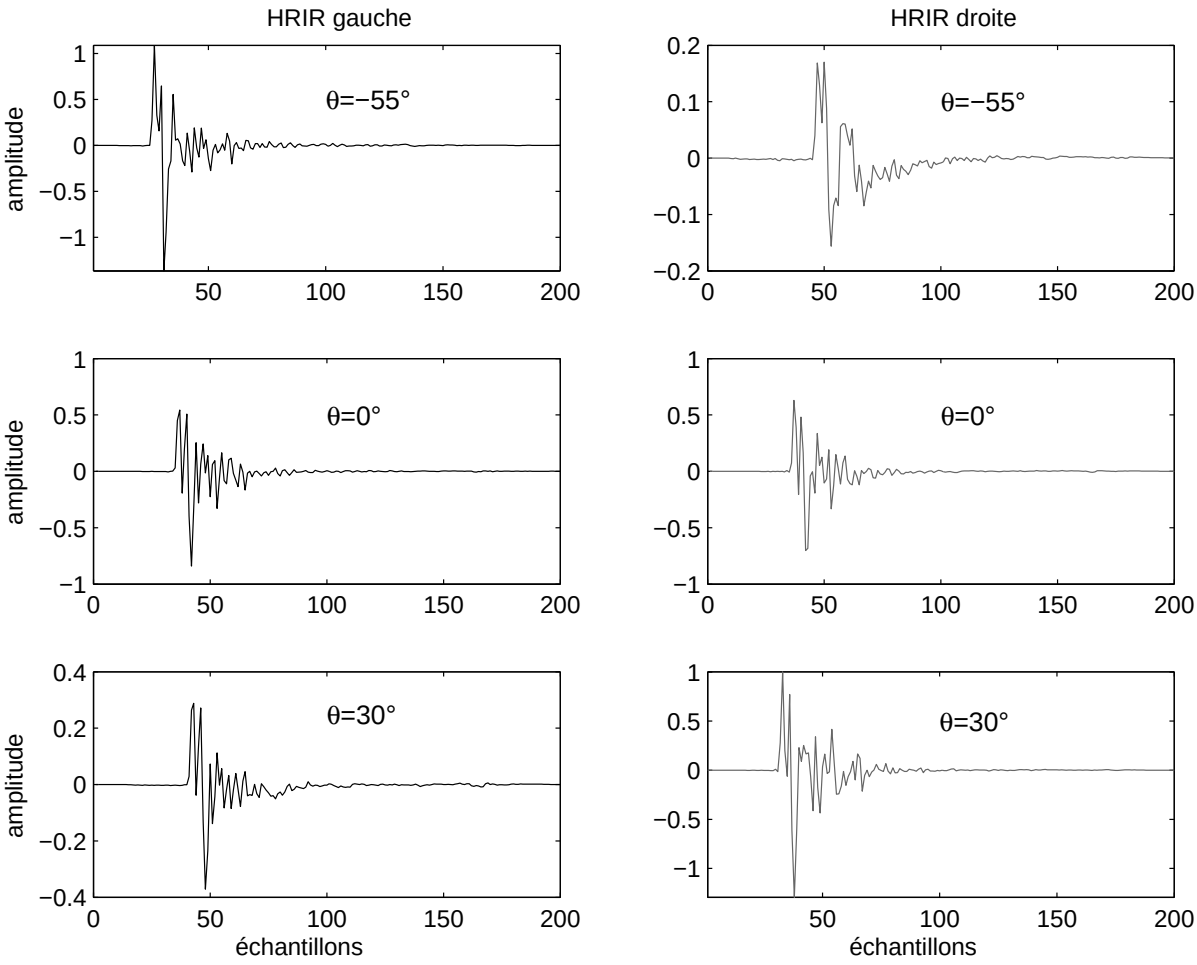


FIG. 15: *HRIR pour le sujet* 12 *de la base de HRIR CIPIC pour plusieurs directions dans le plan horizontal. Oreille gauche (gauche) ; oreille droite (droite).*

où $\alpha(f)$ est le facteur d'échelle moyen qui optimise le modèle sur toute la base CIPIC selon une méthode de moindres carrés pour chaque sujet de la base CIPIC (voir figure 16). L'erreur globale sur la base CIPIC pour tous les sujets, tous les azimuts et toutes les fréquences est d'environ 4.29 dB. L'erreur du modèle moyen et la variance entre les sujets sont illustrées sur la figure 17. L'ILD est une fonction de la fréquence, et sa variance est importante, particulièrement pour les fréquences au-delà de 7 kHz. Plus tard, nous étudierons son impact sur la localisation et la spatialisation de sources sonores.

Ce modèle sinusoïdal d'ILD sera considéré dans la suite de cette thèse.

Modèle de l'ITD

Sur des considérations géométriques sur une tête purement sphérique, Woodworth et Schlosberg [Woo54] ont développé un modèle analytique indépendant de la fréquence. Ce modèle prédit l'ITD en fonction de l'azimut et du rayon de la tête, en analysant le chemin d'une onde plane qui se propage autour de la tête. Sous la condition que la distance de la

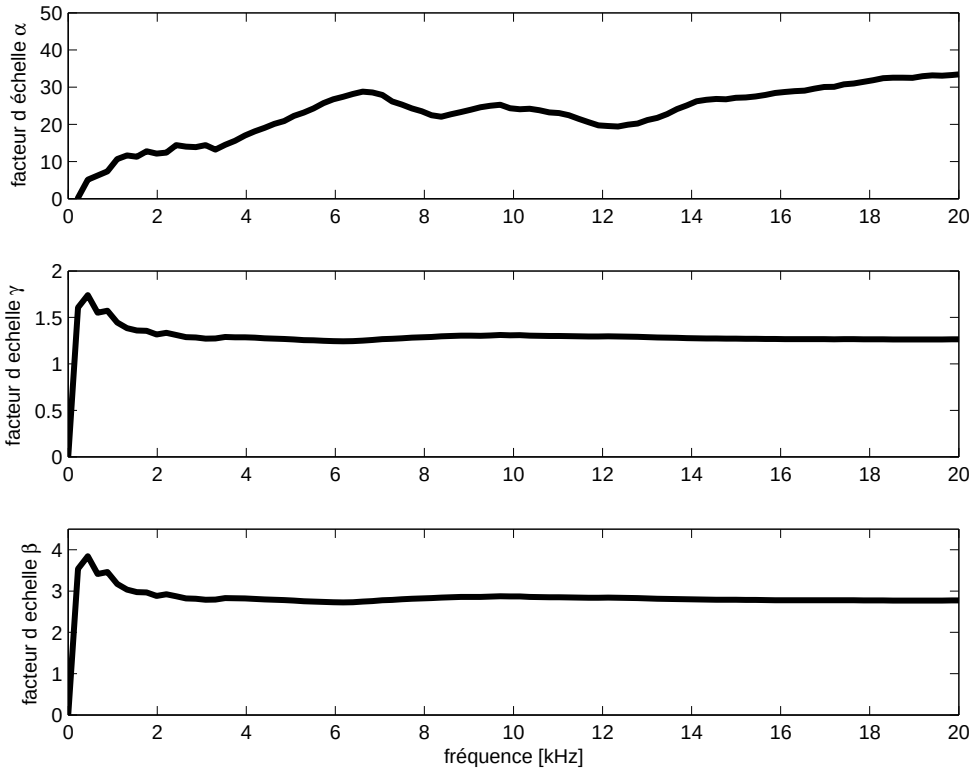


Fig. 16: *Facteurs d'échelle fréquentiels : $\alpha(f)$ (pour modèle ILD Viste), $\gamma(f)$ (pour modèle ITD Viste) et $\beta(f)$ (pour modèle simplifié).*

source est relativement large par rapport au diamètre (d) de la tête de l'auditeur, c'est-à-dire $\rho \gg d$. Les distances de la source et les oreilles gauche et droite sont respectivement $\rho_G = r\theta$ et $\rho_D = r \sin\theta$ (voir figure 18). D'où la différence de temps donnée par :

$$\mathrm{ITD}(\theta) = \frac{\rho_G - \rho_D}{c} = \frac{r(\sin\theta + \theta)}{c}, \tag{36}$$

où r dénote le rayon de la tête, et c est la célérité du son dans le milieu ambiant.

Des tests menés par Nordlund [Nor62] sur ce modèle avec des clics à différentes fréquences montrent une dépendance à la fréquence. Aussi, Wightman et Kistler [WKFA97] constatèrent que l'ITD est plus grande pour les basses fréquences. De plus l'ITD serait plus large, car la tête n'est pas réellement sphérique et les oreilles sont positionnées un peu vers l'arrière [DAA99]. Viste intègre ces précisions et étend le modèle avec un *facteur d'échelle fréquentiel* $\gamma(f)$.

La formule analytique du modèle devient ainsi :

$$\mathrm{ITD}(\theta, f) = \gamma(f)\frac{r(\sin\theta + \theta)}{c}, \tag{37}$$

où $\gamma(f)$ est le facteur d'échelle moyen qui optimise le modèle sur toute la base CIPIC selon une méthode de moindres carrés pour chaque sujet de la base (voir figure 16). L'erreur globale sur la base CIPIC pour tous les sujets, tous les azimuts et toutes les fréquences est d'environ 0.045 ms.

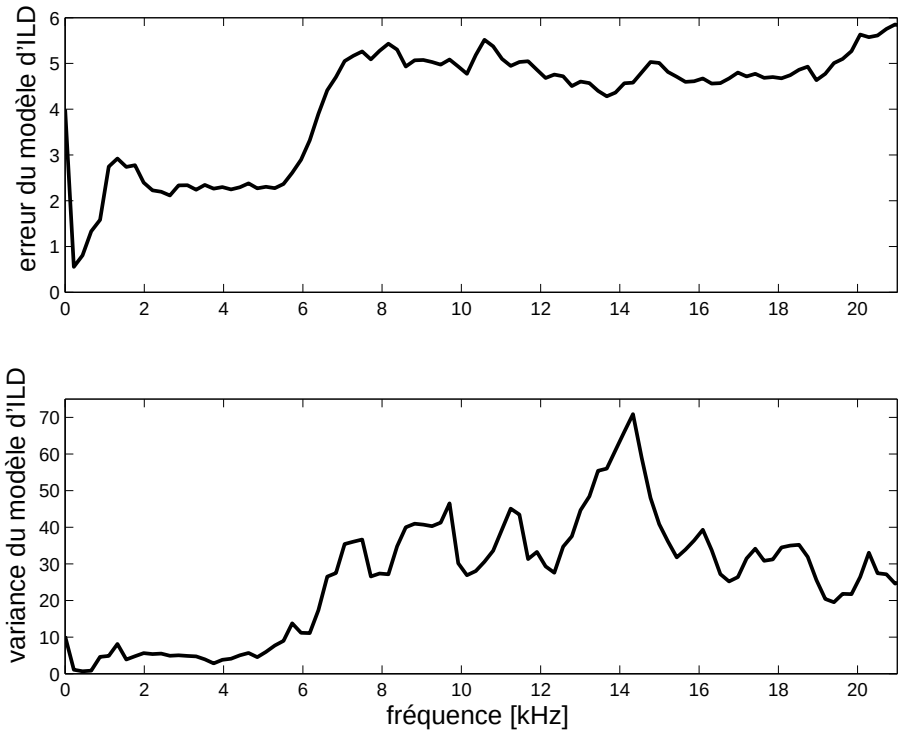


FIG. 17: *Erreur du modèle moyen d'ILD (haut) et variance inter-sujet (bas) sur toute la base CIPIC.*

2.3.2 Modèle d'ITD de George F. Kuhn

En se basant sur des études de la diffraction d'une onde plane sur une sphère [Rzh63], [Sch43], Kuhn [Kuh77] proposa un modèle sinusoïdal de prédiction de l'ITD dépendant de la fréquence ; il s'agit plus précisément de deux modèles indépendant de la fréquence pour les basses et les hautes fréquences. Le modèle répond aux formules suivantes :

$$\mathrm{ITD}(\theta, f) = \begin{cases} 3\frac{r}{c}\sin(\theta) & f < 1.5\mathrm{kHz} \\ \\ 2\frac{r}{c}\sin(\theta) & f > 3\mathrm{kHz} \end{cases}$$

2.3.3 Modèle sinusoïdal simplifié d'ITD

Modèle sinusoïdal simplifié

Nous proposons un modèle sinusoïdal simplifié et algébriquement inversible qui réduit la complexité mathématique et qui prend en compte les variations entre les sujets de la base CIPIC. Ce modèle d'ITD est exprimé par :

$$\mathrm{ITD}(\theta, f) = \beta(f) r \sin(\theta)/c \tag{38}$$

où $\beta(f)$ est le facteur d'échelle moyen qui optimise le modèle sur toute la base CIPIC selon une méthode de moindres carrés pour chaque sujet de la base CIPIC database (voir

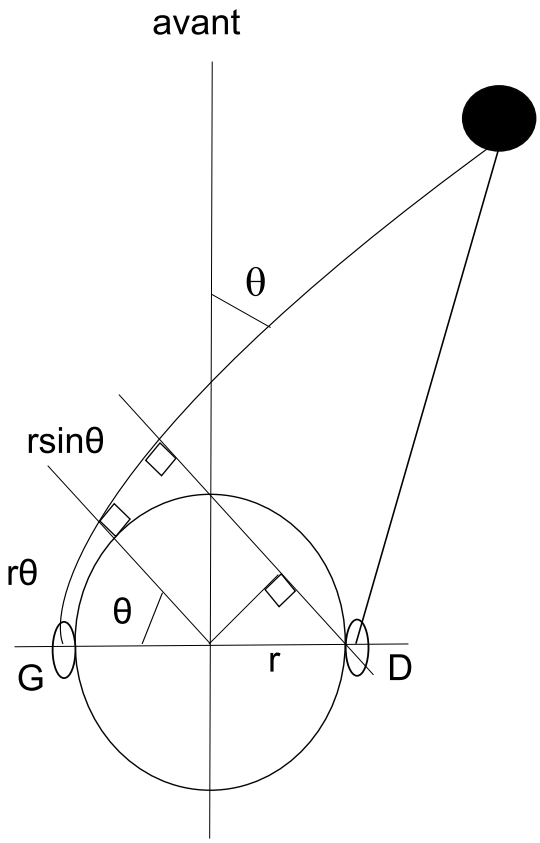


Fig. 18: *La différence en temps d'arrivée (ITD) dépend de l'angle d'incidence de la source. La distance entre l'oreille gauche et le centre de la tête est $r\theta$, de même la distance entre l'oreille droite et le centre de la tête est $r\sin\theta$, la distance entre les deux oreilles est donc $r\theta + r\sin\theta$.*

figure 16). L'erreur globale sur la base CIPIC pour tous les sujets, tous les azimuts et toutes les fréquences est d'environ 0.052 ms. L'erreur du modèle moyen et la variance entre les sujets sont illustrées sur la figure 19.

Discussion sur les modèles binauraux

Une analyse des modèles d'ITD permet de déterminer le modèle le mieux adapté à nos buts scientifiques et informatiques. Contrairement au modèle de Kuhn qui omet la bande de fréquence allant de 1.5kHz à 3kHz dans sa modélisation, le modèle de Viste prend en compte non seulement la dépendance fréquentielle, mais aussi la variation entre les sujets, sur toute la bande fréquentielle audible. Toutefois en terme de complexité mathématique, le modèle proposé par Kuhn est facilement inversible en vue de l'obtention de l'azimut à partir de l'ITD, alors que le modèle de Viste basé sur un noyau $\sin\theta + \theta$ nécessitant une approche analytique plus élaborée. A cet effet, nous avons proposé une approximation polynomiale $\Pi(x)$ d'ordre 5 de l'inverse $\sin(\theta) + \theta$ [MM06] :

$$\theta_{T,p}(t,f) = \Pi\left(\frac{c \cdot \mathrm{ITD}_p(t,f)}{r \cdot \beta_f}\right) \quad \text{avec}$$
$$\Pi(x) = 0.50018\ x + 0.009897\ x^3 + 0.00093\ x^5 + O(x^5). \tag{39}$$

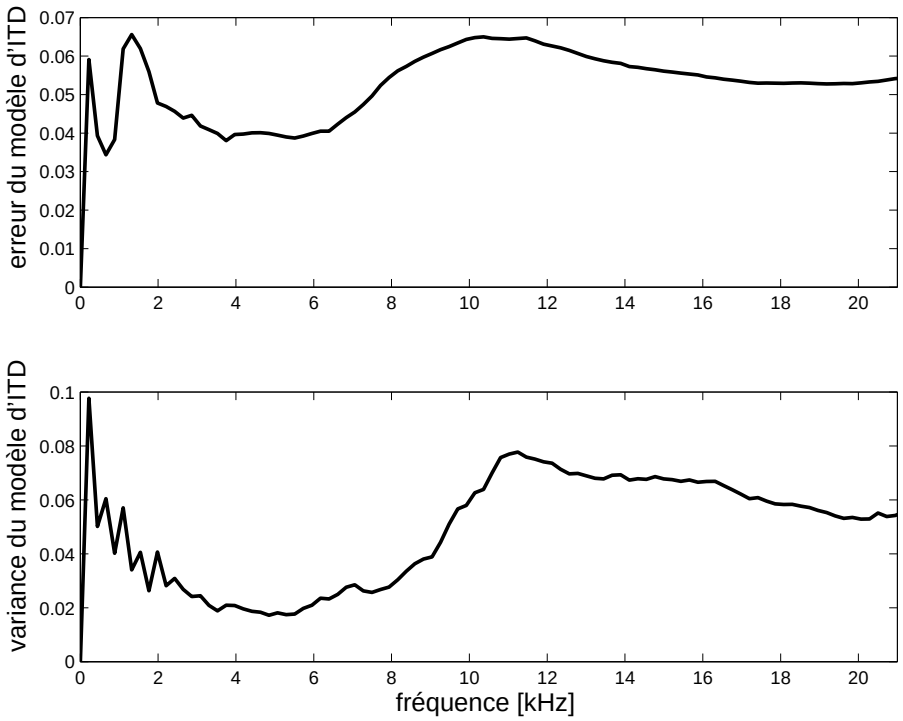


FIG. 19: *Erreur du modèle moyen d'ITD (haut) et variance inter-sujet (bas) sur toute la base CIPIC.*

Modèle de l'ITD	Erreur globale sur la base CIPIC (ms)
Viste – $\sim (\sin\theta + \theta)$	0.045
Simplifié – $\sim \sin\theta$	0.052

TAB. 3: *Erreur globale de différents modèles d'ITD sur la base CIPIC.*

Nous montrons également que l'erreur d'approximation de l'inverse de $\sin\theta+\theta$ croît avec la valeur absolue de l'ITD. Dans [RE08], les auteurs utilisent des séries de Chebyshev et proposent l'expression simplifiée $\Pi(x) = \frac{x}{2} + \frac{x^3}{96} + \frac{x^5}{1280}$ soit environ $0.5000\ x + 0.0104\ x^3 + 0.0008\ x^5$.

Le modèle sinusoïdal simplifié que nous avons proposé répond à plusieurs attentes. Il est plutôt un modèle de Kuhn étendu et rassemble les atouts des modèles de Viste et de Kuhn. En effet, le modèle simplifié considère la disparité entre les sujets, tout en essayant de la minimiser, et il est facilement inversible. Les techniques de localisation basée sur l'ITD estimé seraient plus efficaces. L'erreur globale sur la base CIPIC pour tous les sujets, tous les azimuts et toutes les fréquences est d'environ 0.052 ms, ce qui s'avère comparable à l'erreur globale du modèle de Viste 0.045 ms (voir figure 20). Les facteurs d'échelle du modèle de Kuhn dévient considérablement de la moyenne de la base CIPIC, son erreur est de ce fait supérieure à celle du modèle simplifié. En effet, rien que dans les hautes fréquences, le modèle de Kuhn propose un facteur d'échelle constant valant 2, alors que celui de la base CIPIC vaut environ 2.8.

Le modèle sinusoïdal simplifié d'ITD que nous avons proposé est à juste raison considéré dans la suite de cette thèse.

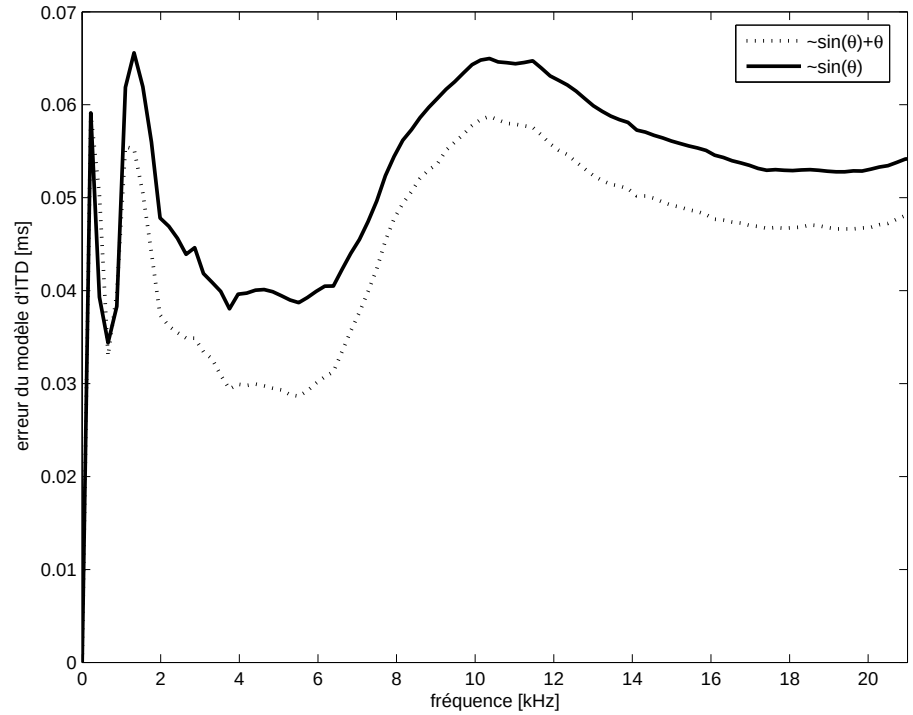


Fig. 20: *Erreurs moyennes des modèles d'ITD en fonction de la fréquence, modèle* $\alpha(f)\sin(\theta)$ *(plein), modèle* $\gamma(f)\sin\theta + \theta$ *(pointillés).*

2.3.4 Détermination des facteurs d'échelle fréquentiels

Estimation des ILD et d'ITD

A partir des spectres des HRTF des canaux gauche (H_G) et droite (H_D) pour une position θ, nous pouvons estimer les indices acoustiques binauraux (ILD et ITD) à chaque composante fréquentielle indexée par f.

L'ILD est donnée par :

$$\mathrm{ILD}(\theta, f) \quad = \quad 20\log_{10}\left|\frac{\mathrm{H}_G(\theta, f)}{\mathrm{H}_D(\theta, f)}\right|. \tag{40}$$

C'est le rapport des amplitudes des points des spectres des canaux gauche et droite exprimé en décibel. La figure 22 montre des mesures d'ILD à différents azimuts.

L'ITD exprimée en secondes est donné par :

$$\mathrm{ITD}_p(\theta, f) \quad = \quad \frac{1}{2\pi f} \cdot \Delta\mathrm{P}_p(\theta, f), \tag{41}$$

où $\Delta\mathrm{P}_p(\theta, f)$ est la différence de phase donnée par :

$$\Delta \mathrm{P}_p(\theta, f) \quad = \quad \angle \frac{\mathrm{H}_G(\theta, f)}{\mathrm{H}_D(\theta, f)} + 2\pi p(f). \tag{42}$$

Le coefficient p met en perspective le fait que la phase est déterminée à un facteur modulo 2π près. Pour $p = 0$, on obtient la position la plus proche de l'azimut zéro. Pratiquement, la phase devient ambiguë au-delà de 1500Hz, lorsque la longueur d'onde devient plus petite que le diamètre de la tête.

De ce fait, la différence de phase n'est pas toujours suffisante pour être transformée en différence réelle de temps. En réalité, l'ITD est quasi-indépendante de la fréquence, car $\beta(t) \simeq 3$ (figure 16), donc $\Delta \mathrm{P}_p(\theta, f) = 2\pi f \cdot \mathrm{ITD}_p(\theta, f)$ est quasi-linéaire en fréquence, mais mesurée modulo 2π dans le spectre, d'où la nécessité de dérouler, pour retrouver la droite.

Il s'agit de s'assurer que les multiples de $2\pi p(f)$ appropriés sont ajoutés à la différence de phase, afin d'obtenir un ITD quasi-indépendant de la fréquence (voir équation 38). La technique que nous avons utilisée consiste à ajouter ou à soustraire 2π à la phase $\Delta \mathrm{P}_p(\theta, f)$ à chaque fois que $|\Delta \mathrm{P}_p(\theta, f + \Delta(f)) - \Delta \mathrm{P}_p(\theta, f)|$ est plus grand que π, avec $\Delta(f)$ la distance entre deux fréquences successives du spectre. La figure 21 montre l'effet du déroulement de la phase sur les réponses de phase des HRTF à différents azimuts. Très souvent, il est nécessaire que la réponse de la phase ait la propriété $\Delta \mathrm{P}_p(\theta, f = 0) = 0$. La différence de phase déroulée est :

$$\Delta \hat{\mathrm{P}}(\theta, f) \quad = \quad deroule(\Delta \mathrm{P}_p(\theta, f)). \tag{43}$$

Sous certaines conditions, la présence de bruit parasite occasionne un décalage de la phase ou "phase offset". Afin de prévoir les cas de offset intrus où $(\Delta \mathrm{P}_p(\theta, 0) = \delta)$ au lieu de $\Delta \mathrm{P}_p(\theta, 0) = 0$, la phase du premier casier du spectre de phase est retranchée à toutes les positions fréquentielles selon :

$$\Delta \tilde{\mathrm{P}}(\theta, f) \quad = \quad \Delta \hat{\mathrm{P}}(\theta, f) - \Delta \hat{\mathrm{P}}(\theta, 0) \tag{44}$$

Cette dernière différence de phase est utilisée afin de déterminer l'ITD (voir équation 38).

La figure 22 montre des mesures d'ITD à différents azimuts. Nous remarquons que les estimations sont suffisamment précises. A titre illustratif, la tendance quelque peu aléatoire de la différence de phase à l'azimut zéro ne semble pas avoir des répercussions significatives sur les estimations de l'ITD pour les hautes fréquences. Pour les autres azimuts, l'ITD est stable, ce qui illustre une fois de plus sa faible variance, alors que l'ILD devient instable, d'où sa large variance.

Facteur d'échelle fréquentiel

Considérons un azimut θ, une fréquence f et un sujet. À partir des modèles d'ILD et ITD des équations 35 et 38, on obtient les facteurs d'échelle fréquentiels $\alpha(f)$ et $\beta(f)$ à partir des équations :

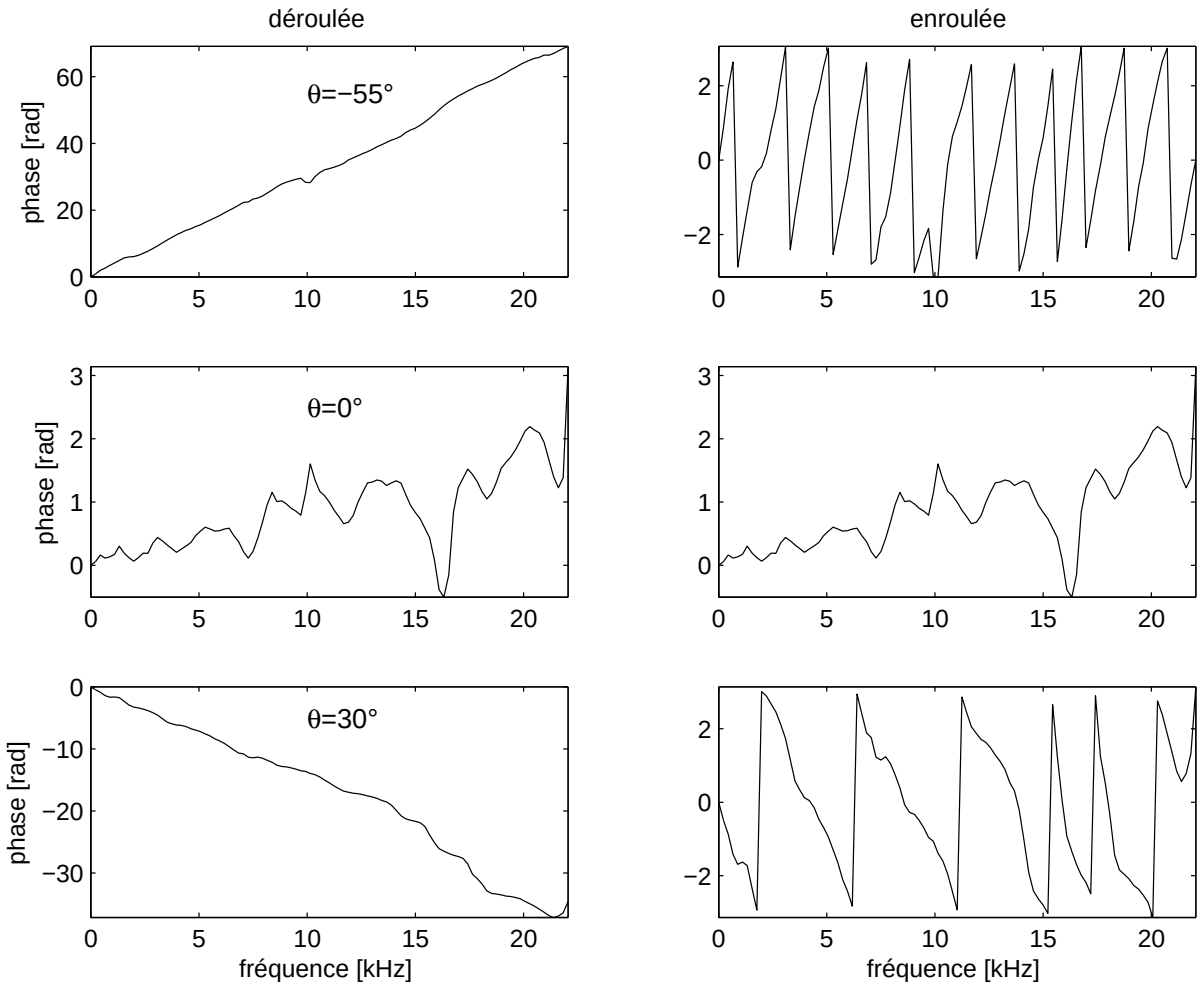


Fig. 21: *Différence de phase pour le sujet* 12 *de la base CIPIC pour plusieurs directions dans le plan horizontal. Différence de phase déroulée avec la fonction unwrap (gauche), différence de phase non déroulée (droite).*

$$\alpha(f) = \mathrm{ILD}(f)/\sin(\theta), \tag{45}$$

$$\beta(f) = \frac{c}{r \cdot \sin(\theta)} \cdot \mathrm{ITD}(f). \tag{46}$$

Ainsi, on obtient une fonction d'échelle sur la bande fréquentielle considérée.

Pour chaque sujet i, des fonctions d'échelle sont obtenues pour chaque azimut.

Facteur d'échelle fréquentiel individuel indépendant de l'azimut

Pour un sujet donné, le but est de trouver la fonction d'échelle fréquentielle optimale pour toutes les positions angulaires mesurées $(\theta_1, \cdots, \theta_N)$. Notre approche consiste en la méthode

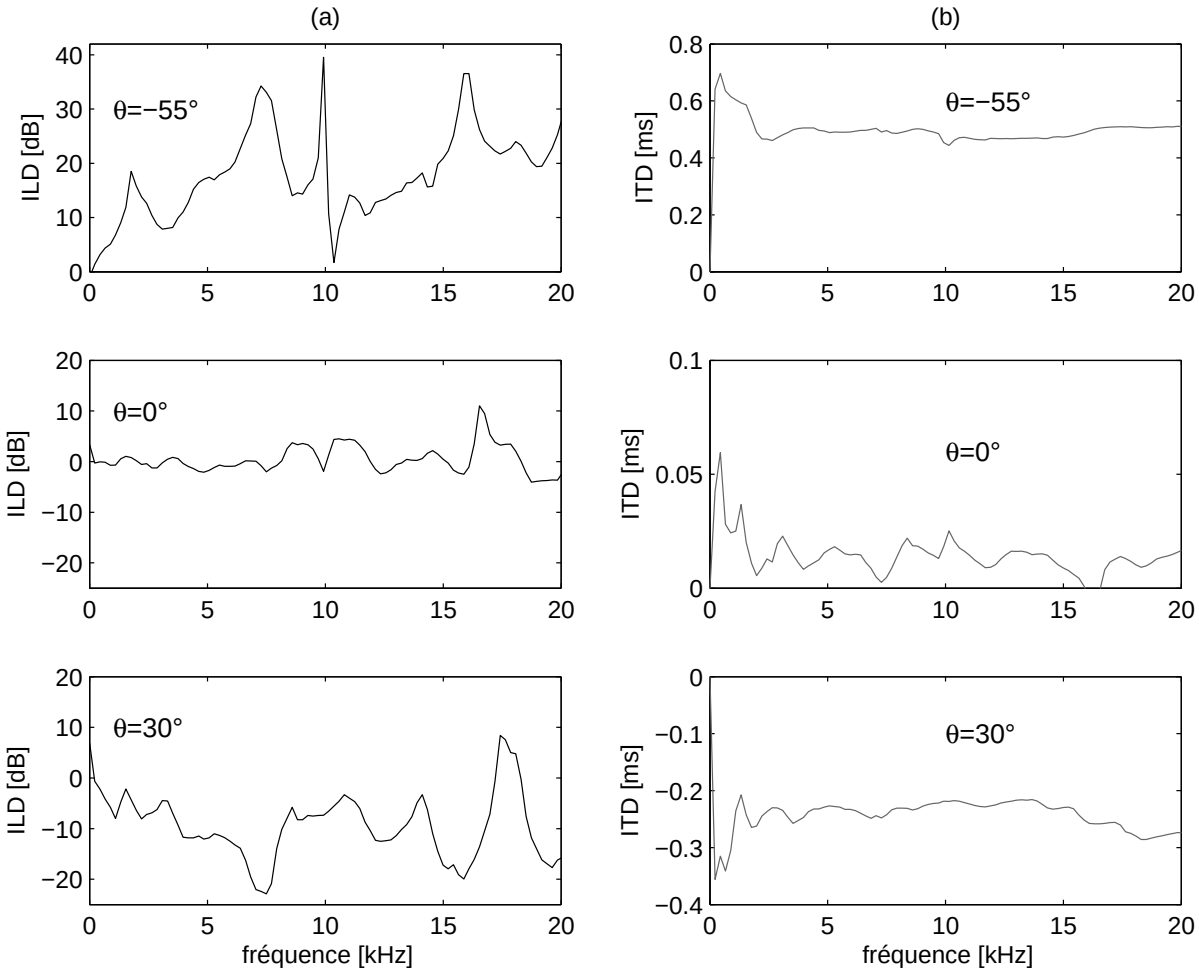


FIG. 22: *Indices acoustiques pour le sujet* 12 *de la base CIPIC pour plusieurs directions dans le plan horizontal. Différence interaurale en amplitude (a), différence interaurale en temps d'arrivée (b).*

des moindres carrés, de sorte que l'erreur quadratique entre les mesures prédites par le modèle et celles mesurées soit minimisée à tous les azimuts. Les quantités à minimiser sont :

$$S(\alpha, f) = \sum_{\theta_i}^{\theta_N} (\mathrm{ILD}(\theta_i, f) - \alpha(f)\sin(\theta))^2, \tag{47}$$

$$S(\beta, f) = \sum_{\theta_i}^{\theta_N} (\mathrm{ITD}(\theta_i, f) - \beta(f) r \sin(\theta)/c)^2. \tag{48}$$

Par une méthode d'approximation de moindres carrés, on obtient ainsi une fonction d'échelle fréquentielle indépendante de l'azimut, optimisée sur le modèle paramétrique pour un sujet donné. Le modèle paramétrique pour tous les 45 sujets de la base de HRIR CIPIC est la moyenne des fonctions d'échelle individuelles.

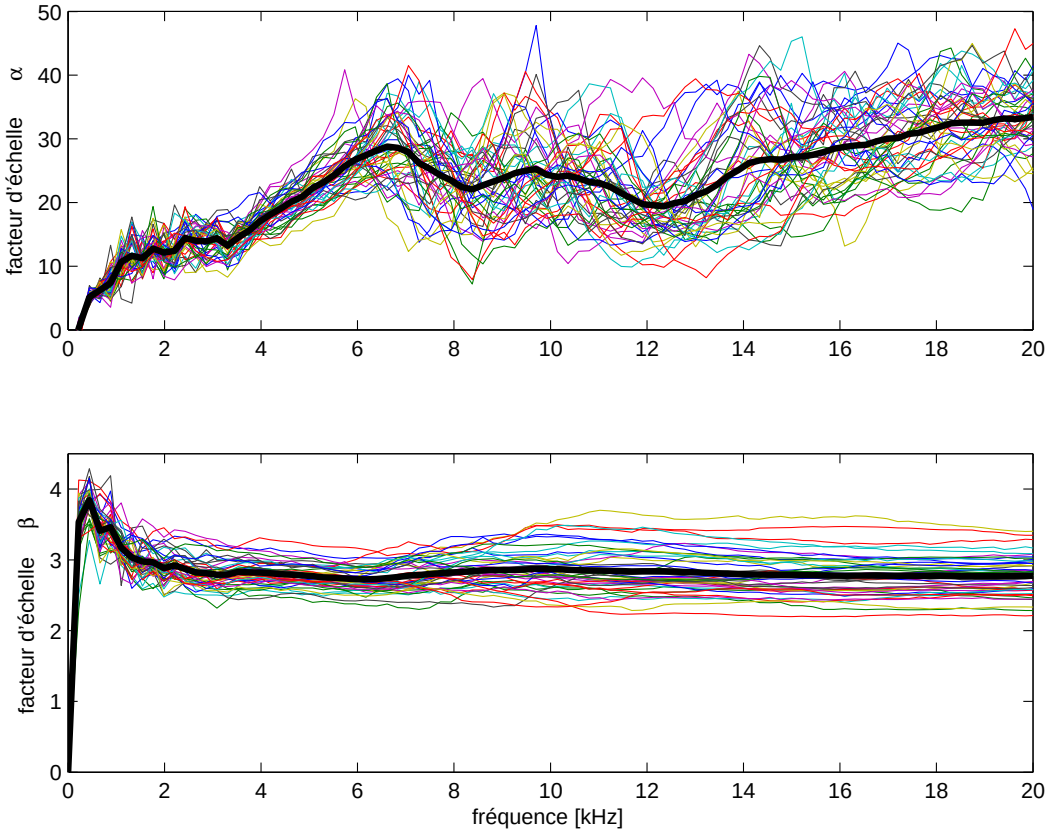


Fig. 23: *Fonctions d'échelle fréquentielles pour tous les sujets, la fonction d'échelle moyenne est en noir épais : $\alpha(f)$ (haut) and $\beta(f)$ (bas).*

La variance inter-sujet pour respectivement $\alpha(f)$ et $\beta(f)$ est représentée dans les figures 17 et 19. On remarque que la variance bondit à partir de 8 kHz. Dans les chapitres suivants, nous analyserons l'impact de la variance sur la localisation et la spatialisation binaurale.

Chapitre 3

Spatialisation de source

Dans la musique électro-acoustique, la spatialisation (ou mise en espace) des sonorités qui composent la pièce musicale améliore l'expérience et la commodité de l'écoute. Plus précisément, la spatialisation consiste à choisir la position des sources dans l'espace, et à choisir l'acoustique du lieu de projection.

Sans la prétention de rendre la spatialisation à l'aide de la table de mixage obsolète, l'objectif principal de nos travaux est d'identifier et de proposer des techniques appropriées à la diffusion de musique électro-acoustique, de sorte que ces dernières deviennent disponibles sous la forme d'un logiciel d'informatique musicale. Nous nous intéressons d'une part à la spatialisation dans des espaces ouverts, et à la spatialisation dans des emplacements fermés tel qu'un amphithéâtre.

Dans la section 3.1, nous relatons des événements mémorables de l'histoire de la spatialisation naturelle, de la spatialisation électro-acoustique et informatique. Un panorama des techniques existantes est exposé dans la section 3.2, des techniques binaurales aux techniques panoramiques à potentiomètre. Nous mettons également l'accent sur l'adaptation des techniques à différentes configurations de haut-parleurs. Dans la section 3.3.2, nous présentons une nouvelle technique de spatialisation binaurale paramétrique en azimut et en distance, qui s'appuie sur le modèle binaural paramétrique développé dans nos recherches. Il s'agit de reproduire les indices acoustiques des signaux parvenant aux oreilles par le biais d'un casque audio. Dans la section 3.3.3, nous étendons la spatialisation paramétrique binaurale à un système multi-diffusion capable de positionner efficacement plusieurs sources concurrentes dans l'espace. Ce système est destiné à la diffusion dans des salles à l'aide d'un système de plusieurs haut-parleurs, et fonctionne par paire d'enceintes autour de l'audience.

La complexité des méthodes proposées jouent un rôle décisif dans nos modèles, car les algorithmes sont implantés dans le logiciel d'informatique musical (projet RetroSpat) pour des performances en temps réel. Les techniques de spatialisation binaurale paramétrique ont été éprouvées subjectivement et objectivement afin d'évaluer leur réalisme spatial et leur qualité sonore dans la mise en espace (section 3.4).

3.1 Événements historiques clefs de la spatialisation

La spatialisation est aujourd'hui accessible au grand public, notamment avec le système Dolby 5.1 à 6 haut-parleurs dans le salon du particulier. Cet enthousiasme est le résultat d'une longue marche historique. Cette section, sans la prétention d'être exhaustive, retrace les événements marquants de l'intégration de l'espace dans les œuvres musicales [PB04] [Dut06]. Nous distinguons principalement deux types de spatialisation. Premièrement, la spatialisation acoustique naturelle qui est directement contrainte par la structure du lieu, l'effet de salle et la disposition physique des sources. Deuxièmement, la spatialisation acousmatique virtuelle qui est soutenue par les technologies sonores, et permet de synthétiser des sons ne coïncidant pas à des emplacements de sources réelles.

3.1.1 Spatialisation réelle (acoustique)

Les musiciens ont généralement essayé d'exploiter des phénomènes en échos, des plans distants ou encore la profondeur sonore. La première idée de mise en perspective du son dans les œuvres musicales date du IV^{e} siècle avec l'antiphonie, qui consistait à faire dialoguer deux groupes vocaux. Cette empreinte a marqué les liturgies chrétiennes. Dans le motet à quarante voix (1573), Thomas Tallis dispose en fer à cheval huit groupes de cinq voix à savoir soprano, alto, ténor, baryton, basse. Il remplit l'espace de chanteurs et plonge l'audience dans une fabuleuse apothéose avec un son qui se déplace du premier chœur au huitième.

Toujours au XVI^{e} siècle, les Gabrieli réussissent à intégrer des réponses en échos grâce aux deux tribunes opposées de la basilique Saint-Marc de Venise [Lou97]. Pour cela, ils abritent deux ensembles "orgue et chœur" sur chaque tribune. Ils introduisent un effet stéréo qui marque l'époque vénitienne.

Jean-Sébastien Bach dans *La passion selon Saint Matthieu* (1789) occupe les deux tribunes positionnées sur la longueur à Saint Thomas de Leipzig. Il utilise deux chœurs d'adultes pour la gauche et la droite, et en plus un chœur d'enfants à l'arrière du public. Il crée ainsi une sensation d'enveloppement de l'audience.

Dans son *Requiem*, Berlioz dispose quatre groupes de cuivre aux quatre points cardinaux, en plus d'un orchestre et d'une chorale impressionnante [Lou97]. Il ouvre ainsi une porte à un système de quadriphonie dès 1837. Il intègre la vitesse de rotation du son en parvenant à des effets de rotation en spirale avec des fanfares. Plus encore dans *Les troyens*, le compositeur expérimente la spatialisation par la distance en disposant trois orchestres à différentes distances. Wilhelm Richard Wagner disposa ses instruments sur 6 estrades pour simuler la distance.

Dans son morceau intitulé *Gruppen* (1955), Karlheinz Stockhausen divise ses 109 musiciens en 3 orchestres de même taille, et les dispose en fer à cheval. Il produit l'illusion de sons en mouvement dans l'espace. L'interprète se retrouve comme au milieu du dispositif.

Certains musiciens vont plus loin, et rompent avec la frontalité des orchestres et mêlent les musiciens au public. En 1966, Xenakis dans *Terretêktorh* dissémine 88 musiciens en huit groupes dans le public [Bra03]. En 1974, dans *Rituel in memoriam Bruno Maderna*, Pierre Boulez éclate huit groupes d'instruments circulairement autour du public. Emmanuel Nunes pousse l'audace jusqu'à placer les instrumentistes au milieu des auditeurs dans l'œuvre *Quodlibet* (1989-1991).

La spatialisation a permis aux œuvres musicales d'accéder à une nouvelle dimension d'enveloppement de l'audience, et ravît un grand public ; remarquons toutefois que la spatialisation dans les exemples précités est contrainte par l'acoustique de la salle, et la position des sources, mais aussi des positions atteignables par les sources. Les sons proviennent uniquement des sources réelles et la spatialisation s'en trouve limitée. L'électro-acoustique et l'informatique vont révéler les dimensions des sources virtuelles, qui ne correspondent pas avec les positions des sources réelles. Alors, une nouvelle page de la spatialisation s'ouvre.

3.1.2 Spatialisation virtuelle (acousmatique)

Au XX^{e} siècle, avec le développement de moyens électro-acoustiques (microphones, haut-parleurs, tables de mixage), les musiciens explorent de nouvelles géométries spatiales. Les acousmaticiens exploitent un orchestre de haut-parleurs, chacun avec des couleurs différentes tels que l'Acousmonium de Paris élaboré en 1974 par François Bayle. Contrairement à des musiciens ou des instruments (sources réelles), les haut-parleurs peuvent facilement être déplacés et même être accrochés sur les murs et au plafond. Ainsi les positions peuvent être démultipliées et variées. Des nouvelles configurations de haut-parleurs deviennent possibles pour le grand plaisir des créateurs et des spectateurs.

Dans le film *Fantasia* de Walt Disney en 1939, une trentaine de microphones fût utilisée afin d'enregistrer l'orchestre pour une diffusion spatiale à base de quatre-vingt dix haut-parleurs autour des spectateurs. Plus encore, Edgard Varèse dans son *poème électronique* mit en œuvre plus de quatre cents haut-parleurs à l'exposition universelle de Bruxelles (1958).

Dans *Gesang der Jünglinge* en 1956, Stockhausen utilise cinq groupes de haut-parleurs distribués autour et au-dessus des auditeurs pour exécuter des mouvements en rotation. Et dans *Kontakte* (1958-1960), il utilise une table tournante munie de quatre microphones qui captent des sources dynamiques circulaires et reproduit astucieusement ces sources avec quatre haut-parleurs pour créer un mouvement circulaire autour des auditeur. Les sources dynamiques fascinent, avec *Turenas* (1977), John Chowning parvient à intégrer l'effet Doppler pour simuler la position et le mouvement de sources réelles [Cho71].

Avec le développement technologique, de nombreuses techniques de spatialisation permettent de simuler des sources dynamiques en utilisant un nombre de haut-parleurs limité, et même un casque audio.

3.2 Techniques de spatialisation

Les techniques de spatialisation modernes s'organisent autour d'une structure commune qui s'étend de la source audio au récepteur final (oreille). Cette chaîne source vers récepteur est constituée de plusieurs éléments qui influencent la qualité de la spatialisation. En amont, nous avons divers sons enregistrés (ou synthétisés) à l'aide de microphones dans un environnement naturel ou artificiel. Les sons sont ensuite mixés à l'aide d'une table de mixage ou d'un ordinateur avec des algorithmes de traitement du signal. C'est à cette étape que des effets de salle peuvent être intégrés artificiellement, et surtout que les algorithmes de spatialisation sont appliqués à plusieurs sources monophoniques. Le but directeur est de reproduire les conditions d'écoute naturelle afin de reconstituer les indices de localisation, et donner l'illusion spatiale

souhaitée au système auditif humain. Parmi les premiers systèmes de diffusion spatiale, on pourrait noter le *Multiplex Gramophone Grand* à trois cors de Columbia (1800) qui jouait trois airs différents. Cependant, ces derniers étaient monophoniques et n'exploitaient que les indices de distance et de profondeur. La première transmission stéréo fut réalisée à l'exposition de Paris en 1881 par Clément Ader. Ce dernier utilisa deux microphones fixés à la rampe de l'opéra de Paris et raccorda les sorties à une paire de microphones pour téléphone, et permit ainsi au public de l'écouter en direct avec un effet stéréo.

C'est dans les années 1930 aux laboratoires Bells que la recherche dans le domaine de la spatialisation prend une autre dimension. Notamment avec la synthèse binaurale dont la diffusion est généralement adaptée à un casque audio (section 3.2.1). Alors que les techniques transaurales veulent étendre l'auditoire en visant une reproduction sur une paire d'enceintes [SS34] (section 3.2.2). La synthèse panoramique, généralement connue sous l'appellation stéréophonie permet de spatialiser un son entre deux haut-parleurs en intégrant des différences d'amplitude et de temps aux signaux diffusés [Bau61] [Ber75] [Mak62] (section 3.2.3). Cette approche a permis de développer des systèmes à plusieurs haut-parleurs comme le fameux Vector-Base Amplitude Panning (VBAP) [Pul99] qui permettent la diffusion dans de grandes salles avec plusieurs enceintes autour de l'auditoire (section 3.2.4).

Nous analysons un certain nombre de techniques qui ont marqué la spatialisation, et qui furent le point de départ de nos réflexions et de nos propositions. La recherche d'un système adéquat est centré autour de facteurs déterminants tels que sa complexité, sa transportabilité, son adaptabilité et sa mise en œuvre. Nous allons aussi explorer des systèmes multi-canal aux approches physiques tel que l'Ambisonic et l'holophonie (sections 3.2.5 et 3.2.6).

Dans la section 3.3.2, nous proposons une méthode paramétrique de synthèse binaurale, et dans la section 3.3.3, cette technique binaurale est étendue à une technique de diffusion multi-canal qui s'adapte à des environnements différents.

3.2.1 Synthèse binaurale

Un moyen simple de reproduction spatiale consiste à enregistrer une source monophonique à l'entrée des oreilles d'une tête humaine ou d'une tête artificielle. Les microphones sont enfoncés jusqu'aux tympans ou à l'entrée des oreilles [Mol92]. Les enregistrements binauraux recèlent les indices de localisation du sujet. Ils peuvent être mixés, et leur rediffusion dans un casque d'écoute assure une fidélité spatiale excellente (Jean-Michel Rivet, SCRIME). Une méthode d'encodage des indices de direction a été proposée par Jot : le Binaural B [JWL98]. Cette dernière nécessite des microphones à une capsule omnidirectionnelle et trois capsules directionnelles, ainsi que l'enregistrement de huit canaux. Une approche plus flexible déjà effleurée dans le chapitre précédent consiste à convoluer un signal monophonique avec les HRIR gauche et droite pour une position donnée, nous appelerons cette méthode SHRIR (Spatialization with HRIR). On peut alors synthétiser artificiellement un son 3D binaural. Une source dynamique est synthétisée en recourant continuellement à la paire d'HRIR appropriée à la nouvelle position. La discrétisation des mesures impose un arbitrage entre les positions à mesurer et les transitions. Il existe des solutions basées sur l'interpolation des HRTF [HBS99] et/ou la modélisation des HRTF [MM77].

La simplicité et la qualité sonore de la synthèse binaurale sont des atouts pour la diffusion dans un casque audio ; en revanche, le coût des mesures et des calculs pour la re-spatialisation

ont un impact direct sur la complexité de la mise en œuvre, car il faudrait enregistrer les HRTF pour toutes les positions cibles.

3.2.2 Synthèse transaurale

La diffusion binaurale est appropriée à une écoute individuelle à l'aide d'un casque audio. Dans le cas d'un public, il est souhaitable d'étendre la zone d'écoute, en transformant les signaux binauraux en ondes acoustiques diffusées sur des haut-parleurs (classiquement deux haut-parleurs en face de l'auditeur), de sorte que les signaux semblables à ceux issus de la perception naturelle soient reproduits au niveau des oreilles.

Dans une configuration stéréophonique avec deux haut-parleurs en face de l'auditeur, le signal binaural peut être transmis après quelques manipulations. Le son émis par chaque haut-parleur est perçu par chaque oreille, les indices spatiaux binauraux s'en trouvent modifiés. En effet, le son est filtré et mixé par deux paires de fonctions de transfert au cours de son voyage vers l'auditeur. Afin de délivrer les signaux corrects à chaque oreille, ce processus de mélange doit être inversé. A cet effet, une approche bien connue est l'*élimination des canaux croisés* ou *cross-talk canceller*, elle consiste à anticiper et à compenser les chemins croisés par un décodage matriciel. Les sons gauche et droit sont alors donnés par :

$$\begin{bmatrix} X_G \\ X_D \end{bmatrix} = \begin{bmatrix} H_{DD} & H_{GD} \\ H_{DG} & H_{GG} \end{bmatrix} \begin{bmatrix} G_{DD} & G_{GD} \\ G_{DG} & G_{GG} \end{bmatrix} \begin{bmatrix} H_G \\ H_D \end{bmatrix} X. \tag{49}$$

soit

$$\begin{bmatrix} X_G \\ X_D \end{bmatrix} = H \cdot G \begin{bmatrix} H_G \\ H_D \end{bmatrix} X. \tag{50}$$

où les H_j désignent les fonctions de transfert idéales de la source vers l'oreille j pour la position de spatialisation souhaitée (voir figure 24).

La matrice d'élimination des chemins croisés est idéalement donnée par :

$$G = \frac{1}{H_{GG}H_{DD} - H_{DG}H_{GD}} \begin{bmatrix} H_{DD} & -H_{GD} \\ -H_{DG} & H_{GG} \end{bmatrix}, \tag{51}$$

où les H_{ij} désignent les fonctions de transfert du haut-parleur i vers l'oreille j (voir figure 24). Il est connu que cette méthode n'est pas très robuste aux mouvements de la tête [WW99].

3.2.3 Synthèse panoramique à deux haut-parleurs

La configuration transaurale est la plus courante, elle consiste à spatialiser un son entre une paire de haut-parleurs. La reproduction se fait en contrôlant les différences en amplitude et en temps d'arrivée entre les deux haut-parleurs. Nous rappelons que dans une configuration stéréophonique, le son émis par chaque haut-parleur est perçu par les deux oreilles. Il est tout à fait possible de produire un effet de spatialisation en jouant uniquement sur les différences en amplitude. Jusqu'à environ 700 Hz, la somme des signaux des haut-parleurs sur chaque oreille résulte en un signal binaural dont l'ITD est proportionnel à l'ILD. Aux hautes fréquences, un contrôle perspicace des amplitudes des deux canaux peut permettre de simuler des différences

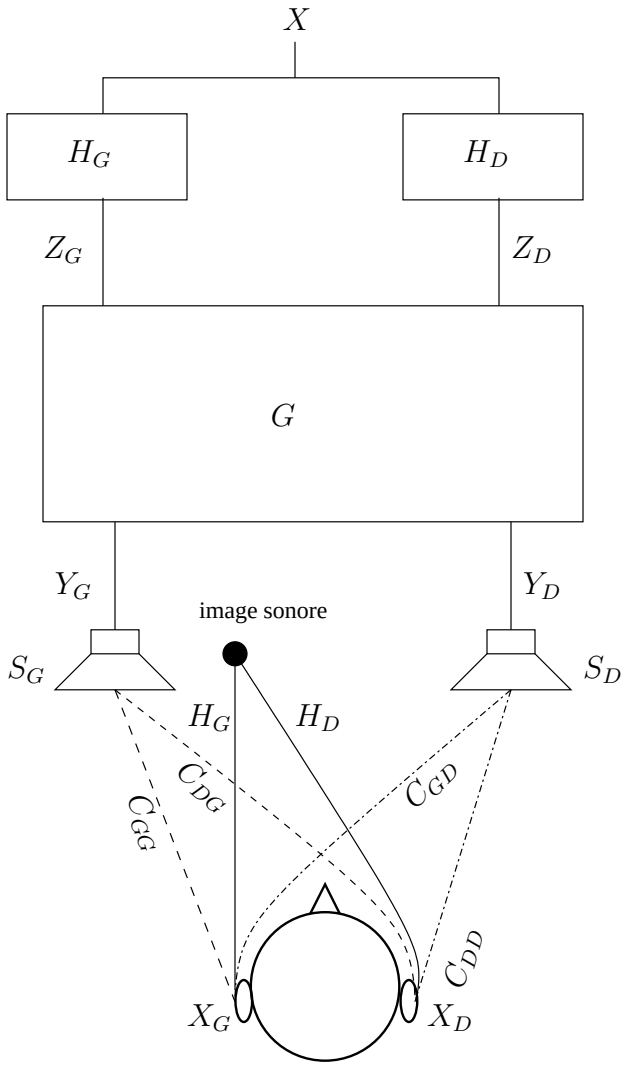


FIG. 24: *Configuration transaurale avec* cross-talk.

de phase similaires à celles de la perception naturelle. C'est le but de quelques techniques *panoramiques d'intensité* [The91].

Dans le cas de deux haut-parleurs, il s'agit de synthétiser pour une position θ, les signaux gauche et droite à diffuser sur le haut-parleur gauche et sur le haut-parleur droit. A partir d'un signal monophonique x, les signaux gauche et droit sont donnés par :

$$y_G = g_G \cdot x \quad (52)$$

$$y_D = g_D \cdot x \quad (53)$$

où g_G et g_D sont les coefficients de spatialisation pour les canaux gauche et droite.

Un précurseur de référence de ces techniques est la *loi du sinus* de Blumlein [Blu31] selon laquelle :

$$\frac{\sin\theta}{\sin\theta_0} = \frac{g_G - g_D}{g_G + g_D}, \tag{54}$$

où g_G et g_D sont les coefficients de spatialisation pour les canaux gauche et droit, θ est une estimation de l'azimut perçu de la source virtuelle, et θ_0 est l'angle de base entre l'enceinte et l'axe central, typiquement θ_0 vaut 30° (voir figure 25). Au risque de produire des anti-phases nuisibles à la continuité du signal, il est recommandé que l'angle de spatialisation (*panning angle*) soit limité selon $|\theta| \leq \theta_0$ et $0° < \theta_0 < 90°$. Blumlein considère les trajets comme des lignes droite. En considérant des trajets courbes autour de la tête, Bennet *et al.* proposèrent une amélioration de la *loi du sinus*, qui est la *loi de la tangente* [Moo90] selon laquelle :

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{55}$$

où θ est une estimation de l'azimut perçu de la source virtuelle.

Une technique de spatialisation bien connue est la loi de conservation des énergies, selon laquelle la somme des carrés des coefficients est égale à 1 quel que soit l'azimut. Ainsi, l'intensité du son est la même quel que soit l'angle, permettant ainsi de créer des sons dynamiques d'amplitude constante. De manière générale, afin d'assurer une sonie constante, les gains de spatialisation sont généralement normalisés :

$$\sqrt[p]{\sum_{n=1}^{n=N} g_i^p = 1.} \tag{56}$$

Typiquement, on choisit $p = 2$.

Les techniques transaurales ont connu un franc succès car elles sont adaptées aux petits espace de maison et elles sont facilement accessibles au grand public. Ces techniques de spatialisation ont permis de développer des systèmes panoramiques multi-canal.

3.2.4 Synthèse panoramique à plusieurs haut-parleurs

Dans les systèmes panoramiques à deux dimensions, les techniques transaurales de spatialisation sont étendues à plusieurs enceintes en utilisant le paradigme par paire classique [Cho71], qui consiste à choisir pour une source donnée, uniquement les deux haut-parleurs les plus proches de l'azimut cible : un sur la gauche et un sur la droite. . Par contre, lorsque la position escomptée correspond à la localisation d'un haut-parleur, ce haut-parleur est utilisé pour la spatialisation. C'est le cas dans des systèmes tels que la quadriphonie avec quatre enceintes à ±45° et ±135°. Dans les systèmes $m - n$ stéréo, on dispose de m enceintes à l'avant et de n enceintes sur les côtés et/ou à l'arrière, comme dans le Dolby 5.1 : 0°, ±30° et ±110°.

Dans des systèmes panoramiques un problème particulier est la zone de diffusion optimale réduite. En revanche, de tels systèmes ont la commodité d'être assez pratiques et mobiles.

Vector-Base Amplitude Panning

Vector-Base Amplitude Panning (VBAP) [Pul97] est une méthode de spatialisation multi-canal de source dans un espace 2D ou 3D. VBAP utilise un, deux ou trois haut-parleurs selon

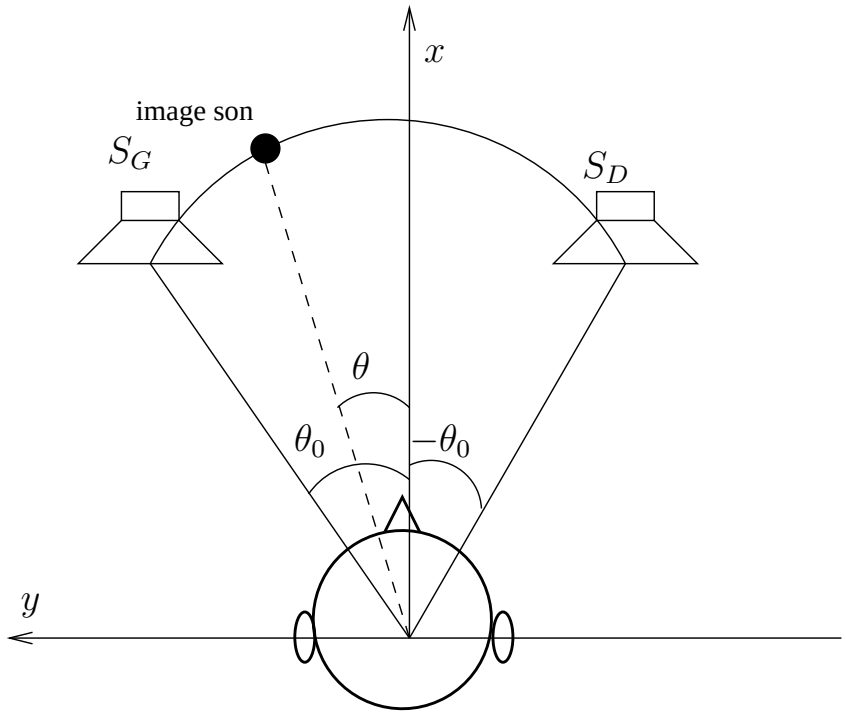


FIG. 25: *Configuration stéréophonique standard.*

que la source est localisée à la position correspondant à un haut-parleur, dans le plan horizontal ou dans un plan élevé. Le calcul des coefficients de spatialisation est fondé sur une reformulation vectorielle de la loi de la tangente (équation 55). Le vecteur de la direction de spatialisation souhaitée $\mathbf{p}$ est exprimé comme une combinaison linéaire des vecteurs de direction des deux haut-parleurs pointés $\mathbf{l}_G, \mathbf{l}_D$; les projections de $\mathbf{p}$ sur les vecteurs directionnels du haut-parleur droit ($\mathbf{I}_D$) et gauche ($\mathbf{I}_G$) sont respectivement $g_D\ \mathbf{I}_D$ et $g_G\ \mathbf{I}_G$ (voir figure 26).

$$\mathbf{p} = g_D\ \mathbf{I}_D + g_G\ \mathbf{I}_G, \tag{57}$$
$$\mathbf{p}^T = \mathbf{g}\mathbf{L}_{DG}. \tag{58}$$

Les facteurs de gain gauche et droite $\mathbf{g} = [g_D\ \ g_G]$, $\mathbf{I}_D = [l_{Dx}\ l_{Dy}]$ et $\mathbf{I}_G = [l_{Gx}\ l_{Gy}]$ sont les vecteurs en coordonnées cartésiennes pointant vers les haut-parleurs gauche et droite utilisés, et $\mathbf{L}_{DG} = [\mathbf{I}_D\ \ \mathbf{I}_G]^T$. Les gains sont obtenus par résolution du système linéaire selon $\mathbf{g}$ avec :

$$\mathbf{g} = \begin{bmatrix} p_D & p_G \end{bmatrix} \begin{bmatrix} l_{Dx} & l_{Dy} \\ l_{Gx} & l_{Gy} \end{bmatrix}^{-1}. \tag{59}$$

Les coefficients sont généralement normalisés. VBAP a été développé sous l'hypothèse que les ondes sonores entre la gauche et la droite sont différentes seulement en amplitude, ce qui est valide pour les fréquences en dessous de 700 Hz.

La loi de la tangente et VBAP sont équivalents [Pul97].

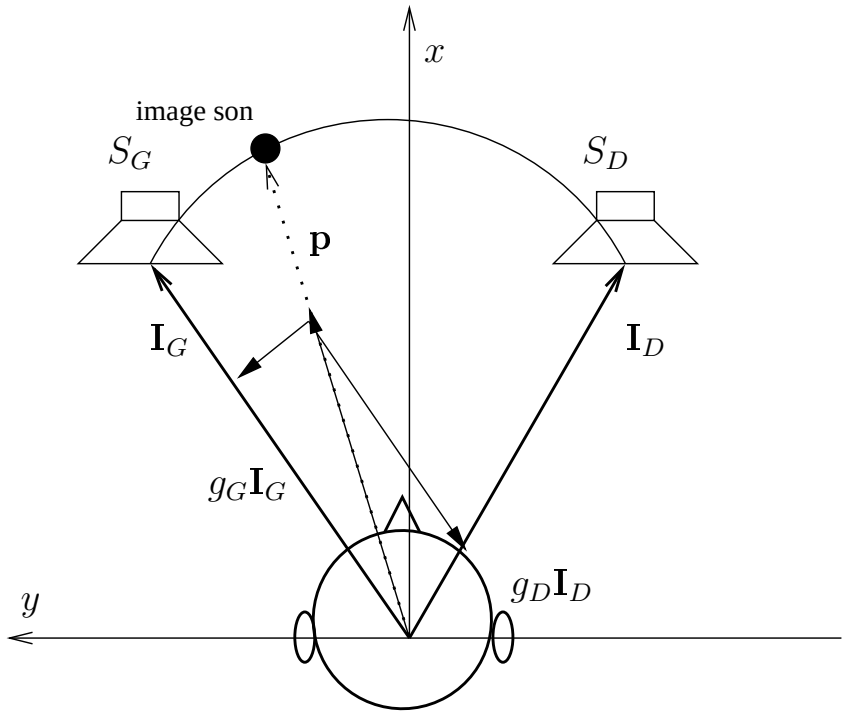


Fig. 26: *Paire de haut-parleurs dans VBAP.*

3.2.5 Synthèse par Ambisonic

Les systèmes Ambisonic tiennent leur origine des travaux de Cooper et Shiga [CS72], dont les principes clefs furent établis par Gerzon *et al.* [Ger73, Ger74, Ger77]. Les systèmes Ambisonic consistent à capturer un champ sonore complet (d'un son directionnel) et à le reproduire en décomposant ses caractéristiques directionnelles en composantes harmoniques sphériques. Le nombre de haut-parleurs dépend de la qualité de rendu souhaitée, qui est liée à l'ordre des harmoniques sphériques. A l'ordre un, l'Ambisonic nécessite 4 haut-parleurs, et à l'ordre deux, 9 haut-parleurs sont d'usage. La figure 27 affiche des ondes sphériques d'ordre zéro, un et deux. La reconstitution de l'onde sonore dépend de la direction et du nombre de haut-parleurs autour de l'auditeur. Tous les haut-parleurs sont actifs et diffusent le même son avec des gains différents. Les gains sont obtenus par un décodage matriciel qui s'adapte à la quantité d'enceintes [NE99].

L'enregistrement nécessite en général des capsules cardioïdes. Rien qu'à l'ordre un, un microphone à quatre cardioïdes est de règle (figure 27). Il existe quatre formats classiques pour les sons Ambisonic de premier ordre : A, B, C et D. A partir de l'azimut θ et de l'élévation ν, le format B est constitué des quatre signaux suivants :

$$W = 1 \tag{60}$$
$$X = \sqrt{2}\cos\theta\cos\nu \tag{61}$$
$$Y = \sqrt{2}\sin\theta\cos\nu \tag{62}$$
$$X = \sqrt{2}\sin\nu. \tag{63}$$

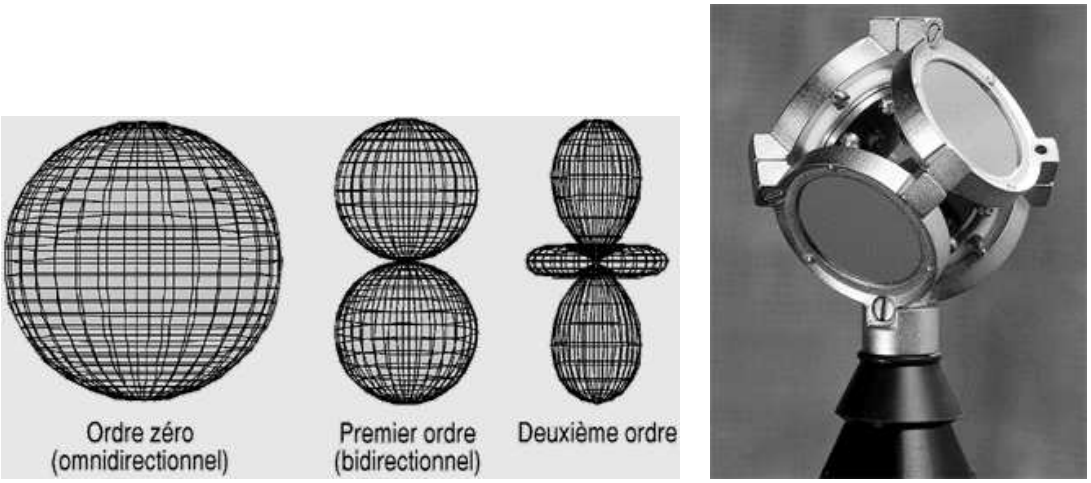


FIG. 27: *Ondes harmoniques d'ordre zéro, un, deux (gauche). Microphone cardioïde (droite).* `Référence : http://www.er.uqam.ca/nobel/k24305/images/3_harmoniques.gif.`

Les systèmes Ambisonic permettent d'atteindre une large région d'écoute avec une haute qualité, car ils permettent une représentation efficace de l'onde de pression sonore 3D autour de l'auditeur [NE99]. Toutefois, la région se réduit avec les hautes fréquences. L'Ambisonic prône l'usage simultané de tous les haut-parleurs. Cela engendre des artefacts spatiaux. Notamment la qualité spatiale se dégrade rapidement lorsqu'on s'éloigne de la position optimale et lorsque la configuration n'est pas régulière. De plus, l'effet de précédence joue un rôle majeur car la localisation est fortement influencée par le haut-parleur le plus proche. Aussi, si le haut-parleur le plus proche a une amplitude d'environ 15 dB de moins que les autres, le son est localisé à un autre endroit. Cet effet est moins prononcé lorsqu'on fait usage du paradigme par paire comme dans le cas de VBAP.

Des systèmes Ambisonic d'ordre supérieur permettraient de réduire des défauts et d'améliorer la qualité sonore, toutefois ces derniers sont limités par la construction de nouveaux microphones qui prendraient en compte des modèles polaires, ainsi que des décodeurs appropriés. Malgré sa pertinence technique, les systèmes Ambisonic n'ont pas connu un enthousiasme commercial, du fait de nombreux brevets, et que l'industrie musicale et cinématographique n'y aient pas porté une considération particulière. Des études plus exhaustives sur les systèmes Ambisonic sont disponibles dans [Rum01].

3.2.6 Synthèse par holophonie

Les systèmes holophoniques sont reconnus comme ceux qui permettent une reconstruction excellente des champs acoustiques tridimensionnels dans un lieu, seulement leur mise en œuvre est très complexe [Ber88]. L'holophonie se base sur le principe de Huygens, selon lequel le champ sonore d'une source primaire dans un volume Ω_1 est reproduit parfaitement dans un espace Ω_2 par les champs sonores de sources secondaires distribuées sur la surface S [Jes73] (voir figure 28). Il est quantifié par l'intégrale de Kirchhoff et Helmoltz [Bru83] avec :

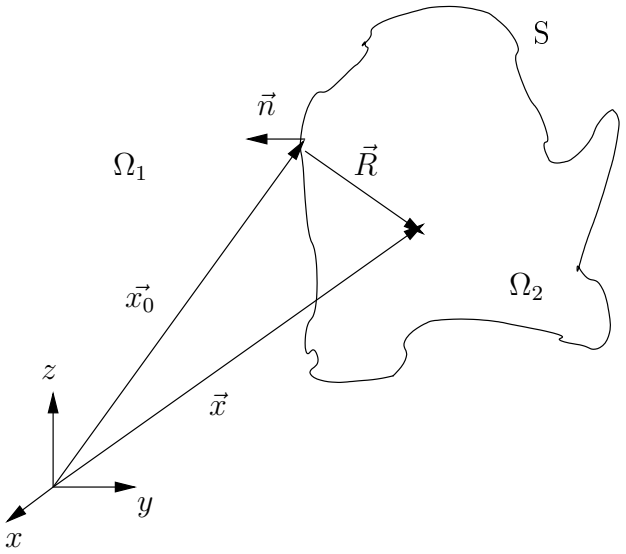


Fig. 28: *Géométrie associée à l'intégrale de Kirchhoff [PB04].*

$$p(\vec{x}, f) = \iint_S \left[\vec{\nabla} p_0(\vec{x}_0, f) \cdot \vec{n} - \frac{\vec{R}}{R} \cdot \vec{n}(1 + jkR) \frac{p_0(\vec{x}_0, f)}{R} \right] \frac{e^{-jkR}}{4\pi R} \, \mathrm{d}S \qquad \forall \vec{r} \in \Omega \qquad (64)$$

où p est le champ de pression dans Ω restitué par les sources secondaires, p_0 est le champ de pression induit par les sources primaires, c est la célérité du son dans l'air, $\vec{x}$ est la position dans l'espace de diffusion, $\vec{x_0}$ est la position sur la surface S, $\vec{R} = \vec{x} - \vec{x_0}$ est le trajet de $\vec{x}$ vers $\vec{x_0}$, $\vec{n}$ est la normale unitaire aux limites (figure 28), et $k = \frac{2\pi f}{c}$ est le nombre d'onde. La diffusion consiste en un enregistrement permanent du champ sur la surface à l'aide de microphones (de pression et de gradient), et en la diffusion de ces signaux avec des haut-parleurs.

L'holophonie assure une reconstitution du champ sonore sur une large zone, c'est-à-dire que l'auditeur peut se déplacer dans la zone de restitution en percevant la même onde, donc les mêmes effets spatiaux. La réalité est que l'holophonie devient impraticable dans la plupart des situations. La continuité des arrangements des microphones et des haut-parleurs est pratiquement impossible, la contrainte de Shannon exige que les transducteurs soient distants de la moitié de plus petite longueur d'onde au risque d'*artefacts spatiaux* [PB04]. Ainsi, les systèmes holophoniques discrets ont besoin d'un grand nombre de haut-parleurs, qui exigent une puissance de calcul énorme. Seulement dans le plan horizontal, des exemples de dispositifs ont nécessité au moins une centaine de haut-parleurs, et permettent une bonne reproduction spatiale jusqu'à environ 1000 Hz. Pour des systèmes assumant une fréquence limite d'environ 1500 Hz, la qualité spatiale n'est pas fortement affectée. Dans le plan horizontal, des microphones de gradient distribués sur une courbe sont suffisants [Ver98, NE98]. De nombreuses recherches se sont portées sur la troncature du réseau de transducteurs.

Techniques	Type de transducteurs	Taille zone d'écoute
technique binaurale	casque d'écoute	écoute individuelle
technique transaurale	2 haut-parleurs	écoute individuelle
technique panoramique	2 ou 3 haut-parleurs	écoute réduite
holophonie	grand nombre de haut-parleurs	large jusqu'à une fréquence limite
Ambisonic	à partir de 4 haut-parleurs	diminue lorsque la fréquence augmente

TAB. 4: *Quelques avantages comparatifs des techniques de spatial de sons.*

3.3 Contributions à la reproduction spatiale

Dans cette section, nous décrivons les apports théoriques et pratiques de nos recherches afin d'approcher un spatialisateur universel qui s'adapte à toute salle de diffusion (section 3.3.1). Dans le contexte binaural, le but principal est de proposer une méthode simple et numériquement efficace de spatialisation, qui permette de s'affranchir des mesures exhaustives de HRTF à toute localisation, tout en maintenant un réalisme et une qualité spatiale conséquente (section 3.3.2). Dans le domaine de la reproduction multi-diffusion, nos visées sont de proposer une méthode pratique et adaptable pour le logiciel d'informatique musicale RetroSpat (section 3.3.3).

3.3.1 Spatialisateur universel et la loi des compromis spatiaux

Au vue de la multitude de systèmes de spatialisation, quel serait le système idéal pour nos objectifs ? Quelles seraient les caractéristiques d'un système de spatialisation universel ? Le spatialisateur universel serait un système qui s'adapte facilement au nombre de haut-parleurs disponibles, et qui produit la meilleure qualité audio avec toute nouvelle configuration (uniforme ou non). Le rendu spatial devrait être assez fin et identique pour chaque localisation spatiale. C'est-à-dire pour la même source, la perception de volume et la qualité de la perception angulaire devraient être similaires. Aussi, la zone d'écoute devrait être suffisamment large de manière à reproduire le même effet en chaque endroit de la salle de diffusion. Le système universel devrait être facilement transportable avec des temps de mise en œuvre courts, notamment un nombre de haut-parleurs réduit afin d'être accessible à un plus grand nombre à un coût raisonnable.

Dans la pratique, certaines propriétés évoluent dans des sens opposés. Par exemple, la largeur de la zone d'écoute est croissante avec le nombre de haut-parleurs ou décroissante lorsque la fréquence maximale à spatialiser augmente. Les systèmes panoramiques par paire ou triplet utilisent moins de haut-parleurs et sont très pratiques. En revanche leur zone de rendu spatial est réduite et est adaptée aux basses fréquences. Le tableau 4 montre que l'holophonie couvre la plus large zone d'écoute au prix d'un grand nombre de haut-parleurs.

Pour le rendu dans un casque, il est clair que l'écoute est individuelle. Dans la section 3.3.2, nous proposons un système paramétrique de spatialisation binaurale, ce système est étendu à un système multi-diffusion par paire de haut-parleurs, également adapté à la spatialisation des hautes fréquences (section 3.3.3). Le but est de construire un système simple et efficace adapté aux hautes fréquences pour un rendu spatial fidèle.

3.3.2 Spatialisation binaurale paramétrique

Reproduction binaurale de l'azimut

L'écoute binaurale est aujourd'hui la plus répandue avec la prolifération d'appareils mobiles et portables munis d'écouteurs. Dans l'écoute binaurale, les sons gauche et droit sont perçus respectivement et exclusivement par l'oreille gauche et l'oreille droite. Ainsi, les indices acoustiques interauraux ne sont pas parasités par des signaux croisés comme dans le cas d'une configuration stéréophonique.

Analyse approfondie des modèles paramétriques

Nous nous proposons d'explorer davantage les modèles paramétriques que nous avons proposés (équation 69). Pour le modèle paramétrique d'ILD, nous considérons un bruit blanc d'amplitude 0 dB spatialisé au centre comme référence, et nous calculons les différences de niveau pour chacune des oreilles sur l'ensemble de la base CIPIC. Les facteurs d'échelle pour les deux oreilles sont très similaires à un signe près, et s'élèvent à environ la moitié du facteur d'échelle global : $\alpha_G \simeq -\alpha_D \simeq \alpha/2$.

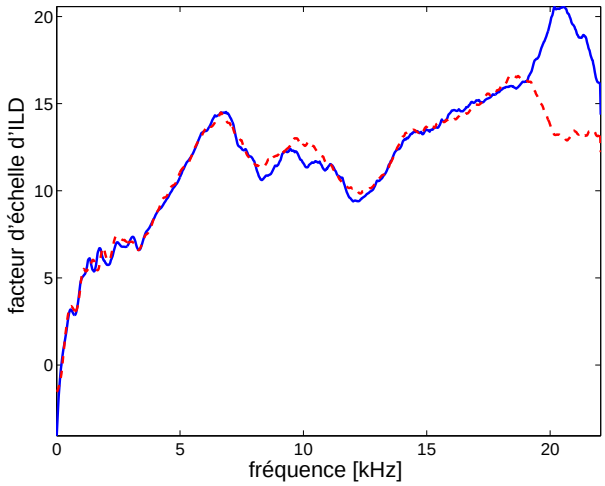
Cela signifie que les canaux gauche et droit peuvent être considérés de façon indépendante. Chaque canal est descriptible par un facteur d'échelle partiel d'une fonction sinusoïdale de l'azimut. Une méthode de spatialisation intuitive consiste à partir d'un signal monophonique, à affecter la moitié de la différence d'amplitude au canal de gauche et de soustraire la même quantité au canal droit (équations 65, 66). Les facteurs d'échelle des canaux gauche et droite sont affichés sur la figure 29. Le facteur d'échelle globale $\alpha(f)$ n'est autre que la différence de ces facteurs, soit $\alpha_G - \alpha_D$.

De la même manière, nous explorons davantage le modèle paramétrique d'ITD. Nous considérons une source de phase 0 comme référence, et dérivons les facteurs d'échelle de chaque oreille. Similairement à l'ILD, nous obtenons $\beta_G \simeq -\beta_D \simeq \beta/2$. Aussi, une méthode de spatialisation naturelle consiste à attribuer à chaque canal monophonique la moitié du délai d'ITD (ou phase) avec des signes contraires. Les facteurs d'échelle des canaux gauche et droite sont affichés sur la figure 29. Le facteur d'échelle globale $\beta(f)$ n'est autre que la différence de ces facteurs, soit $\beta_G - \beta_D$.

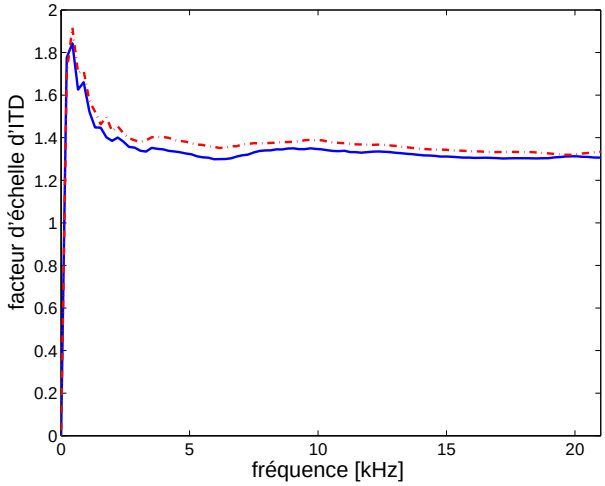
De par ces analyses, nous proposons une technique de spatialisation binaurale paramétrique, nommée SSPA (Source SPAtialization). SSPA s'applique dans le domaine temps-fréquence, ouvrant la voie à des manipulations spatiales à chaque case fréquentielle. Afin de spatialiser une source sonore monophonique s à la position azimutale θ, d'abord, le signal est projeté dans le plan temps-fréquence avec la transformée de Fourier à court terme. A partir de chaque spectre à court terme X, nous calculons la paire de spectres gauche (X_G) et droit (X_D) à partir des indices binauraux correspondant à l'azimut θ (équations 69, 70). Ce processus est résumé par la figure 30. En considérant la tête symétrique et la symétrie axiale des oreilles, on propose les spectres gauche et droit suivants :

$$X_G(t,f) = X(t,f) \cdot 10^{+\Delta_a(f)/2} e^{+j\Delta_\phi(f)/2}, \quad (65)$$

$$X_D(t,f) = X(t,f) \cdot 10^{-\Delta_a(f)/2} e^{-j\Delta_\phi(f)/2}, \quad (66)$$



(a)



(b)

FIG. 29: *Facteurs d'échelle fréquentiels pour la différence d'amplitude (a) et la différence de phase (b). Oreille gauche (pointillés), oreille droite (trait plein).*

où Δ_a et Δ_ϕ sont donnés par :

$$\Delta_a(f) = \mathrm{ILD}(\theta, f)/20, \quad (67)$$

$$\Delta_\phi(f) = \mathrm{ITD}(\theta, f) \cdot 2\pi f, \quad (68)$$

avec les modèles paramétriques :

$$\mathrm{ILD}(\theta, f) = \alpha(f)\sin(\theta), \quad (69)$$

$$\mathrm{ITD}(\theta, f) = \beta(f) r \sin(\theta)/c. \quad (70)$$

Les versions temporelles des signaux x_G, x_D sont obtenues par transformée de Fourier inverse des spectres X_G, X_D, et diffusées aux oreilles à travers un casque audio. Contrairement

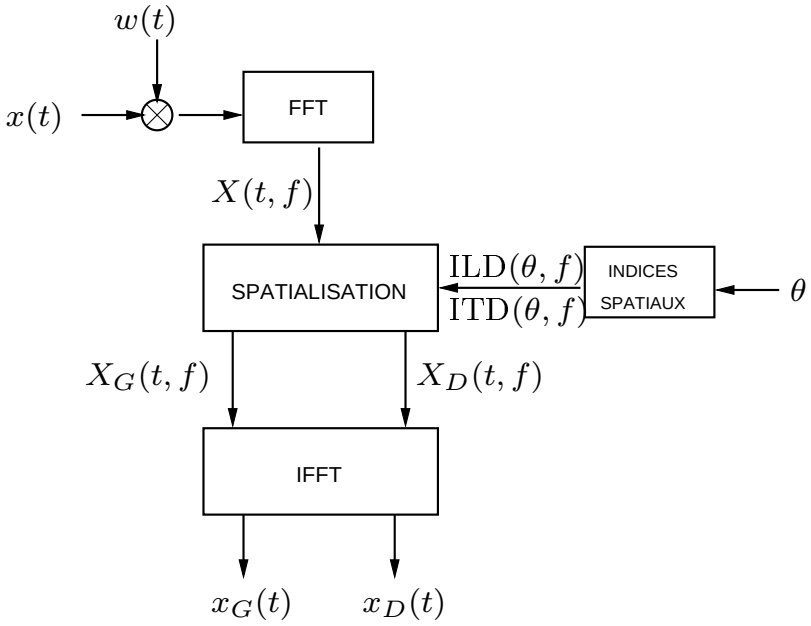


Fig. 30: *Spatialisation binaurale d'une source x à l'azimut θ.*

aux méthodes classiques, notre technique contrôle l'amplitude et la phase, ce qui entraîne un gain en qualité audio par rapport aux méthodes contrôlant uniquement l'amplitude [TF06]. Dans la section 3.4, des tests de réalisme spatial sont menés avec un casque d'écoute AKG K240 Studio [1].

Plusieurs sources virtuelles binaurales peuvent être générées et diffusées simultanément afin de créer un environnement sonore (figure 31). Les signaux binauraux mixés sont donnés sous forme matricielle par :

$$\begin{bmatrix} X_G(\omega) \\ X_D(\omega) \end{bmatrix} = \begin{bmatrix} H_{G_1} & \dots & \dots & H_{G_k} & \dots & H_{G_K} \\ H_{D_1} & \dots & \dots & H_{D_k} & \dots & H_{D_K} \end{bmatrix} \begin{bmatrix} X_1(\omega) \\ \vdots \\ X_k(\omega) \\ \vdots \\ X_K(\omega) \end{bmatrix}, \tag{71}$$

avec les filtres pour la position θ_i construits par :

$$H_{G_i}(t,f) = 10^{+\Delta_a(f,\theta_i)/2} e^{+j\Delta_\phi(f,\theta_i)/2}, \tag{72}$$
$$H_{D_i}(t,f) = 10^{-\Delta_a(f,\theta_i)/2} e^{-j\Delta_\phi(f,\theta_i)/2}. \tag{73}$$

Cette technique a le mérite d'être simple, pratique et paramétrique. Elle permet de maîtriser facilement les manipulations spatiales, même aux positions virtuelles non mesurées.

[1] voir l'URL : `http://dept-info.labri.fr/~sm/SMC08/`

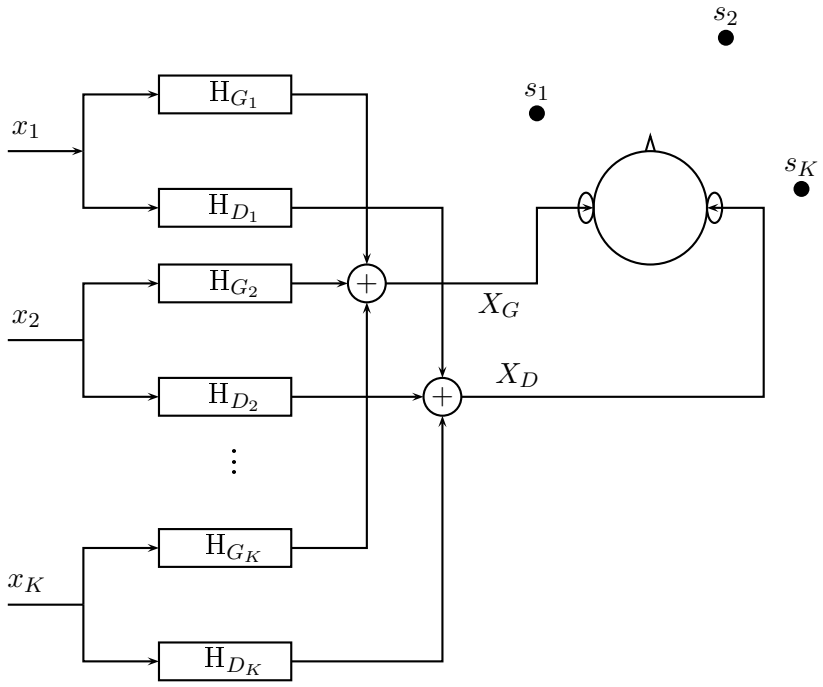


FIG. 31: *Synthèse de multiples sources binaurales.*

3.3.3 Spatialisation transaurale paramétrique

Dans une configuration stéréophonique, le son propagé par chaque haut-parleur est entendu par les deux oreilles. Le signal de chaque haut-parleur est filtré par une paire de filtres ; le son stéréo est alors filtré par une matrice de quatre fonctions de transfert $(C_{ij}(f,\theta))$ entre les haut-parleurs et les oreilles (voir figure 32). Ici, nous générons ces trajets artificiellement à l'aide du modèle binaural paramétrique (équations 69, 70).

Nous utilisons des filtres fixes et nous supposons que la salle de diffusion n'est pas très réverbérée. Généralement, dans les configurations stéréophoniques, on utilise un ensemble de HRTF combinées avec un ensemble de filtres afin de réduire ou d'éliminer les trajets croisés, ce sont les techniques appelées couramment *cross-talk canceller* [YBC07]. Ici, nous ne faisons pas usage d'un système adaptatif. Pour chaque azimut, nous fixons les canaux entre les oreilles et les haut-parleurs. Les filtres de propagation sont déduits de notre modèle binaural paramétrique.

L'objectif est alors de trouver les coefficients complexes optimaux pour chaque canal, de manière à ce que les sommes des signaux au niveau de chaque oreille correspondent aux signaux binauraux.

Les meilleurs coefficients de spatialisation dans des conditions de la base CIPIC pour une paire de haut-parleurs, afin d'approcher les signaux binauraux au niveau des tympans (voir

équations (65) et (66)) sont donnés par :

$$K_G(t,f) = C \cdot (C_{DD}H_G - C_{GD}H_D), \quad (74)$$
$$K_D(t,f) = C \cdot (-C_{DG}H_G + C_{GG}H_D), \quad (75)$$

avec le déterminant calculé avec

$$C = 1/(C_{GG}C_{DD} - C_{DG}C_{GD}). \quad (76)$$

Dans des cas extrêmes où $|C| = 0$ (ou proche de zéro) à chaque fréquence, la matrice C est mal conditionnée ; dans de tels cas, la solution devient instable.

Les cas d'instabilité de la matrice peuvent être évités en prenant garde à la disposition des haut-parleurs lors de la configuration avant toute diffusion. Dans la pratique des cas d'instabilité ne furent pas observés.

Pendant la diffusion, les haut-parleurs gauche et droit sont alimentés avec les signaux gauche et droit y_G, y_D obtenus par transformée de Fourier inverse des spectres Y_G et Y_D. Ces derniers sont obtenus en multipliant chaque spectre à court-terme X avec les coefficients spatiaux gauche et droit K_G and K_D selon :

$$Y_G(t,f) = K_G(t,f) \cdot X(t,f), \quad (77)$$
$$Y_D(t,f) = K_D(t,f) \cdot X(t,f). \quad (78)$$

Dans une configuration avec plus de deux haut-parleurs, nous appliquons le paradigme par paire classique.

3.4 Résultats de spatialisation

3.4.1 Stratégie d'évaluation

Nous avons conduit des tests objectifs et subjectifs afin d'évaluer la qualité des sources virtuelles créées dans le cas binaural et dans le cas de techniques multi-diffusion. Dans le cas binaural, des sons binauraux sont générés par convolution avec les HRTF de la base CIPIC et par la méthode binaurale paramétrique. Nous avons utilisé des sons vocaux et des sons d'instruments. Les sons sont normalisés de manière à avoir le maximum à 0 dB. Pour les méthodes transaurales des tests objectifs sont basés sur la localisation des signaux binauraux enregistrés avec le phonocasque (un casque audio où les haut-parleurs sont remplacés par des microphones).

Les paramètres étudiés sont premièrement le réalisme spatial, en d'autres mots, la capacité de la méthode à positionner la source à la position cible, ensuite la qualité globale du son spatialisé par rapport à une référence, en jugeant par l'écoute la dégradation éventuelle du son.

3.4.2 Résultats de la méthode binaurale paramétrique

Nous comparons d'une part les sons binauraux d'une même méthode entre eux, et d'autre part des sons binauraux de méthodes différentes, à savoir la méthode SHRIR qui consiste

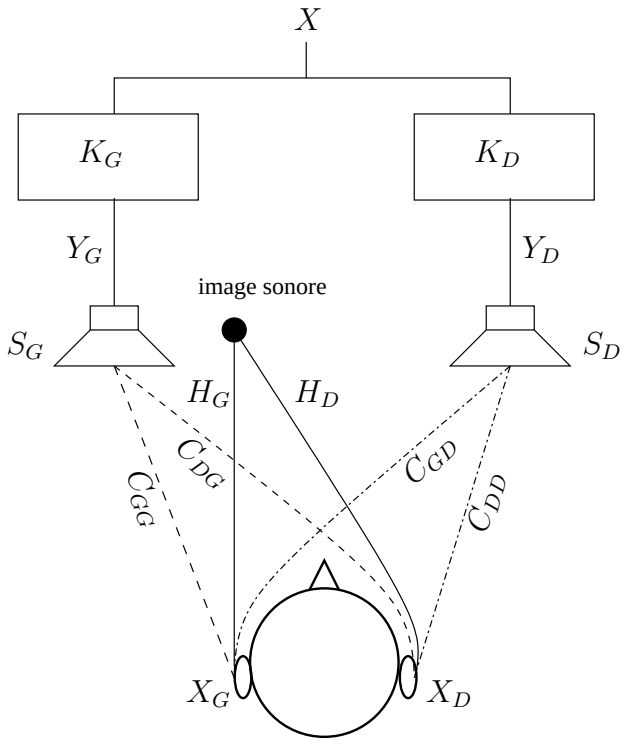


FIG. 32: *Configuration transaurale.*

à convoluer un signal monophonique avec les HRIR et la méthode binaurale paramétrique (SSPA). Les tests ont pour but de comparer quelques propriétés de spatialisation : le réalisme spatial et la qualité sonore. Le réalisme spatial consiste à qualifier la précision subjectif de la projection ; et la qualité sonore est en rapport avec la sensation globale (contenu fréquentiel, sonie ...).

La procédure de tests est simple : le sujet écoute les échantillons spatialisés avec SSPA et SHRIR, en gardant le volume inchangé pendant tout le test. Un questionnaire à 4 questions (à choix simple) a été établi :

- écouter chaque échantillon et indiquer si la source est positionnée à GAUCHE, à DROITE ou au CENTRE. Au total 9 échantillons furent éprouvés.
- écouter un couple d'échantillons, et indiquer si le second est positionné relativement à GAUCHE, à DROITE ou à la même position que le premier. Au total 12 couples d'échantillons de 3 types furent éprouvés : (SSPA, SSPA), (SHRIR, SSPA), (SHRIR, SHRIR).
- écouter l'échantillon original et l'échantillon spatialisé, et juger de la qualité de ce dernier sur une échelle de qualité à 5 niveaux (1 : parfait, 2 : artefacts mineurs, 3 : déformé mais intelligible, 4 : très déformé et 5 : pas intelligible). Au total 12 couples d'échantillons furent éprouvés.
- écouter l'échantillon et indiquer sa position absolue parmi 9 positions proposées (0°, ±5°, ±30°, ±65°, ±90°).

Les tests furent conduits sur 10 sujets (scientifiques et musiciens) du LaBRI et du SCRIME, tous familiers avec le son et ayant une certaine connaissance et pratique musicale. Les résultats

sont analysés et résumés dans les sections suivantes.

Tests subjectifs

Premièrement, la qualité sonore des sons spatialisés sur l'échelle de 5 proposé, a atteint un score de 2 (artefacts mineurs) pour les deux méthodes (SSPA et SHRIR), avec un léger avantage au dixième près pour SSPA (2.1 contre 2.2). La préférence pour le SSPA découlerait du fait que la moyenne des modèles se rapprocherait d'un modèle moins réverbéré que les HRIR d'origine. La réverbération plus présente pour les sons SHRIR pourrait être éprouvée comme des artefacts.

Deuxièmement, la détection de la source dans le plan gauche ou droit s'est faite sans ambiguïté. Les sources au centre furent toutes bien localisées, toutefois pour une résolution de 5°, 90% de sujets ne purent pas différencier spatialement deux sources adjacentes.

Pour des paires croisées (SHRIR, SSPA) à la même position, 15% des sujets perçurent la source SSPA plus excentrique que la la source SHRIR, et 85% les jugèrent à la même position. Ce constat met en évidence que notre modèle paramétrique s'adapte aux têtes du plus grand nombre et ne modifie pas la véritable localisation de la source. Les tests de localisation absolue sont généralement très difficiles, mais les résultats sont positivement surprenant. La tâche fut quelque peut faciliter en proposant des localisations assez distantes, excepté au centre. Les participants purent sélectionnés en moyenne 90% de bonnes réponses pour les positions ±30°, ±65°, ±90°, 100% pour la position 0°. L'erreur la plus fréquente fut celle de la confusion prévisible entre ±5° et 0°, où nous avons relevé un taux d'erreur de l'ordre de 40%.

Tests objectifs

Réalisme spatial et qualité sonore

La précision de la spatialisation et le contenu fréquentiel de la source spatialisée jouent un rôle dans la qualité de la spatialisation. Dans un premier temps, nous avons comparé des signaux binauraux paramétriques entre eux. La localisation absolue étant difficile à déterminer par écoute, nous avons conduit des tests sur la location relative. Durant l'écoute, l'auditeur devrait identifier la source la plus à gauche, à droite ou si les deux sources sont au même emplacement. La méthode paramétrique SSPA ne pose pas d'ambiguïté entre le plan gauche et le plan droit. Aussi, elle est jugée comme étant précise en terme de spatialisation. Toutefois, lorsque l'offset en azimut était moins de 5°, les deux sons étaient souvent jugés comme venant de la même position.

Pour des paires croisées (SHRIR-SSPA) et pour une même position ; les sons issus de SHRIR sont perçus plus à gauche que le son issu de SSPA. Toutefois, la SHRIR nécessite la mesure de HRIR pour toutes les positions, tandis que le SSPA fait une interpolation angulaire perceptivement correct. Les fonctions d'intercorrélation des signaux issus de SSPA et de SHRIR ont été mesurées afin de mesurer la qualité de la spatialisation (figure 34). Les résultats montrent une précision spatiale correcte, en effet nous observons un pic largement dominant aux environs du bon délai interaural. Aucune ambiguïté n'est présente. Les fonctions d'intercorrélation issues de SSPA sont plus lisses et présentent moins de pics parasites que celles issues de signaux de SHRIR. Aussi, la hauteur des pics parasites pour les signaux SSPA est

moins élevée que pour les signaux issus de SHRIR.

La spatialisation avec la méthode SSPA paraît plus précise et plus stable que la méthode SHRIR. La méthode SSPA permet donc de spatialiser avec précision toute source monophonique (voix, son d'instrument). Aussi, nous remarquons que les signaux de parole présentent un pic plus large que ceux issus de signaux d'instruments (figure 34).

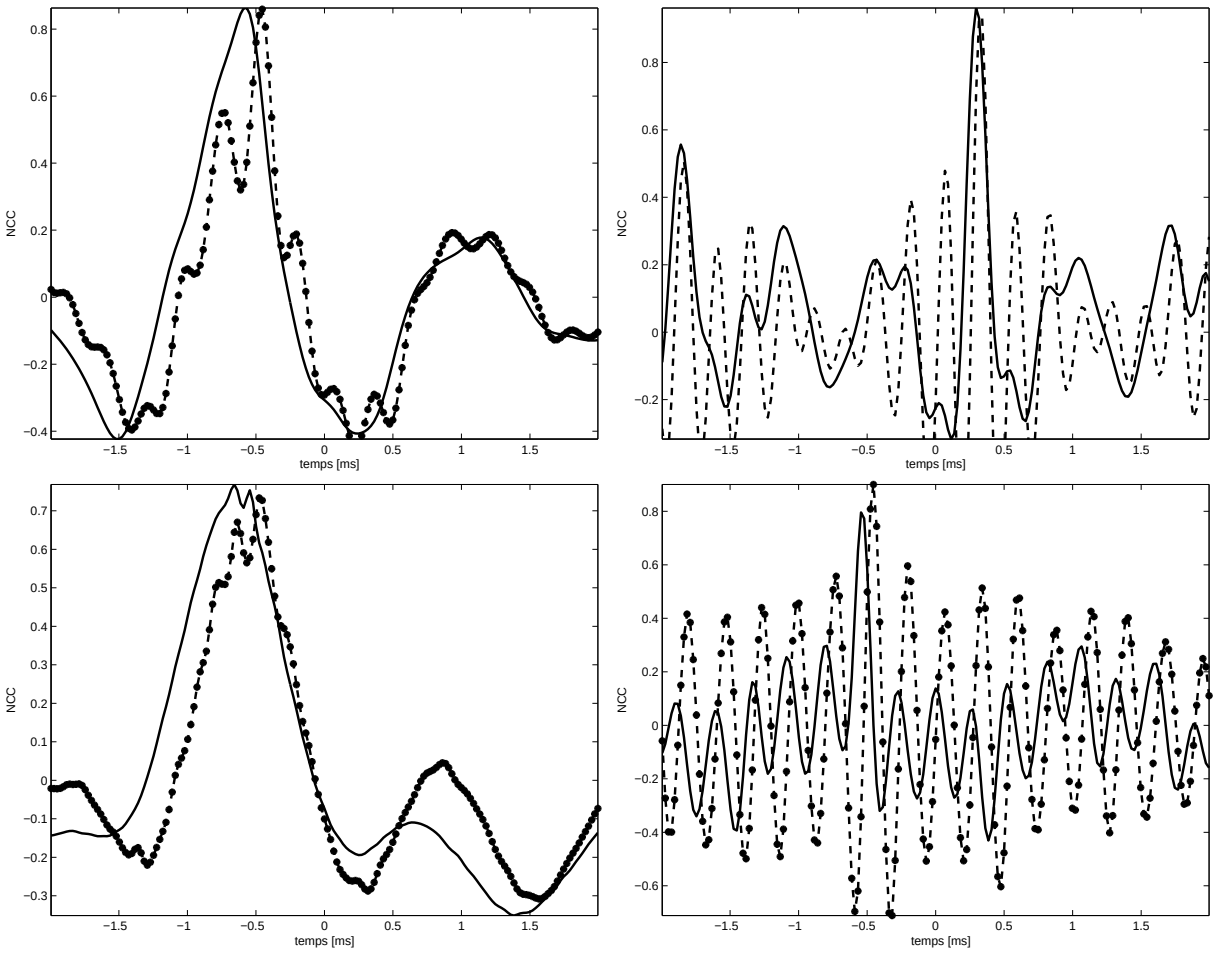


FIG. 33: *Fonctions d'intercorrélation à partir de signaux binauraux issus de SSPA (trait plein) et ceux issus de SHRIR (tiret). De haut en bas et de gauche à droite :* $-65°$ *(parole),* $+30°$ *(trompette),* $-65°$ *(voix chantée)* ,$-30°$ *(xylophone).*

La qualité sonore globale de sons binauraux issus de SHRIR et SSPA ont été comparés perceptivement. Des tests préliminaires montrent que les sons binauraux issus de la méthode paramétrique sont d'une qualité correcte [2].

Deuxièmement, la détection de la source dans le plan gauche ou droit s'est faite sans ambiguïté. Les sources au centre furent toutes bien localisées, toutefois pour une résolution de 5°, 90% de sujets ne purent pas différencier spatialement deux sources adjacentes.

Pour des paires croisées (SHRIR, SSPA) à la même position, 15% des sujets perçurent la source SSPA plus excentrique que la la source SHRIR, et 85% les jugèrent à la même position.

[2]see URL : `http://dept-info.labri.fr/~sm/SMC08/`

Ce constat met en évidence que notre modèle paramétrique s'adapte aux têtes du plus grand nombre et ne modifie pas la véritable localisation de la source. Les tests de localisation absolue sont généralement très difficiles, mais les résultats sont positivement surprenant. La tâche fut quelque peut faciliter en proposant des localisations assez distantes, excepté au centre. Les participants purent sélectionnés en moyenne 90% de bonnes réponses pour les positions $\pm 30°$, $\pm 65°$, $\pm 90°$, 100% pour la position $0°$. L'erreur la plus fréquente fut celle de la confusion prévisible entre $\pm 5°$ et $0°$, où nous avons relevé un taux d'erreur de l'ordre de 40%.

Tests objectifs

Nous avons également comparé objectivement les signaux issus de SSPA et de SHRIR. A cet effet, nous avons utilisé la méthode d'intercorrélation PHAT-GCC pour la localisation.

Les résultats montrent une bonne précision spatiale, en effet nous observons un pic largement dominant aux environs du bon délai interaural. Aucune ambiguïté n'est présente. Les fonctions d'intercorrélation issues de SSPA sont plus lisses et présentent moins de pics parasites que celles issues de signaux de SHRIR. Aussi, la hauteur des pics parasites pour les signaux SSPA est moins élevée que pour les signaux issus de SHRIR. Ainsi, La spatialisation avec la méthode SSPA paraît plus précise et plus stable que la méthode SHRIR. La méthode SSPA permet donc de spatialiser avec précision toute source monophonique (voix, son d'instrument). Toutefois, nous remarquons que les signaux de parole présentent un pic plus large que ceux issus de signaux d'instruments (figure 34). Cela résulte en partie du contenu spectral de la voix.

Pour des couples de sons (SHRIR-SSPA) à la même position, nous avons la confirmation que les sons issus de SSPA sont plus excentriques spatialement que ceux générés par SHRIR. Cette observation est illustrée sur la figure 34 où le pic de la corrélation de SHRIR est placé à droite de celui de SSPA pour les angles négatifs et inversement pour les angles positifs.

Aussi, la SHRIR exige des mesures de HRIR pour toutes les positions cibles, tandis que la SSPA fait une interpolation angulaire correcte au vue des fonctions d'intercorrélation. ainsi, la méthode SSPA permet de spatialiser précisément des sources monophoniques (voix, instrument) et représente un bon candidat pour la réduction de la complexité des systèmes basés sur les HRIR.

Réalisme spatial et qualité sonore

La précision de la spatialisation et le contenu fréquentiel de la source spatialisée jouent un rôle dans la qualité de la spatialisation. Dans un premier temps, nous avons comparé des signaux binauraux paramétriques entre eux. La localisation absolue étant difficile à déterminer par écoute, nous avons conduit des tests sur la location relative. Durant l'écoute, l'auditeur devrait identifier laquelle est la plus à gauche, à droite ou au même emplacement. La méthode paramétrique SSPA ne pose pas d'ambiguïté entre le plan gauche et le plan droit. Aussi, elle est jugée comme étant précise en terme de spatialisation. Toutefois, lorsque l'offset en azimut était moins de $5°$, les deux sons étaient souvent jugés comme venant de la même position.

Pour des paires croisées (SHRIR-SSPA) et pour une même position ; les sons issus de SHRIR sont perçus plus à gauche que le son issu de SSPA. Toutefois, la SHRIR nécessite la mesure de HRIR pour toutes les positions, tandis que le SSPA fait une interpolation angulaire

perceptivement correct.

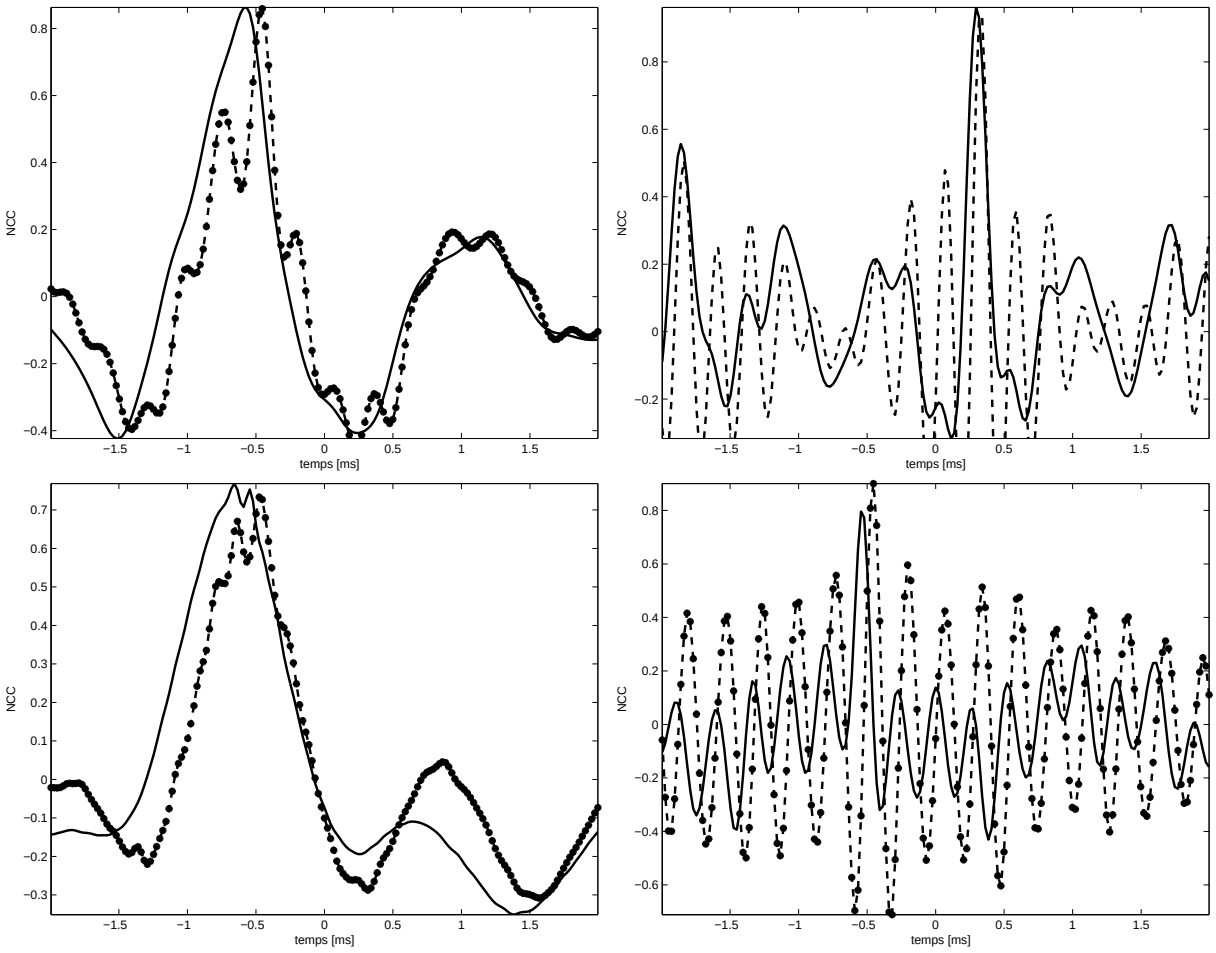


FIG. 34: *Fonctions d'intercorrélation à partir de signaux binauraux issus de SSPA (trait plein) et ceux issus de SHRIR (tiret). De haut en bas et de gauche à droite :* $-65°$ *(parole),* $+30°$ *(trompette),* $-65°$ *(voix chantée) ,*$-30°$ *(xylophone).*

3.4.3 Résultats de la méthode transaurale paramétrique

Dans cette section, nous comparons la méthode multi-diffusion transaurale paramétrique MSPA à la méthode classique VBAP [Pul97]. Dans le plan horizontal, ces deux méthodes travaillent par paire de haut-parleurs selon le même paradigme. La méthode VBAP est bien connue et fonctionne bien dans de nombreuses situations. Dans l'approche VBAP, la spatialisation est contrôlée uniquement par la différence en amplitude avec des coefficients indépendants de la fréquence. Nonobstant, VBAP est théoriquement adapté pour les fréquences inférieures à 700 Hz, ce qui peut s'avérer suffisant dans la mesure où l'ITD qui est un indice important de la localisation domine jusqu'à environ 1.5 kHz. Les coefficients de spatialisation de VBAP sont fixes pour chaque azimut et quelle que soit la salle de diffusion. MSPA propose également des coefficients de spatialisation statiques quel que soit l'environnement, par contre ces derniers sont de nature complexe, et leur calcul théorique est défini sur toute la bande de fréquence

audible [MMMR08].

VBAP est une technique assez simple à mettre en œuvre, elle a contribué à la spatialisation dans des salles, d'où son intégration réussie comme outil de spatialisation dans MAX/MSP [Pul00]. VBAP est le meilleur candidat qui se rapproche de nos attentes, donc par rapport auquel nous pouvons mieux évaluer MSPA. Du fait du paradigme par paire, la comparaison des coefficients de spatialisation pour une paire de haut-parleurs est largement suffisante. Des tests objectifs de localisation sont ensuite menés sur des enregistrements réels. C'est-à-dire la paire de haut-parleurs diffuse un son à partir de VBAP et MSPA, ce dernier est enregistré avec le phonocasque afin d'enregistrer le signal binaural. La localisation à partir du signal binaural permet ainsi d'évaluer la performance des méthodes de spatialisation.

Analyse des coefficients de spatialisation

Nous avons utilisé la paire de haut-parleurs localisés à $(-30°, +30°)$ pour le calcul de coefficients de spatialisation pour tout azimut entre les haut-parleurs avec les deux techniques : notre approche MSPA et le classique VBAP. Pour les fréquences jusqu'à 700 Hz, pour une différence en amplitude donnée entre les haut-parleurs, la phase change approximativement de manière linéaire à la fréquence [Blu31]. En maîtrisant correctement les amplitudes des deux canaux, il est possible de produire des différences en phase et différences en amplitude pour des sons continus, qui seraient proches de ceux expérimentés dans les sons naturels [Rum01]. Nous restreignons alors nos comparaisons dans la bande de fréquence $[0, 700]$ Hz.

Comparaisons des coefficients de spatialisation

Nous remarquons que les coefficients de spatialisation des deux approches sont très similaires jusqu'à 600 Hz (voir figure 35). Ces derniers peuvent se différencier significativement ensuite. En effet, nos coefficients sont des nombres complexes, et la partie imaginaire peut apporter une contribution significative (voir figure 36).

Comparaison des rapports des coefficients de spatialisation

Généralement, les différences inter-canal sont perceptivement plus notables (par exemple ILD, ITD) par rapport aux valeurs absolues [Pul01].

À partir des coefficients de spatialisation gauche et droit, K_G and K_D, nous définissons la différence en amplitude des coefficients de spatialisation (ou *panning level difference* (PLD)) :

$$\mathrm{PLD} = 20 \log_{10} \left| \frac{K_G}{K_D} \right|. \tag{79}$$

Nous calculons les différences absolues entre les PLD des deux approches, VBAP et la notre. Le maximum de la différence entre les PLD, dans la bande de fréquence considérée n'excède pas les 3 dB. Ainsi, les approches semblent consistantes dans la bande $[0, 700]$ Hz (voir figure 38).

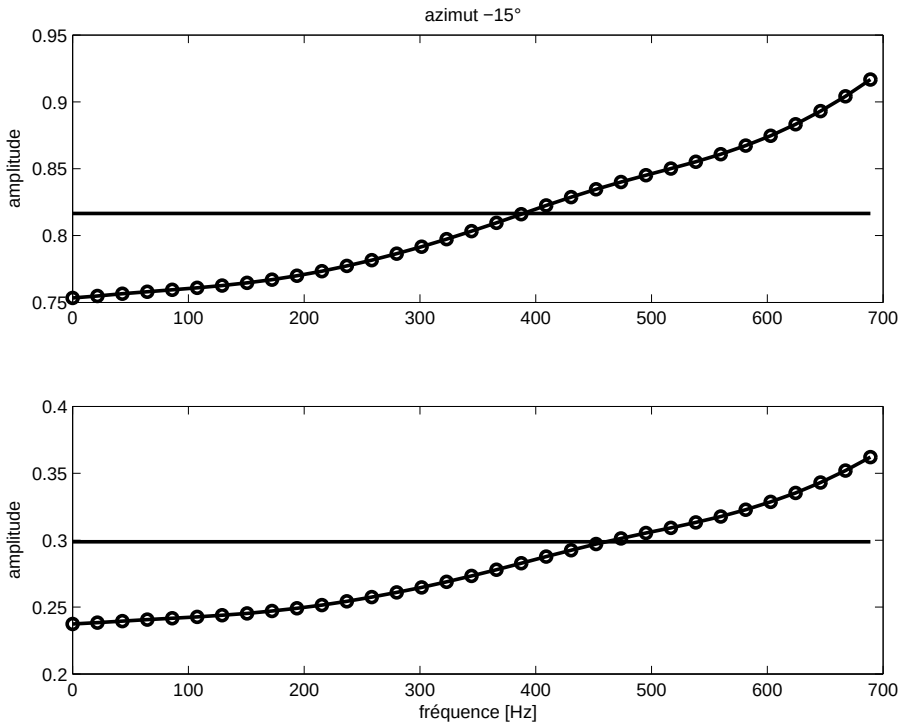


FIG. 35: *Amplitude des coefficients de spatialisation de VBAP (trait plein) et de notre approche (pointillés) dans la bande* $[0, 700]$ *Hz, pour les canaux gauche (haut) et droit (bas) de la paire de haut-parleurs, pour simuler une source à* $-15°$.

Tests subjectifs

Les tests subjectifs ont pour but d'accéder à la perception naturelle des auditeurs. Les attributs que nous désirons sont la précision de la localisation relative, c'est-à-dire la différence entre deux localisations, la qualité du son spatialisé avec VBAP ou MSPA par rapport au son naturel. A cet effet, nous avons employé des signaux vocaux. L'auditeur devait pointer du doigt la localisation perçue. Ces méthodes plutôt descriptives sont souvent sujettes à erreur. Dans tous les cas, la source réelle procurait un meilleur plaisir à l'écoute et sa localisation est sans équivoque.

Des tests d'écoute révèlent que la qualité spatiale des sources virtuelles issues de MSPA et VBAP sont de précision similaire. Toutefois, le contenu spectral est différent. La source virtuelle issue de MSPA semble contenir plus de hautes fréquence (plus brillante). Observation qui laisse présager un meilleur contrôle de la spatialisation des composantes large bande, en effet, du point de vue mathématique, la figure 37 montre que les coefficients de spatialisation optimaux ne sont pas constants. Tout comme dans le cas binaural, la localisation relative n'entraîne aucune ambiguïté pour les méthodes MSPA et VBAP. La source est jugée correctement plus à gauche ou à droite de la précédente pour une résolution de $5°$.

Des sources dynamiques ont été créées pour un système octophonique avec VBAP et MSPA. Les sons issus de VBAP semblent avoir une intensité sonore plus constante quand la source se déplace (tourne autour de l'auditeur). Cette différence de constance de l'intensité peut s'expliquer par la normalisation des coefficients de spatialisation évidente avec VBAP

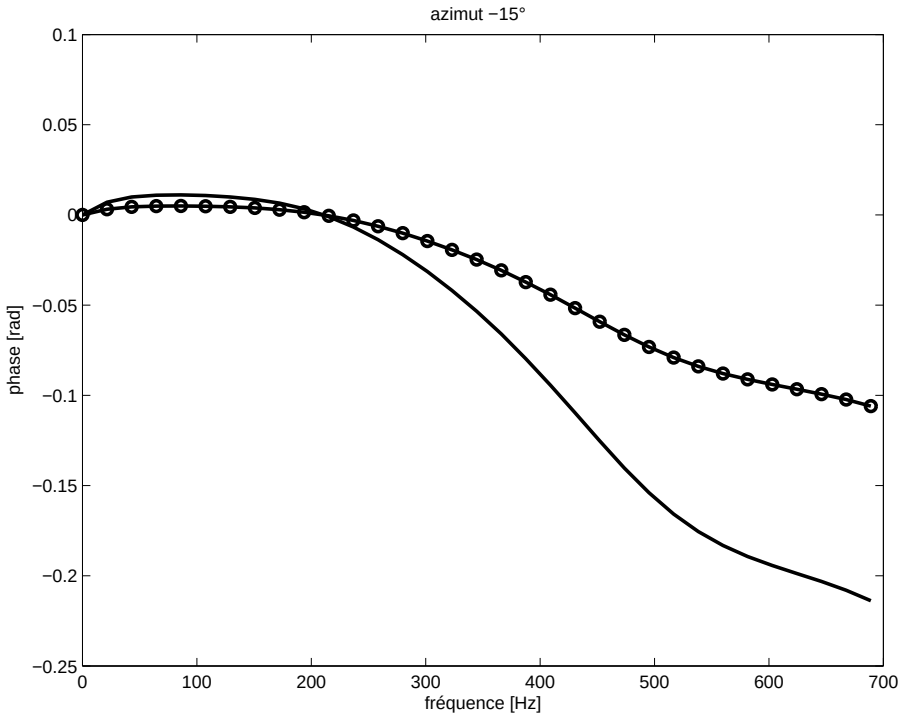


Fig. 36: *Phase des coefficients de spatialisation de MSPA, pour les canaux gauche (pointillés) et droit (trait plein) dans la bande* $[0, 700]Hz$*, pour simuler une source à* $-15°$.

(avec des coefficients indépendants de la fréquence) alors que sous MSPA, la dépendance fréquentielle semble gêner quelque peu la normalisation. Toutefois pour les deux approches, des accélérations entre certains haut-parleurs ont été constatées, avec un biais vers le haut-parleur le plus proche de la localisation souhaitée. Cet effet est modéré avec un nombre croissant de haut-parleurs et une réduction de l'arc angulaire entre chaque paire de haut-parleurs. Un avantage des méthodes par paire est que l'erreur de spatialisation est bornée entre les deux haut-parleurs. Les sons virtuels issus de MSPA paraissent moins diffus que ceux issus de VBAP. Avec MSPA les hautes fréquences sont mieux contrôlées et suivies (voir figure 37). En effet, les coefficients de MSPA sont modélisés de manière à ce que dans des environnements peu réverbérés, la perception naturelle soit respectée, au moins dans les sons directs. Les tests objectifs confortent cette constatation.

Tests objectifs

Dans cette section, nous commentons les résultats issus des tests objectifs. Un son est spatialisé avec MSPA ou VBAP ensuite diffusé sur le système de haut-parleurs. Le son diffusé est enregistré à l'aide d'un phonocasque pour former un signal binaural test.

A partir de l'enregistrement binaural, il est possible d'obtenir des mesures objectives sur la précision de la localisation. Dans la salle réverbérée Bonnefont de l'université de Bordeaux 1, avec un phonocasque, nous avons réalisé des enregistrements binauraux de bruits blancs spatialisés à différents azimuts. Les bruits blancs sont prisés du fait de leur spectre constant et large, ici nous utilisons une fréquence d'échantillonnage de 44.1 kHz. Le sujet portant le

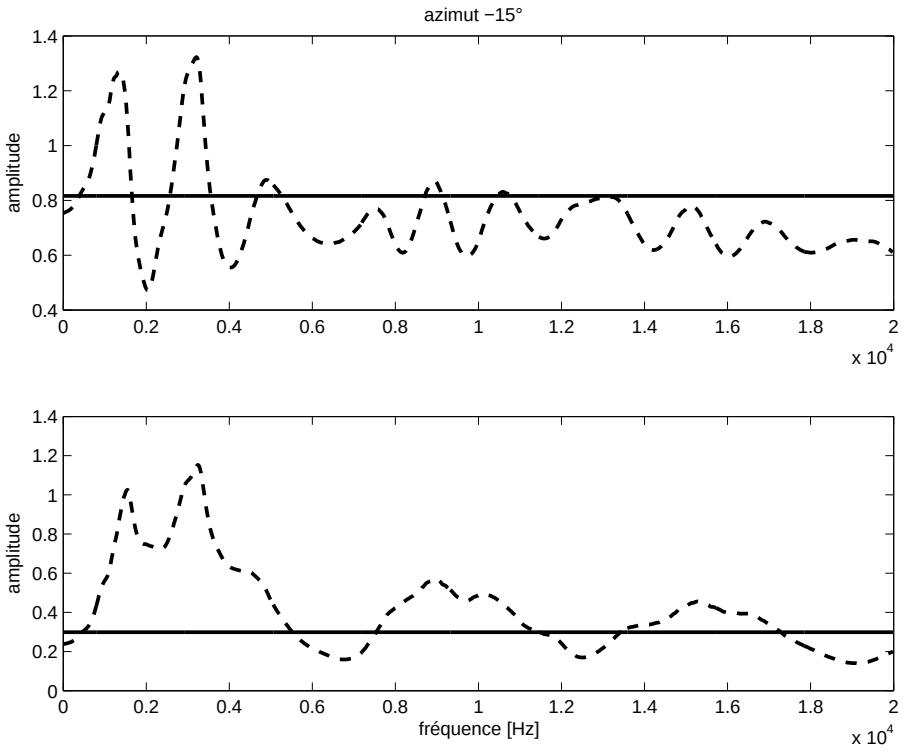


FIG. 37: *Amplitude des coefficients de spatialisation de VBAP (trait plein) et de notre approche (pointillés) dans la bande* $[0, 20000]Hz$*, pour les canaux gauche (haut) et droit (bas) de la paire de haut-parleurs, pour simuler une source à* $-15°$.

phonocasque fixe dans la direction de l'azimut zéro. Les signaux enregistrés sont de trois types, à savoir ceux issus de la diffusion d'une source monophonique réelle, c'est-à-dire un haut-parleur à la localisation cible, ceux issus de la méthode MSPA et ceux issus de la méthode VBAP. Compte tenu de la complexité pratique des prises avec le matériel disponible, plusieurs enregistrements ont été nécessaires. La spatialisation VBAP et MSPA se réalisant par paire d'enceintes, nous avons analysé les azimuts $-30°, -15°, 0°, +15°, +30°$ issus de la paire $(-30°, +30°)$, les autres localisations se faisant par translation de cette paire.

Les ILD ne sont pas toujours des indices acoustiques très fiables dans les milieux réverbérés à cause des superpositions multiples, nôtre critère de précision fut l'ITD et l'azimut déduit de ce dernier. L'ITD est en effet un indice appréciable de ce point de vue, nous l'estimons à partir d'une méthode d'intercorrélation généralisée [BKBF07] (chapitre suivant). La localisation estimée correspond au maximum de la fonction d'intercorrélation.

Les estimations de l'ITD et de l'azimut correspondant peuvent être négatifs ou positifs selon que la source fut positionnée vers l'oreille gauche ou l'oreille droite. Les figures 39 et 40 présentent les fonctions d'intercorrélation obtenues pour les signaux réels et les sources virtuelles issues de VBAP et de MSPA.

Nous remarquons aisément que les fonctions d'intercorrélation PHAT des sources réelles sont plus précises et sont localisées à la bonne position, alors que les fonctions d'intercorrélation des méthodes virtuelles présentent des pics parasites un peu plus prononcés. En revanche le pic dominant demeure un bon estimateur de la localisation. Les parasites s'expliquent par les

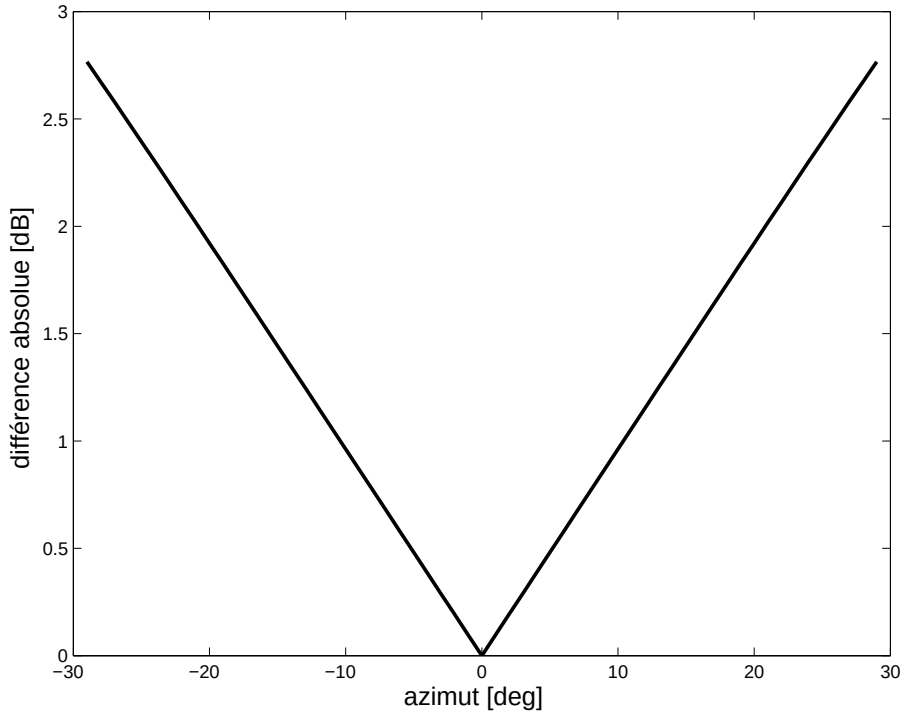


Fig. 38: *Différence maximale par azimut entre les PLD de VBAP et ceux de MSPA dans la bande* $[0, 800] Hz$.

interactions plus complexes issues de l'utilisation de deux enceintes. L'estimation des ITD de sources réelles sont plus précises et ces sources sont localisées correctement avec une légère déviation, la table 5 résume les résultats de localisation pour les trois types de sources. En effet, la localisation déduite de la diffusion de la source monophonique est la meilleure qu'on puisse espérer dans les conditions acoustiques de la salle. Elle sert ainsi de référence. Nous savons que dans les techniques par paire de haut-parleurs, la source la plus proche à tendance à l'emporter dû à l'effet de précédence. La comparaison des résultats des trois approches confirme la supériorité de la localisation de la source réelle, avec des localisations qui correspondent de près à la réalité.

Une étude poussée des résultats numériques annoncent que pour les angles négatifs, VBAP et MSPA manifestent un biais vers l'enceinte de gauche, de même pour les angles positifs, VBAP et MSPA manifestent un biais vers l'enceinte de droite, donc vers l'enceinte la plus proche de la localisation cible. Toutefois, nous notons un avantage de localisation avec les signaux binauraux issus de MSPA, cet avantage est en moyenne d'environ 2° (voir table 5). Ces constatations confirmeraient que MSPA renforce la corrélation entre l'ITD et l'ILD dans les signaux binauraux, permettant ainsi à ces derniers de mieux approcher les signaux issus de la perception naturelle.

3.4.4 Discussion

Dans ce chapitre nous avons fait des comparaisons de différentes techniques de spatialisation, tant dans la diffusion binaurale que dans la multi-diffusion. Nous avons proposé une

azimut θ	source réelle $\hat{\theta}$	MSPA $\hat{\theta}$	MSPA ITD ms	VBAP $\hat{\theta}$	VBAP ITD ms
$-30°$	$-25°$	$-27°$	-0.22	$-27°$	-0.24
$-15°$	$-12°$	$-20°$	-0.18	$-22°$	-0.20
$0°$	$+1°$	$+1°$	0.01	$+2°$	0.02
$+15°$	$+13°$	$+19°$	-0.18	$+22°$	0.20
$+30°$	$+27°$	$+27°$	-0.24	$+27°$	0.24

TAB. 5: *Estimations de l'azimut avec la méthode d'intercorrélation à partir de signaux binauraux issus d'une diffusion de bruit blanc avec MSPA avec la paire de haut-parleurs ($-30°$, $+30°$), enregistrés avec le phonocasque au studio Bonnefont.*

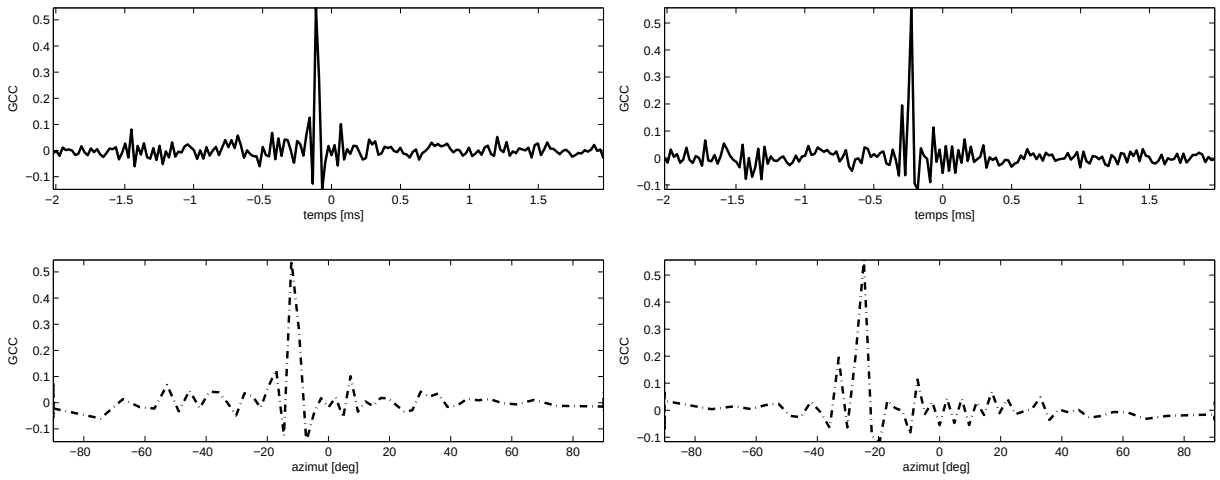


FIG. 39: *Histogrammes obtenus par inter-corrélation généralisée (PHAT) à partir de signaux binauraux issus de source réelles dans un environnement réel (studio Bonnefont). De haut en bas et de gauche à droite : $-15°$, $-30°$.*

méthode paramétrique binaurale qui nécessite uniquement un son monophonique et la localisation désirée. Cette dernière permet de s'affranchir de la contrainte de mesure de HRTF à toutes les localisations cibles, avec une bonne qualité sonore et une précision spatiale compétitives. Dans le cas de la multi-diffusion, notre but principal fut de proposer un système simple, approprié, flexible, facilement configurable pour un nombre variable de haut-parleurs. À cet effet, l'extension de la méthode binaurale à un système multi-diffusion fut retenue.

Cette dernière fonctionne par paire de haut-parleurs (MSPA) comme VBAP. La comparaison de ces méthodes montre que le son réel est d'une qualité sonore et spatiale supérieures aux deux. Les tests subjectifs et objectifs ont montré un avantage pour MSPA du point de vue de la précision spatiale de la source virtuelle. Ces deux approches souffrent toutefois d'un biais vers le haut-parleur qui contribue le plus en énergie. Cet effet n'est pas forcément observé dans les méthodes comme l'holophonie, qui malheureusement a besoin d'un nombre important de microphones et de haut-parleurs. L'erreur peut être diminuée en réduisant l'écart entre des haut-parleurs adjacents. MSPA est flexible dans la mesure où il s'adapte automatiquement pour la diffusion binaurale et transaurale.

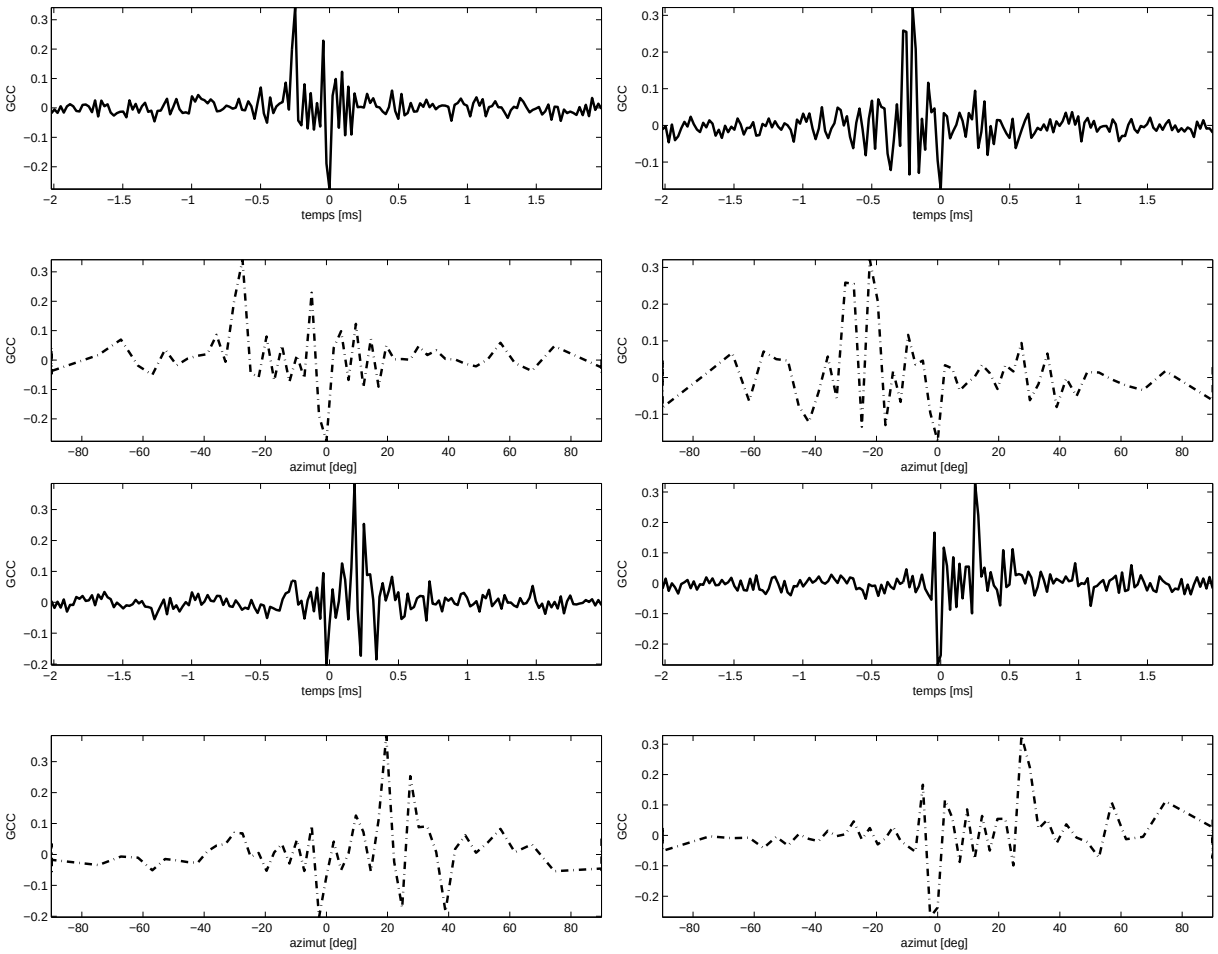


Fig. 40: *Histogrammes obtenus par inter-corrélation généralisée (PHAT) à partir de signaux binauraux enregistrés suite à la diffusion avec MSPA dans un environnement réel (studio Bonnefont). De haut en bas et de gauche à droite :* $-30°$, $-15°$, $+15°$,$+30°$.

Chapitre 4

Localisation binaurale mono-source

Ce chapitre présente nos résultats sur la localisation binaurale dans le plan horizontal. La localisation binaurale consiste à estimer la position d'une source en azimut et en distance, en se basant sur les signaux perçus par les deux oreilles. Nous nous concentrons sur la cas d'un scénario à une source unique. La localisation multi-source sera abordée dans le chapitre suivant. Dans la section 4.1, nous étudierons la localisation en azimut. Il existe une relation entre l'azimut et la différence en temps d'arrivée entre les deux oreilles ; de ce fait la localisation en azimut et en temps d'arrivée sont équivalentes.

Premièrement, nous nous intéresserons aux modèles classiques de localisation fondés sur une localisation temporel par intercorrélation entre les signaux perçus par les oreilles gauche et droite, notamment le modèle de Jeffress (1948) et le modèle de Durlach (1963). L'intercorrélation s'applique aussi bien dans le domaine temporel (section 4.1.1) que dans le domaine spectral (section 4.1.2). Sous l'inspiration de ces modèles classiques, une famille de méthodes appelées intercorrélation généralisée ont été proposées dans la littérature (section 4.1.2), ces dernières ont pour objectif directeur d'améliorer le rapport signal-bruit par filtrage des signaux d'entrée avant de procéder à l'intercorrélation. De ce fait, ces méthodes sont prisées pour la localisation dans des environnements bruités et réverbérés.

Deuxièmement, nous présentons un aperçu des méthodes de correspondance entre les indices acoustiques et l'azimut (section 4.1.3), et de vérifier leur efficacité. Partant du fait qu'à tout azimut correspondent des indices acoustiques uniques référencés dans une base, la minimisation de l'écart entre ces derniers et ceux mesurés sur un signal binaural permettrait de remonter à un azimut fiable par correspondance.

Troisièmement, nous nous pencherons sur une variante de la méthode de localisation paramétrique précédemment proposée par Viste (section 4.1.4), cette dernière s'appuie sur le modèle binaural simplifié (voir chapitre 2), et repose sur une utilisation conjointe des indices binauraux afin de déterminer un azimut à chaque point temps-fréquence et la construction d'un histogramme spatial de puissance.

Les méthodes de localisation par intercorrélation et par évaluation conjointe sont éprouvées par des simulations d'environnements synthétiques (cas idéal), et dans des environnements réels (bruités ou réverbérés) à partir de signaux mesurés dans des salles anéchoïques. Les analyses et les illustrations offrent une vue en profondeur de leurs performances et de leurs limitations.

La localisation complète dans le plan horizontal implique la localisation en distance. Dans la section 4.3, nous recourons à un modèle d'absorption de l'énergie en fonction de la fréquence et de la distance. Une étude de l'atténuation spectrale sur un bruit blanc nous a permis de suggérer un estimateur de la distance à partir d'une mesure de la brillance qui est la centroïde (section 4.4). L'estimation de la distance par la centroïde est éprouvée dans le cas idéal.

4.1 Localisation en azimut

La localisation en azimut d'une source unique est généralement aisée, car les signaux binauraux ne sont pas composés par des mélanges de signaux provenant de sources concurrentes. Cette condition est avantageuse pour la plupart de méthodes classiques qui opèrent dans l'espace temporel ou dans de larges bandes fréquentielles.

De nombreux modèles de localisation sont reconsidérés par Colburn et Durlach dans [CD78], on distingue principalement deux théories sur les modèles de localisation binaurale : le modèle de coïncidence de Jeffress [Jef48] et le modèle d'égalisation-annulation (equalisation-cancellation) de Durlach [Dur63, Dur72] (section 4.1.1). De nombreuses méthodes modernes de localisation se fondent sur l'un et/ou l'autre de ces archétypes. Notamment les méthodes d'intercorrélation généralisée dans le domaine spectral (section 4.1.2). Un autre type de localisation par une méthode de correspondance entre HRTF mesurées et azimuts est étudiée dans la section 4.1.3. Dans la section 4.1.4, nous proposons une adaptation et des améliorations numériques de l'algorithme de localisation binaurale précédemment exposé par Viste [VE03, VE04].

4.1.1 Utilisation de l'intercorrélation temporelle

Modèle de Jeffress

Le modèle de Jeffress aurait le plus influencé la localisation binaurale. Son principe est illustré dans la figure 41. Il repose sur un réseau de lignes de retard et de détecteurs de coïncidence, les détecteurs de coïncidence correspondent à des cellules nerveuses qui sont excitées lorsque les signaux retardés présents à ses deux entrées sont cohérents. La présence d'un réseau similaire a été mis en évidence par des découvertes anatomiques et physiologiques chez de nombreuses espèces [GB69], notamment chez le chat [SJY93] .

Jeffress suppose que les signaux binauraux se propagent le long du réseau dans des sens opposés, et se croisent au niveau de détecteurs de coïncidence correspondant chacun à un décalage temporel ; le nombre d'excitations au niveau d'un détecteur permet ainsi d'extraire une information de différence de temps interne entre les deux oreilles, transformant ainsi l'information neurale en information spatiale.

Ce mécanisme est assimilable à une fonction d'intercorrélation temporelle décrit par :

$$R_{gd}(\tau) = \int_{-\infty}^{+\infty} x_G(t) x_D(t-\tau) dt, \tag{80}$$

où $x_G(t)$ et $x_D(t)$ représentent les signaux de gauche et de droite, τ représente le délai interaural (ITD). Le délai interaural de tout humain est inclus dans l'intervalle $-1 \leq \tau \leq< 1$

ms.

Sous certaines conditions, à l'exemple d'un environnement réverbéré, il est souhaitable d'employer uniquement le son direct pour la précision de l'intercorrélation. Un fenêtrage des signaux d'entrée est un moyen avantageux pour limiter l'influence de la réverbération. Blauert et Cobben [BC78] ont recouru à une fonction de fenêtrage exponentielle et ont proposé une intercorrélation à court terme entre des signaux issus de filtres passe-bande, afin de prendre en compte le modèle d'analyse fréquentielle dans la cochlée. L'intercorrélation à court terme est définie par :

$$R_{gd}(\tau) = \int_{-\infty}^{+\infty} w(t-\tau) x_G(t) x_D(t-\tau)\ dt, \tag{81}$$

où $w(t)$ est une fonction de fenêtrage. La fenêtre exponentielle $w(t) = Ae^{-t/t'}$ appliquée par Blauert possède une constante temporelle t' de quelques millisecondes ($4-5$ ms). La constante temporelle doit être choisie de manière à réduire la présence des réflexions. Nous remarquons qu'il est délicat d'appliquer un tel processus sur un long signal, car le décalage temporel de la fenêtre ne concorde pas forcément avec la présence du son direct. L'arrière de la fonction pourrait bien coïncider avec l'arrivée du son direct de la prochaine trame et l'atténuer au lieu de le favoriser.

Le délai τ qui maximise la fonction d'intercorrélation (équation 82) fournit une estimation du délai interaural réel avec :

$$\mathrm{ITD} = \mathrm{argmax}_\tau \left| R_{gd}(\tau) \right|. \tag{82}$$

Le temps d'observation étant limité, l'intercorrélation est estimée en calculant l'espérance mathématique $E[x_G(t)x_D(t-\tau)]$.

Dans la pratique, l'intercorrélation normalisée (équation 83) produit généralement de meilleurs résultats, et sa valeur est comprise entre 0 et 1. L'idéal est d'obtenir la valeur maximale 1 à la localisation réelle, et une valeur nulle pour toutes les autres valeurs interaurales. L'intercorrélation généralisée est calculée selon la relation :

$$R_{gd}(\tau) = \int_{-\infty}^{\infty} \frac{x_G(t,f) x_D(t-\tau)}{\sqrt{|x_G(t)|^2 |x_D(t-\tau)|^2}}\ dt. \tag{83}$$

Le modèle de Jeffress a été étendu par Lindemann [Lin86a, Lin86b], qui inclut un modèle de l'effet de précédence et la différence interaurale en amplitude pour traiter les signaux non-stationnaires. Ce dernier fut une nouvelle fois étendu par Gaik [Gai93] qui considère une combinaison naturelle de la différence en amplitude et de la différence en temps d'arrivée.

Localisation par Durlach

Le modèle d'Egalisation-Suppression (ES) de Durlach a pour but principal d'éliminer les sources concurrentes. Les signaux provenant des deux oreilles sont égalisés en amplitude et en temps d'arrivée, puis soustraits. L'égalisation en amplitude s'effectue par atténuation afin de compenser la différence interaurale en amplitude ou ILD de la source concurrente, et l'égalisation en temps d'arrivée s'effectue en compensant l'ITD de la source concurrente (voir équation 84) :

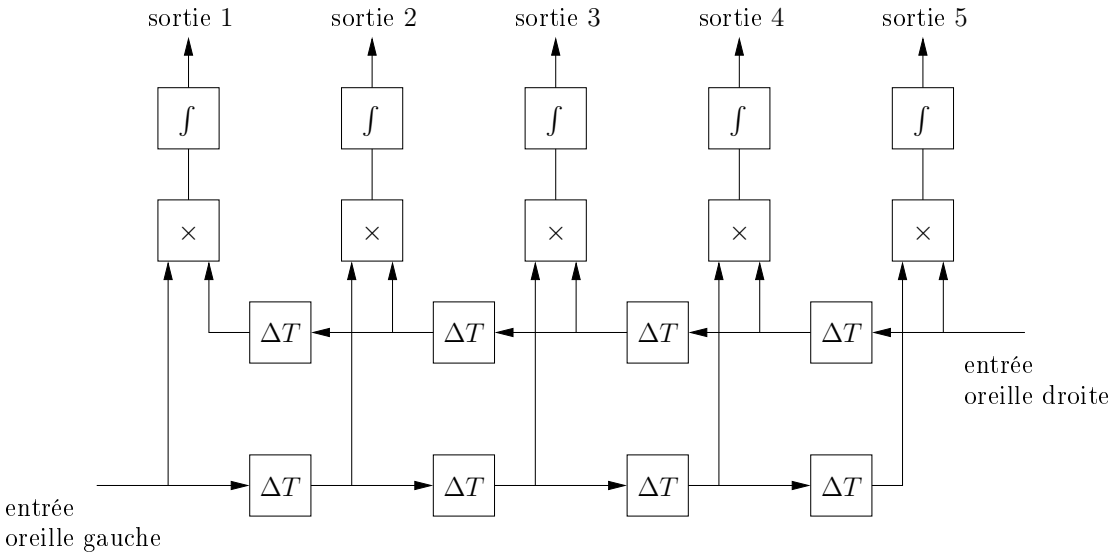


FIG. 41: *Modèle de Jeffress. Les multiplicateurs sont des corrélateurs (×) qui enregistrent les coïncidences de l'activité neuronale des deux oreilles après les délais (ΔT).*

$$R_{gd}(\tau) = \int_{-\infty}^{+\infty} (x_G(t) - \alpha x_D(t-\tau))^2 dt, \tag{84}$$

où α est le facteur d'atténuation.

Dans le modèle de Durlach, il s'agit de déterminer les paramètres α et τ qui maximisent la fonction d'intercorrélation, dans certains cas le rapport signal-bruit. Le reste de la soustraction est composé de sources dont l'ILD et/ou l'ITD diffèrent de la localisation de la source supprimée. La suppression permet la détection de la source cible après égalisation. Pratiquement, on utilise un seuil de détection binaural basé sur une combinaison du couple d'ILD/ITD qui produit la meilleure égalisation. L'effet de l'ITD domine dans la combinaison des indices binauraux.

L'ES approche ainsi le processus de suppression neuronal qui n'implique pas de mouvement de la tête, contrairement à la suppression acoustique. Ce modèle a également été modifié par Culling afin de procéder dans des canaux spectraux différents, assurant une suppression simultanée de plusieurs sources.

Toutefois, la recherche du couple (ILD, ITD) approprié est une opération exhaustive et coûteuse. Aussi, la combinaison optimale d'ILD et d'ITD n'est pas objectivement définie par les psycho-acousticiens. Dans le cas de ce travail de recherche appliqué, nous visons des applications temps réel rapides et efficaces.

4.1.2 Utilisation de l'intercorrélation spectrale

L'estimation de l'ITD peut être améliorée par un pré-filtrage des signaux gauche et droit avant l'opération d'intercorrélation. On parle d'intercorrélation généralisée. Comme indiqué sur la figure 42, les signaux gauche et droite sont filtrés respectivement par les filtres H_G et H_D ; les résultats y_G et y_D sont ensuite multipliés, intégrés, élevés au carré pour un intervalle

de délai τ. Le délai pour lequel un pic maximum est observé est une estimation du délai interaural.

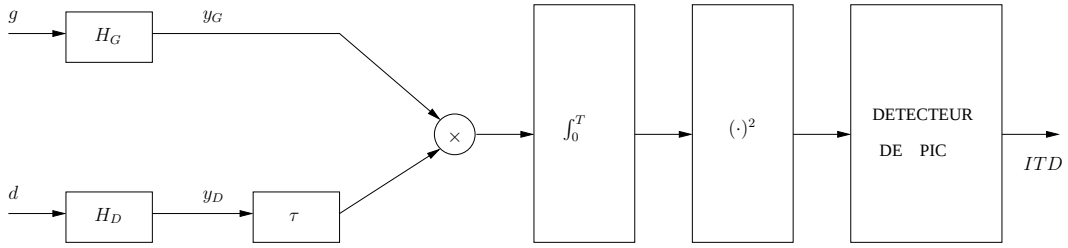


Fig. 42: *Schéma du mécanisme de l'intercorrélation généralisée.*

L'intercorrélation dans le domaine temporel est liée à la densité inter-spectrale dans le domaine spectral. Il est ainsi possible d'estimer les localisations dans différentes canaux fréquentiels conformément au système auditif humain qui décompose les signaux d'entrée en cochlégramme. La figure 43 montre un modèle simple de 10 filtres Gamma utilisés par la cochlée entre 100 Hz et 8000 Hz. Le rapport entre la fréquence centrale et la bande fréquentielle du filtre est constant. L'intercorrélation généralisée est alors définie par :

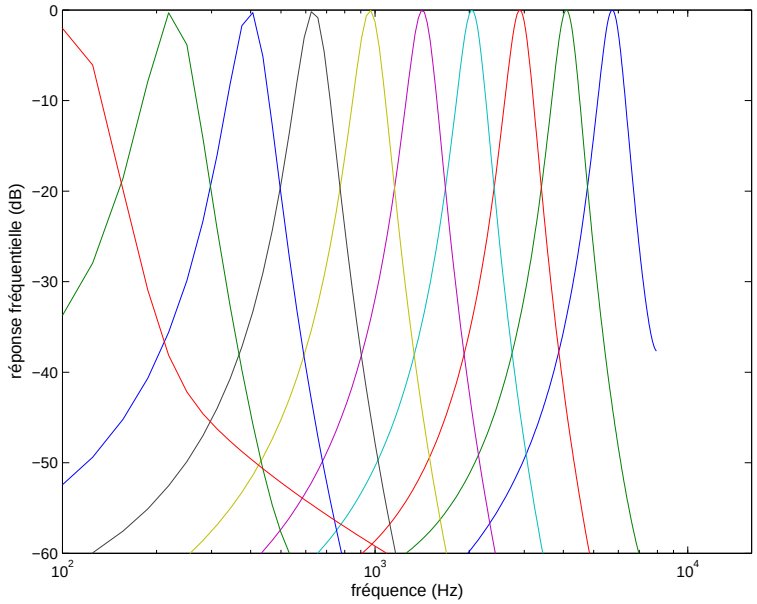


Fig. 43: *Modèle d'un banc de filtres Gamma pour la décomposition temps-fréquence par la cochlée.*

$$R_{GD}(\tau) = \int_{-\infty}^{\infty} \Phi(t,f) X_G(t,f) X_D^*(t,f)\ e^{j2\pi f\tau}\ df, \tag{85}$$

où * dénote le conjugué complexe, $\Phi(t,f)$ représente le pré-filtrage destiné à améliorer l'estimation de l'ITD. De nombreux pré-processeurs existent, parmi les plus significatifs on distingue celui de Scott [CNC73] et sa version modifiée [RRD+96], celui de Hannan-Thomson [HT73] basé sur le maximum de vraisemblance. Ces pré-filtres assignent généralement les poids en s'appuyant sur le rapport signal-bruit (RSB). Plus le RSB est élevé, plus l'estimation à

un point fréquentiel aura de l'importance dans la pondération. Nous porterons un intérêt particulier pour le pré-filtrage par *Phase Transform* ou PHAT [KC76]. En effet, ce dernier affiche moins d'étalement du pic observé que les autres pré-processeurs ; toutefois dans la pratique, une attention est à porter dans les régions où la puissance du signal est très faible. Particulièrement, si l'intercorrélation vaut zéro à un point fréquentiel, alors la phase n'étant pas définie, son estimation est erronée. Ces erreurs sont souvent uniformément distribuées dans l'intervalle $[-\pi, \pi[$.

Le pré-processeur PHAT est donné par l'équation suivante :

$$\Phi_{\mathrm{PHAT}}(t,f) = 1/\sqrt{|X_G(t,f)|^2 |X_D(t,f)|^2}, \tag{86}$$

$$\tag{87}$$

L'ITD est alors calculée par l'équation suivante :

$$\mathrm{ITD} = \mathrm{argmax}_\tau \left| R_{GD}(\tau) \right|. \tag{88}$$

Le passage de l'ITD à un azimut peut être réalisé par des relations de correspondance ITD-azimut. La section suivante expose une méthode de correspondance basée sur les indices interauraux estimés à partir des HRTF de la base CIPIC.

4.1.3 Utilisation des HRTF

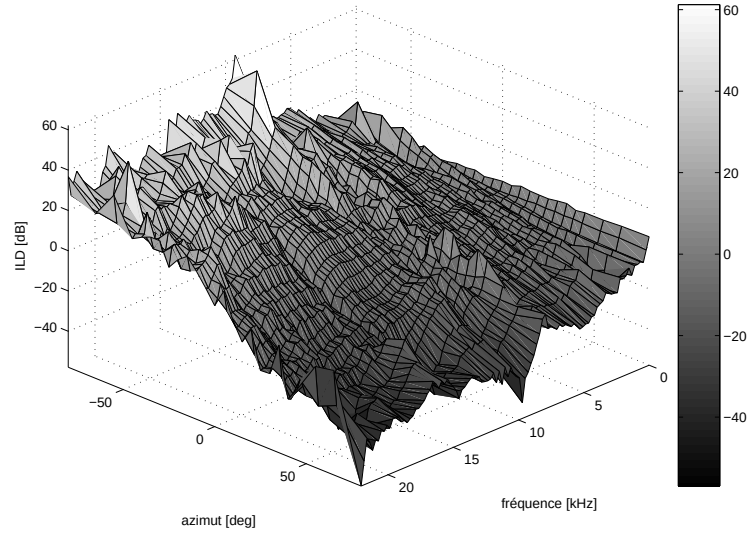
Les réponses impulsionnelles relatives à la tête (HRIR) pour plusieurs azimuts sont mesurées pour les deux oreilles, et ce pour 45 sujets dans la base CIPIC. Les HRIR sont d'une longueur d'environ 5 ms afin de réduire la présence d'écho et de ne conserver que le son direct. Du fait de la durée très courte des ces filtres, on les suppose stationnaires. Leur fonction de transfert ou HRTF est alors dépendante de la fréquence et non du temps, de plus la HRTF dépend de l'angle d'azimut et de la morphologie de l'auditeur.

Pour une paire de HRTF, $\mathrm{H}_G(\theta, f)$ et $\mathrm{H}_D(\theta, f)$, on estime les $\mathrm{ILD}_{\mathrm{HRTF}}(\theta, f)$ et les $\mathrm{ITD}_{\mathrm{HRTF}}(\theta, f)$ de référence comme fonctions de l'azimut et de la fréquence (équations 89 et 90). Similairement, à partir de signaux binauraux enregistrés, les ILD et ITD correspondantes sont estimables. La comparaison de ces dernières avec les indices acoustiques issus des HRTF permet d'aboutir à une estimation de l'azimut, par simple recherche de données [HPP05]. A cet effet, nous pouvons minimiser la distance entre les indices acoustiques de reférence et ceux mesurés (équations 92 et 91). On a :

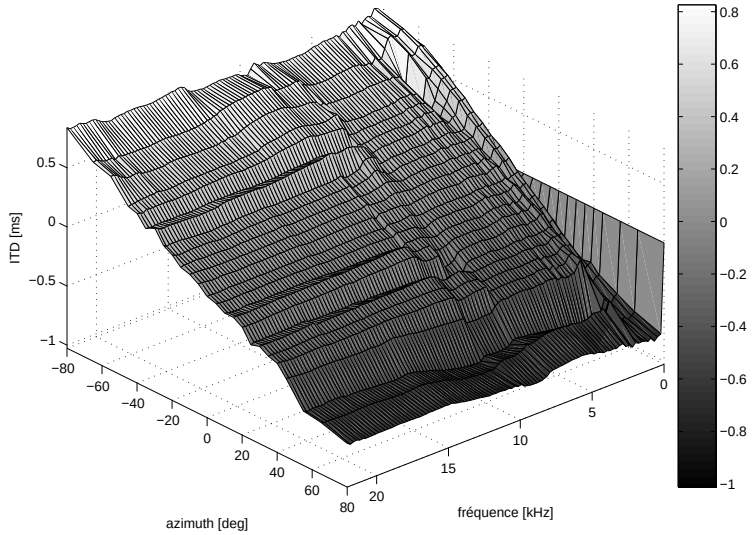
$$\mathrm{ILD}(\theta, f) = 20\log_{10} \left| \frac{\mathrm{H}_G(\theta, f)}{\mathrm{H}_D(\theta, f)} \right|, \tag{89}$$

$$\mathrm{ITD}_p(\theta, f) = \frac{1}{2\pi f} \left(\angle \frac{\mathrm{H}_G(\theta, f)}{\mathrm{H}_D(\theta, f)} \right) + 2\pi p(f). \tag{90}$$

Les erreurs quadratiques sont calculées avec :



(a)



(b)

Fig. 44: *Sujet numéro* 1 *de la base CIPIC : ILD*(θ, f) *(a), ITD*(θ, f) *(b).*

$$e_{\mathrm{ITD}} = \int_f \left\{ \mathrm{ITD}(\hat{\theta}, f) - \mathrm{ITD}_{\mathrm{HRTF}}(\theta, f) \right\}^2 df, \tag{91}$$

$$e_{\mathrm{ILD}} = \int_f \left\{ \mathrm{ILD}(\hat{\theta}, f) - \mathrm{ILD}_{\mathrm{HRTF}}(\theta, f) \right\}^2 df, \tag{92}$$

où $\mathrm{ITD}_{\mathrm{HRTF}}$ et $\mathrm{ILD}_{\mathrm{HRTF}}$ sont les ILD et ITD obtenues à partir des HRTF à la position θ. ITD et ILD sont les indices binauraux obtenus à partir des signaux binauraux en cours de

traitement ; et $\hat{\theta}$ est l'estimation de l'azimut θ pour lequel les erreurs $e_{\text{ILD}}, e_{\text{ITD}}$ sont minimales.

La figure 44 représente des tables d'ILD et d'ITD, dans toute la bande fréquentielle audible, obtenues pour un sujet de la base CIPIC à différents angles. Nous observons que des fréquences différentes et des azimuts différents peuvent disposer de mêmes indices acoustiques, donc une ambiguïté. Des expérimentations menées par Hwang *et al.* montrent que cette approche fournit de performances appréciables dans les milieux anéchoïques [HPP05]. Cette approche a le mérite d'être simple, mais la recherche est exhaustive, car au pire, tous les angles devraient être testés. De plus elle nécessite la mesure de HRTF à toutes les positions envisageables, et pour chaque sujet individuel. Il est ainsi nécessaire de recourir à une méthode paramétrique suffisamment précise et généralisable pour plusieurs sujets et aux azimuts non mesurés. Viste s'est basé sur un modèle paramétrique afin de proposer une méthode de localisation adaptée dans une certaine mesure aux sujets représentatifs de la base CIPIC [VE04]. Intuitivement, le système auditif humain ne procéderait pas séparément dans l'évaluation de l'ILD et de l'ITD, mais arbitrerait les deux indices dans le même processus.

4.1.4 Utilisation conjointe des indices binauraux

Le système auditif humain dispose de plusieurs indices acoustiques afin de localiser les sources, en plus des indices acoustiques binauraux, des indices de plus haut niveau sont impliqués. Ces indices ne sont pas facilement modélisables et leur lien avec la localisation ne paraît pas évident et traçable. Intéressons-nous à l'ILD et à l'ITD. Les modèles génériques proposés de l'ILD et de l'ITD (équations 35 et 38) dépendent de paramètres communs à savoir le rayon de la tête et l'angle d'azimut. Ces modèles sont optimisables pour chaque individu, toutefois dans cette section nous considérons les modèles moyens, qui sont utilisables pour différents sujets.

Évaluation conjointe de l'ILD et de l'ITD

L'estimation de l'azimut par HRTF indique que l'ILD et l'ITD sont toutes les deux reliées à l'azimut, par déduction ces deux indices acoustiques peuvent être joint entre eux. Viste a proposé une méthode d'évaluation conjointe de l'ILD et de l'ITD afin d'estimer l'angle d'azimut [VE03], [VE04]. L'azimut estimé à partir du modèle de l'ILD ne souffre d'aucune ambiguïté, toutefois cette estimation a une large variance (allure très rugueuse sur la figure 44) ; cependant l'azimut estimé à partir du modèle d'ITD a une faible variance (allure lisse sur la figure 44) mais est entachée d'ambiguïté par l'ajout d'un multiple de 2π au-delà de 1500 Hz. L'idée consiste à considérer l'azimut basé sur l'ILD comme référence, et de dérouler l'azimut basé sur l'ITD de sorte que ce dernier soit le plus proche possible de l'azimut basé sur l'ILD.

Nous avons adapté cette technique à notre modèle de la tête suivant le processus suivant.

Une estimation de l'azimut à partir de l'ILD (équation 40) est obtenue simplement par inversion de l'équation (35) :

$$\theta_L(t,f) = \arcsin\left(\frac{\text{ILD}(t,f)}{\alpha(f)}\right). \tag{93}$$

L'argument de la fonction arcsin doit être compris entre -1 et 1 afin que l'inverse soit un

angle réel.

Similairement, un candidat de l'azimut pour chaque entier p à partir de l'ITD (équation (41)), est obtenu par inversion de l'équation (38) :

$$\theta_{T,p}(t,f) = \arcsin\left(\frac{c \cdot \mathrm{ITD}_p(t,f)}{r \cdot \beta(f)}\right). \tag{94}$$

Les estimations $\theta_L(t,f)$ sont plus dispersées, cependant elles ne subissent aucune ambiguïté à chaque fréquence, ainsi ces estimations sont exploitées objectivement afin de retrouver le facteur de modulo p qui permet un déroulement correct de la phase. Ensuite, l'estimation $\theta_{T,p}(t,f)$ la plus proche de $\theta_L(t,f)$ est validée comme l'estimation finale $\hat{\theta}(t,f)$ pour la composante fréquentielle traitée. L'estimateur final assure un déroulement de la phase et affiche un faible écart-type :

$$\theta(t,f) = \theta_{T,m}(t,f), \tag{95}$$

$$\text{avec} \quad m = \mathrm{argmin}_p \left|\theta_L(t,f) - \theta_{T,p}(t,f)\right|.$$

Pratiquement, nous avions dans un premier temps fait usage d'une boucle sur les valeurs entières possibles, une valeur de p allant de -12 à 12 par pas de 1 pour couvrir toutes les ITD possibles. Mais la complexité est grande car le traitement est exécuté par indice fréquentiel. Pour un spectre de 2048 points on s'impose déjà $25 * 1025$ opérations et tests. Dans un soucis d'optimisation, nous avons montré que le choix de l'entier p peut se réduire à un choix ($\lceil p_r \rceil$, $\lfloor p_r \rfloor$), avec

$$p_r = f \cdot \mathrm{ITD}(\theta_L, f) - \frac{1}{2\pi}\angle\frac{X_G(t,f)}{X_D(t,f)}. \tag{96}$$

La complexité de l'algorithme est alors réduite à $2 * 1025$ opérations par indice fréquentiel, ce qui représente un gain conséquent pour de longs signaux.

Théoriquement, dans le cas d'une source unique, toutes les fréquences devraient afficher la même estimation de l'azimut, qui correspondrait exactement à la localisation de la source. Cependant, dans la pratique, les bruits dus à l'acquisition des signaux, aux traitements et aux observations induisent des écarts ; on observe généralement un nuage de points autour de la bonne valeur. Nous obtenons une estimation de l'azimut de la source comme le pic culminant d'un histogramme pondéré en énergie [MM06]. Plus précisément, pour chaque composante temps-fréquence, l'azimut est estimé et l'énergie correspondante est cumulée à la position dans l'histogramme. La construction de notre histogramme est détaillée dans la suite.

Construction de l'histogramme spatial

Nous définissons un axe d'abscisse discret pour les valeurs possibles de l'azimut, avec une résolution Δ_θ. À chaque estimation $\theta(t,f)$, l'énergie du point fréquentiel est répartie sur l'axe spatial selon une distribution Gaussienne g, selon :

$$h(\theta) = \sum_f \left|g(\theta, \sigma^2 = 1)(t,f) X_G(t,f) X_D(t,f)\right|, \tag{97}$$

où $g(\theta)$ est une fonction Gaussienne de moyenne θ, et de variance $\sigma^2 = 1$, qui assure une répartition lisse de l'énergie sur l'axe spatial θ.

Notons que l'énergie cumulée est approchée selon $X_G X_D = X^2$ (équations 35 et 38), qui serait globalement l'énergie de la source, lorsqu'elle est jouée à l'azimut $\theta = 0$ (cas monophonique avec $x_G = x_D$). à un décalage de l'axe des abscisses θ de l'histogramme, et l'énergie serait conservée. La figure 45 montre un exemple d'histogramme obtenu pour une source synthétique à la position $-45°$. Pratiquement, nous constatons la présence de pics parasites qui sont dus à des valeurs extrêmes de l'ILD et à la présence de bruit. En effet, si la valeur mesurée de l'ILD est supérieure au maximum de $\alpha(f)$, la valeur absolue du sinus devient supérieure à 1 (voir équation 93), de même dans les bandes fréquentielles où seul le bruit ambiant est présent, les ratios ne sont plus exactement conformes aux modèles binauraux. Plusieurs expérimentations ont été menées afin d'évaluer les performances des modèles de localisation. Les résultats sont présentés dans la section 4.2.

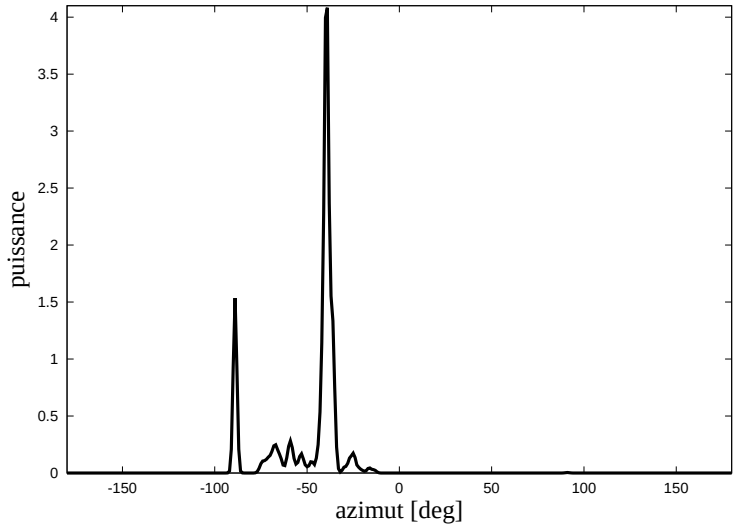


FIG. 45: *Histogramme obtenu avec une source à l'azimut $-45°$. On peut observer clairement les deux maxima dominants (pics) : L'un autour de l'azimut $-45°$, l'autre à l'azimut $-90°$. Le plus élevé correspond à la source sonore et le second est un pic fantôme résultant des ILD extrêmes.*

4.2 Résultats de la localisation en azimut

Les méthodes de localisation présentées et améliorées ont été implantées sous Matlab et testées avec des données issues des espaces anéchoïques (synthétiques) et des données enregistrées dans une salle réelle (réverbérée). Des signaux sources à différents contenus fréquentiels sont utilisés, afin de mettre en exergue une éventuelle dépendance de la localisation au contenu fréquentiel de la source. Les méthodes sont ainsi évaluées sous différents aspects. Dans cette section, nous rapportons les résultats des simulations conçues ; la mesure principale de l'efficacité des méthodes est l'erreur absolue entre la localisation réelle et la localisation estimée. Nous présentons d'abord les objectifs et la méthodologie des tests de localisation menés. Ensuite, nous caractérisons les données test utilisées, en mettant en perspective leur acquisition (section 4.2.1). Les analyses des résultats sont appuyées par de nombreuses illustrations (sections

4.2.2, 4.2.3 et 4.2.4).

4.2.1 Données de tests

Nous utilisons différents stimuli, à savoir des bruits blancs, des signaux de parole, des signaux de musique, des signaux d'instruments (figure 46). La diversité des stimuli permet de vérifier l'effet du contenu temporel et spectral sur la localisation. Les algorithmes de localisation binaurale emploient des signaux binauraux (avec un canal gauche et avec un canal droit) issus de différents environnements. Les trois types d'environnements rencontrés dans la pratique sont des environnements synthétiques (sans écho) comme dans une salle sourde, un environnement bruité dans lequel un bruit ambiant quasi-constant persiste, et un environnement réel réverbéré avec des effets de réflexions, de réverbération, d'atténuation. La difficulté de localisation dans cette dernière situation est généralement plus difficile.

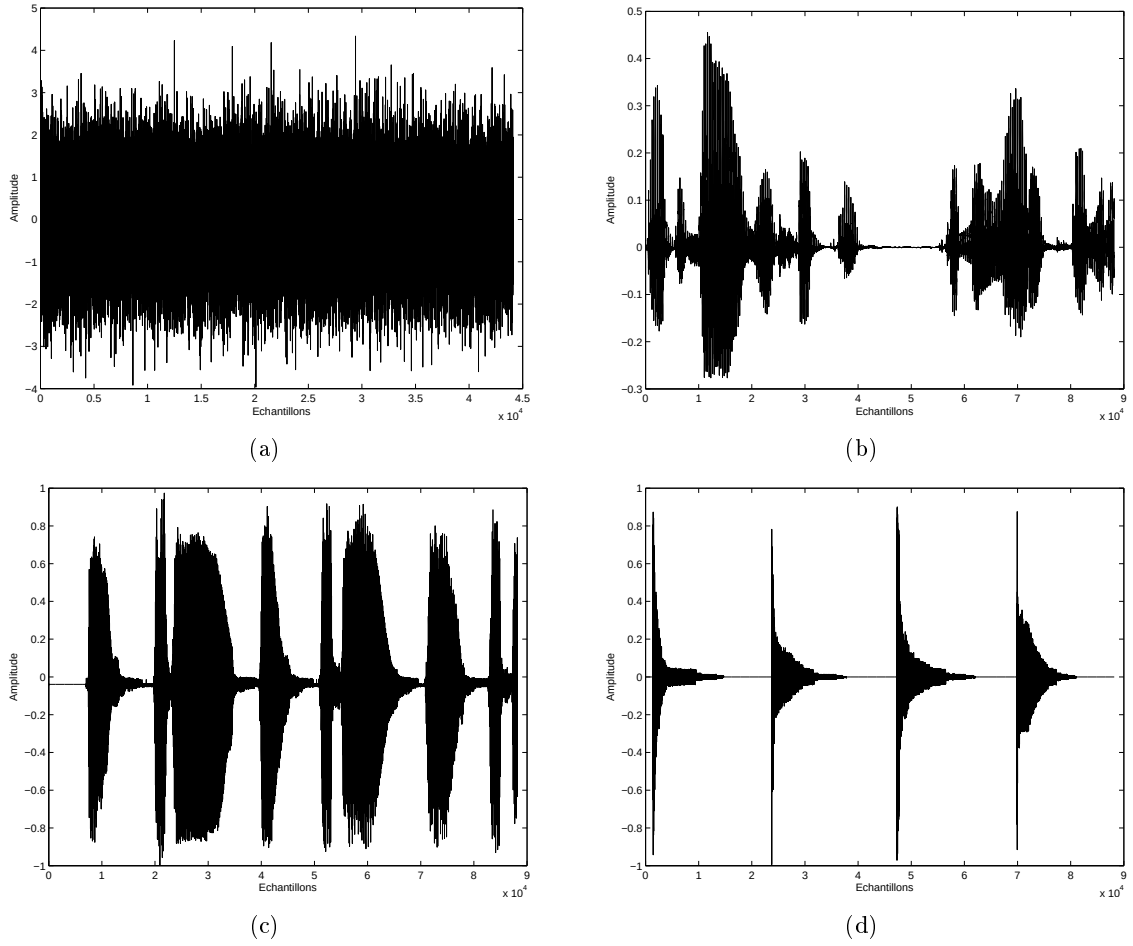


Fig. 46: *Signaux temporels : bruit blanc (a), parole (b), trompette (c) et xylophone (d).*

Signaux synthétiques

Les signaux binauraux qui nous servent de matériel de tests sont de deux catégories : les signaux binauraux synthétiques (environnement anéchoïque) et les signaux binauraux réels (environnements réverbérés).

Un signal binaural synthétique à la position θ est construit par convolution d'un signal monaural s avec les HRIR gauche $\mathrm{HRIR}_G(\theta)$ et droite $\mathrm{HRIR}_D(\theta)$. Les signaux gauche x_G et droit (x_R) sont donnés par :

$$x_G = s \otimes \mathrm{HRIR}_G(\theta), \quad (98)$$
$$x_D = s \otimes \mathrm{HRIR}_D(\theta), \quad (99)$$

où $*$ représente l'opération de convolution entre des signaux temporels, les HRIR étant des réponses impulsionnelles. Les HRIR sont longues de 200 échantillons à une fréquence d'échantillonnage de 44.1 kHz. Dans un tel intervalle de temps le son direct domine généralement les réflexions.

Signaux en environnement bruité

Nous allons également évaluer les méthodes de localisation sur des signaux binauraux bruités. Il existe différents types de bruit possible, toutefois ici, le bruit ambiant est modélisé par un bruit blanc Gaussien additif. Le signal binaural bruité est donné par :

$$x_G = s \otimes \mathrm{HRIR}_G(\theta) + n_G, \quad (100)$$
$$x_D = s \otimes \mathrm{HRIR}_D(\theta) + n_D, \quad (101)$$

où n_G et n_D sont des bruits blancs Gaussiens décorrélés de s.

Signaux réverbérés

Nous disposons de deux types de signaux réverbérés, les uns sont produits avec convolution de signaux sources avec les fonctions de transfert des milieux réverbérés (Binaural Room Impulse Response–BRIR), les autres sont obtenus par enregistrements direct au niveau des oreilles à l'aide de microphones acoustiques.

La première catégorie de signal réverbéré est obtenue par convolution de signaux sources avec les BRIR gauche et droite de la localisation θ avec :

$$x_G = s \otimes \mathrm{BRIR}_G(\theta), \quad (102)$$
$$x_D = s \otimes \mathrm{BRIR}_D(\theta). \quad (103)$$

La seconde catégorie de signal réverbéré est un signal binaural réel enregistré dans une salle réverbérée. Un auditeur porte un casque dans lequel sont encastrés des microphones miniatures (voir figure 47, où des microphones Sennheiser KE4-211-2 ont été insérés dans les

écouteurs standard). L'auditeur oriente sa tête vers la position azimut zéro. Le signal diffusé par une source à la position réelle θ est capturé au niveau du casque et forme le signal binaural réel.

Fig. 47: *Le "phonocasque" utilisé pour l'enregistrement des signaux binauraux : casque standard avec des capsules de microphone insérés.*

Les signaux tests formés sont injectés à l'entrée de chaque modèle de localisation, générant des localisations pour différentes positions, et différentes bandes fréquentielles.

4.2.2 Localisation en environnement anéchoïque

Dans ces exemples, les signaux binauraux comportent une seule source. Nous étudions la localisation de signaux de bruit, de signaux vocaux et de signaux d'instruments. Tous les signaux utilisés ont une durée de 1 seconde. Nous mettons en concurrence la localisation paramétrique conjointe et la localisation par intercorrélation.

La localisation paramétrique utilise une fenêtre glissante de Hann de 2048 échantillons, avec un taux de recouvrement de 50%. Les simulations sont exécutées sous Matlab 7.

- **Bruit blanc**
  Le premier signal étudié est un bruit blanc Gaussien (voir figure 46). Un bruit blanc a la particularité d'avoir une densité spectrale de puissance constante, qui garantit une activité dans chaque bande fréquentielle. Plusieurs bruits blancs ont été spatialisés à différents azimuts dans le plan horizontal, entre $-80°$ et $+80°$, et nous les localisons avec la méthode paramétrique conjointe et à la méthode d'intercorrélation normalisée.
  Intéressons-nous d'abord à la méthode de localisation paramétrique conjointe proposée et basée sur un histogramme de puissance. Théoriquement, notre histogramme afficherait une impulsion unique à l'azimut recherché. Dans la pratique, un bon estimateur engendrera un regroupement de points autour de la bonne valeur avec une faible variance

[Kro96]. Sur les histogrammes obtenus (figure 48), on observe différents pics, particulièrement un pic dominant très proche de la valeur idéale et des pics indésirables. Les pics parasites prennent de l'amplitude lorsque la source se déplace vers les extrêmes. On note que la source à $-15°$ présente moins de dispersion que la source à $+55°$ et que le rapport le plus élevé entre le pic culminant de l'histogramme et le plus haut pic parasite se situe à l'azimut zéro. Les pics superflus à $\pm 90°$ résultent de l'implantation de l'algorithme, qui alloue l'énergie des azimuts au-delà de $|80|°$ aux extrémités, y compris celle issue du bruit. Ces pics pourraient être ignorés car ils sortent de l'espace cible des simulations, soit $-80°$ et $+80°$.
Il est important d'observer que dans tous les cas, le pic dominant est un bon estimateur de la position réelle de la source. Dans la plupart des cas, un seuillage à 20% du pic maximum permettrait aisément de minimiser de nombreux pics parasites, aussi une opération de lissage de l'histogramme serait également bénéfique afin de fondre les pics très proches. Ces pré-traitements seront explorés davantage dans la cas de mélanges multi-source (chapitre 5). Les résultats de localisation pour la méthode conjointe sont résumés dans la figure 49. La méthode conjointe devient de moins en moins précise au fur et à mesure que la source se rapproche des extrémités latérales. Ce phénomène est également présent dans les tests d'écoute, les sons latéraux sont plus difficiles à localiser. On constate que l'erreur absolue est inférieure à $5°$ dans l'intervalle $[-65, +65]°$.
Dans un second temps nous comparons la méthode paramétrique conjointe et la méthode d'intercorrélation généralisée. La figure 50 montre les azimuts cibles contre les azimuts estimés pour les deux approches. Comme nous pouvons le constater, les deux approches sont quasi parfaites dans l'intervalle $[-45, +45]°$ avec une erreur maximale de $3°$, au-delà de $|45°|$, les deux méthodes deviennent progressivement instables. La méthode d'évaluation conjointe annonce une erreur inférieure à son protagoniste. Cette performance est qualitativement acceptable comparée au système auditif humain, qui détecterait des différences de $1°$. En pratique, la source n'est pas un point (mais étend son activité autour d'un ensemble de points), la taille de la membrane du haut-parleur et l'intensité de la source influencent ainsi la différence d'angle détectable. Nous discernons que l'erreur introduite par le modèle de localisation conjointe sur ces signaux est systématiquement positive ($-15° \rightarrow -11°, +55° \rightarrow +62°$).
Des tests similaires ont été menés sur des sources ayant un contenu spectral différent, notamment les signaux de parole et de musique dont la trompette est un représentant. Les résultats sont exposés dans les sections suivantes.

- **Signal vocal**
  Considérons maintenant des signaux de parole. De manière général, nous remarquons que les histogrammes formés mettent en évidence moins de crêtes parasites, ce qui facilite la recherche de maximum global dans le processus de localisation (figure 48). Au vu des histogrammes, les résultats de localisation avec des signaux de paroles sembleraient meilleurs que dans le cas de bruits blancs. L'embellie des histogrammes a deux raisons principales. La première est le contenu spectral de la voix, dont la fréquence maximale n'excède pas 10kHz. Précisément, dans les hautes fréquences les erreurs de localisation auraient lieu sur des zones de très faible énergie. Le contenu fréquentiel influencerait ainsi la précision de la localisation. Toutefois, remarquons que la précision ne semble pas bénéficier d'un gain supplémentaire dans le cas d'une source unique, en dehors de l'aspect visuel. En revanche dans le cas multi-source, la forme de l'histogramme pourrait être déterminant,notamment pour des sources spatialement proches.

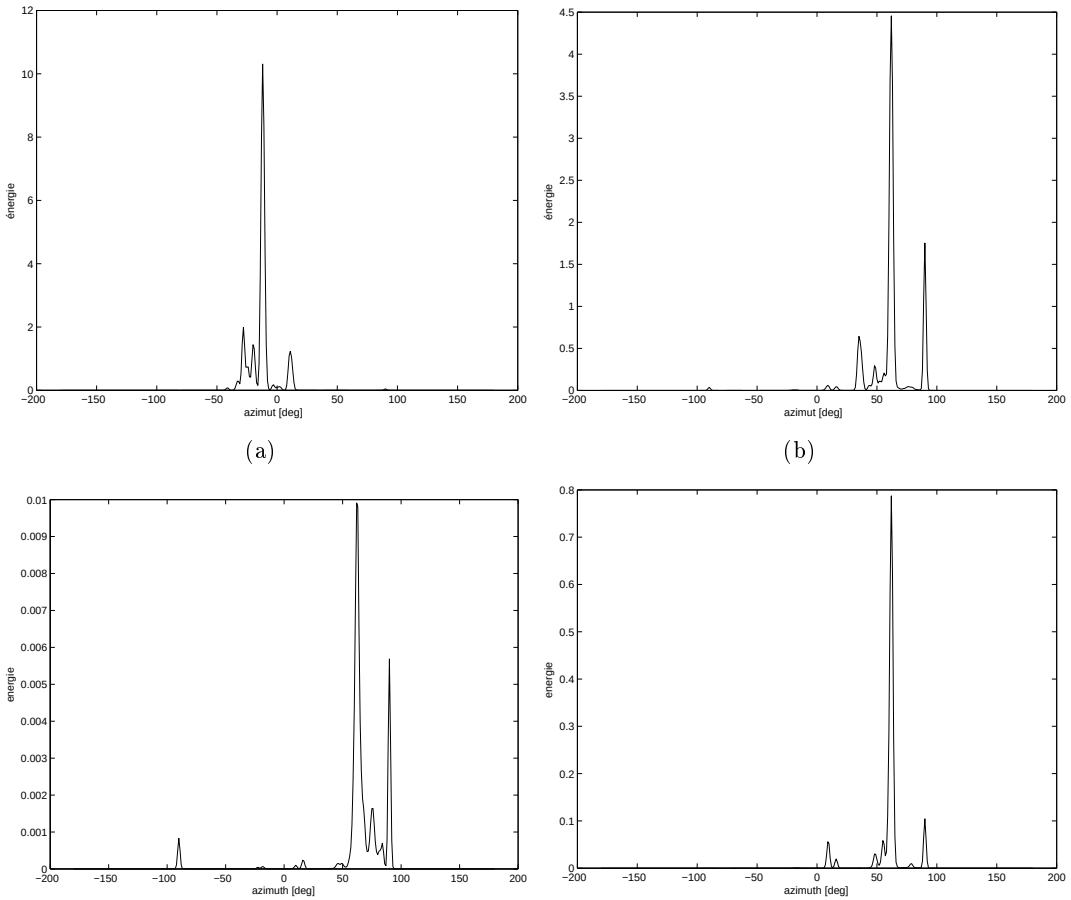


Fig. 48: *Histogrammes obtenus avec différents signaux synthétiques : bruit blanc spatialisé à l'azimut* $-15°$ *(), le pic le plus élevé est à* $-12°$*, bruit blanc spatialisé à l'azimut* $+55°$ *(b), le pic le plus élevé est à* $+62°$*. Signal vocal spatialisé à l'azimut* $+55°$ *(c), le pic le plus élevé est à l'azimut* $+62°$*. Signal de trompette spatialisé à l'azimut* $+55°$ *(d), le pic le plus élevé est à l'azimut* $+62°$

Les résultats de localisation de source de parole spatialisée avec la méthode paramétrique conjointe et la méthode d'intercorrélation généralisée sont affichés dans la figure 50. Les deux méthodes ont des précisions équivalentes avec un léger avantage pour la méthode d'intercorrélation.

– **Signal de trompette**

Le prochain signal test est un son de trompette ; il a la caractéristique principale de produire de longues attaques à forte amplitude. Nous constatons une amélioration visuelle des histogrammes (figure 48) qui pourrait s'avérer bénéfique dans le cas multi-source. Cette amélioration est due premièrement au contenu fréquentiel limité, deuxièmement aux fortes attaques qui favorisent la localisation. Les zones de forte énergie dominent dans la pondération. Les résultats de localisation pour la méthode conjointe et la méthode d'intercorrélation généralisée ont les mêmes caractéristiques que ceux des signaux vocaux résumés dans la figure 50.

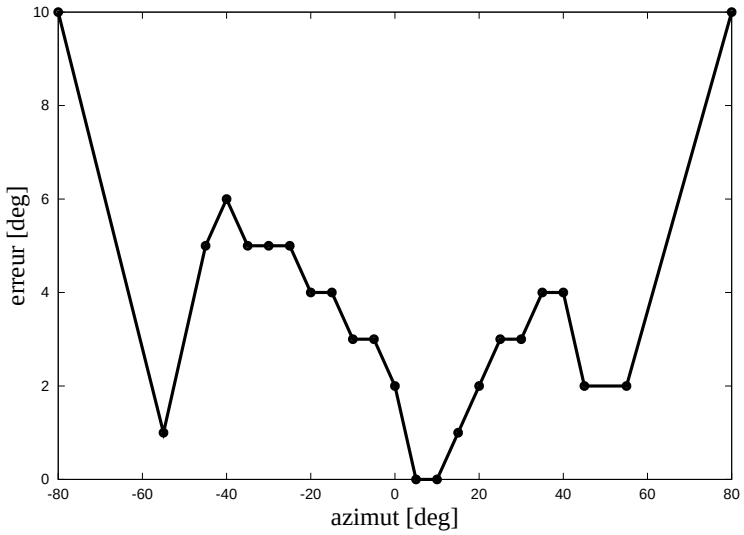


FIG. 49: *Erreur absolue de l'estimation de l'azimut à partir de signaux de bruit blanc Gaussien spatialisés à différent azimuts par convolution avec des HRIR du mannequin KEMAR.*

Dans le cadre d'une seule source synthétique présente, la méthode de localisation par évaluation conjointe et les méthodes d'intercorrélation ont des résultats similaires quel que soit le type de signal source (bruit, parole, instrument). La question ouverte est celle de leur adaptation et leur efficacité en présence de bruit et de réverbération.

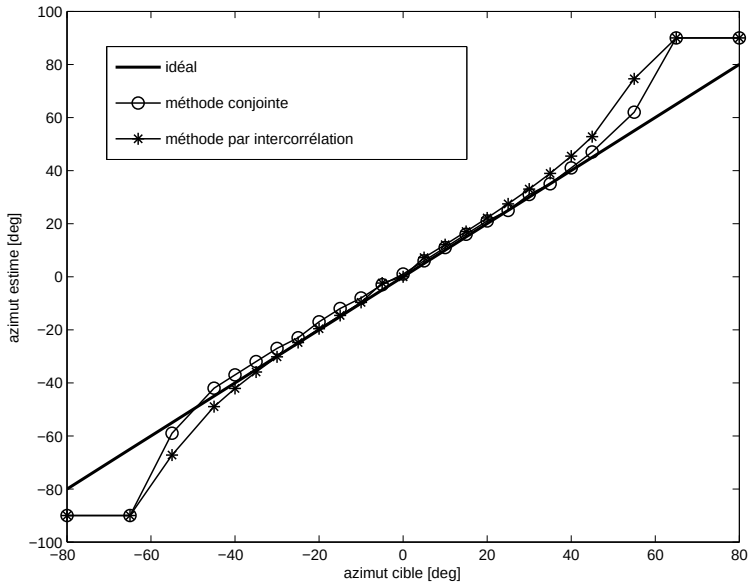
4.2.3 Localisation en environnement bruité

Dans cette section, nous étudions la robustesse de la méthode conjointe face au bruit ambiant. Le bruit ambiant est modélisé par un bruit blanc additif (section 4.2.1), et le rapport signal-bruit (RSB) correspond à une mesure quantitative de la qualité de l'environnement. Plus le RSB est élevé, moins perturbé est le canal de transmission. Pratiquement, le bruit additif est ajouté de sorte que le RSB moyen soit le même sur les deux canaux du signal binaural.

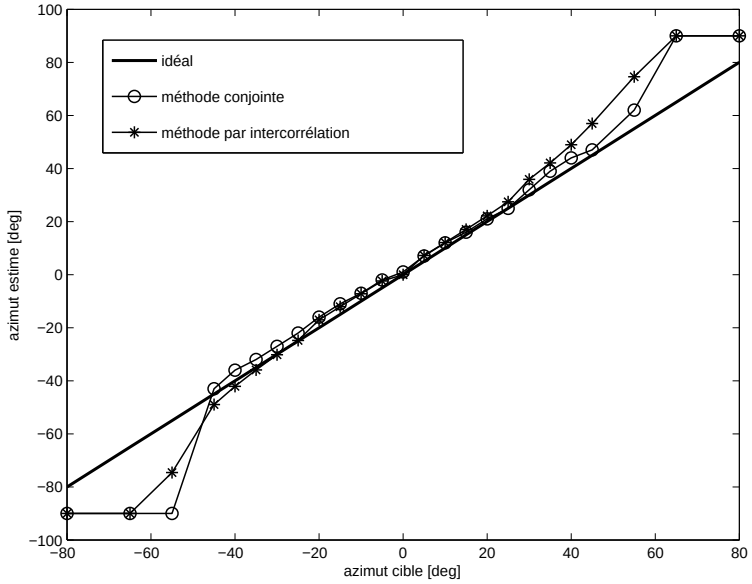
Le bruit a tendance à décorréler les signaux gauche (x_G) et droit (x_D), avec pour conséquence de dénaturer la paire (ILD, ITD), et donc une altération des estimations à chaque fréquence. La figure 51 montre des histogrammes pour différentes valeurs de RSB (15, 5, 0, -5) en dB à partir d'un signal de parole à l'azimut 30°. La localisation de la source n'est pas modifiée, toutefois, les pics parasites prennent de l'amplitude, particulièrement aux extrémités et au centre. L'affichage de l'histogramme se détériore avec la détérioration du RSB. Le pic parasite au centre prend de l'amplitude au point de dominer le pic souhaité pour un RSB $= -5$ dB. La position du pic parasite dominant dépend du rapport entre les points du spectre de puissance du bruit dans les canaux gauche et droit, conformément aux modèles binauraux paramétriques.

Nous remarquons qu'en présence de bruit ambiant la méthode de localisation conjointe est robuste, même à 0 dB. En effet, la localisation en présence d'une seule source est similaire au cas synthétique.

Le cas réel implique, en plus du bruit, la réverbération de la salle. Dans la section suivante nous comparons les performances de l'approche conjointe et à celle de l'intercorrélation en



(a)



(b)

Fig. 50: *Localisation du signal bruité (a) et signal de parole (b) avec différentes approches : idéal (plein), méthode conjointe (rond), méthode d'intercorrélation normalisée (astérisque).*

milieu réverbéré.

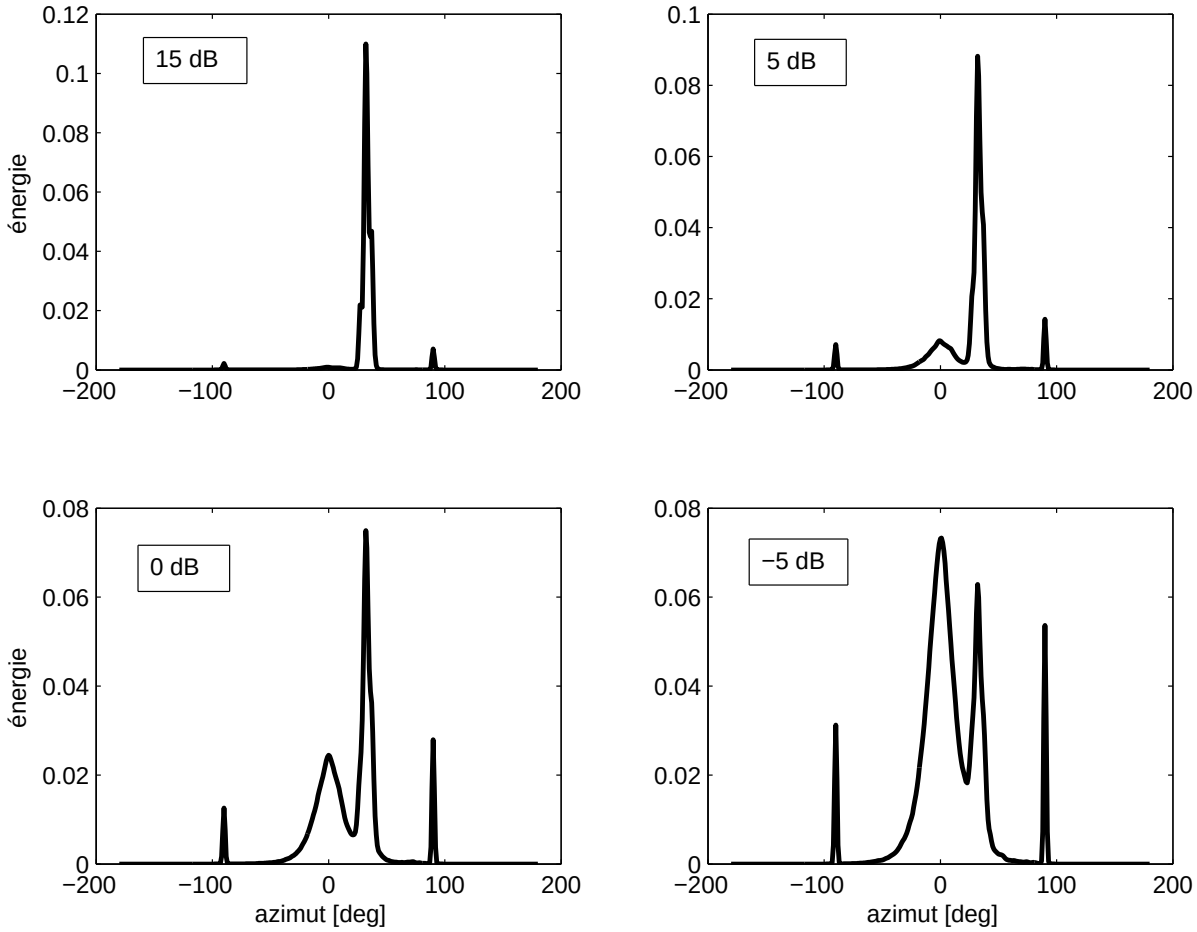


FIG. 51: *Histogrammes obtenus avec une source de parole positionnée à l' azimut* 30° *dans un environnement bruité à différents rapports signal-bruit (RSB= 15, 5, 0, -5 dB).*

4.2.4 Localisation en environnement réverbéré

Dans un environnement réverbéré le son direct émanant de la source est accompagné de réflexions sur les surfaces physiques du milieu environnant. Le système auditif humain utilise ainsi plusieurs mécanismes afin de détecter la source active et de supprimer les effets des réflexions concurrentes. Les réflexions influencent l'estimation des indices acoustiques, notamment l'ILD et l'ITD. Les méthodes de localisation doivent être robustes en présence de réverbération. De nombreux auteurs se sont intéressés aux scénarios où un ou deux distracteurs particuliers sont présents (indépendants ou stationnaires) [Bra02, GG96]. En effet, la superposition des signaux directs et des réflexions au niveau des oreilles engendre des rapports d'amplitude moins appropriés au modèle [BSCM05]. De ce fait, la méthode conjointe de localisation basée sur l'information d'ILD comme référence pourrait s'avérer moins appropriée dans certains cas. En revanche, l'ITD semble être un indice plus robuste aux interférences. Certains auteurs ont proposé des méthodes qui s'appuient sur une modélisation de l'effet de précédence [FM04], [Zur80]. Toutefois, ces méthodes nécessitent une quantité d'information *a*

priori, notamment la forme de la fenêtre utilisée, la longueur de la fenêtre, la déviation des indices acoustiques. Et ces méthodes sont généralement appliquées dans les basses fréquences. Toutefois, les méthodes basées sur l'intercorrélation généralisée sont reconnues pour être de bons candidats dans les environnements réverbérés. Dans cette section nous présentons les résultats obtenus par la méthode d'évaluation conjointe et la méthode PHAT, sur des signaux enregistrés dans des salles, dont le studio Bonnefont (8m × 10m × 4m). D'autre part, nous utilisons des fonctions de transfert d'une salle de classe réverbérée (5m × 9m × 3.5m) pour générer des signaux binauraux réels [BSCM05] conformément aux équations 102 et 103.

Studio Bonnefont

Un auditeur porte un phonocasque et tient la tête orientée vers l'azimut zéro. Un haut-parleur est positionné dans le même plan que l'auditeur, à l'azimut θ, et à au moins 1m (champ lointain). Le haut-parleur diffuse un bruit blanc, le signal binaural enregistré sert de matière première aux algorithmes de localisation. La méthode de localisation conjointe semble assez robuste pour les azimuts proches du centre.

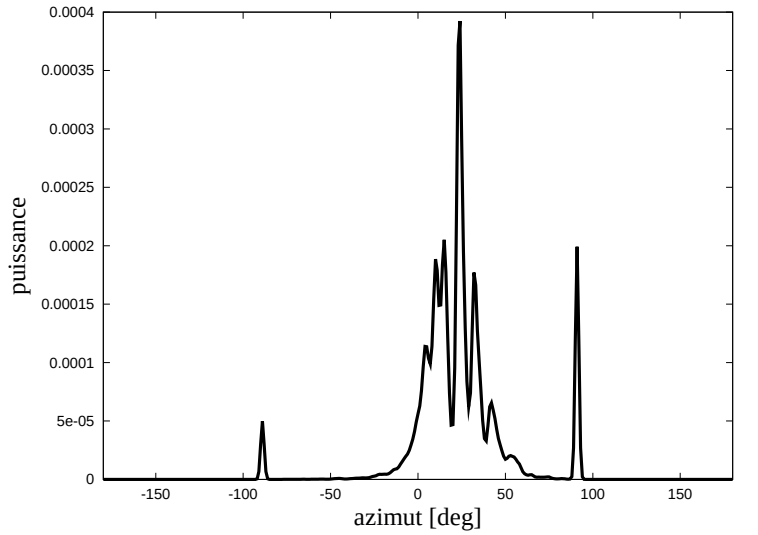
La figure 52 affiche un exemple représentatif de localisation d'un bruit blanc à 30°. La méthode conjointe et PHAT identifient la source respectivement à 31° et 33°. La précision semble similaire. Toutefois, nous constatons que la qualité de l'histogramme est fortement affectée par des pics secondaires significatifs ; en effet, les différences en amplitude sont corrompues par les superpositions provenant des réflexions multiples. En revanche, la fonction d'intercorrélation PHAT semble plus stable avec une plus faible variance. La difficulté de mesures ne nous a pas permis de faire plusieurs tests de localisation au studio Bonnefont. Nous avons constaté sur les quelques tests opéré que la précision de la localisation est similaire pour les deux approches. Dans la section suivante, la comparaison des deux méthodes est approfondie dans le cas d'une salle de classe.

Salle de classe

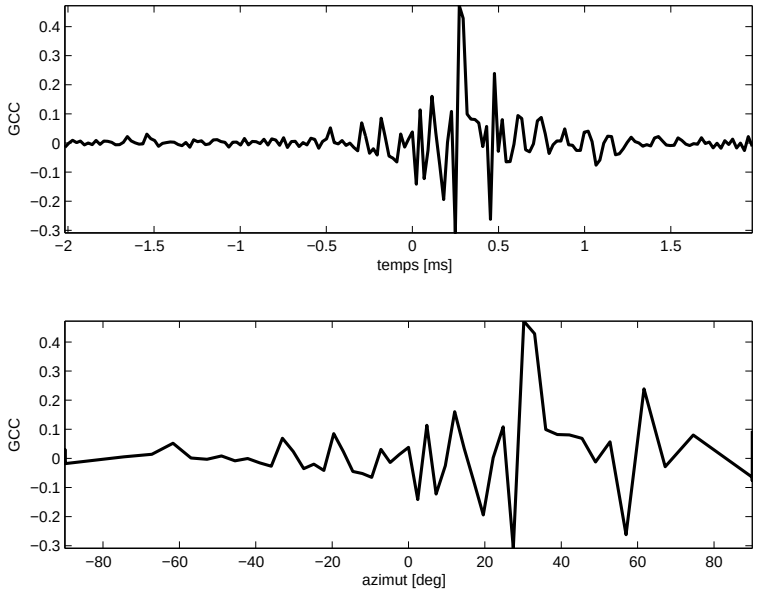
Les fonctions de transfert des milieux réverbérés sont nommées *Binaural Room Impulse Response* (BRIR). Nous utilisons les BRIR mesurées à partir du KEMAR par Shinn-Cunningham [BSCM05] dans une salle de classe de dimensions 5m × 9m × 3.5m pour simuler les signaux binauraux réels avec les équations 102 et 102 aux azimuts $0°, 15°, 30°, 45°, 60°, 90°$. Les BRIR ont toutes été mesurées dans le plan horizontal à partir d'une séquence de longueur maximale (*maximum-length sequence*), qui est une séquence pseudo-aléatoire binaire [RV89]. Ils ont une longueur de 32767 échantillons et peuvent être vus comme une combinaison de son direct, de premiers échos et de réverbération tardive. Le temps de réverbération de la salle T_{60} est compris entre 580 et 619 ms avec les algorithmes de Brown et Schroder [Bro02], [Sch65].

Des bruits blancs en milieu réverbéré ont été localisés avec la méthode conjointe. Les estimations montrent que la méthode conjointe est robuste en milieu réverbéré car on distingue un pic révélateur proche de la bonne localisation. La performance est décroissante avec les angles proches des extrêmes (figure 53).

La figure 54 montre que la méthode d'intercorrélation généralisée semble être meilleure pour la localisation dans un environnement réverbéré. En effet, nous constatons que la méthode conjointe devient progressivement instable au fur et à mesure que la source se déplace vers les



(a)



(b)

FIG. 52: *Localisation d'une source réelle à +30° jouant un bruit blanc dans une salle réverbérée (studio Bonnefont), à partir d'enregistrements binauraux mesurés au niveau des oreilles du musicien : histogramme obtenu par méthode conjointe avec bruit blanc (a) ; fonction d'intercorrélation obtenue par intercorrélation généralisée (PHAT) (b).*

côtés. À partir de 60° l'erreur atteint déjà 10°. Pour des conditions adverses de cette intensité de réverbération, la méthode s'en sort plutôt bien. On a l'impression dans le cas de cette salle que la méthode conjointe sous-estime la localisation.

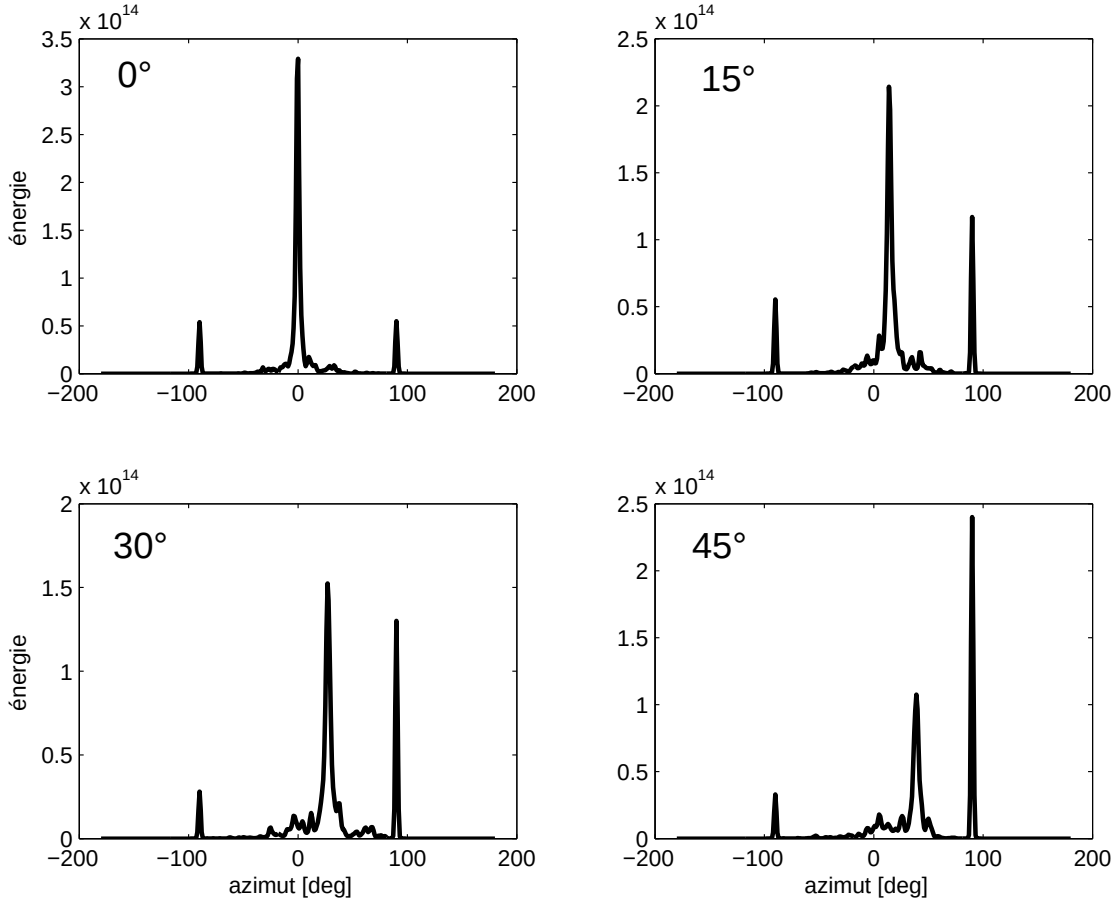


Fig. 53: *Histogrammes obtenus par méthode conjointe à différentes positions dans une salle de classe réverbérée, à* $(0, +15, +30, +45)°$.

4.3 Localisation en distance

Après l'inspection de l'estimation de l'azimut, la distance constitue la seconde dimension de localisation importante dans le cadre de nos recherches. En effet, dans les salles de concert, la distance est souvent simulée en positionnant les haut-parleurs près ou loin de l'audience, contrainte qui physiquement peut être limitative dans des salles de dimensions modestes. Assurément, l'architecture de la salle joue un rôle important, l'interprétation de la pièce électroacoustique peut ainsi être plus ou moins modifiée. En définitif, l'estimation de la distance et le positionnement virtuel des sources par la distance est nécessaire. De nombreux indices acoustiques ont été identifiés pour leur implication notoire dans l'estimation de la distance : l'énergie globale, la différence en amplitude (pour les distances de moins d'un mètre), la réverbération et le changement spectral. Toutefois, la mesure et la combinaison de ces indices demeurent complexes, de plus la détermination de la distance absolue nécessite des indices de haut niveau comme la familiarité aux sons perçus [NS04]. Nos recherches se focalisent sur la détermination d'une distance relative avec pour objectif une simulation plus réaliste de la distance, qui prend en charge les changements spectraux et l'énergie globale.

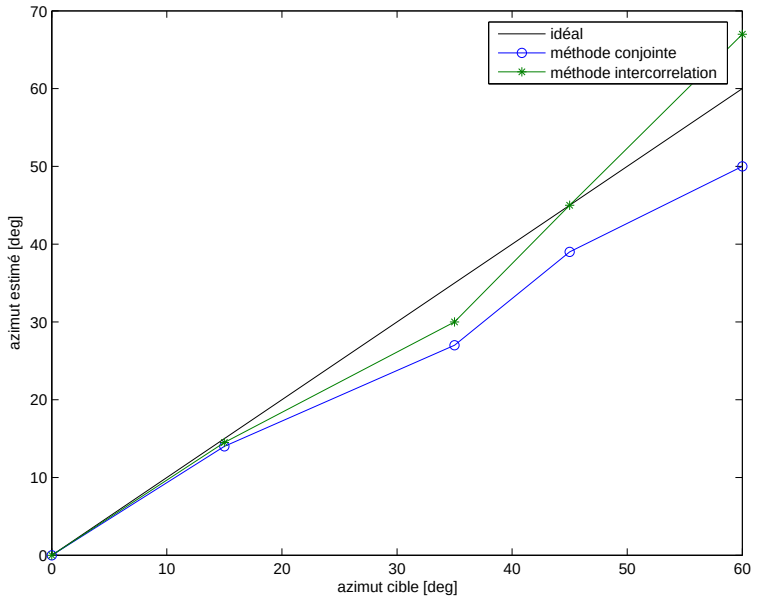


FIG. 54: *Localisation de bruits spatiaux enregistrés dans une salle de classe réelle avec différentes approches : idéal (plein), méthode de localisation conjointe (rond), méthode d'intercorrélation (astérisque).*

Nous proposons un estimateur de la distance basé sur la brillance. La brillance est une caractéristique forte d'un son, qui permet de différencier un son (ou un instrument) d'un autre. C'est un attribut quelque peu subjectif. Seulement, la brillance est approchée objectivement par la quantité de basses fréquences et de hautes fréquences contenues dans un son. On la quantifie par la mesure de la centroïde spectrale ou centre de gravité du spectre du son.

4.3.1 Positionnement relatif par la distance à partir du spectre

Tel que nous l'avons abordé au chapitre sur la localisation. La distance est fortement liée au centre de gravité spectral. Nous avons proposé une relation de quantification de la variation d'énergie avec la distance à partir de la norme ISO 9613-1, qui stipule que l'absorption de l'intensité fréquentielle (de la sonie) est à grand traits proportionnelle à la fréquence. L'équation 33, nous permet d'estimer la différence de la puissance du spectre d'une source X à chaque distance ρ, soit $D(f, \rho)$. Par conséquent, la relation nous permet de positionner virtuellement une source à la distance ρ. A cette distance, le spectre en échelle linéaire de la source est donné par :

$$\hat{X}(\rho, t, f) = X \cdot 10^{(-D(f,\rho))/20}, \tag{104}$$

où $D(f, \rho)$ est le facteur d'atténuation en décibels.

Cette méthode permet de simuler une distance relative. Nous procédons à une modification individuelle de chaque fréquence en approchant une loi naturelle d'absorption. Le réalisme en effet peut-être accentué en ajoutant des effets de Doppler à des sources mobiles. La distance absolue demeure un problème réel tant dans la localisation que dans la projection. En réalité la distance est une caractéristique difficile à estimer, après l'azimut et l'élévation, pour le système

auditif humain, car elle nécessite souvent la familiarité à l'environnement et à la source perçus.

4.3.2 Positionnement relatif par la distance à partir du spectre

Notre objectif est un positionnement relatif en distance d'une source, car l'estimation absolue de la distance dépend d'une connaissance *a priori* sur le lieu de diffusion et les paramètres de la source. Il s'agit donc de juger la distance de la source par rapport à une position précédente ; et de vérifier que le changement est proche de la réalité perceptive.

A partir de l'équation 33, nous pouvons chiffrer la variation de la puissance du spectre d'une source X à chaque distance ρ, soit $D(f, \rho)$. Nous déduisons que cette relation permet de positionner virtuellement une source à la distance ρ. Comme vu précédemment, à cette distance, le spectre en échelle linéaire de la source est donné par l'équation 104.

Deux caractéristiques du spectre sont contrôlées, à savoir la puissance globale et le contenu spectral. Cette approche présente des possibilités de création de sons virtuels en distance. La simulation par la loi en carré inverse (constante à toutes les fréquences) ne permet pas de prendre en compte la modification du contenu spectral de la source, des tests préliminaires sur des signaux de parole et d'instruments marquent un gain de réalisme à notre approche. Les sons dynamiques en distance sont préférés par rapport à ceux générés simplement par la loi en carré inverse (`http://dept-info.labri.fr/~sm/SMC08/`).

4.3.3 Signal de référence et estimation de la densité spectrale

La centroïde se déplace vers les basses fréquences lorsque la source s'éloigne de l'observateur. La brillance représente un indice acoustique capital. Comme signal de référence pour l'estimation de la distance, nous utilisons un bruit blanc Gaussien. Les sons purs ne sont pas adéquats pour les jugements de la distance, contrairement aux sons complexes [Mol73]. L'estimation de la distance repose sur la quantification de l'évolution spectrale du son au cours de la propagation dans l'air. Ainsi, nous estimons d'abord les densités spectrale du bruit émis et du bruit perçu. Une méthode efficace est la méthode de Welch [Wel67, Hay96]. Selon Welch, l'estimation du spectre d'amplitude est donnée par la moyenne des puissances spectrales de L blocs temporels, ensuite nous considérons la racine carrée, ainsi :

$$|X| = \sqrt{\frac{1}{L} \sum_{l=-(L-1)/2}^{l=+(L-1)/2} |X_l|^2}. \tag{105}$$

L'estimateur de Welch est un estimateur de la puissance spectrale asymptotiquement non biaisé. Dans nos expériences, nous considérons $L = 21$ blocs du signal de taille N = 2048 échantillons, avec un chevauchement de 50 % (et avec une qualité CD, taux d'échantillonnage de 44.1 kHz, ce qui correspondant à un segment du son d'une durée inférieure à 0.5 s).

Ensuite, nous utilisons cette amplitude pour calculer la centroïde spectrale avec :

$$\mathcal{C} = \frac{\sum_f f \cdot |X(f)|}{\sum_f |X(f)|}. \tag{106}$$

Pour un long signal, la centroïde du signal est estimée comme moyenne des centroïdes des périodogrammes.

4.3.4 Méthode de localisation par la distance

Dans la méthode de mesure de données de la base CIPIC, nous savons que les sources ont été placées sur un cercle de 1 m de rayon autour de l'auditeur. Pour chaque distance $0 < \rho < 100$ m, nous estimons le spectre atténué $X(f, \rho)$ à partir de l'équation 104 et nous déterminons la centroïde (voir équation 106). La figure 55 montre la centroïde en fonction de la distance pour un bruit blanc. Par approximation polynomiale du logarithme de la fonction de cette courbe obtenue grâce à la norme ISO 9613-1 et les équations 33, 34 et 106, nous proposons une fonction d'estimation de la distance à partir de la centroïde :

$$\begin{aligned} \rho(\log(\mathcal{C})) &= -38.89044\mathcal{C}^3 + 1070.33889\mathcal{C}^2 \\ &- 9898.69339\mathcal{C}^{+}30766.67908 \end{aligned} . \tag{107}$$

L'algorithme que nous avons mis en place prend en paramètres les conditions atmosphériques. La relation de l'équation 107 est calculée pour les conditions atmosphérique suivantes : température de 20° Celsius, humidité relative de 50%, pression de 1 atmosphère.

4.4 Résultats de la localisation en distance

Nous avons mené plusieurs simulations dans un cadre théorique afin d'évaluer l'erreur théorique de l'estimateur sur des bruits blancs. Les applications que nous visons s'appliquent dans des salles de taille moyenne. Jusqu'à 25 m, l'erreur maximale de la distance est théoriquement inférieure à 4 mm, si la densité spectrale du bruit est connue.

D'autres expérimentations basées sur l'estimation préalable du spectre d'amplitude à l'aide de l'équation 105 ont fourni une erreur plus grande d'environ 3 m, bien que très raisonnable jusqu'à 50 m. La figure 56 montre les résultats globaux de nos simulations sur des bruits blancs Gaussiens positionnés à différentes distances dans l'intervalle de $[0, 100]$ m. Le réalisme de la méthode de positionnement de source en distance peut être amélioré, notamment en incorporant des informations relatives à la vélocité du son, par exemple l'effet de Doppler en relation avec la distance. Pour simuler la distance, Chowning [Cho71] essaie de contrôler le rapport du son direct à la réverbération. La réverbération permet de prendre en compte la taille et la géométrie de la salle ainsi que le matériel constitutif. L'estimation et la maîtrise de la réverbération pendant une performance en direct demeure un défi.

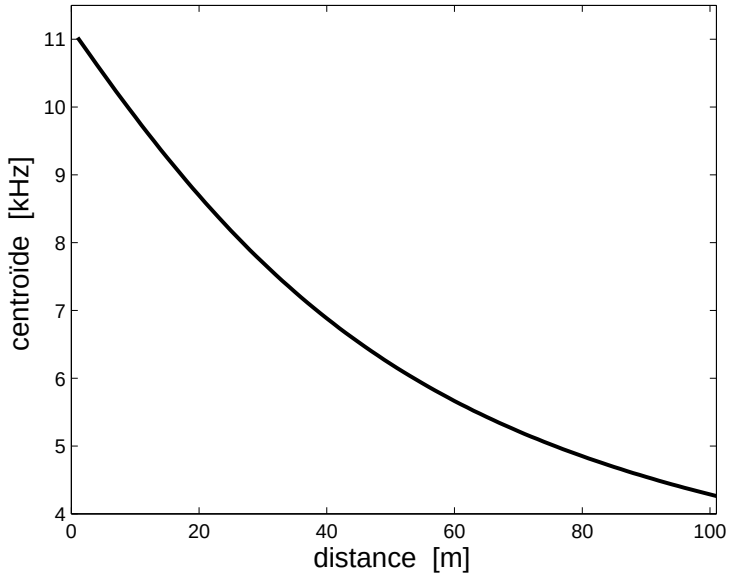


Fig. 55: *Centroïde spectrale (liée à la brillance perceptive) comme fonction de la distance à une température de 20° Celsius, une humidité relative de 50%, et une pression atmosphérique de 1 atm (pour un bruit blanc joué à la qualité CD).*

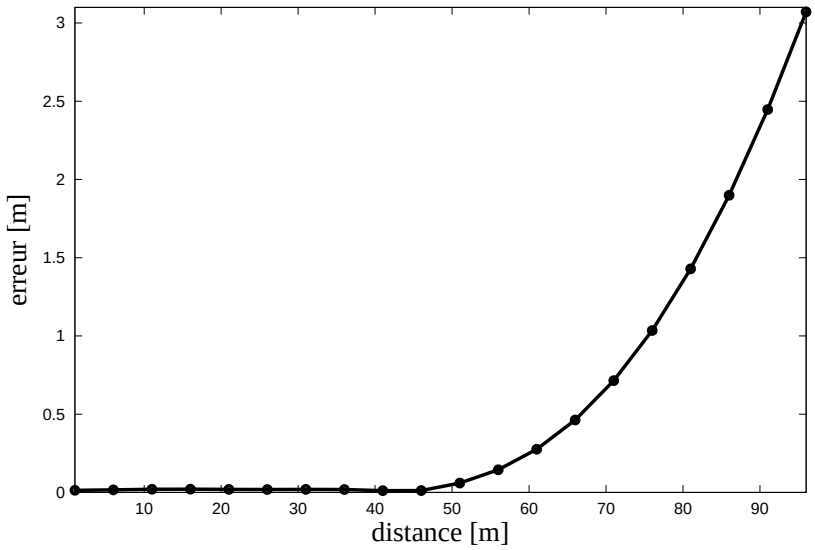


Fig. 56: *Erreur absolue de la localisation par la distance d'un bruit blanc Gaussien positionné à différentes distances.*

Chapitre 5

Séparation binaurale de sources

Dans les pièces acousmatiques, les compositeurs mélangent des sons naturels et électroniques. La création prend toute sa valeur lorsqu'elle est jouée en concert avec un orchestre de haut-parleurs. Pendant la diffusion, il est nécessaire d'interagir avec les sources afin de contrôler leur localisation. De ce fait, il est vivement désirable de manipuler individuellement les sources écoutées dans le mélange dans un but créatif ou ludique, afin d'éliminer, d'isoler, d'amplifier, de changer la localisation ou même le timbre d'une source précise.

C'est un grand défi de séparer des sources dans notre cas où nous ne disposons que de deux mélanges (au niveau des oreilles), et surtout, nous ne nous imposons aucune restriction sur le nombre de sources présentes dans la pièce. Dans la littérature, ce cas est dit *dégénéré* ou *sous-déterminé*. Cette disposition est de plus complexe car peu de techniques de séparation de sources développées jusqu'à lors traitent de ce cas. Communément fondées sur une inversion de matrice, les approches classiques sont fréquemment conduites à l'échec.

Tout d'abord, nous présentons le modèle de mélange convolutionnel de sources (section 5.1), car les sources perçues par les oreilles sont filtrées par le canal de propagation. Dans la section 5.2, nous exposons quelques travaux antérieurs déterminants sur des approches de séparation de sources basées sur les indices spatiaux. Premièrement, les méthodes d'analyse numérique de la scène sonore ou *computational auditory scene analysis* (CASA) qui imitent les processus du système auditif humain, deuxièmement un aperçu des méthodes statistiques de séparation aveugle de source avec l'analyse en composantes indépendantes (ACI). Troisièmement, le classique *Beamforming* qui se fonde sur un filtrage directionnel. Quatrièmement des techniques reposant sur des masques temps-fréquences. Ces techniques s'alignent sur un traitement non-linéaire des mélanges. Elles reposent principalement sur l'hypothèse d'orthogonalité dans le plan temps-fréquence. Ces méthodes concentrent notre plus grand intérêt, non seulement parce qu'elles sont récentes mais aussi parce qu'elles sont adaptées et prometteuses pour le cas dégénéré de sources audio. Un de ces précurseurs est la technique DUET (*Degenarate Unmixing Estimation Technique*) de Rickard *et al.* [RY02] qui utilise un masque binaire ; Avendano [Ave03] propose un masque Gaussien afin de prendre en compte les cas d'interférence, une approche Bayésienne est proposée par Master [Mas04] afin d'identifier les sources qui interfèrent.

Dans la section 5.3, nous proposons une modélisation du mélange comme un mélange de Gaussiennes. À partir d'une adaptation de l'algorithme de Maximisation de la Vraisemblance

(MV), nous effectuons un apprentissage des paramètres du mélange. Nous déduisons alors un masque temps-fréquence Gaussien orienté sur la probabilité *a posteriori*. Dans la section 5.4, les algorithmes de séparation sont évaluées sur des mélanges contenant de 2 à 5 sources, et la qualité des estimations est quantifiée par des mesures de distorsions et d'interférence.

5.1 Modèle de mélange de sources sonores

Dans de nombreuses scènes sonores, les sons se propagent des sources vers des capteurs qui peuvent être des oreilles humaines dans le cas d'une écoute binaurale, ou une série de microphones dans le cas d'enregistrements. De manière générale, pour M capteurs et K sources, le processus de mélange se modélise avec :

$$x_m(n) = \sum_{k=1}^{K} h_{mk}(n, \theta_k) * s_k(n), \tag{108}$$

où $(x_m(n))_{1 \leq m \leq M}$ est le mélange au capteur m, $(s_k(n))_{1 \leq k \leq K}$ est le signal source monophonique k et $h_{mk}(n, \theta_k)$ représente le filtre de propagation de la source k positionnée à θ_k au capteur m. l'opérateur $*$ est l'opération de convolution.

Sous la forme matricielle, le modèle de mélange temporel est donné par $\mathbf{x} = \mathbf{h} * \mathbf{s}$ avec :

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & \cdots & h_{1K}(n) \\ \vdots & \ddots & \vdots \\ h_{M1}(n) & \cdots & h_{MK}(n) \end{bmatrix} * \begin{bmatrix} s_1(n) \\ \vdots \\ s_K(n) \end{bmatrix}. \tag{109}$$

La transformée de Fourier de la convolution temporelle de deux fonctions correspond au produit de leurs transformées de Fourier. Dans le domaine spectral, le modèle de mélange de l'équation 108 s'écrit :

$$X_m(t, f) = \sum_{k=1}^{K} H_{mk}(t, f, \theta_k) \cdot S_k(t, f). \tag{110}$$

Sous la forme matricielle, le modèle de mélange spectral est donné par $\mathbf{X} = \mathbf{HS}$ avec :

$$\begin{bmatrix} X_1(t, f) \\ \vdots \\ X_M(t, f) \end{bmatrix} = \begin{bmatrix} H_{11}(t, f, \theta) & \cdots & H_{1K}(t, f, \theta) \\ \vdots & \ddots & \vdots \\ H_{M1}(t, f, \theta) & \cdots & H_{MK}(t, f, \theta) \end{bmatrix} \cdot \begin{bmatrix} S_1(t, f, \theta) \\ \vdots \\ S_K(t, f, \theta) \end{bmatrix}, \tag{111}$$

où $X_m(t, f)$, $S_k(t, f)$ et $H_{mk}(t, f, \theta)$ sont respectivement les spectres à court terme des sources $s_k(n)$, des signaux des capteurs $x_m(n)$ et des filtres de propagation $h_{mk}(n, \theta_k)$. Rappelons que les filtres sont considérés invariant, la notation $H_{mk}(t, f, \theta)$ est équivalente à $H_{mk}(f, \theta)$.

La séparation de sources consiste à retrouver les signaux sources s_k à partir des mélanges disponibles x_m, avec le moins de distorsions et d'interférences possible. Dans la pratique, des hypothèses sur les sources et les capteurs sont nécessaires afin de restreindre les solutions possibles et la complexité des algorithmes [CCPL95]. La difficulté de la séparation dépend

du type de source, du nombre de sources, de l'environnement de mélange et bien d'autres conditions. En effet, plus on a des sources dans le mélange moins les sources deviennent séparables ; aussi la difficulté est plus grande dans le cas sous-déterminé ($M < K$) que dans les cas sur-déterminé ($M > K$) et déterminé ($M = K$) [VJA+05] ; de plus la séparation est plus difficile en milieu réverbéré qu'en milieu anéchoïque.

La séparation de source consiste en général en quatre étapes fondamentales :

1. transformation dans le domaine de traitement (exemples : spectre, ondelettes) ;
2. estimation des paramètres de mélanges (exemples : position spatiale, harmonicité) ;
3. allocation de l'énergie à partir des signaux de mélanges en se basant sur les paramètres de mélange ;
4. transformation inverse pour obtenir des signaux temporels.

La qualité des sources séparées est évaluée perceptivement par des tests d'écoute et objectivement par des critères de qualité. Il s'agit en général d'évaluer la distorsion entre la source originale monophonique et la source estimée en calculant le rapport d'énergie de la source à celle des distorsions d'une part, et à celle des interférences d'autre part. Ces critères sont analysés en rapport avec des critères qui évaluent la difficulté de séparabilité des mélanges, à l'exemple du degré d'orthogonalité des sources [YR04].

Plusieurs paramètres de mélange peuvent être combinés pour une séparation efficace. Dans la suite, nous étudions quelques approches de filtrage de signaux. Les méthodes CASA (section 5.2.1), les méthodes d'ACI (section 5.2.2), les méthodes de filtrage directionnel (section 5.2.3) et les méthodes de masque fréquentiel (section 5.2.4).

5.2 Méthodes de détection et de séparation de sources

5.2.1 Séparation par analyse numérique de scène auditive

Analyse de Scène Auditive Computationnelle (CASA) est une approche de séparation de source qui se réfère aux principes d'analyse du son par le système auditif humain [SL90]. À partir d'une décomposition temps-fréquence du signal motivée par la sélection fréquentielle du système auditif, l'approche CASA s'intéresse à des paramètres sonores spécifiques tels que l'harmonicité [Pea76], la continuité, les temps d'onset et offset [HW04], la fréquence fondamentale [Lic51], le multi-pitch [Wei85], le timbre [KNKT95], l'enveloppe spectrale, les modulations en amplitude [Ber95] et en fréquence [Mel91], et bien d'autres ainsi que leurs combinaisons [BC94a].

A partir des paramètres considérés, des régions temps-fréquence sont groupées en objets par des règles de similarité, de proximité, de contour et de destin commun [Bre90][DC95]. Ces objets qui peuvent être des signaux purement harmoniques, des transitoires ou du bruit sont ensuite transformés en flux audio [BC94b]. Les méthodes CASA ont été utilisées dans la séparation d'instruments à partir de l'enveloppe spectrale et de la fréquence fondamentale ; toutefois elles nécessitent généralement un apprentissage sur de larges bases, ainsi qu'une réduction des cibles afin de réduire la complexité des algorithmes.

L'approche CASA a le mérite d'être utilisable sur un canal unique et permet la séparation de sonorités à partir d'un signal polyphonique. Mais l'efficacité de cette approche sur des

problématiques réelles comme la détection d'instruments reste limitée [WB06]. Néanmoins, les méthodes CASA qui ont tendance à exploiter des indices haut niveau peuvent être utilisées de concert avec d'autres méthodes comme l'analyse en composante indépendantes (ACI) qui exploitent directement d'orthogonalité des échantillons du signal.

5.2.2 Séparation aveugle et analyse en composantes indépendantes

L'analyse en composantes indépendantes (ACI) consiste à séparer les sources en assumant que ces dernières sont statistiquement indépendantes [HJ86]. De nombreux algorithmes d'ACI ont été proposés dans la littérature [HKO01], certains maximisent les statistiques d'ordre supérieure [Com94], d'autres minimisent l'information mutuelle des sources [BS95]. Ces méthodes se fondent sur des propriétés de sources comme la non-Gaussianité des sources [Car98], la non-stationarité des sources [MOK95] [CCA02] et d'autres font usage d'informations *a priori* sur les distributions des sources [Mas04]. Afin de cibler différents types de signaux, des méthodes hybrides sont possibles.

Le but de cette section n'est pas d'exposer chacun des algorithmes, mais plutôt de donner un aperçu des performances générales de l'ACI dans notre cas d'intérêt qui est le cas sous-déterminé. Il s'agit en général d'estimer la matrice pleine $\mathbf{W}_{(N\times M)}$, l'inverse de la matrice de mélange $\mathbf{H}_{(M\times N)}$ tel qu'un critère d'indépendance statistique soit maximisé. Les estimations des signaux sources sont alors données par :

$$\hat{\mathbf{S}} = \mathbf{WX}. \tag{112}$$

L'inversion de la matrice est appropriée pour les cas déterminé et sur-déterminé ($K \leq M$). Dans le cas sous-déterminé, une infinité de solutions inverses sont possibles. Il demeure possible de calculer une pseudo-inverse, qui représentera l'inverse de norme minimale, même si elle n'est pas forcément valide.

Dans tous les cas, l'inversion de la matrice n'assure pas l'allocation des composantes séparées aux mêmes sources (problème de permutation) dans les différentes trames, et les solutions sont acceptées à un facteur d'échelle près.

Une approche adaptée au cas sous-déterminé revient à considérer les sources comme des distributions statistiques, la séparation est alors fondée sur la maximisation de la probabilité *a posteriori*. Ces approches permettent une caractérisation simple des sources, un paramètre pour une distribution Laplacienne et deux paramètres (moyenne et variance) pour une Gaussienne. Le choix des distributions influence également la complexité analytique des estimateurs.

Dans la plupart de scènes musicales, nous avons plus de sources que de capteurs. Les musiciens ont tendance à démarrer au même moment. L'indépendance des sources s'en trouve réduite, et les paramètres harmoniques de sources différentes auront tendance à se superposer. La Toolbox ICALAB (figure 57) implémente plus d'une vingtaine de méthodes de séparation de sources et d'ACI. La figure 57 montre un exemple de séparation de deux sources mixées avec une matrice de colonnes $[1\ 2]^T$ et $[2\ 1]^T$ à l'aide de la méthode Fixed-Point ICA. Des tests montrent que ces méthodes sont mal adaptées au cas sous-déterminé, alors que dans le cas déterminé des performances appréciables peuvent être atteintes, notamment le problème de permutation de sources, où des composantes séparées peuvent être attribuées à la mauvaise source, la calcul de statistiques d'ordre élévé, l'inversion d'énormes matrices et leurs

diagonalisations sont contraignants pour des systèmes temps réels.

Des méthodes autant efficaces et moins complexes numériquement sont réalisables à partir de la localisation spatiale. Le filtrage spatial a été utilisé pour isoler des sources à partir de leur localisation. Dans la section suivante, nous étudierons une technique de filtrage directionnel appelée beamforming.

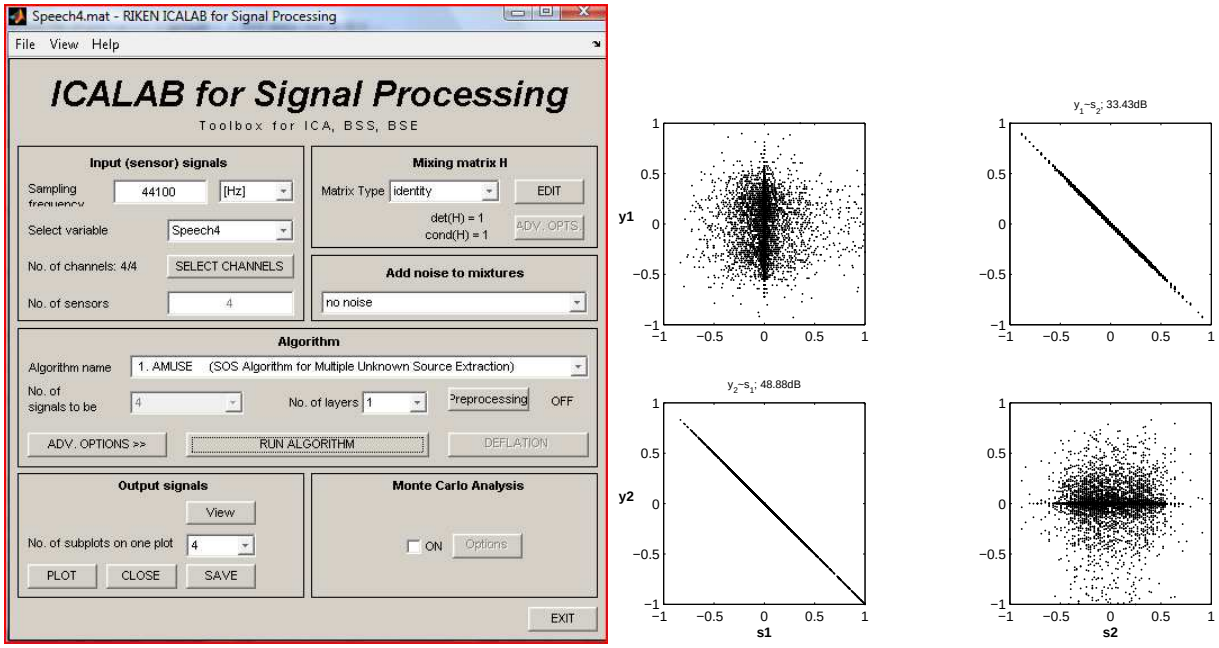


Fig. 57: *ICALAB Toolboxes implémente de nombreuses techniques de séparation basées sur l'ACI (gauche). Nuages de points obtenus par la méthode Fixed-Point ICA sources sonores estimées contre leurs originaux mixés avec une matrice de colonnes* $[1\ 2]^T$ *et* $[2\ 1]^T$ *(droite). Les points du mélange sont projetés sur les deux axes indépendants identifiés par Fixed-Point ICA.*

5.2.3 Séparation par filtrage directionnel : Beamforming

Le *beamforming* est une technique de filtrage spatial qui opère sur les signaux issus d'un arrangement de plusieurs microphones. Le but est de focaliser le faisceau de l'ensemble des capteurs vers une direction précise, et de rejeter les contributions issues des directions des sources concurrentes. Cette technique tient ses origines des domaines du radar et du sonar.

Généralement, les capteurs sont placés uniformément sur une ligne. Selon la direction cible, les sorties des capteurs sont décalées dans le temps de manière à obtenir des signaux en phase pour tous. Les versions décalées des sorties sont additionnées et forment ainsi le signal spatial filtré (voir figure 58). Selon la géométrie de la disposition des capteurs, par projection des capteurs sur une ligne perpendiculaire à la direction pointée ou *Maximum Response Angle* (MRA), on obtient une distance pour chaque capteur. Cette distance divisée par la célérité du son donne le délai requis pour le capteur afin de contribuer au faisceau orienté vers la direction cible [BM07].

Le beamformer peut être modélisé comme un filtre à réponse impulsionnelle finie (FIR).

L'observation des filtres spatiaux selon la fréquence du signal met en exergue les contraintes fréquentielles du beamforming. Selon le théorème de Shannon, des artefacts spatiaux apparaissent lorsque le signal contient une fréquence supérieure à la fréquence de Nyquist. En d'autres termes, la longueur d'onde de la fréquence doit être supérieure ou égale à deux fois la distance entre les capteurs. Des espacements non-uniformes entre les capteurs permettent d'étendre la bande fréquentielle [GE93].

Dans la pratique, le lobe central du filtre est parasité par des lobes secondaires ; à cet effet, des méthodes de beamforming qui combinent plusieurs beamformers ont été proposées. Le *Griffiths-Jim adaptive beamformer* [GJ82][YHHD92] combine un beamfomer fixe avec un filtrage adaptatif afin de réduire au minimum les lobes secondaires. La figure 59 montre des réponses spatiales pour le *delay-and-sum beamformer*, on remarque que le nombre croissant de capteurs permet une meilleure localisation.

Les beamfomers performants nécessitent en général plusieurs capteurs [Muc06] et ne considère que des délais purs ; de ce fait cette méthode ne nous intéresse que très peu car nous ne disposons que de deux capteurs, comme le système auditif humain, qui malgré tout atteint des performances impressionnantes. Au centre de nos intérêts, on distingue une autre famille de méthodes spatiales qui opèrent dans le cas sous-déterminé. À l'image de la technique DUET [RY02], ces dernières reposent principalement sur les indices spatiaux et des masque temps-fréquence afin d'exploiter la diversité fréquentielle et spatiale des sources musicales.

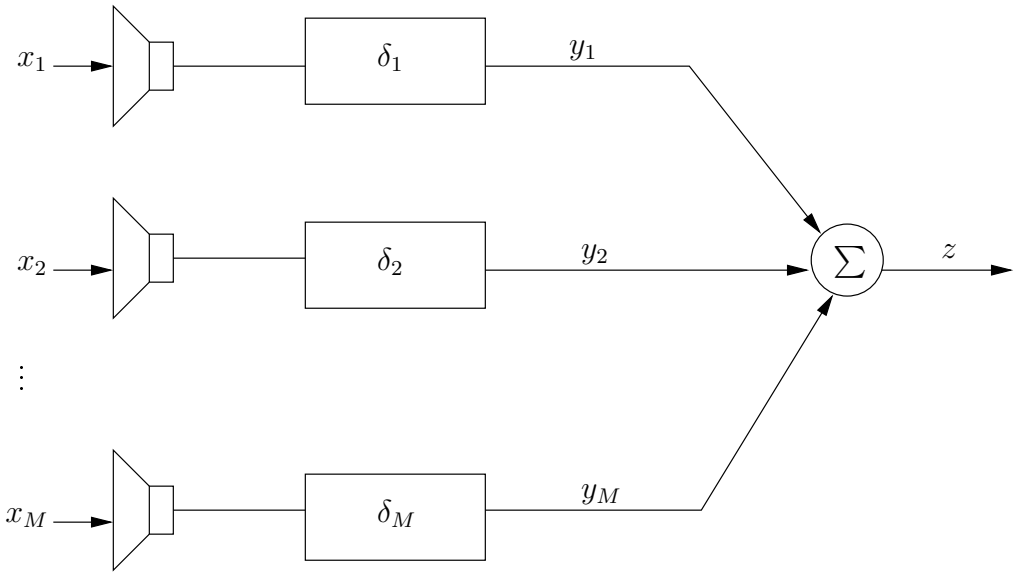


FIG. 58: *Diagramme du Beamformer delay-and-sum.*

5.2.4 Séparation basée sur un masque de séparation

Depuis quelques années une famille de méthodes de séparation de source basées sur des masques temps fréquence ont été proposées. En effet, des expérimentations menées par Rickard et Yilmaz sur des signaux de parole et de musique [RY02][SRR03] montrent que les représentations temps fréquence de ces derniers sont approximativement orthogonales. Précisément, on considère qu'une seule source domine à un indice temps fréquence donné. Une méthode de

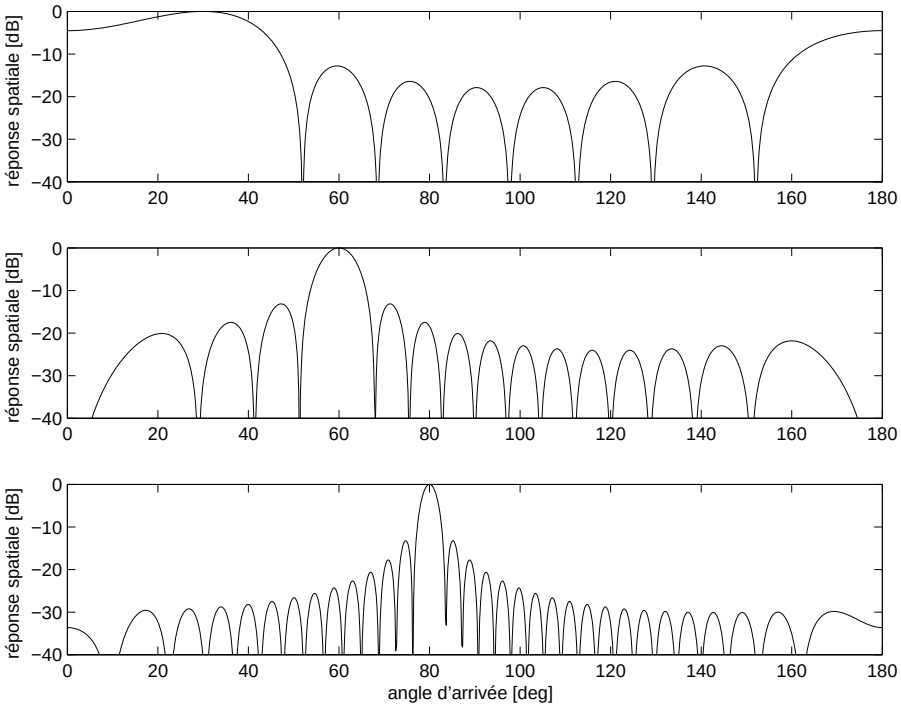


Fig. 59: *Réponses spatiales du delay-and-sum beamformer avec différents nombres de capteur (8, 16, 32) pour différents angles d'arrivée (30, 60, 80) degrés.*

séparation consisterait à attribuer le point temps fréquence à la source dont les paramètres sont le plus proche des paramètres du point temps-fréquence courant. De plus leurs représentations temps fréquence de signaux audio ont une forte densité, c'est-à-dire qu'un nombre de point réduit renferme la quasi-totalité de l'énergie de la source [YR04]. Cette approche est favorable à la séparation d'un grand nombre de sources sonores à partir de deux mélanges.

Récemment une méthode appelée DUET (Degenerate Unmixing Estimation Technique) [RY02] a été proposée, cette dernière utilise un histogramme à 2-dimensions (différences inter-canal) et un masque binaire pour séparer les sources (section 5.2.4). Carlos Avendano [Ave03] a proposé un filtrage Gaussien adaptatif afin de séparer les sources (section 5.2.4). Aaron Master [Mas04] propose un algorithme de séparation avec une approche Bayesienne en prenant en compte la superposition possible de deux sources au niveau d'une case temps-fréquence (section 5.2.4). Dans nos recherches, nous expérimentons un masque probabiliste basé sur la probabilité *a posteriori*, nous considérons que toutes les sources du mélange peuvent être actives sur le même point du plan.

Séparation par masque binaire de Rickard (DUET)

Approche DUET

Considérons les enregistrements d'une paire de microphones et les indices spatiaux (ITD, ILD). En négligeant les effets de réflexion, les mélanges sont des sommes de versions atténuées et retardées de signaux sources. Nous pouvons absorber l'atténuation et le délai relatif à

chaque source du premier microphone dans la définition des sources. Le modèle de l'équation 110 devient :

$$X_G(t,f) = \sum_{k=1}^{K} S_k(t,f), \tag{113}$$

$$X_D(t,f) = \sum_{k=1}^{K} a_k e^{-j\omega\delta_k} \cdot S_k(t,f), \tag{114}$$

où δ_k représente la différence de délai de la source k entre les deux microphones et a_k l'amplitude relative de la source k sur le microphone D, en rapport avec son amplitude sur le microphone G.

DUET suppose que les sources sont orthogonales deux à deux dans le plan temps-fréquence [YR04]. L'orthogonalité temps-fréquence (W-Disjoint Orthogonality–WDO) est exprimée par :

$$S_u(t,f) \cdot S_v(t,f) = 0 \qquad u,v = 1,\cdots,K \qquad u \neq v. \tag{115}$$

Tous les termes de la somme sont théoriquement nuls hormis celui de la seule source J active, par conséquent le mélange devient :

$$X_G(t,f) = S_J(t,f), \tag{116}$$

$$X_D(t,f) = a_J e^{-j\omega\delta_J} \cdot S_J(t,f). \tag{117}$$

$$\tag{118}$$

À chaque position (t,f) du plan, les mélanges sont des fonctions d'une seule source. De ce fait, la paire de paramètres (a_J,δ_J) de la source S_J active est approchée avec :

$$(a_J,\delta_J) = \left(\frac{|X_D|}{|X_G|}, \angle\left(\frac{X_G}{X_D}\right)/\omega\right). \tag{119}$$

On obtient ainsi un nuage de points disséminés dans le plan. Une technique de classification permettrait de constituer des classes dont le nombre représenterait le nombre de sources, et les points centraux représenteraient les paramètres spatiaux (amplitude relative et délai) associés à chaque source. La séparation consiste à attribuer chaque point (t,f) à la source la plus proche. Le critère de proximité dépend de la méthode spécifiée. L'ensemble des points (t,f) ainsi regroupés forment le spectre de la source.

Dans l'approche DUET [JRY00], un histogramme de puissance à 2 dimensions est construit (figure 60). Après lissage de l'histogramme, le nombre de pics dominants représente l'ordre du mélange. Sur la figure nous voyons deux sources à $(\ln(a_1) = 0.7, \delta_1 = -5.5)$ et $\ln(a_2) = -0.7, \delta_2 = 5.5)$. Nous observons aussi des pics parasites de faible amplitude qui sont causés par des interférences et par des estimations erronées aux hautes fréquences. La séparation de chaque source consiste à trouver le masque temps-fréquence optimal pour chaque source. Les sources se disputent un point temps-fréquence sans partage. Le masque binaire de DUET est donné par le maximum de vraisemblance à un point temps-fréquence, soit pour la source s_J :

$$M_J(t,f) = 1_{\{J=\text{argmax}_k L_k(t,f)\}}, \tag{120}$$

avec

$$L_J(t,f) = \frac{1}{2\pi\sigma^2} e^{-(\frac{1}{2\sigma^2})|a_J e^{\delta_J} f\omega_0 X_G(t,f) - X_D(t,f)|^2/(1+a_J^2)}. \quad (121)$$

Soit dans notre cas de notre modèle paramétrique :

$$L_J(t,f) = \frac{1}{2\pi\sigma^2} e^{-(\frac{1}{2\sigma^2})|\alpha(f)\sin(\theta_i) e^{\beta(f)\sin(\theta_J)} f\omega_0 X_G(t,f) - X_D(t,f)|^2/(1+a_J^2)}. \quad (122)$$

L'énergie du point (t,f) du mélange est assignée au point (t,f) de la source ayant la valeur maximale L_J. Les autres sources sont privées d'énergie à ce point de leurs spectres. En pratique une énergie minimale différente de zéro est attribuée afin d'éviter le bruit musical. La source temporelle est estimée par transformée de Fourier inverse et par une technique *overlap-add* à partir des spectres à court terme obtenus. Pour le canal droit par exemple on a : $\hat{S}_J = M_J(t,f)X_D$. Les auteurs ont également proposé une estimation par maximum de vraisemblance du canal monophonique [YR04], dans le cas où le masque est égal à 1.

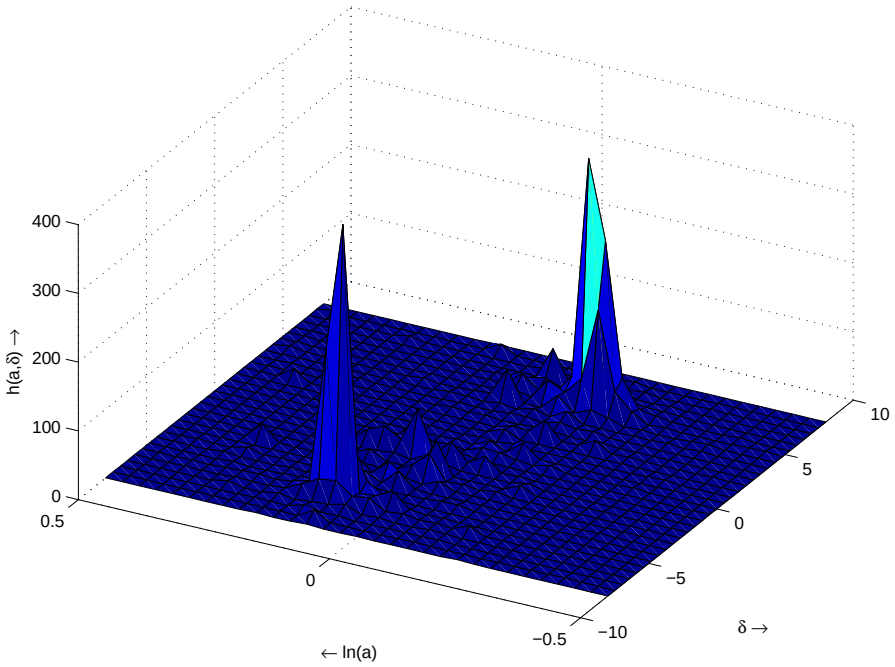


Fig. 60: *Histogramme à deux dimensions $h(a,\delta)$ de deux sources en fonction du délai et de l'amplitude relative.*

Limites de DUET

Bien que l'approche DUET obtient des performances appréciables, elle connaît quelques limites. D'abord l'estimation du délai devient ambigue au-delà de 1500Hz. Pour une source large bande, les estimations des délais peuvent être erronées. Ensuite, une violation de la condition d'orthogonalité des sources entraîne une dégradation croissante des performances [BZ02]. En effet, lorsqu'au moins deux sources contribuent à l'énergie d'une case fréquentielle, les paramètres estimés correspondent à une source inexistante. Dans de tels cas, il n'est pas

évident de retrouver les sources originales qui ont contribué à l'énergie du point fréquentiel. De ce fait, attribuer le point exclusivement à une source engendrerait un taux d'interférence élevé. Avendano [Ave03] propose un masque basé sur une fonction Gaussienne spatiale, dont la largeur de la fenêtre permet de faire un compromis entre les distorsions et les interférences possibles.

Séparation par masque Gaussien de Avendano

Partant d'un modèle de mélange simple, cette méthode fait usage d'une mesure de similarité entre les spectres à court terme des signaux binauraux. Elle permet d'identifier les régions occupées par chaque source dans le mélange, en se basant sur les "coefficients de spatialisation" assignés à chaque source à la synthèse du mélange. Une classification de coefficients temps-fréquence possédant le même "coefficient de spatialisation" permet ensuite l'identification de chaque source. Le modèle utilisé est :

$$X_G(t) = \sum_{k=1}^{K} \alpha_{Gk} S_k(t,f), \tag{123}$$

$$X_D(t) = \sum_{k=1}^{K} \alpha_{Dk} S_k(t,f), \tag{124}$$

où les $(\alpha_{Gk}, \alpha_{Dk})$ représentent les coefficients de spatialisation, dont le rapport d'amplitude permet de positionner les sources en assumant la loi sinusoïdale de conservation de l'énergie (*sinusoidal energy preserving panning law*) avec $\alpha_{Dk} = \sqrt{1-\alpha_{Gk}^2}$.

Afin d'identifier la proximité des composantes temps-fréquence, Avendano se sert d'une mesure de similarité qu'il dénomme *panning index*. Elle est définie par :

$$\psi(t,f) = 2\frac{|X_G(t,f)X_D^*(t,f)|}{|X_G(t,f)|^2 + |X_D(t,f)|^2}, \tag{125}$$

où * dénote le conjugué complexe. Dans le cas de sources orthogonales dans le plan temps-fréquence et en négligeant tout effet d'échos, l'équation (125) peut s'écrire :

$$\psi(t,f) = 2\alpha\sqrt{1-\alpha^2}.$$

Cette fonction est bornée entre 0 et 1 contrairement à d'autres métriques [FB02]. La dépendance quadratique de α entraîne une ambiguïté. L'ambiguïté quadratique est résolue en définissant des mesures de similarité partielles et leurs différences :

$$\psi_G(t,f) = 2\frac{|X_G(t,f)X_D^*(t,f)|}{|X_G(t,f)|^2}. \tag{126}$$

Leur différence est donnée par :

$$\triangle(t,f) = \psi_G(t,f) - \psi_D(t,f), \tag{127}$$

où les régions avec des valeurs positives de $\triangle(t,f)$ correspondent à des signaux spatialisés vers la gauche, et les valeurs négatives à des signaux spatialisés vers la droite et des valeurs nulles au centre. La métrique de similarité devient :

$$\Psi(t,f) = [1 - \psi(t,f)]\text{sign}(\triangle(t,f)), \tag{128}$$

La fonction de similarité est alors bornée entre -1 et 1. Les sources spatialisées à différentes directions sont identifiées à partir de la valeur de la similarité de chaque composante temps-fréquence. En pratique, la condition d'orthogonalité des spectres n'est pas idéale, ainsi on observe différents nuages de points regroupés, dont le nombre représenterait le nombre de sources, et les paramètres de similarité centraux ceux de la source identifiée. L'ensemble des points d'une classe forme le spectre de la source correspondante. Dans la méthode de séparation, Avendano utilise une fenêtre Gaussienne adaptative, de sorte que les valeurs correspondant à celles de la source soient moins filtrées que celles qui en sont éloignées. Ainsi, des interférences peuvent être limitées et contrôlées par la largeur de la fenêtre. La fonction de filtrage spatiale proposée dans [Ave03] est :

$$\Theta(t,f) = \nu + (1-\nu)e^{-\frac{1}{2\xi}(\Psi(t,f)-\Psi_d)^2}, \tag{129}$$

où Ψ_d est la valeur de similarité cible, ξ contrôle la largeur de la fenêtre et ν est une valeur minimale, afin d'éviter les valeurs fréquentielles nulles qui suscitent des artefacts, particulièrement le bruit musical. La reconstruction des signaux temporels à partir des spectres modifiés se fait par transformée de Fourier inverse. Des méthodes d'estimation de signal particulières permettent d'obtenir un gain supplémentaire dans le reconstruction du signal à partir de spectres modifiés [LDA07]. Contrairement à l'approche de DUET, Avendano considère indirectement les possibilités d'interférence et surtout leur contrôle. Toutefois, il ne s'agit pas de l'azimut mais plutôt d'un coefficient inter-canal d'amplitude, ce qui rend cette méthode très sensible au bruit et à la réverbération (voir section 5.4).

Dans notre cas, la relation entre les sources à partir du modèle paramétrique (voir chapitre 2) peut s'écrire :

$$X_D(t,f) = X_G(t,f) \cdot 10^{\Delta_a} e^{j\Delta_\phi}. \tag{130}$$

Le modèle d'Avendano ne dépend pas de la phase, le facteur $e^{j\Delta_\phi}$ ne joue alors aucun rôle. De plus son module est toujours égal à 1. Le paramètre *panning index* devient :

$$\psi(t,f) = 2\frac{10^{\Delta_a}}{1+10^{+2\Delta_a}}, \tag{131}$$

avec

$$\begin{aligned} \Delta_a &= \text{ILD}(\theta,f)/20 && (132)\\ &= \alpha(f)\sin(\theta)/20, && (133) \end{aligned}$$

$$\begin{aligned} \psi_G(t,f) &= 2\cdot 10^{\Delta_a}, && (134)\\ \psi_D(t,f) &= \frac{2}{10^{\Delta_a}}. && (135) \end{aligned}$$

La métrique de similarité devient :

$$\Psi(t,f) = \left[1 - 2\frac{10^{\Delta_a}}{1+10^{+2\Delta_a}}\right] \text{sign}\left(10^{\Delta_a} - \frac{1}{10^{\Delta_a}}\right). \tag{136}$$

Soit l'expression complexe suivante :

$$\Psi_i(t,f) = \left[1 - 2\frac{10^{\alpha(f)\sin(\theta_i)/20}}{1+10^{+\alpha(f)\sin(\theta_i)/10}}\right] \text{sign}\left(10^{\alpha(f)\sin(\theta_i)/20} - \frac{1}{10^{\alpha(f)\sin(\theta_i)/20}}\right). \tag{137}$$

Contrairement aux paramètres d'Avendano qui sont indépendants de la fréquence, les nôtres dépendent de la fréquence à l'instar de la fonction d'échelle $\alpha(f)$. Ainsi le centre de la fonction Gaussienne pour le même azimut θ_i est fonction de la fréquence.

Dans cette approche, la largeur de la fenêtre doit être donnée à l'entrée, elle est fixée pour chaque source. Ce qui suppose un apprentissage préalable du mélange. Un apprentissage automatique de la largeur de la fenêtre est souhaitable pour chaque source. Master propose une méthode de séparation pour DUET basée sur la détection de collision de deux sources, et l'identification des deux sources concurrentes afin de mieux répartir l'énergie du point temps-fréquence.

Séparation par masque spatial de Master (DASSS)

L'approche *Delay And Scale Subtraction Scoring (DASSS)* proposée par Master [Mas04] a pour dessein d'améliorer les performances de séparation de DUET. La détection des paramètres des sources avec DASSS est la même que DUET avec un histogramme à deux dimensions. Mais pour l'étape de filtrage, DASSS envisage que deux sources au plus peuvent participer à l'énergie d'une composante fréquentielle, alors que DUET attribue le point exclusivement à la source voisine la plus proche. DASSS construit des signaux Y_k (équation 138) en éliminant exactement la source S_k du mélange ; ces signaux sont ensuite comparés à leurs prédictions en cas de présence d'une unique source. Une fonction de performance ou *scoring function* (équation 146) permet de décider de la présence d'une ou deux sources au niveau de point (t,f).

Les signaux Y_k sont donnés par :

$$Y_k = X_G - \frac{1}{a_k} e^{+j\omega\delta_k} X_D. \tag{138}$$

Pour la source active S_J, ce modèle prédit :

$$\hat{Y}_{k=J} = 0 \tag{139}$$

$$\hat{Y}_{k\neq J} = a_{l,k} S_l \tag{140}$$

$$= a_{l,k} X_G, \tag{141}$$

où

$$a_{u,v} = (1 - \frac{a_v}{a_u} e^{+j\omega(\delta_u - \delta_v)}). \tag{142}$$

Et pour deux sources actives S_u et S_v on a :

$$\hat{Y}_{k=u} = a_{uv}S_v \quad (143)$$
$$\hat{Y}_{k=v} = a_{vu}S_u \quad (144)$$
$$\hat{Y}_{k\neq(u|v)} = a_{ku}S_v + a_{kv}S_v. \quad (145)$$

À partir d'une fonction de coût, on peut alors déterminer la source active ou les deux sources actives. Au-delà d'un coût seuil, on considère la présence d'une deuxième source. Dans [Mas03] la fonction de coût suivante a été utilisée :

$$f(J) = \frac{\sum_{k=1}^{K} |\hat{Y}_k^J - Y_k|}{\sum_{k=1}^{K} |Y_k|}, \quad (146)$$

où $\hat{Y}_k^J$ est la prédiction Y_k lorsqu'on assume que seule la source S_J est active. Dans [Mas04], une approche Bayésienne a été proposée. Les distributions des sources à chaque point temps-fréquence $|S_k|$ et $|Y_k|$ sont modélisées comme des Gaussiennes centrées [Mas06] selon :

$$|\hat{Y}_{k=u}| \sim N(0, \sigma_v^2 \cdot |a_{uv}|^2) \quad (147)$$
$$|\hat{Y}_{k=v}| \sim N(0, \sigma_u^2 \cdot |a_{vu}|^2) \quad (148)$$
$$|\hat{Y}_{k\neq(u|v)}| \sim N(0, \sigma_u^2 \cdot |a_{ku}|^2 + \sigma_v^2 \cdot |a_{kv}|^2) \quad (149)$$
$$(150)$$

où σ_k^2 représente la variance de la source k à une fréquence donnée. Le but est alors de déterminer les deux sources (u, v) qui auraient contribué à l'énergie du point. Cette probabilité d'avoir la donnée D sous forme de Y_k est donnée par :

$$p(D|u, v) = \prod_{k=1}^{K} p(|Y_k| \ |u, v). \quad (151)$$

En présence de deux sources, les contributions de chaque sources peuvent être déterminées par inversion linéaire d'après le modèle linéaire de l'équation 114 :

$$\begin{bmatrix} S_u \\ S_v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ a_u e^{-j\omega\delta_u} & a_v e^{-j\omega\delta_v} \end{bmatrix} \cdot \begin{bmatrix} X_G \\ X_D \end{bmatrix}. \quad (152)$$

Après avoir assigné tous les points temps-fréquence aux sources, une transformée de Fourier inverse permet d'obtenir les estimations des sources.

La méthode DASSS présente des avantages par rapport à DUET car elle signale la présence de plus d'une source au niveau d'un point, et permet une répartition plus intelligente de l'énergie entre plusieurs sources. Dans la pratique, on observe des superpositions, notamment avec les sources harmoniques (instruments) qui ont tendance à jouer ensemble. DASSS se base sur le même histogramme que DUET, de même il est limité par l'ambiguïté de la phase au-delà de 1500 Hz. De plus, un apprentissage des sources est typique des approches Bayésiennes [Mas06].

Nous proposons une approche qui considère que toute case fréquentielle est le résultat de la contribution de toutes les sources du mélange. La distribution de l'énergie se fait par rapport à la probabilité de la présence de chaque source. Cette dernière est estimée à partir des données du point temps fréquence. Aussi, nous utilisons un seul paramètre qui est l'azimut estimé conjointement à partir de l'ILD et de l'ITD. Ce dernier est plus robuste au niveau des hautes fréquences (chapitre 4). Notre approche de séparation considère les distributions de sources dans l'histogramme comme des Gaussiennes, par une méthode d'apprentissage automatique à partir des données du mélange, nous déterminons les paramètres des Gaussiennes.

5.3 Séparation par masque probabiliste

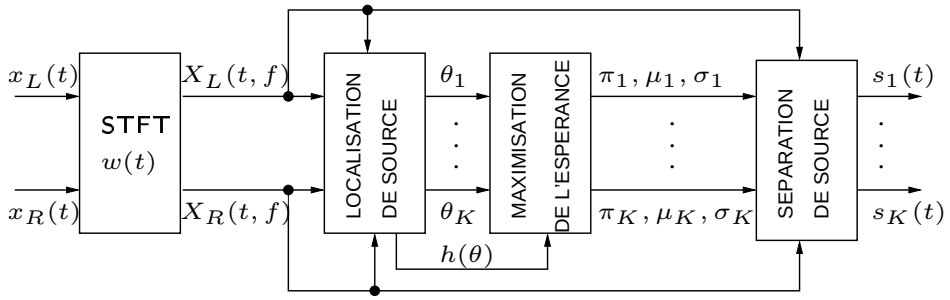


FIG. 61: *Vue d'ensemble du système analyse-localisation-séparation de sources basé sur une classification non supervisée des indices interauraux (ILD, ITD) par un algorithme EM modifié.*

5.3.1 Mélange de Gaussiennes

Étant donné que les sources ne sont pas exactement orthogonales, pour chaque source on obtient une distribution autour de la valeur réelle. Comme mentionné précédemment, nous avons choisi d'approcher l'accumulation de l'énergie autour de chaque point avec une distribution Gaussienne. Dans le cas de K sources, nous introduisons un modèle de K Gaussiennes spatiales (K-GMM, mélange de Gaussiennes d'ordre K) :

$$P_K(\theta|\Gamma) = \sum_{k=1}^{K} \pi_k \ \phi_k(\theta|\mu_k, \sigma_k) \text{ avec } \pi_k \geq 0 \text{ et } \sum_{k=1}^{K} \pi_k = 1 \tag{153}$$

où Γ est un ensemble multiple de K triplets $(\pi_k, \mu_k, \sigma_k^2)$ qui représentent tous les paramètres du modèle ; π_k, μ_k, et σ_k^2 indiquent respectivement le poids, la moyenne et la variance de la k-ième composante Gaussienne décrite mathématiquement par :

$$\phi_k(\theta|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\theta-\mu_k)^2}{2\sigma_k^2}\right). \tag{154}$$

Nous sommes intéressés par l'estimation de l'architecture du K-GMM, qui est le nombre de sources K et l'ensemble des paramètres Γ, qui permettront de configurer automatiquement les filtres de séparation.

Première estimation

Dans l'histogramme, on observe des maxima locaux dominants dont le nombre fournit une estimation de l'ordre du nombre de sources dans le mélange. L'abscisse du k-ième maximum local révèle l'emplacement θ_k de la k-ième source. En général, un réglage plus fin de l'histogramme permet de distinguer visuellement des sources très proches spatialement. Toutefois, nous détecterons aussi de maxima locaux parasites. Alors, nous proposons de lisser l'histogramme à l'aide d'un opérateur binomial $\mathcal{B}$, afin de fusionner les disparités négligeables au niveau des sommets rencontrés. L'opérateur binomial permet une bonne conservation de pics saillants de l'histogramme tout en assurant un lissage correct (voir figure 62).

$$\mathcal{B}(n) = \frac{1}{2^{D-1}\binom{D-1}{n}} \qquad n = 0, \cdots, D-1, \tag{155}$$

où D est la dimension de l'opérateur. Des tests préliminaires recommandent d'utiliser un ordre impair supérieur à 3. L'histogramme lissé $\tilde{h}$ est construit par convolution de l'histogramme courant avec le noyau de l'opérateur $\mathcal{B}$ selon :

$$h_s(\theta) = \sum_{n=0}^{D-1} \mathcal{B}(n) h\left(\theta - \frac{D-1}{2} + n\right). \tag{156}$$

Les interférences entre les sources de repères causent des indices acoustiques corrompus ; ainsi l'histogramme produit indiquerait une position qui ne correspond pas à celle d'une source existante. Lorsque les sources ne coïncident pas, généralement les hauteurs des pics indésirables sont très faibles. Ainsi, nous appliquons un seuil de détection, de telle façon que seuls les sommets désirés subsistent. Le plancher du seuil est fixé relativement à un niveau de bruit de l'histogramme. Par des tests empiriques, nous avons fixé le seuil de détection de pic au tiers du maximum de l'histogramme lissé, soit $\frac{1}{3}\max(h_s(\theta))$. Ce niveau de seuil n'est pas robuste à tout type de mélange, il dépend du taux des spectres des sources et du degré de réverbération de l'environnement (amplitude des pics parasites est élevée en présence de réverbération). La figure 62 montre que le nombre de sources détectées est amélioré par l'application d'un seuil de détection, on passe de 10 pics parasites à 1.

Des expérimentations préliminaires montrent que le nombre de sources estimé et les emplacements sont corrects, comme étudié dans le chapitre 4. Nous obtenons le modèle d'ordre K et une première estimation des moyennes des Gaussiennes (μ_k en Γ), qui servent à l'initialisation de l'algorithme EM. Cette estimation peut être affinée et complétée - avec σ_k^2 et le poids π_k.

5.3.2 Maximisation de l'Espérance du mélange de sources

Chaque source dans le mélange se caractérise par une représentation Gaussienne dans l'histogramme. Pour la discrimination et la séparation des sources, le poids, la moyenne, et la variance de chaque source sont essentiels pour notre algorithme de filtrage spatial Gaussien.

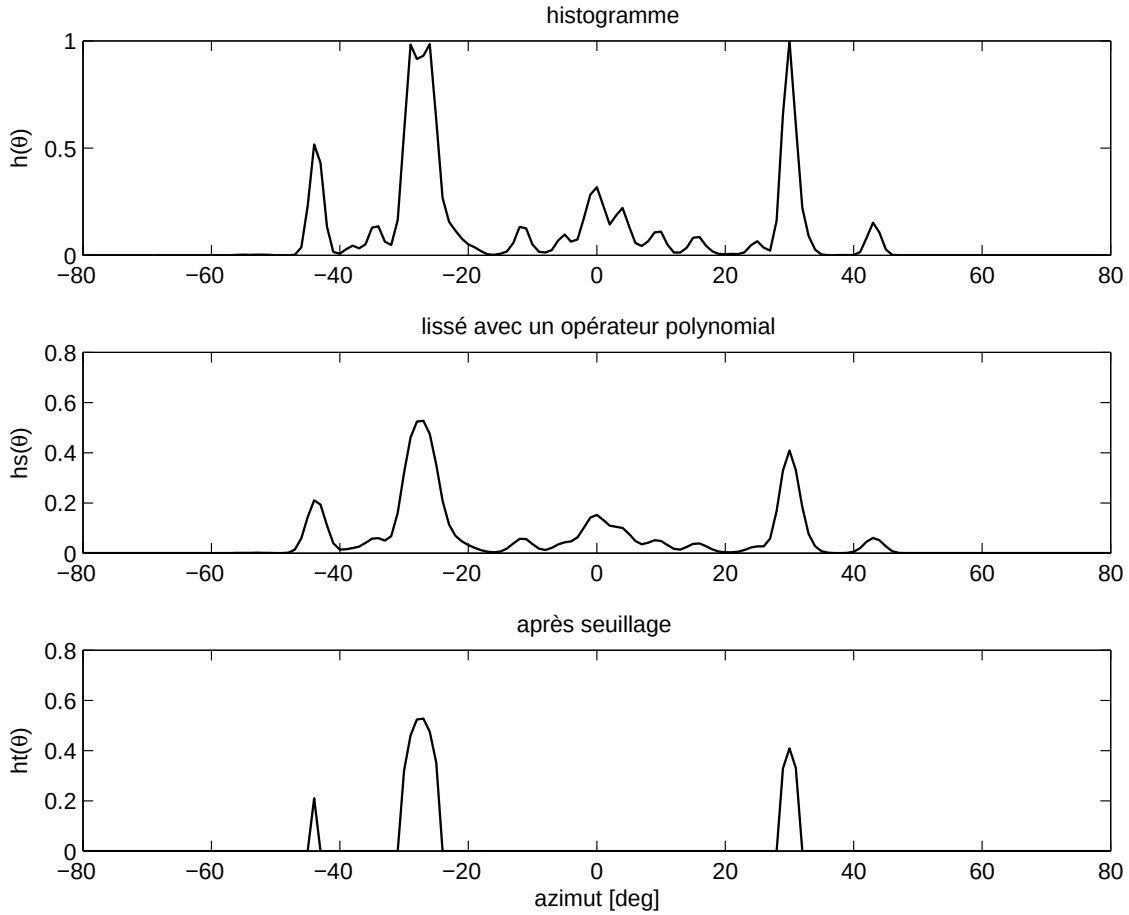


FIG. 62: *Mélange de deux sources à* $(-30°, 30°)$. *Successivement histogramme originel* $(h(\theta))$, *histogramme lissé* $h_s(\theta)$ *et histogramme après application d'un seuil de détection égale au tiers du maximum de l'histogramme lissé* $(h_t(\theta))$. *Le nombre de pics décroît de* $K = 13$ *à* $K = 3$, *seul un pic parasite subsiste.*

La méthode de Maximum de l'Espérance (EM) est une approche classique pour estimer les paramètres de chaque densité dans un mélange de densités à partir d'un ensemble de données x. L'idée est de compléter les données observées x avec une variable non observée y pour former des données complètes (x, y), où y indique l'indice de la composante Gaussienne de laquelle x a été tiré. Ici, le rôle de x est joué par l'azimut θ, en prenant des valeurs dans l'ensemble de tous les azimuts discrets couverts par l'histogramme. Nous associons θ à sa fonction d'intensité $h_s(\theta)$ (histogramme lissé). Le rôle de y est joué par $k \in \{1, \cdots, K\}$, qui est l'ensemble des indices des composantes Gaussiennes.

L'algorithme EM s'exécute de manière itérative, à chaque itération on calcule les paramètres optimaux qui entraînent une augmentation locale du logarithme de la vraisemblance du mélange $(P_K(\theta|\Gamma))$. En d'autres termes, on augmente la différence du log de vraisemblance actuelle avec des paramètres Γ et la vraisemblance suivante, avec des paramètres Γ'. On peut

l'exprimer sous forme de fonction logarithmique notée $Q(\Gamma', \gamma)$ comme :

$$\begin{aligned} Q(\Gamma', \Gamma) &= \sum_{\theta} h_s(\theta) \left(\mathcal{L}(\theta|\Gamma') - \mathcal{L}(\theta|\Gamma) \right) \quad \text{avec} \\ \mathcal{L}(\theta|\Gamma) &= \log \left(P_K(\theta|\Gamma) \right). \end{aligned} \tag{157}$$

Nous pouvons reformuler $\mathcal{L}(\theta|\Gamma)$ avec :

$$\begin{aligned} \mathcal{L}(\theta|\Gamma) &= \log \left(\sum_{k} P_K(\theta, k|\Gamma) \right) \quad \text{avec} \\ P_K(\theta, k|\Gamma) &= \pi_k \ \phi_k(\theta|\mu_k, \sigma_k). \end{aligned} \tag{158}$$

La concavité de la fonction log permet de minorer la fonction $Q(\Gamma', \Gamma)$ avec l'inégalité de Jensen. En introduisant l'énergie de l'histogramme, nous pouvons alors écrire :

$$Q(\Gamma', \Gamma) \geq \sum_{\theta} \sum_{k} h_s(\theta) P_K(k|\theta, \Gamma) \log \left(\frac{P_K(\theta, k|\Gamma')}{P_K(\theta, k|\Gamma)} \right), \tag{159}$$

où $P_K(k|\theta, \Gamma)$ est la probabilité *a posteriori*, qui représente le degré de confiance avec lequel nous croyons que les données ont été générées par la k-ième Gaussienne à partir des données. Nous la calculons à partir de la règle de Bayes :

$$P_K(k|\theta, \Gamma) = \frac{P_K(\theta, k|\Gamma)}{P_K(\theta|\Gamma)}. \tag{160}$$

Les nouveaux paramètres sont ensuite estimés en maximisant la limite inférieure par rapport à Γ :

$$\Gamma' = \text{argmax}_{\gamma} \sum_{\theta} \sum_{k} h_s(\theta) P_K(k|\theta, \Gamma) \log \left(P_K(\theta, k|\gamma) \right). \tag{161}$$

Augmenter cette limite inférieure entraîne automatiquement une augmentation du logarithme de la vraisemblance ; cette méthode de résolution est mathématiquement plus facile. Enfin, la maximisation de l'équation 161 fournit les relations de mise à jour suivante, à appliquer dans l'ordre, car ils modifient la valeur actuelle avec des effets de bord, donc la mise à jour de valeur doit être prise en compte dans les relations suivantes :

$$\pi_k \leftarrow \frac{\sum_{\theta} h_s(\theta) \ P_K(k|\theta, \Gamma)}{\sum_{\theta} h_s(\theta)}, \tag{162}$$

$$\mu_k \leftarrow \frac{\sum_{\theta} h_s(\theta) \ \theta \ P_K(k|\theta, \Gamma)}{\sum_{\theta} h_s(\theta) \ P_K(k|\theta, \Gamma)}, \tag{163}$$

$$\sigma_k^2 \leftarrow \frac{\sum_{\theta} h_s(\theta) \ (\theta - \mu_k)^2 \ P_K(k|\theta, \Gamma)}{\sum_{\theta} h_s(\theta) \ P_K(k|\theta, \Gamma)}. \tag{164}$$

Les performances de l'EM dépendent généralement des paramètres initiaux. La première estimation des paramètre devraient permettre de contourner les pièges des maxima locaux. La procédure de notre approche EM modifiée est la suivante :

1. étape d'initialisation
   • initialiser K avec l'ordre obtenu de la première estimation
   • initialiser les poids de manière équiprobable, les moyennes en accord avec la première estimation, et les variances avec la variance des données (de sorte que les Gaussiennes initiales couvrent l'ensemble des données) :
   $$\pi_k = 1/K, \quad \mu_k = \theta_k, \text{ et } \sigma_k^2 = \text{var}(\theta)$$
   • définir un seuil de convergence ε
2. étape d'estimation de l'Espérance
   • calculer $P_K(k|\theta, \Gamma)$ avec l'équation 160
3. étape de Maximisation de la Vraisemblance
   • calculer Γ' à partir de Γ avec les équations 162, 163, et 164
   • if $P_K(\theta|\Gamma') - P_K(\theta|\Gamma) > \varepsilon$
   alors $\Gamma \leftarrow \Gamma'$ et retourner à l'étape d'Espérance
   sinon terminer (l'algorithme EM a convergé).

5.3.3 Algorithme de filtrage de source

Afin de restituer chaque source k, nous avons sélectionné et regroupé les points temps-fréquence proches d'un même azimut θ. Nous utilisons les paramètres issus de la composante numéro k de l'EM. L'énergie des canaux gauche et droit du mélange est attribuée au canaux gauche et droit de la source proportionnellement à la probabilité *a posteriori*. Plus précisément, pour chaque source, nous définissons le masque MAP ci-après :

$$M_k(t, f) = P_K(k|\theta(t, f), \Gamma) \tag{165}$$

si $10 \log_{10} |\phi_k(\theta(t, f)|\mu_k, \sigma_k)| > L_{\text{dB}}$, et 0 sinon.

Cette contrainte limite le fait que la queue d'une distribution Gaussienne tend asymptotiquement vers 0 à l'infini. Au-dessous du seuil L_{dB} (exprimé en dB, et réglé à -20 dans nos simulations), nous supposons qu'une source d'intérêt ne contribue plus. Pour chaque source k, la paire de spectres à court terme est construite en fonction de :

$$S_L(t, f) = M_k(t, f) \cdot X_L(t, f), \tag{166}$$
$$S_R(t, f) = M_k(t, f) \cdot X_R(t, f). \tag{167}$$

La version temporelle de chaque source k est finalement obtenue par le biais d'une transformée de Fourier inverse avec une procédure de *overlap-add*.

5.4 Résultats de séparation de sources

5.4.1 Métriques

Comme dans toute application de séparation, l'oreille est le premier juge de la qualité des sources estimées. Toutefois, il est souvent souhaitable de saisir la qualité des sons par des critères mesurables, non seulement pour des raisons de commodités, mais aussi afin de comparer objectivement des systèmes différents.

Les critères de mesure de qualité pour les systèmes de séparation de source ne font pas l'unanimité, il est donc important de préciser la méthode d'évaluation. Nous nous alignons à quelques consignes de tests fréquemment observées dans la littérature. Pour l'évaluation nous utilisons le signal original, le mélange et l'estimation du canal où le signal original est le plus fort [RWB03] ; en d'autres termes, pour une source spatialisée vers la gauche, le canal de mélange gauche sera utilisé, ainsi que le canal gauche du signal d'origine, inversement pour une source spatialisée à droite, alors que pour une source au centre, tout canal est approprié. Les sources peuvent être évaluées dans le domaine temporel ou dans le domaine spectral. Le domaine spectral est utilisé sous l'hypothèse qu'il se rapproche de la perception par le système auditif, qui procède à une décomposition fréquentielle de sons perçus. En général, dans le domaine spectral, il est considéré que la phase ne joue pas un rôle dans les évaluations. Nous présentons les critères d'évaluation.

Les critères de qualité doivent adresser séparément les problèmes de bruit musical et celui des interférences. Des critères pour des systèmes basés sur des masques temps-fréquence ont été proposés.

Le rapport signal-interférence RSI mesure la présence des autres sources (interférences) dans le signal estimé. A cet effet, la quantité d'interférence à l'entrée RSIin_j et à la sortie RSIout_j est évaluée, ainsi que le gain obtenu RSII_j.

Pour une source J, la quantité des interférences est donnée par :

$$Y_J = \sum_{\substack{k=1 \\ J \neq k}}^{K} S_k \tag{168}$$

soit

$$Y_J(t,f) = X(t,f) - S_J(t,f). \tag{169}$$

Le RSI représente le rapport de l'énergie du signal S_J à l'énergie des interférences Y_J. Le RSI en décibels est donné par :

$$\begin{aligned}
\mathrm{RSIin}_J &= 10\log_{10}\frac{\parallel S_J(t,f) \parallel^2}{\parallel Y_J(t,f) \parallel^2}, && (170)\\
\mathrm{RSIout}_J &= 10\log_{10}\frac{\parallel M_J(t,f)S_J(t,f) \parallel^2}{\parallel M_J(t,f)Y_J(t,f) \parallel^2}, && (171)\\
\mathrm{RSII}_J &= \mathrm{RSIout}_J - \mathrm{RSIin}_J. && (172)
\end{aligned}$$

La valeur du RSI peut aller jusqu'à l'infini en cas d'absence totale d'interférence.

Le RSB ou rapport signal-bruit évalue l'écart entre le signal originelle et la version estimée [AMSM05]. Le RSB en décibels est donné par :

$$\begin{aligned}
\mathrm{RSBin}_J &= \mathrm{RSBin}_J, && (173)\\
\mathrm{RSBout}_J &= 10\log_{10}\frac{\parallel S_J(t,f) \parallel^2}{\parallel S_J(t,f) - M_J(t,f)S_J(t,f) \parallel^2}, && (174)\\
\mathrm{RSBI}_J &= \mathrm{RSBout}_J - \mathrm{RSBin}_J. && (175)
\end{aligned}$$

Dans nos cas, nous ne faisons pas de distinction entre les différentes catégories de bruit (bruit de microphone, artefacts ...).

Ces métriques existent similairement dans le domaine temporel pour les cas sous-déterminés. Ainsi nous procédons à l'évaluation dans le domaine temporel avec des critères équivalents à ceux énumérés plus haut. Soit pour les interférences :

$$\mathrm{RSBin}_J = 10\log_{10}\frac{\sum_n s_J(n)^2}{\sum_n y_J(n)^2}, \quad (176)$$

$$\mathrm{RSBout}_J = 10\log_{10}\frac{\sum_n \hat{s}_J(n)^2}{\sum_n \hat{y}_J(n)^2}, \quad (177)$$

$$\mathrm{RSB}_J = 10\log_{10}\frac{\sum \hat{s_J}(n)^2}{\sum_n(\hat{s}(n) - s_J(n))^2}. \quad (178)$$

Notons que le RSB ne considère pas séparément le bruit et les interférences ; il inclut en réalité l'ensemble des distorsions (interférences, bruit, artefacts) [VGF06].

Nous évaluerons les résultats de séparation pour des mélanges de 2 à 4 sources vocales et/ou instrumentales. La séparation des sources se fera principalement de deux manières. On utilisera d'une part un masque binaire basé sur le Maximum de Vraisemblance et un masque basé sur la probabilité *a posteriori* à l'aide d'un masque Gaussien. L'ajustement des modèles des sources se fera automatiquement par une approche EM.

5.4.2 Stratégie des tests

Le but des simulations est non seulement d'évaluer les performances du système de séparation, mais aussi de déterminer les facteurs qui affectent les capacités du système. Dans les chapitres précédents, nous avons pu constater que plus on a de sources dans un mélange, plus la condition d'orthogonalité des sources est réduite, et la localisation spatiale est également perturbée par des pics parasites. De même, les sources moins espacées spatialement entraîneraient des difficultés supplémentaires de localisation selon la résolution de l'histogramme. De ce fait, l'impact du nombre de sources dans le mélange et la distance spatiale entre les sources mérite d'être inspecté. Nous explorons des mélanges de 2 à 4 sources, avec des écarts spatiaux allant de 5° à 45°.

Le type de sources joue également un rôle à cause de la possible répercussion sur l'orthogonalité temps-fréquence. Les signaux de parole sont en général plus disjoint dans le domaine temps-fréquence que les sons instrumentaux [VE04] qui sont en général harmoniques, et certains instruments ont tendance à jouer au même moment, donc à se superposer [Mas04]. Afin de construire les exemples de mélanges, nous considérons des signaux de parole tirés de la base TIMIT [JSG93] et des signaux instrumentaux [VGF05]. Les signaux instrumentaux sont des signaux de basse, de percussion, de guitare et de piano.

Les mélanges tests seront générés en mélangeant additivement des versions spatialisées de signaux source. La spatialisation s'effectue avec la méthode de spatialisation binaurale paramétrique. La localisation est effectuée à partir de la méthode conjointe, qui assure une meilleure robustesse dans les hautes fréquences.

Pour la séparation de source, nous considérons deux stratégies. D'une part un masque binaire, en attribuant le point fréquentiel à la source la plus proche spatialement, d'autre part

en utilisant un filtrage spatial Gaussien paramétré automatiquement par une méthode EM modifiée. Dans ce cas, le masque est déterminé par la probabilité *a posteriori*.

Il n'est pas évident de comparer notre système avec les systèmes DUET, car leurs performances sont limitées à une bande fréquentielle inférieure à 1500Hz (à cause de l'ambiguïté de la phase), de même que la méthode de DASSS qui procède à la même détection que DUET, et nécessite un apprentissage Bayésien, et aussi, DASSS considère au maximum deux sources concurrentes par point temps-fréquence. Alors que la méthode d'Avendano se base sur un indice d'amplitude avec un ajustement de la largeur de la fenêtre, dans le cas de notre modèle paramétrique, l'indice d'ILD a une variance élevée, fait qui handicape l'approche d'Avendano qui se base sur une différence d'amplitude indépendante de la fréquence.

Nous analysons ainsi les performances de notre approche en cas de séparation par Maximum de Vraisemblance (masque binaire) et par Maximum *a posteriori* (masque par probabilité *a posteriori*) pour différents types de signaux et des mélanges de différents ordres.

5.4.3 Signaux sources

Les signaux sources monophoniques manipulés dans nos simulations sont des sons d'instruments généralement utilisés pour l'évaluation de méthodes d'analyse en composantes indépendantes et de séparation aveugle de sources [VGF05]. Il s'agit de cinq sources monophoniques (deux basses, une percussion, une guitare et un piano). Les deux sources de basses étant similaires, nous ne les mixeront pas dans un même exemple.

Ces quatre sources seront utilisés pour composer les mélanges binauraux nécessaires pour nos tests. Les signaux sont spatialisés par la méthode de spatialisation paramétrique avant d'être mélangés.

5.4.4 Mélanges à 2 sources

Le cas à deux sources est un cas déterminé car on a autant de mélanges que de sources, nous utilisons deux sources à savoir les signaux de basse et de percussion [VGF05] comme mélange représentatif. Il existe trois combinaisons possibles de superposition de sources s_1, s_2 ou s_1 et s_2 simultanément, c'est-à-dire au maximum deux sources contribuent à l'énergie du point (t, f). La séparation par masque binaire assigne le point fréquentiel exclusivement à s_1 ou s_2, sans prendre en compte les cas où les deux sources sont actives. Tandis que la séparation par masque MAP envisage qu'à chaque point fréquentiel, les deux sources sont actives. Dans le cas idéal, le MAP est binaire similairement à l'approche ML. En cas de superposition, le point d'intersection correspond à une position entre les deux sources, la position issue de la collision se rapproche de la source ayant le plus fort apport. La méthode EM modifiée prend en compte l'énergie moyenne de la source dans le calcul de la probabilité *a posteriori*. De ce fait, les contributions peuvent être considérées égales pour les points à mi-chemin entre les deux sources. Ainsi, une source ne se verrait pas attribuer toute l'énergie de la source d'interférence. L'approche par MAP à pour but de limiter les interférences, tout en limitant les distorsions dans les signaux estimés.

La figure 64 montre un résultat typique de séparation. La *basse* est à 0° et la *percussion* à 15°. Visuellement, nous remarquons que les estimations par masque MAP présentent moins

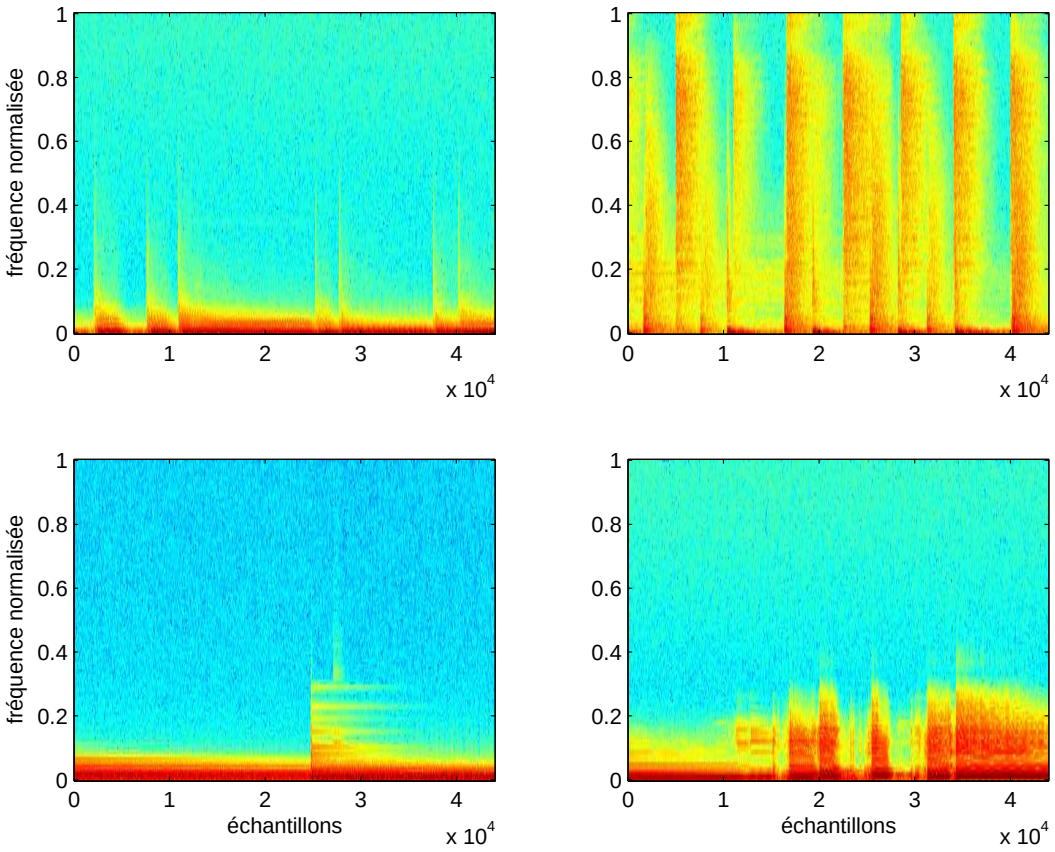


FIG. 63: *Spectrogrammes des sources : basse, percussion, guitare et piano (de gauche à droite et de haut en bas).*

d'interférence, mais elles présentent des niveaux d'énergie inférieur à l'original, ce qui entraînerait des distorsions. Les estimations issues du masque binaire par ML ont généralement une énergie supérieure à l'original, elles sont entachées d'interférences. Ces observations sont confirmées par les mesures de RSI avant et après la séparation exposées dans le tableau 6. Perceptivement, les sons issus du masque MAP sont préférés à ceux issus du masque ML. De manière générale, plus les sources sont éloignées les unes des autres, plus commode est la séparation ; en effet, une source plus à gauche aura un canal gauche dominant d'autant plus que sa source concurrente est spatialisée à droite, et vice versa. Il est également nécessaire de notifier que dans le cas de séparation par masque, on peut reconstruire un signal binaural stéréo. Des écoutes préliminaires montrent que l'effet spatial demeure pour la source stéréo estimée.

La table 6 confirme que l'approche par MAP est appréciable dans le cas de deux sources de paroles ou de deux sources musicales, les gains en RSI sont en moyenne supérieurs à 12 dB pour un niveau de distorsion d'environ 7 dB. Les tests ont montré que le niveau de RSI par MAP est largement inférieur à la séparation par masque binaire, pour un niveau de distorsion équivalent.

source (θ)	RSI in (dB)	RSI out (dB)	RSI gain (dB)	RSB out (dB)	RSBI (dB)
basse (0°)	8.34	21.36	13.02	9.36	1.02
percussion (+45°)	-7.19	11.53	18.71	1.58	8.77
speaker 1 (0°)	1.72	13.32	11.59	7.53	5.81
speaker 2 (+15°)	-0.07	14.55	14.61	7.00	7.07
basse (−20°)	-6.89	4.18	11.07	1.16	12.73
percussion (0°)	-2.93	16.55	19.48	9.10	12.03
piano (+15°)	-0.69	15.85	16.53	7.38	8.07
speaker 1 (−15°)	-2.32	8.17	10.49	3.35	5.67
speaker 2 (0°)	-2.82	11.47	13.39	2.59	5.41
speaker 3 (+15°)	-2.27	11.20	13.46	3.83	6.1
basse (−15°)	-3.36	13.05	16.41	5.24	8.6
percussion (0°)	-2.15	10.19	12.34	5.25	7.4
guitare (15°)	-3.07	10.98	14.05	4.68	7.75
piano (+30°)	-11.35	3.91	15.26	0.72	12.07
speaker 1 (−20°)	-3.86	14.17	18.04	-0.59	4.45
speaker 2 (−10°)	-4.78	13.54	18.32	2.04	6.82
speaker 3 (+5°)	-4.00	8.61	12.61	2.55	6.55
speaker 4 (+25°)	-4.02	6.48	10.50	1.03	5.05

Tab. 6: *Performances de séparation par masque MAP pour des mélanges binauraux à 2, 3 ou 4 sources. On a mesuré les niveaux d'interférence (RSI) et de bruit (RSB), ainsi que les gains apportés par la séparation de sources. On a des gains en RSI supérieurs à 10 dB, et des gains en RSB supérieurs à 5 dB.*

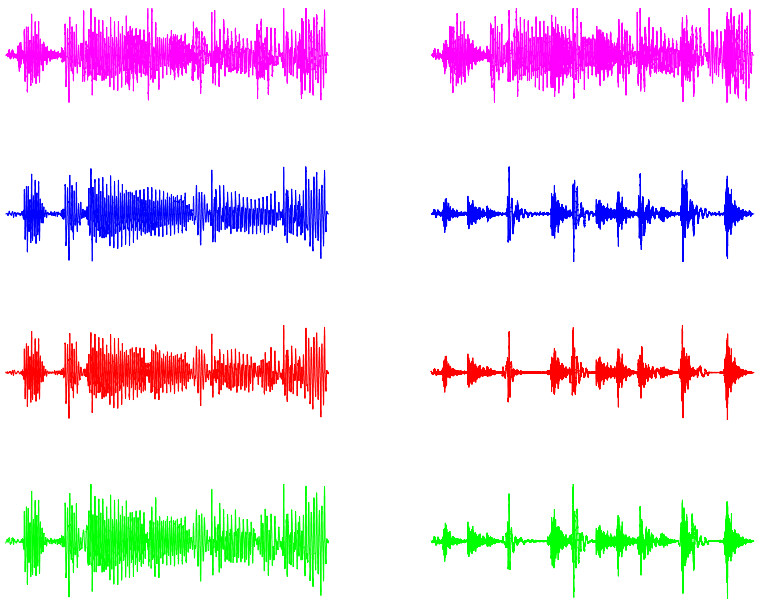


FIG. 64: *(ligne 1) Mélanges binauraux de deux sources instrumentales (basse, 0°) et (percussion, −15°), (ligne 2) originaux des deux sources, (ligne 3) estimations par masque MAP, (ligne 4) estimations par masque ML.*

5.4.5 Mélanges à 3 sources

Nous considérons des mélanges à trois sources (paroles et d'instruments). Dans le cas de trois sources, on distingue 11 combinaisons de sources actives possibles. Il est évident que les collisions dans le plan spectral sont plus fréquentes que le cas de deux sources.

La figure 65 montre un exemple typique de séparation pour un mélange binaural de trois sources, à savoir (basse, −60°), (percussion, +15°), (piano, +30°), elle affiche les versions temporelles des signaux originaux et des estimations par le masque MAP. La table 6 atteste par les niveaux de RSI que le niveau d'interférence croît lorsque le nombre de sources augmente et que la qualité des estimations est également affectée. Le gain en niveau d'interférence demeure supérieur à 10 dB, mais, on observe que le niveau de RSB se dégrade légèrement par rapport au cas de mélange de 2 sources. Dans le cas de trois sources, les estimations par masque MAP sont préférées à celles par masque binaire ML. En effet les spectrogrammes montrent que les sources ont une probabilité de collision importante pour les fréquences [0, 10] kHz. Les tests montrent que le masque ML devient inapproprié. Pour des positions spatiales proches, des points d'interférence sont passibles d'être attribués à une source quelconque. Réellement, dans la cas de trois sources, une superposition des deux sources latérales peut donner naissance à un point entre les deux sources, qui correspond à la source située entre les deux ; fait qui rend la séparation à plus de deux sources encore plus complexe. Objectivement, le niveau le gain en interférence est bon, toutefois la qualité globale en terme de RSB (distorsions) se dégrade par rapport au cas à trois sources. On a un gain en RSB au-dessus de 4 dB. Les tests d'écoute démontrent une fois de plus une préférence significative pour l'approche de séparation par masque Gaussien basé sur le MAP (table 6).

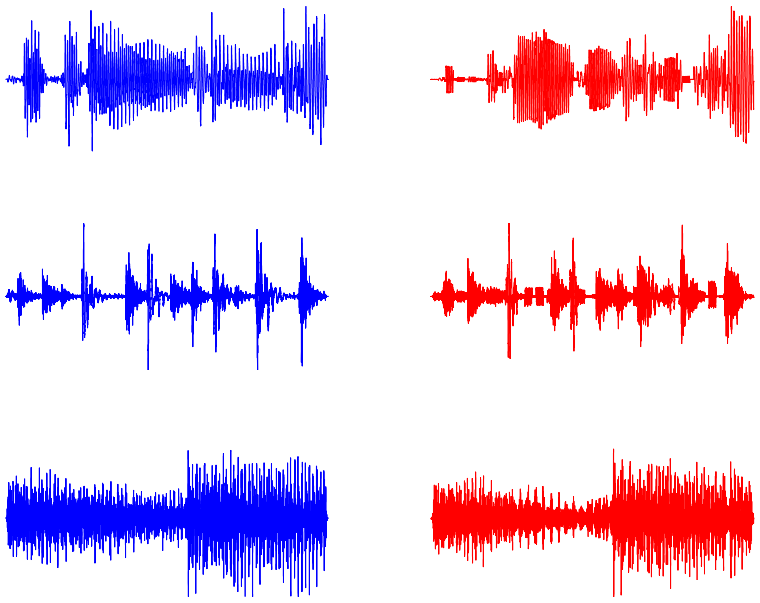


Fig. 65: *Séparation de trois sources instrumentales (basse, $-60°$), (percussion, $+15°$), (piano, $+30°$). Originaux (gauche) et estimations par masque MAP (droite).*

5.4.6 Mélanges à 4 sources

Dans cette section, nous explorons un mélange à quatre sources (basse, $-60°$), (percussion, $-15°$), (piano, $0°$), (guitare, $+25°$). Plus on a de sources, plus les collisions sont présentes. On obtient les mêmes observations que dans les cas du mélange à trois sources pour les approches de séparation par masques MAP et ML, c'est-à-dire que le gain en interférence est supérieur à 10 dB et une légère dégradation du RSB, environ 2 dB par rapport au cas à 3 sources mélangées. Avec ces quatre sources musicales cibles, la séparation semble avoir atteint sa limite de performance, alors qu'avec des signaux de paroles la séparation dispose encore d'une marge de manœuvre (table 6). La qualité globale des estimations s'est significativement dégradée, toutefois, la source de percussion est estimée raisonnablement, en effet son spectre à des hautes fréquences qui ne sont pas dans les autres sources. Ces signaux ont été choisis afin de montrer la pertinente de l'approche par masque MAP en cas de collisions multiples.

5.5 Discussion

Les algorithmes de séparation de sources sonores par masque de séparation sont prometteurs dans le cas sous-déterminé. Dans le plan temps-fréquence, l'orthogonalité idéale des sources n'est pas garantie. La distribution de l'énergie entre les différentes sources est de moins en moins évidente lorsque le nombre de sources croît. L'algorithme de localisation et séparation de sources DUET attribue l'énergie d'un point temps fréquence exclusivement à la source dominante en ce point. Déjà dans le cas de deux sources musicales, on a remarqué des interférences accrues (voir figure 64). Principalement, du fait que les sources sont généralement synchronisées et harmoniques, elles se superposent assez fréquemment. Il était donc nécessaire de déterminer les sources actives afin de répartir intelligemment l'énergie du point

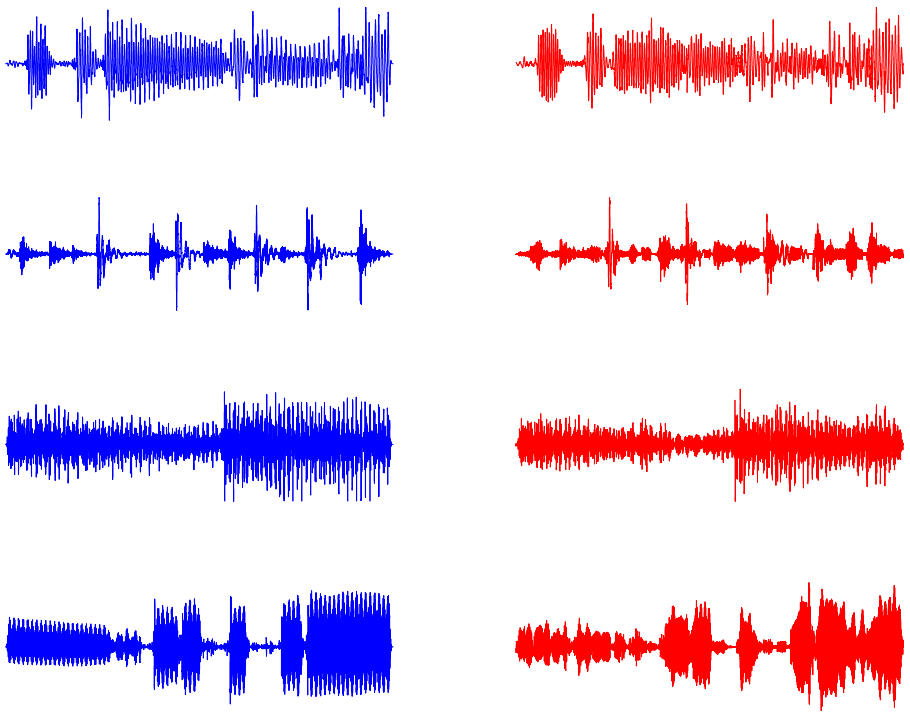


FIG. 66: *Séparation de quatre sources instrumentales (basse, −15°), (percussion, 0°), (guitare, +15°), (piano, +30°). Originaux (gauche) et estimations par masque MAP (droite).*

temps-fréquence. L'approche DASSS a proposé une solution où les deux sources les plus probables sont identifiées, l'énergie est alors répartie entre les deux sources. DASSS souffre du fait de son approche Bayésienne, c'est-à-dire qu'un apprentissage statistique de la distribution de la source avant le mélange est nécessaire. Aussi DUET et DASSS sont tous les deux limités par la précision de l'estimation de l'ITD, qui est valable jusqu'à environ 2000 Hz, au-delà l'estimation des ITD dans le spectre devient erronée. Alors nous utilisons pour notre approche de séparation une méthode de localisation conjointe qui résout les problèmes d'ambiguïté dans les hautes fréquences.

Notre approche par masque probabiliste prend tout de suite en compte la possibilité de mélange de toutes les sources, l'énergie du point temps-fréquence est alors répartie selon la probabilité de présence de chaque source du mélange. Les résultats obtenus par notre approche montrent un gain en terme de réduction d'interférence d'au moins 10 dB pour des mélanges de 2 à 4 sources (table 6), les gains en distorsion sont supérieurs à 5 dB. Il est à remarquer que la performance de la séparation se réduit avec l'augmentation de l'ordre du mélange et la proximité entre les sources. En effet, dans ces cas la probabilité de collision est plus grande. L'utilisation d'indices monoraux comme l'harmonicité des sources peut aussi aider à affiner la répartition de l'énergie.

Conclusion et perspectives

Cette thèse s'inscrit dans un contexte d'écoute active. Nous avons proposé un système complet qui permet de manipuler individuellement la position des sources sonores présentes dans un mélange binaural. Ce système analyse, spatialise, localise, sépare, re-spatialise et synthétise un nombre arbitraire de sources sonores à partir de signaux binauraux, dans le but de les diffuser sur un casque d'écoute (écoute individuelle) ou un ensemble de haut-parleurs (écoute publique).

Motivées par la perception naturelle du système auditif humain, nos contributions se positionnent dans les domaines de la modélisation et de la spatialisation, de la détection, de la localisation binaurale et de la séparation de sources sonores. Les résultats sont encore perfectibles scientifiquement. Ainsi, nous mentionnons des pistes et des orientations pour des travaux futurs.

Contributions

Modélisation et spatialisation de source

Nous avons vu que les indices primordiaux dans la perception spatiale sont la différence en amplitude (ILD) et la différence en temps d'arrivée (ITD), qui dépendent de la position azimutale θ de la source. Dans le chapitre 2, après une étude des ITD sur la base CIPIC, nous avons affiné le modèle sinusoïdal paramétrique de Kuhn avec une fonction d'échelle fréquentielle, et nous l'avons complété sur la bande de $[1-3]$ kHz.

Nous avons aussi proposé une répartition d'énergie paramétrique pour les canaux gauche et droit afin de simuler toute source monophonique à la position θ (chapitre 3). Cette technique permet de spatialiser efficacement toute source monophonique en combinant l'ILD et l'ITD et de l'adapter au rayon de la tête. En plus, elle permet de s'affranchir des mesures exhaustives des HRTF à toute position cible. Nous avons ensuite étendu la spatialisation binaurale à la spatialisation multi-diffusion en utilisant une matrice d'adaptation statique pour chaque azimut. En intégrant la distance, cette technique donne lieu à une adaptation à différentes configurations de haut-parleurs. Nous avons mis en évidence les capacités de ces techniques de spatialisation par des tests d'écoute et des comparaisons avec des techniques classiques comme VBAP [Pul97].

Après quantification de l'absorption par la distance, nous avons suggéré une relation entre la centroïde spectrale et la distance. Cette dernière sert autant pour la spatialisation que pour la localisation par la distance.

Localisation et séparation de source

Pour la localisation, nous avons adapté et optimisé la méthode de localisation conjointe décrite par Viste [Vis04] à notre modèle binaural paramétrique ; cette méthode combine l'azimut issu du modèle paramétrique d'ILD et l'azimut issu du modèle paramétrique d'ITD, afin de dériver un azimut robuste pour les hautes fréquences (chapitre 4). L'histogramme spatial de puissance issu des estimations à chaque point fréquentiel assure une détection de plusieurs sources dans le mélange binaural.

Dans le chapitre 5, nous avons modélisé l'histogramme de puissance comme un mélange de Gaussiennes, une méthode efficace de maximisation de vraisemblance nous a rendu possible la caractérisation de chaque source en estimant leur nombre, leurs azimuts, leurs variances et leurs probabilités *a priori*. À partir de ces paramètres, un banc de filtres spatiaux est paramétré automatiquement, et permet d'extraire les sources par un masque basé sur la probabilité *a posteriori*. Une comparaison avec les méthodes binaires a mis en évidence une maîtrise des interférences dans les estimations tout en assurant un niveau de bruit musical tolérable. Des résultats sur des mélanges de 2 à 4 sources musicales ou vocales soutiennent ces observations.

Perspectives

Localisation et re-spatialisation

Jusqu'alors, nous avons considéré les cas de localisation où l'auditeur et l'émetteur sont statiques. Nous nous intéressons à la localisation dans des scènes dynamiques, à cet effet des méthodes de suivi à l'exemple de ceux utilisés dans le suivi de partiels [MLR05] représentent une voie d'exploration. Il s'agirait aussi de déterminer la localisation et la trajectoire des sources en vue d'une éventuelle re-spatialisation et séparation dynamiques de sources. Aussi, nous avons ainsi commencé des travaux d'estimation de la source originelle monophonique à partir des signaux binauraux, et nous avons constaté que l'énergie de la source originelle estimée dépend de l'angle d'azimut. Des travaux futurs consisteraient à modéliser le biais azimutal et de rechercher une possible correction spatiale.

Aussi, nos méthodes s'appliquent dans le plan horizontal, l'intégration de l'élévation constitue également une perspective pour assurer une localisation et une spatialisation dans l'espace.

Séparation de sources audio

La séparation de sources demeure un défi face aux performances du système auditif humain. En effet, ce dernier combine naturellement différents indices acoustiques. Afin d'améliorer les performances de notre algorithme de séparation, il est nécessaire de combiner les indices spatiaux aux indices spectraux. Un suivi de l'évolution des caractéristiques d'harmonicité, d'enveloppe spectrale et de l'évolution de la localisation dans chaque bande spectrale [RE08] aideraient à affiner la répartition correcte de l'énergie du mélange. Inévitablement, la complexité des algorithmes augmentera aussi. Il s'avère ainsi important d'avoir un minimum d'informations *a priori* sur le modèle du mélange (mélange CD par exemple). Aussi, l'identification de sources actives dans une bande fréquentielle permettrait d'éviter des compétitions inutiles, par exemple entre une source à basses fréquences et une source à hautes fréquences

dans les hautes fréquences.

Projet RetroSpat

Une perspective beaucoup plus technique est de poursuivre le développement du logiciel d'informatique musicale RetroSpat (chapitre A). La mise en œuvre des techniques proposées permettrait aussi de les éprouver facilement dans divers environnements acoustiques et de mieux évaluer leur l'utilité pratique pour les acousmaticiens.

Annexe A

Le logiciel RetroSpat

Le système RetroSpat est mis en œuvre comme un logiciel temps réel d'informatique musicale sous licence publique GNU General Public License (GPL). La version actuelle est basée sur C++, Qt4 [BS06], JACK [JAC08], FFTW [FFT08] et fonctionne sur Linux et MacOS X.

Actuellement, RetroSpat met en œuvre les méthodes décrites dans les chapitres 3 et 4 (spatialisation et localisation). Les méthodes de séparation de sources sont en cours de développement afin d'être intégrées. RetroSpat s'organise en deux modules principaux : *RetroSpat Localizer* pour la localisation et la détection automatique de l'arrangement de haut-parleurs et *RetroSpat Spatializer* pour la spatialisation des sources. Le module à venir *RetroSpat Separator* sera destiné aux processus de recouvrement des sources proposées dans le chapitre 5.

Le développement du logiciel RetroSpat poursuit deux objectifs principaux, premièrement servir de plateforme d'expérimentation pour la recherche en localisation/spatialisation/separation de source audio, deuxièmement devenir un véritable outil de diffusion semi-automatique de musique acousmatique.

A.1 RetroSpat Localizer

RetroSpat Localizer (voir figure 67) est en charge de la détection automatique de la configuration des haut-parleurs. Il permet également à l'utilisateur de modifier interactivement la configuration, qui a été détectée ou chargée à partir d'un fichier XML.

La détection automatique des positions (en azimut et en distance) des haut-parleurs connectés à la carte son est d'une grande importance. Elle permet de s'adapter à toute nouvelle configuration de haut-parleurs. En effet, cette configuration sera l'une des premières actions à exécuter par un interprète dans un nouvel environnement.

Pour la configuration de la salle, l'interprète porte un casque avec des microphones encastrés dans des emplacements d'écouteurs sur les conseils du compositeur Jean-Michel Rivet (SCRIME), nous avons inséré des capsules Sennheiser KE4-211-2 dans un casque standard (phonocasque). L'interprète oriente la tête vers l'azimut zéro. Ensuite, chaque haut-parleur joue tour à tour un bruit blanc Gaussien échantillonné à 44.1 kHz. Le signal binaural enregistré à partir du phonocasque aux oreilles du musicien est transféré à l'ordinateur qui exécute

RetroSpat Localizer. Ensuite, chaque haut-parleur est localisé en azimut et en distance. La configuration proposée pourrait être ajustée ou modifiée par l'interprète en fonction des caractéristiques de la salle. En effet, l'histogramme de localisation permet de juger de la qualité de la localisation par la précision de l'estimation de l'azimut sur des signaux tests et par l'intensité des pics parasites.

A.2 RetroSpat Spatializer

Pour la spatialisation de son, des sources monophoniques sont chargées dans RetroSpat. Chaque source est paramétrée, puis diffusée. Les paramètres comprennent le volume, la localisation, des choix spéciaux tels que des trajectoires en cercle, arc, etc. La configuration de haut-parleurs est l'élément de base pour une spatialisation réussie, il est important d'en avoir le contrôle (voir la section A.1).

L'illustration de la figure 68 représente un mélange de 7 instruments et de voix (icônes de notes), dans une configuration frontale de 6 haut-parleurs (icônes de haut-parleur). Cette configuration a été paramétrée par RetroSpat Localizer.

Au cours de la diffusion, le musicien peut interagir individuellement avec chaque source de la pièce, il peut changer les paramètres de chaque source (azimut et distance) ; il peut même supprimer ou insérer une source dans la scène sonore. Dans la version actuelle, l'interaction avec RetroSpat est réalisée par le biais d'une souris.

Grâce à une implantation efficace en C++, avec le serveur de son JACK, RetroSpat peut diffuser simultanément plusieurs sources localisées à l'intérieur de la même paire de haut-parleurs (voir figure 68), où nous distinguons trois sources dans la paire (2,3)). Toutes les paires de haut-parleurs doivent rester synchronisées afin d'éviter des artefacts sonores. L'interface utilisateur basée sur Qt s'exécute dans un processus séparé moins prioritaire que le processus de traitement de signal.

RetroSpat a été testé sur un MacBook Pro, connectée à 8 haut-parleurs par le biais d'une carte son MOTU 828 MKLL. Il était en mesure de jouer plusieurs sources sans problèmes.

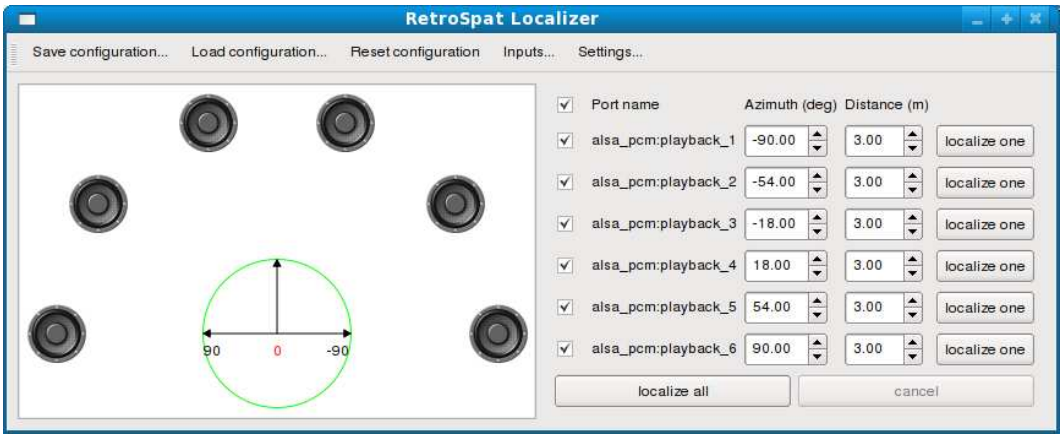


FIG. 67: *RetroSpat Localizer : interface graphique avec une configuration de 6 haut-parleurs.*

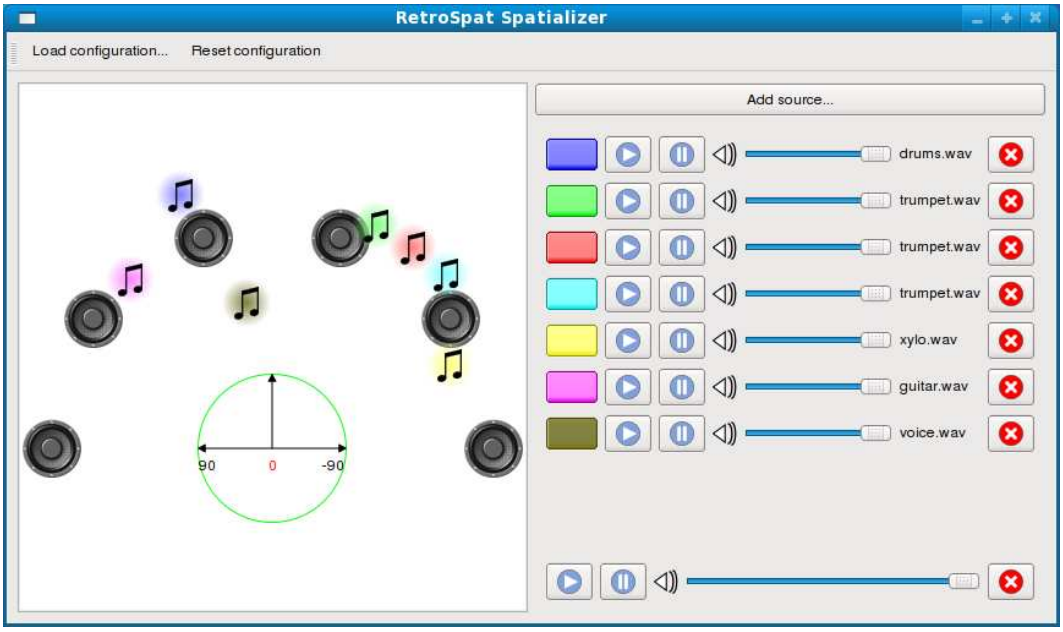


FIG. 68: *RetroSpat Spatializer : interface graphique avec 7 sources spatialisées à partir de la configuration de haut-parleurs présentée sur la figure 67.*

A.3 Applications musicales

Dans un concert, l'acousmaticien interagit avec la scène par le biais d'une console de mixage. Avec RetroSpat, le musicien dispose de plus de degrés de liberté sur un seul contrôleur (souris) :

- mouvements de la souris pour contrôler simultanément l'angle d'azimut et la distance de la source ;
- sources monophoniques possibles en entrée du système. Avec le module RetroSpat Separator, RetroSpat sera capable d'extraire les sources à partir d'une source CD, et permettra de modifier chaque source en direct dans un contexte d'écoute active ;
- de nombreuses sources peuvent être spatialisées à différents endroits en même temps ;
- visualisation globale dynamique de toute la scène sonore (apparition de source, mouvement, vitesse, etc).

Nous pensons que RetroSpat devrait largement simplifier les interactions de l'interprète avec la scène, et devrait donc lui permettre de se concentrer davantage sur la performance artistique que technique.

A.4 Architecture et processus principaux

RetroSpat implémente actuellement dans 57 classes une interface graphique et une chaîne d'analyse/transformation/synthèse de son. L'étape d'analyse consiste à lire le fichier audio (format .wav) et à le transformer dans le domaine spectral à l'aide de la transformée de Fourier rapide (FFT). La FFT constitue la base du traitement du signal dans RetroSpat, à

cet effet nous utilisons l'implantation efficace de FFT issue de la bibliothèque FFTW[1].

L'étape de transformation inclut les algorithmes de localisation/spatialisation sonore proposés dans le cadre de cette thèse. Ainsi, RetroSpat sert également de plateforme de tests. Enfin, l'étape de synthèse se résume en la resynthèse de versions temporelles de signaux spectraux transformés. La transformée de Fourier inverse (IFFT) de la FFTW est utilisée. La FFT et la IFFT sont mises en œuvre en combinaison avec une méthode *overlap-add*.

RetroSpat fonctionne dans un mode *multithreading* pour gérer plusieurs processus en parallèle, actuellement on a deux processus (threads) principaux : celui de l'interface graphique (GUITHREAD) pour l'initialisation de l'interface graphique et les mises à jour graphique et celui de traitement audio (AUDIOTHREAD) pour les opérations de traitement du signal. En plus, l'AUDIOTHREAD communique avec le serveur JACK. La fonction de cette communication est principalement de transférer les résultats des traitements à JACK. Le serveur JACK s'occupe alors de la gestion bas niveau des haut-parleurs (gestion des ports de sortie, ordonnancement, synchronisation, diffusion).

En présence de plusieurs sources, l'AUDIOTHREAD peut être lourdement chargé ; pour cela, nous l'avons fixé prioritaire par rapport aux mises à jour du GUITHREAD.

JACK permet de contrôler facilement les haut-parleurs connectés à la carte son MOTU par un principe de ports d'entrée associés à des ports de sortie, mais son fonctionnement ne favorise pas vraiment les performances ; dans l'implantation actuelle JACK appelle la fonction *callback-jack()* de l'AUDIOTHREAD, dans laquelle les fonctions de traitement de signal s'exécutent (FFT, localisation, spatialisation, IFFT) (voir figure 69).

Pour rémedier à cette situation, une solution consisterait à générer des processus pour des traitements complexes spécifiques comme la localisation et la spatialisation ; dans ce cas, il faut s'assurer de la synchronisation entre les processus et le serveur JACK, par exemple en utilisant un tampon circulaire (ringbuffer) en l'associant au principe producteur/consommateur.

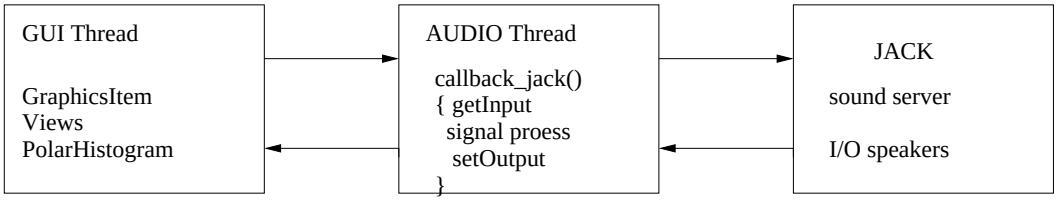


FIG. 69: *Interactions entre les processus principaux de RetroSpat.*

A.4.1 Processus graphique

L'interface graphique est programmée sous QT, le GUI Thread est donc le premier processus appelé, c'est le processus d'exécution de l'interface graphique. La fenêtre principale initialise tous les objets graphiques, les contrôles et le plan polaire (voir figure 68). La configuration

[1] voir URL : `http://www.fftw.org`

de haut-parleurs peut être paramétrée à l'aide d'un fichier XML (classe *SpeakerConfiguration*). Le fichier XML correspondant à la configuration de la figure 68 est affiché ci-après :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE speaker_configuration>
<speaker_configuration version="1.0">
    <speaker port="alsa_pcm:playback_1" azimuth="-90" distance="3"></speaker>
    <speaker port="alsa_pcm:playback_2" azimuth="-54" distance="3"></speaker>
    <speaker port="alsa_pcm:playback_3" azimuth="-18" distance="3"></speaker>
    <speaker port="alsa_pcm:playback_4" azimuth="18" distance="3"></speaker>
    <speaker port="alsa_pcm:playback_5" azimuth="54" distance="3"></speaker>
    <speaker port="alsa_pcm:playback_6" azimuth="90" distance="3"></speaker>
</speaker_configuration>
```

L'interface graphique permet aussi dans un dialogue (classe *SourceDialog*) de définir une trajectoire à la source. RetroSpat propose différents types d'animations : trajectoire en cercle, en arc, en segment (voir figure 70). Les arcs et les segments sont définis par une localisation de départ et une localisation d'arrivée. Une extension de cette fonctionnalité serait de dessiner la trajectoire désirée à main levée.

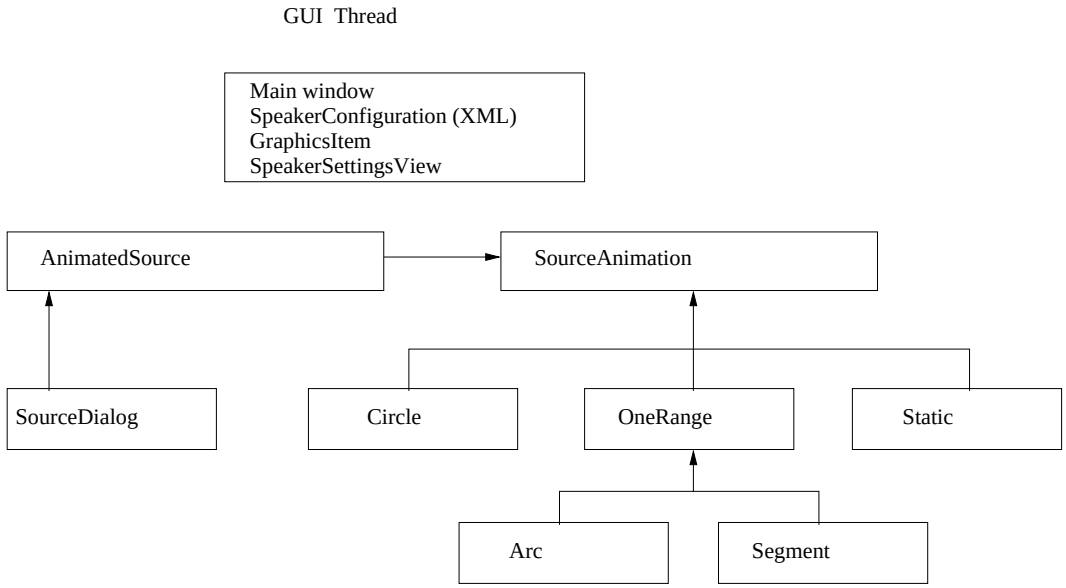


Fig. 70: *Architecture et fonctionnement de l'interface graphique dans RetroSpat.*

A.4.2 Processus audio

Le processus Audio s'occupe de toutes les opérations de traitement du signal (analyse/transformation/synthèse). Les traitements s'appuient sur les informations du contexte binaural (classe *BinauralContext*, voir figure 70) composé de la fenêtre de pondération (fenêtre de Hann), des facteurs d'échelle fréquentiels des modèles d'ITD et d'ILD, de la fréquence d'échantillonnage (44.1 kHz) et de la longueur de la FFT (2048 échantillons).

Les processus de localisation et de spatialisation accèdent au contexte binaural et au module TemporalSpectralTransformer (figure 71) qui contient les fonctions nécessaires à la transformation de signaux spectraux dans le domaine temporel. La classe *LocalizeProcess* appelle toutes les classes utiles aux fonctions de la classe *Localizer* qui implante l'algorithme de localisation ; similairement la classe *SpatializeProcess* regroupe toutes les données nécessaire à la bonne exécution de l'algorithme de spatialisation dans la classe *Spatializer*. La classe *Distancer* permet de positionner une source à distance ou de la localiser. La classe *Histogram* s'occupe du calcul de l'histogramme nécessaire à la localisation par recherche de maxima, ainsi que de la représentation graphique de l'histogramme spatial.

Pour l'instant, le processus de spatialisation nécessite la recherche de la paire de haut-parleurs encadrant l'angle d'azimut cible. La recherche de la paire dans RetroSpat est linéaire car nous cherchons uniquement le haut-parleur de gauche ou de droite, le deuxième étant facilement identifiable à partir de la configuration de haut-parleurs.

Pour la diffusion, à chaque haut-parleur (classe *Speaker*) est associé sa position en azimut, sa distance et un port de sortie sur le serveur JACK. Les données à diffuser par le haut-parleur sont accumulées par le biais de la classe *DataSpeaker* dans un buffer qui lui est associé (voir figure 72). Les données des spectres à court terme (classe *Spectrum*) sont superposées dans la classe *OverlapAccumulater* pour garantir le principe *overlap-add*, elle accède à deux trames successives.

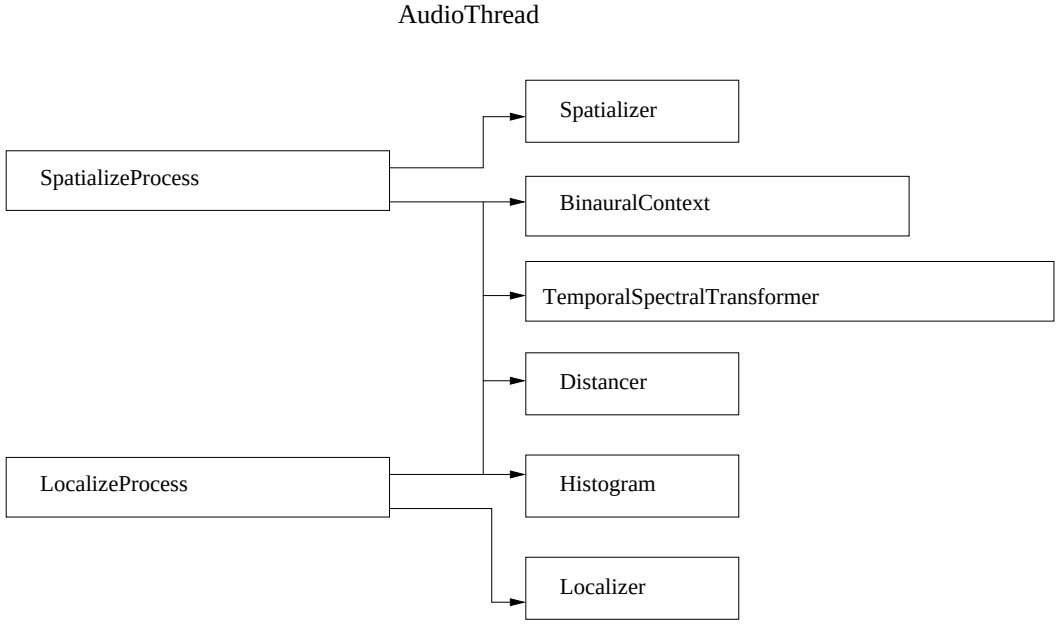


FIG. 71: *Architecture et fonctionnement du traitement audio dans RetroSpat.*

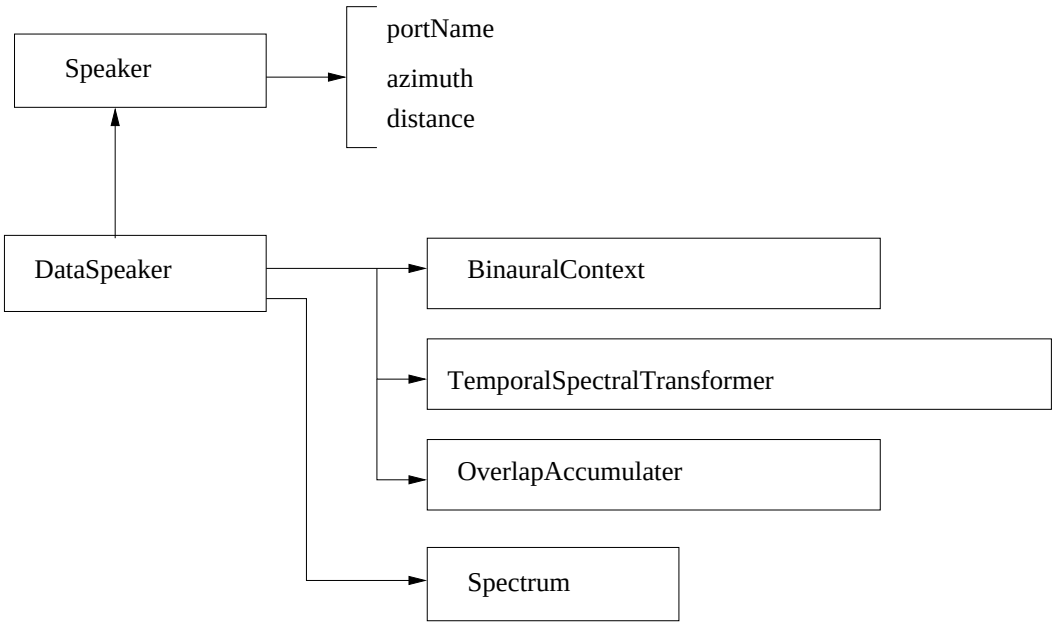


Fig. 72: *Architecture de l'environnement d'un haut-parleur et gestion des données à diffuser dans RetroSpat.*

Bibliographie

[ADTA01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database. In *Proc. IEEE WASPAA*, pages 99–102, New Paltz, New York, USA, October 2001.

[AMSM05] S. Araki, S. Makino, H. Sawada, and R.C. Mukai. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In *Proc. IEEE ICASSP*, volume 3, pages 81–84, 2005.

[Ave03] C. Avendano. Frequency-Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression, and Re-Panning Applications. In *Proc. IEEE WASPAA*, New York, 2003.

[Bau61] B. Bauer. Phasor analysis of some stereophonic phenomena. *J. Acoust. Soc. Am.*, 33(11), 1961.

[BC78] J . Blauert and W. Cobben. Some consideration of binaural cross correlation analysis. *Acustica.*, 39(2) :96–104, 1978.

[BC94a] G. J. Brown and M. P. Cooke. Computational auditory scene analysis. In *Computer speech and language*, volume 8, pages 297–336, 1994.

[BC94b] G. J. Brown and M. P. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8 :297–336, 1994.

[Beg92] D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. of Audio engineering Society*, 40 :895–904, 1992.

[Ber75] B. Bernfeld. Simple equations for multichannel stereophonic sound localization. *J. Audio Eng. Soc.*, 23(7), 1975.

[Ber88] A. J. Berkhout. A holographic approach to acoustic control. In *Journal of the Audio Engineering Society*, volume 36, pages 977–995, December 1988.

[Ber95] F. Berthommier. Source separation by functional model of amplitude demodulation. In *Proc. EUROSPEECH*, volume 4, pages 135–138, 1995.

[BH99] A.W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397 :517–520, 1999.

[BKBF07] E. Bates, G. Kearney, F. Boland, and Dermot Furlog. Localization accuracy of advanced spatialization techniques in small concert halls. In 153*rd meeting of the Acoustical Society of America*, June 2007.

[Bla97] J. Blauert. *Spatial Hearing.* MIT Press, Cambridge, Massachusetts, revised edition, 1997. Translation by J. S. Allen.

[Blu31] A. Blumlein. Improvements in and relating to sound trnasmission, sound recording and sound reproducing systems. Technical Report 394325, British patent Specification, 1931.

[BM07] J. Bourgeois and W. Minker. *Time-domain beamforming and blind source separation.* Springer, 2007.

[BR99] R. Brungart and W. Rabinowitz. Auditory localization of nearby sources. *J. Acoust. Soc. Am.*, 106 :1465–1479, 1999.

[Bra02] J . Braasch. Localization in the presence of a distracterans reverberation in the frontal horizontal plane. *Acust. Acta Acust.*, 88 :956–969, 2002.

[Bra03] Jean Yves Bras. *Les courants musicaux du XX^e siècle ou la musique dans tous ses états.* Papillon, 2003.

[Bre90] S. Bregman. *Auditory Scene Analysis : the perceptual organization of sound.* MIT Press, Cambridge, Massachusetts, first edition, 1990.

[Bro02] C. Brown. T60 Matlab function. 2002.

[Bru83] M. Bruneau. *Introduction aux théories de l'acoustique.* Université du Maine, Le Mans, France, 1983.

[BS94] R. E. Berg and D. G. Stork. *The physics of sound.* Prentice hall, second edition, 1994.

[BS95] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural computation*, 7 :1129–1159, 1995.

[BS06] J. Blanchette and M. Summerfield. *C++ GUI Programming with Qt4.* Prentice Hall, 2006.

[BSC00] N. Kopco B.G. Shinn-Cunningham, S. Santarelli. Tori of confusion : binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, 107(3) :1627–1636, 2000.

[BSCM05] N. Kopco B.G. Shinn-Cunningham and T Martin. Localizing nearby sound sources in a classroom : Binaural room impulse responses. *J. Acoust. Soc. Am.*, 117 :3100–3115, 2005.

[BSZ+95] H. Bass, L. Sutherland, A. Zuckerwar, D. Blackstock, and D. Hester. Atmospheric Absorption of Sound : Further Developments. *Journal of the Acoustical Society of America*, 97(1) :680–683, 1995.

[BZ02] Matthias Baeck and Udo Zölzer. Performance analysis of a source separation algorithm. In *Proc. Digital Audio Effects (DAFx) Conf.*, Hamburg, Germany, September 2002.

[Car98] J.-F. Cardoso. Blind source separation : statistical principles. *Proceeedings of the IEEE*, 9(10) :2009–2025, 1998.

[CCA02] S. Choi, A. Cichocki, , and S. Amari. Equivariant nonstationary source separation. *Neural Networks*, 15 :121–130, 2002.

[CCPL95] S. Choi, A. Cichocki, H-M. Park, and S-Y. Lee. Blind source separation and independent component analysis : A review. *Neural information processing*, 6(1) :1–57, 1995.

[CD78] H. S. Colburn and N. I. Durlach. *Models of binaural interaction.* Academic Press, 1978. In Handbook of perception.

[Cho71] John M. Chowning. The Simulation of Moving Sound Sources. *Journal of the Acoustical Society of America*, 19(1) :2–6, 1971.

[CNC73] G. C. Carter, A. H. Nuttall, and P. G. Cable. The smoothed Coherence Transform. *IEEE Signal Processing Letters*, 61 :1497–1498, 1973.

[Com94] P. Common. Independent component analysis- a new concept ? *Signal processing*, 36(3) :287–314, 1994.

[CS72] D. Cooper and T. Shiga. Discrete matrix multichannel stereo. *Journal of the Audio Engineering Society*, 20 :346–360, 1972.

[DAA99] R. Duda, C. Avendano, and V. Algazi. An adaptable ellipsoidal head model for the interaural time difference. In *Proc. ICASSP*, 1999.

[DC95] C. J. Darwin and R. P. Carlyon. *Auditory grouping in "Hearing"*. Academic Press, 1995.

[DM97] R. O. Duda and W. L. Martens. Range-Dependence of the HRTF for a Spherical Head. In *Proc. IEEE WASPAA*, New York, 1997.

[Dur63] N. I. Durlach. Equalization and cancellation theory of binaural masking level difference. *J. Acoust. Soc. Am.*, 35(8) :1206–1218, 1963.

[Dur72] N. I. Durlach. Binaural signal detection : Equalization and cancellation theory. *Foundations of Modern Auditory Theory*, 2 :369–462, 1972.

[Dut06] A. Dutto. http ://pedagogie.ac-montpellier.fr/musique/pedagogie/ecoute/spatialisation/histoire.htm Technical report, Académie Montpellier, 2006.

[FB02] C. Faller and F. Baumgarte. Binaural cue coding : A novel and efficient representation of spatial audio. In *Proc. IEEE ICASSP*, volume 2, pages 1841–1844, 2002.

[FFT08] `http ://www.fftw.org`. 2008.

[FM04] C. Faller and J. Merimaa. Source Localization in Complex Listening Situations : Selection of Binaural Cues Based on Interaural Coherence. *J. Acoustic. Soc. Am.*, 116(5) :3075–3089, 2004.

[Gai93] Werner Gaik. Combined evaluation of interaural time and intensity differences : Psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94(1) :98–110, 1993.

[Gar97] W. G. Gardner. *3-D Audio using loudspeakers*. PhD thesis, Massachusetts institut of technology, 1997.

[GB69] Goldberg and Brown. Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli : Some physiological mechanisms of sound localization. *Journal of Neurophysiology*, 32 :613–636, 1969.

[GE93] M.M. Goodwin and G.W. Elko. Constant beamwidth beamforming. In *Proc. IEEE ICASSP*, volume 1, pages 169–172, 1993.

[Ger73] M. Gerzon. Periphony : with-height soud reproduction. *Journal of the Audio engineering Society*, 21 :2–10, 1973.

[Ger74] M. Gerzon. Surround sound psychoacoustics. *Wireless world*, 80 :483–486, 1974.

[Ger77] M. Gerzon. Criteria for evaluating surround sound systems. *J. of Audio engineering Society*, 25 :400–408, 1977.

[GG96] M. D. Good and R. H. Gilkey. Sound localization in noise : The effect of signal to noise ratio. *J. Acoust. Soc. Am.*, 99 :1108–1117, 1996.

[GJ82] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. In *IEEE Trans. Ant. Prop.*, pages 27–34, 1982.

[Har] William M. Hartmann. `http ://www.aip.org/pt/nov99/locsound.html`.

[Har78] Frederic J. Harris. On the Use of Windows for Harmoic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 60(1) :51–83, 1978.

[Har97] W. M. Hartmann. *Listening in a room and the precedence effect, in binaural and spatial hearing in real and virtual environments*. R. H. Gilkey and T.R Anderson, Lawrence, 1997. ISBN 0-8058-1654-2.

[Hay96] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, New Jersey, 1996.

[HBS99] K. Hartung, J. Braasch, and S. Sterbing. Comparison of different methods for the interpolation of HRTFs. In 16*th conference of the Audio Engineering Society*, Helsinki, Finland, April 1999.

[HJ86] J . Hérault and C. Jutten. Space or time adaptive signal processing by neural network models. In *American institute of physics (AIP) Conf.*, volume 1, pages 206–211, 1986.

[HKO01] A. Hyvärinen, J. Karhunen, and E. Oja. *Independant component analysis*. John Wiley & Sons, 2001.

[HPP05] S. Hwang, Y. Park, and Y. Park. Source localization using HRTF databases. In *Proc. Int. conference on control, automation and sytems*, Gyeong gi, Korea, June 2005.

[HT73] E. J. Hannan and P. J. Thomson. Estimating Group Delay. *Biometrika*, 60 :241–253, 1973.

[HW04] G. Hu and D. L. Wang. Monoral speech segregation based on pitch tracking and amplitude modulation. In *Proc. IEEE Trans. Neural Networks*, volume 15, pages 1135–1150, 2004.

[Int93] International Organization for Standardization, Geneva, Switzerland. *ISO 9613-1 :1993 : Acoustics – Attenuation of Sound During Propagation Outdoors – Part 1 : Calculation of the Absorption of Sound by the Atmosphere*, 1993.

[JAC08] `http ://jackaudio.org`. 2008.

[Jef48] L. A. Jeffress. A place theory fo sound localization. *Journal of Comparative and Physiological Psychologie.*, 61 :468–486, 1948.

[Jes73] M. Jessel. *Acoustique théorique, propagation et holophonie*. Masson, Paris, France, 1973.

[JRY00] A. Journine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals : Demixing N sources from 2 mixtures. In *Proc. IEEE ICASSP*, pages 5–9, 2000.

[JSG93] et al. John S. Garofolo. TIMIT Acoustic-Phonetic Continuous Speech Corpus. In *Linguistic Data Consortium, University of Pennsylvania*, Philadelphia, USA, 1993.

[JWL98] J. M. Jot, S. Wardel, and V. Larcher. Approaches to binaural synthesis. In *Presented at the* 105*th convention of the Audio Engineering Society*, San Francisco, California, USA, September 1998.

[KC76] C. H. Knapp and G. C. Carter. The Generalized Correlation Method for the Estimation of Time Delay. *IEEE Trans. on Sig. Proc.*, 24(4) :320–327, 1976.

[KNKT95] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of bayesian probability network to music scene analysis. In *Proc. of IJCAI Workshop on CASA*, pages 115–137, 1995.

[Kro96] Kristian Kroschel. *Statistische Nachrichten theorie 2nd. ed.* Springer-Verlag Berlin, 1996.

[Kuh77] George F. Kuhn. Model for the Interaural Time Differences in the Azimuthal Plane. *J. Acoust. Soc. Am.*, 62(1) :157–167, 1977.

[LDA07] Kuldip K. Paliwal Leigh D. Alsteris. Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra. *Computer Speech and Language*, 21 :174–186, 2007.

[Lic51] J. C. R. Licklider. A duplex theory of pitch perception. In *Experimentia*, volume 7, pages 128–133, 1951.

[Lin86a] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of laterateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6) :1608–1622, 1986.

[Lin86b] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front. *J. Acoust. Soc. Am.*, 80(6) :1623–1630, 1986.

[Lou97] Alain Louvier. *L'orchestre.* Editions, Combre, Paris, 1997.

[Ma95] H. Moller and al. Head-related transfert functions of human subjects. *J. of Audio engineering Society*, 43(5) :300–321, 1995.

[Mak62] Y. Makita. Localisation directionnelle du son dans un champ sonore stéréophonique. *Revue de l'UER*, 1962.

[Mas03] Aaron S. Master. Sound source separation of n sources from stereo signals via fitting to n models each lacking one source. Technical Report EE391, Stanford University, 2003.

[Mas04] Aaron S. Master. Bayesian two source modeling for separation of N sources from stereo signals. In *Proc. IEEE ICASSP*, 2004.

[Mas06] Aaron S. Master. *Stereo Music Source Separation via Bayesian Modeling.* PhD thesis, Stanford University, 2006.

[Mel91] D. Mellinger. *Event formation and separation in musical sound.* PhD thesis, Standford, 1991.

[MLR05] S. Marchand M. Lagrange and J-B. Rault. Tracking of partials for the sinusoidal modeling of polyphonic sounds. In *Proc. IEEE ICASSP*, volume 3, pages 229–232, 2005.

[MM77] S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *J. Acoust. Soc. Am.*, 61(6) :1567–1576, 1977.

[MM06] J. Mouba and S. Marchand. A Source Localization/Separation/Respatialization System Based on Unsupervised Classification of Interaural Cues. In *Proc. Digital Audio Effects (DAFx) Conf.*, pages 233–238, Montreal, Quebec, Canada, September 2006.

[MMMR08] J. Mouba, S. Marchand, B. Masencal, and J. M. Rivet. RetroSpat : a perception-based system for semi-automatic diffusion of acousmatic music. In *Proc. of the sound and music computing conference*, pages 33–40, Berlin, Germany, July–August 2008.

[MOK95] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3) :411–419, 1995.

[Mol73] John Molino. Perceiving the Range of a Sound Source When the Direction is Known. *J. Acoust. Soc. Am.*, 53(5) :1301–1304, 1973.

[Mol92] H. Moller. Fundamentals of binaural technology. *Applied acoustics*, 36(5) :171–218, 1992.

[Moo90] F. R. Moore. *Elements of computer music.* Prentice Hall, 1990.

[Muc06] R.A. Mucci. A Comparison of Efficient Beamforming Algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(3) :548–558, 2006.

[NE98] R. Nicol and M. Emerit. Reproducing 3d-sound for videoconferencing : a comparison between holophony and ambisonics. In *Proc. Digital Audio Effects (DAFx) Conf.*, Barcelona, Spain, November 1998.

[NE99] R. Nicol and M. Emerit. 3d-sound reproduction over an extensive listening area : A hybrid method derived from holophony and ambisonic. In *Presented at the* 16*st international conference of the Audio Engineering Society*, Helsinki, Finland, april 1999.

[Nor62] B. Nordlund. Physical factors in angular localization. *Acta Oto-Laryngol*, 54 :74–93, 1962.

[NS04] A. Nishimura and M. Sasaki. Absolute cues for auditory distance in front and lateral directions. *Journal of the Acoutical science and technology*, 25(2) :127–135, 2004.

[OSB99] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing.* Prentice Hall, New Jersey, edition edition, 1999.

[PB04] F. Pachet and J.P. Briot. *Informatique musicale : du signal au signe musical.* Lavoisier Paris, 2004.

[Pea76] T. W. Pearsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 50(4) :911–918, 1976.

[Pul97] V. Pulkki. Virtual Sound Source Positioning using Vector Base Amplitude Panning. *Journal of the Acoustical Society of America*, 45(6) :456–466, 1997.

[Pul99] V. Pulkki. Uniform Spreading of Amplitude Panned Virtual Sources. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 17–20, New Paltz, New York, 1999.

[Pul00] V. Pulkki. Generic panning tools for MAX/MSP. In *Proceedings of the International Computer Music Conference*, pages 304–307, Berlin, Germany, August 2000.

[Pul01] V. Pulkki. *Spatial sound generation and perception by amplitude panning techniques.* PhD thesis, Helsinki university of technology, Finland, 2001.

[Ray07] J. W. Strutt (Lord Rayleigh). On our perception of soudn direction. *Phil. Mag.*, 13 :214–302, 1907.

[RE08] M. Raspaud and G. Evangelista. Binaural partial tracking. In *Proc. Digital Audio Effects (DAFx) Conf.*, Espoo, Finland, September 2008.

[RRD+96] D. V. Rabinkin, R. R. Renomeron, A. Dahl, J. C. French, J. L. Flanagan, and M. H. Bianchi. A DSP Implementation of Source Localization using Microphone Arrays. In *Proceedings of the SPIE, 2846*, pages 88–89, 1996.

[Rum01] Francis Rumsey. *Spatial Audio.* Focal Press, Oxford, United Kingdom, first edition, 2001. Reprinted 2003, 2005.

[RV89] D. D. Rife and J. Vaderkooy. Transfer-function measurement with maximum-length sequences. *J. Audio. Eng. Soc.*, 6 :419–444, 1989.

[RWB03] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 22(2) :149–157, 2003.

[RY02] S. Rickard and O. Yilmaz. On the W-disjoint orthogonality of speech. In *Proc. IEEE ICASSP*, pages 13–17, 2002.

[Rzh63] S. N. Rzhevkin. *The theory of sound.* Pergamon, Oxford, 1963.

[Sch43] L. Schwarz. Zur Theorie der Beugung einer ebenen Schallwelle an der Kugel. *Akust. Zeits.*, 8 :91–117, 1943.

[Sch65] M. R. Schroeder. New method of measuring reverberation time. *J. Acoust. Soc. Am.*, 37 :409–412, 1965.

[SJY93] P. H. Smith, P. X. Joris, and T. C. T. Yin. Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat : evidence for delay lines to the medial superior olive. *J. Comp. Neurol.*, 331 :245–260, 1993.

[SL90] M. Slaney and R. F. Lyon. A perceptual pitch detector. In *Proc. IEEE ICASSP*, volume 1, pages 357–360, 1990.

[SRR03] R. Balan S. Rickard and J. Rosca. Blind source separation based on space-time-frequency diversity. In *Symposium on Independent Compoenent Analysis and Blind Signal Separation*, pages 493–498, 2003.

[SS34] J. Steinberg and W. Snow. Auditory perspectives - physical factors. In stereophonic techniques. *Audio engineering Society*, pages 3–7, 1934.

[TF06] C. Tournery and C. Faller. Improved Time Delay Analysis/Synthesis for Parametric Stereo Audio Coding. *Journal of the Audio Engineering Society*, 29(5) :490–498, 2006.

[The91] G. Theile. On the naturalness of two channel stereo sound. *J. of Audio engineering Society*, 39(10) :761–767, 1991.

[TR67] W. R. Thurlow and P. S. Runge. Effects of induced head movements on localization of direct sound. *J. Acoust. Soc. Am.*, 42 :480–487, 1967.

[VE03] Harald Viste and Gianpaolo Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. Digital Audio Effects (DAFx) Conf.*, London, UK, September 2003.

[VE04] Harald Viste and Gianpaolo Evangelista. Binaural source localization. In *Proc. Digital Audio Effects (DAFx) Conf.*, Naples, Italy, September 2004.

[Ver98] E.N. G. Verheijen. *Sound reproduction by Wave Field Synthesis.* PhD thesis, Delft university of technology, 1998.

[VGF05] E. Vincent, R. Gribonval, and C. Févotte. http ://www.irisa.fr/metiss/bass-db. Technical report, Académie Montpellier, 2005.

[VGF06] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. In *IEEE trans. on audio, speech , and language procesing*, volume 14, pages 1462–1469, 2006.

[Vis04] H. Viste. *Binaural Localization and Separation Techniques.* PhD thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[VJA+05] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Queen Mary University of London, 2005.

[Wal40] H. Wallach. The role of head movements and the vestibular and visual cues in in sound localization. *J. Comp. Physiol.*, 61 :339–368, 1940.

[WB06] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis : Principles, Algorithms, and Applications.* Wiley-IEEE Press, 2006.

[Wei85] M. Weintraub. *A theory and computational model of monoral auditory sound separation.* PhD thesis, Stanford, 1985.

[Wel67] D. Welch. The Use of Fast Fourier Transform for the Estimation of Power Spectra : A Method Based on Time-Averaging over Short, Modified Periodograms. *IEEE Trans. on Audio and Electroacoustics*, 15(22) :70–73, 1967.

[WKFA97] F.L. Wightman, D.J. Kisteler, S. H. Foster, and J. Abel. *chapter : Factors affecting the relative salience of sound localization cues in book : Binaural and spatial hearing.* In. R. Gilkey ans T. Anderson, New Jersey, 1997.

[Woo54] R. S. Woodworth. *Experimental Psychology.* Holt, New York, 1954.

[WRTR67] J. W. Mangels W. R. Thurlow and P. S. Runge. Head movements during sound localization. *J. Acoust. Soc. Am.*, 42 :489–493, 1967.

[WW99] D.B. Ward and G. W.Elko. Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation. *IEEE signal processing letter*, 6(5) :106–108, 1999.

[YBC07] Huang Yiteng, Jacob Benesty, and Jingdong Chen. On Crosstalk Cancellation and Equalization With Multiple Loudspeakers for 3-D Sound Reproduction. *IEEE Signal Processing Letters*, 14(10) :649–652, 2007.

[YHHD92] C. Yuan-Hwang and F. Hwai-Der. Frequency-domain implementation of Griffiths-Jim adaptive beamformer. *Journal of the Acoustical Society of America*, 91(6) :3354–3366, 1992.

[YR04] O. Yilmaz and S. Rickard. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, 52(7) :1830–1847, 2004.

[Zur80] P. M. Zurek. The precedence effect and its possible role in the avoidance of interaural ambiguities. *J. Acoust. Soc. Am.*, 67(3) :952–964, 1980.