Etudes d'algorithmes d'extraction des informations de spatialisation sonore : application aux formats multicanaux

Manuel Briand

▶ **To cite this version:**

Manuel Briand. Etudes d'algorithmes d'extraction des informations de spatialisation sonore : application aux formats multicanaux. Traitement du signal et de l'image [eess.SP]. Institut National Polytechnique de Grenoble - INPG, 2007. Français. NNT : . tel-00141862

HAL Id: tel-00141862

https://theses.hal.science/tel-00141862

Submitted on 16 Apr 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : « Signal, Image, Parole, Télécoms »

préparée au laboratoire Speech & Sound Technologies & Processing de France Télécom R&D

dans le cadre de **l'Ecole Doctorale** « Electronique, Electrotechnique, Automatique et Traitement du Signal »

présentée et soutenue publiquement

par

Manuel BRIAND

le 20 Mars 2007

ETUDES D'ALGORITHMES D'EXTRACTION DES INFORMATIONS DE SPATIALISATION SONORE : APPLICATION AUX FORMATS MULTICANAUX

JURY

M.	Nicolas MOREAU,	Professeur, ENST/TSI – Paris,	Rapporteur
M.	Jean-Marc JOT,	Docteur ENST, Creative ATC – Scotts Valley	Rapporteur
Mme.	Nadine MARTIN,	Directeur de Recherche CNRS, LIS – Grenoble	Directrice de thèse
M.	David VIRETTE,	Ingénieur R&D, France Telecom – Lannion	Co-encadrant
M.	Bruno TORRÉSANI,	Professeur, LATP – Marseille,	Examinateur
M.	Gang FENG,	Professeur, ICP – Grenoble,	Examinateur

RÉSUMÉ

Ce travail de thèse s'intéresse à l'extraction d'informations spatiales pertinentes pour la compression du son multicanal. L'approche empruntée par les procédés émergents de codage audio paramétrique, pour la compression de signaux stéréophoniques[1] ou multicanaux (format 5.1, 7.1 etc.), considère une représentation paramétrique du signal basée sur les indices de la localisation auditive. Associée à un codage audio classique du signal somme des canaux originaux, la transmission de ces paramètres permet la reconstruction d'un signal multicanal dont les indices inter-canaux approximent ceux du signal original. Ces méthodes offrent alors une solution de codage efficace pour les applications à débit contraint puisque le débit global est équivalent à celui du codeur audio employé (mono ou stéréo pour la compression d'un signal au format 5.1) moyennant un débit variable pour la transmission des informations spatiales. Néanmoins, la qualité subjective de la reconstruction audio et spatiale est fortement liée au modèle employé qui n'est pas complètement adapté à la nature diverse et variée des signaux audio multicanaux.

La première orientation de ce travail de thèse nous a conduit à approfondir cette approche de codage audio multicanal avec l'objectif d'en améliorer les performances. Alors que les procédés de codage actuels extraient les paramètres spatiaux avec une résolution temps-fréquence fixe, nous avons cherché à adapter l'extraction des paramètres inter-canaux au contenu fréquentiel des signaux. Nous nous sommes appuyés sur un processus de segmentation dans le plan temps-fréquence pour extraire les motifs spectraux porteurs des informations de spatialisation sonore. Cependant, les expérimentations menées ne nous ont pas permis d'obtenir les résultats escomptés étant donné l'interprétation délicate des motifs spectraux issus du processus de segmentation. Nous avons également cherché à mesurer les dégradations introduites par les procédés de codage audio paramétrique pour être en mesure d'affiner la représentation paramétrique des zones temps-fréquence subjectivement critiques. Nos expérimentations montrent finalement que l'erreur de reconstruction établie par différence entre les signaux reconstruits et les signaux originaux (avec prise en compte du seuil de masquage fréquentiel) présente une trop grande répartition énergétique pour permettre l'identification des zones temps-fréquence perceptuellement critiques.

La seconde orientation de nos recherches s'est concentrée sur l'établissement d'un modèle, issu du mélange instantané de sources directionnelles et d'ambiances sonores, pour proposer une alternative au schéma de codage paramétrique actuel. L'analyse en temps et en sous-bandes de fréquences de signaux suivant ce modèle nous a permis de décrire la hiérarchisation des composantes d'après la distribution des valeurs propres. Nous avons ensuite évalué les performances, en termes de concentration d'énergie, de l'Analyse en Composante Principale (ACP), réalisée en temps et en sous-bandes de fréquences, de signaux à deux et trois composantes suivant notre modèle. Pour répondre à l'objectif principal de compression multicanale, nous avons fait le choix d'utiliser une approche paramétrique pour réaliser l'ACP bi- et tridimensionnelle. Nous proposons une méthode pour extraire un ou plusieurs angles de rotation utiles à l'ACP et une interprétation physique de ces angles à partir de la connaissance du système de reproduction sonore. Finalement, nous utilisons cette décomposition paramétrique de la covariance au sein d'une nouvelle méthode de codage qui repose à la fois sur la concentration de l'information dominante et sur l'extraction de paramètres utiles à la reconstruction du signal. Un premier test subjectif nous a permis de définir les paramètres pertinents de la méthode et un second test a été mené pour évaluer les performances de notre implémentation de la méthode de codage stéréo paramétrique. Grâce à l'approche paramétrique utilisée pour réaliser l'ACP tridimensionnelle, nous décrivons également le principe de l'extension de la méthode pour le codage paramétrique de signaux au format 5.0.

Mots clés : codage audio paramétrique, spatialisation sonore, représentation et segmentation temps-fréquence, Analyse en Composante Principale, matriçage adaptatif.

[1] Le terme stéréophonie (stéréo) désigne, dans l'ensemble du document, un signal audio à deux canaux.

ABSTRACT

This thesis deals with the extraction of relevant spatial cues for parametric coding of multichannel audio signals. The classical approach used for the parametric coding of stereo or multichannel signals (5.1, 7.1 audio signals, etc.) considers a parametric representation of the multichannel signal based on the auditory localization cues. Associated with a traditional audio coding of the input sum signal, the transmission of these parameters allows the reconstruction of a multichannel signal whose inter-channel cues approximate those of the original multichannel signal. These coding methods offer an effective solution for the data-rate constrained applications since the overall data rate is equivalent to the bit rate of the audio coder (mono or stereo for the compression of a 5.1 audio signal) plus the variable bit rate of the spatial cues. Nevertheless, the subjective audio and spatial quality of the reconstructed audio signal is strongly related to the parametric model which is not completely adapted to all multichannel audio signals.

The first axis of this thesis is looking further into this multichannel audio coding approach in order to improve its performances. Whereas the current parametric coding methods extract the spatial cues with a fixed time-frequency resolution, we suggest to adapt the extraction of the inter-channel parameters to the spectral contents of the audio signals. This approach is based on a time-frequency segmentation process in order to extract the spectral patterns carrying the spatial cues. However, we have not been able to obtain the desirable results based on our experiments which have given difficult interpretations of the signal regions resulting from the segmentation process. We also sought to measure the perceptual degradations introduced by parametric audio coding to refine the parametric representation of the critical time-frequency regions. Our experiments show that the error signal established by the difference between the reconstructed and the original signals (taking into account the instantaneous masking threshold) presents a too important energy distribution on the time-frequency plane to detect some time-frequenciy critical regions.

The second axis of our research relies on a multichannel audio model resulting from the instantaneous mixture of directional sources and ambiences in order to propose an alternative parametric coding method. The time and subband analysis of signals following our model has allowed us to describe the component hierarchisation within the eigenvalues. Then, we have evaluated the performances, in terms of energy concentration, of the Principal Component Analysis (PCA), carried out both in time and in frequency subbands, of two- and three- channel audio signals. To answer the main issue of multichannel compression, we have chosen a parametric approach to carry out the bi- and three-dimensional PCA. We propose a method to extract rotation angles used for the PCA and, a physical interpretation of these angles based on the knowledge of the reproduction sound system. Finally, we use this parametric decomposition of the covariance within a new parametric coding method which relies on the concentration of the dominant sources and the extraction of useful parameters for the audio signal reconstruction. A first subjective listening test has enabled us to define the relevant parameters for the coding method and the second one was carried out to evaluate the performances of our implementation of the parametric stereo coding method. Thanks to the parametric approach used to carry out the three-dimensional PCA, we also describe the extension of the method for parametric coding of 5.0 audio signals.

Keywords: parametric audio coding, upmix, spatial sound, time-frequency analysis and segmentation, Principal Component Analysis, adaptive matrixing.

REMERCIEMENTS

Ce document présente mes travaux de thèse réalisés à France Télécom R&D (site de Lannion) au laboratoire *Speech & Sound Technologies & Processing* (SSTP) et plus particulièrement au sein de l'équipe Traitement de la Parole et du Son (TPS). Ce travail est issu d'une collaboration avec le Laboratoire des Images et des Signaux (LIS) de l'Institut National Polytechnique de Grenoble (INPG) où j'ai passé six mois sur trois années de recherche en contrat CIFRE.

Je voudrais d'abord remercier Jean-Pierre Petit pour m'avoir admis au sein du laboratoire SSTP. Je remercie sincèrement Dominique Massaloux pour m'avoir accueilli dans l'équipe TPS et avoir pris le temps de suivre mes travaux en apportant, entre autres, son expertise dans le domaine de la compression du son.

L'initiative de cette thèse revient à David Virette et Rozenn Nicol qui ont su me faire confiance et me proposer un sujet d'étude passionnant mélant les technologies de la spatialisation sonore et de la compression du son. Vos qualités personnelles et professionnelles m'ont permis d'avoir un soutien permanent et de grande qualité. David, je tiens à te remercier sincèrement pour avoir dirigé mes travaux. Malgré le fait que tu représentes un élément central de l'équipe TPS et donc que tu sois très sollicité, tu as su te montrer disponible et mener diverses activités en parallèle pour la bonne conduite de mon projet de recherche. Je tiens à souligner le fait que tu as encadré ma thèse (et de multiples stages connexes effectués par Thierry Etamé Etamé, Guillaume LeGaffric et Benjamin Duval) tout en exerçant ton activité d'Ingénieur de Recherche à France Télécom et en reprenant tes propres travaux de thèse… chapeau !

J'exprime toute ma gratitude à Nadine Martin qui a accepté d'être ma Directrice de thèse sur un sujet d'étude qu'elle a su apprivoiser. Je te suis très reconnaissant d'avoir guidé et régulièrement évalué la pertinence de mes recherches. Je pense avoir beaucoup appris à tes côtés notamment pour ce qui est relatif à la gestion d'un projet de recherche. Merci également à Julien Huillery et Fabien Milloz avec qui les réflexions menées ont été très enrichissantes. Je remercie tous les membres du LIS de Grenoble qui ont rendu mon « séjour » professionnel très agréable grâce à un accueil très chaleureux.

Je voudrais aussi exprimer mes plus sincères remerciements à Jean-Marc Jot et Nicolas Moreau qui ont accepté de lire et de juger ces travaux. Je remercie aussi Gang Feng et Bruno Torrésani de leur présence dans le jury d'évaluation.

Un grand merci à toutes les personnes qui ont prodigué des conseils judicieux et à toutes celles qui ont participé activement à l'avancement de ces travaux. Je pense notamment à André Gilloire, Grégory Pallone, Stéphane Ragot, Jérôme Daniel, Alexandre Guérin, Marc Emerit et Arnault Nagle. Je n'oublie pas non plus Stéphanie Bertet, Pierre Guillon et Nicolas Chétry qui ont le courage de relire certaines parties de ce manuscrit dans un délai très court. Tous les résultats présentés dans ces travaux n'auraient pas lieu d'être sans la participation active des experts du laboratoire SSTP aux tests subjectifs d'évaluation de qualité sonore, merci à vous tous.

Je remercie mes parents, mes sœurs, ma famille et mes amis qui m'ont encouragé et fortement soutenu tout au long de ces trois années.

Merci aux joueurs de l'équipe corporative de l'Asptt tennis, aux planchistes de Trégastel, aux VTTistes du Trégor, aux *snowboarders* des 7Laux et aux « zikos » de Bédé-Gévezé pour m'avoir oxygéné l'esprit !

Enfin, un merci très spécial à celle qui m'a donné beaucoup depuis notre rencontre…

TABLE DES MATIÈRES

NOTATIONS

Symboles mathématiques

j	$\sqrt{-1}$
$*$	opération de convolution
$.^T$	transposition d'un vecteur ou d'une matrice
$\|.\|$	valeur absolue ou module d'une quantité complexe
$\angle .$	argument ou phase d'une quantité complexe
$\Re(.)$	partie réelle d'une quantité complexe
$\Im(.)$	partie imaginaire d'une quantité complexe
$\mathrm{E}[.]$	espérance mathématique
$\bar{.}$	quantité, vecteur ou matrice centré(e)
$\sigma^2_{.}$	variance d'un vecteur

Indices et dimensions des variables

n	indice temporel discret
l	indice des fenêtres d'analyse (portions glissantes de signal)
k	indice des fréquences
b	indice des sous-bandes de fréquences
m	indice des canaux (et des ambiances)
d	indice des sources directionnelles
N	longueur (nombre d'échantillons) d'une fenêtre d'analyse (portion glissante)
N_T	nombre d'échantillons discrets (total) du signal
N_C	nombre de coefficients d'une cellule pour l'analyse de la RTF
Z	nombre de zéros en complément du signal pour le calcul de la TF
M	nombre total de canaux d'un signal audio multicanal
D	nombre total de sources directionnelles
K_b	nombre total de sous-bandes
P_d	nombre de coefficients, d'une cellule $\Xi_{l,k}$, qui portent l'énergie du signal déterministe
P	nombre de pôles/zéros qui défini la réponse impulsionnelle d'un filtre en peigne, H^{cf}, ou d'un filtre passe-tout décorrélateur, H^{apl}.

Symboles particuliers, variables et signaux

f	fréquence
ω	$2\pi f$, pulsation
f_S	fréquence d'échantillonnage
T_0	période fondamentale
F_0	fréquence fondamentale
$x[n]$	signal aléatoire discret
$d_t[n]$	signal déterministe à temps discret
$b_g[n]$	bruit blanc gaussien à temps discret
$w[n]$	fenêtre d'analyse spectrale à temps discret
S	variable aléatoire (v.a.) Source sonore directionnelle
$s[n]$	réalisation de la source directionnelle S à l'instant n
$\mathbf{S_D}$	vecteur aléatoire de D v.a. Sources directionnelles de dimension $D \times N_T$
A	v.a. Ambiance sonore

$a[n]$	réalisation de l'ambiance A à l'instant n
$\mathbf{A_M}$	vecteur aléatoire de M v.a. Ambiance (sa réalisation est) de dimension $M \times N_T$
C	v.a. Canal (signal sonore)
$c[n]$	réalisation du canal (signal) C à l'instant n
$\mathbf{C_M}$	vecteur aléatoire de M v.a. Canal soit un signal multicanal de dimension $M \times N_T$
$\mathbf{D_M}$	vecteur aléatoire de M v.a. Canal Décorrélé (issu de l'ACP) soit un signal multicanal de dimension $M \times N_T$
$d_m[n]$	réalisation du canal décorrélé (signal) D_m à l'instant n
g	scalaire (gain réel)
$\mathbf{G_{MD}}$	matrice de gains réels de dimension $M \times D \times N_T$
$\mathrm{V_M}$	matrice des vecteurs propres de dimension $M \times M$
λ_i	$i^{\text{ème}}$ valeur propre (triée par ordre décroissant)
$\mathbf{\Lambda_M}$	matrice (diagonale) des valeurs propres de dimension $M \times M$
$\mathbf{\Gamma}_{\mathbf{C_M}}$	matrice de covariance du signal multicanal $\mathbf{C_M}$ de dimension $M \times M$
$F_x[l,k]$	TFCT du signal $x[n]$ à l'instant l et à la fréquence k
$S_{F_x}[l,k]$	spectrogramme du signal $x[n]$
$\chi^2(p_1,p_2,p_3)$	loi du *chi* deux de coefficient de proportionnalité p_1, à p_2 degrés de liberté et de paramètre de décentrage p_3
$\mathsf{N}\ (p_1,p_2)$	loi normale de moyenne p_1 et de variance p_2.
$\Xi_{l,k}$	cellule pour l'analyse de RTF centrée en l,k
p	proportion de coefficients (d'une cellule $\Xi_{l,k}$) suivant une loi du χ^2 décentré
p_{fa}	probabilité de fausse alarme
$r[l,k]$	rapport signal sur bruit local (sur la cellule)
$f_{S_{F_x}[l,k]}(x)$	loi de probabilité du coefficient de la RTF à l'instant l et à la fréquence k
$\overline{S}_{F_d}[l,k]$	moyenne des coefficients de la RTF de la composante déterministe du signal sur la cellule centrée en (l,k)
$M_1[l,k]$	moyenne des coefficients de la cellule de la RTF centrée en (l,k)
$M_2[l,k]$	moment non-centré d'ordre deux des coefficients de la cellule de la RTF centrée en (l,k)
Cl_i	$i^{\text{ème}}$ classe qui regroupe un ensemble de coefficients temps-fréquence
R_b	région de confinement du bruit
R_y	rayon du cercle ayant pour centre le point de l'EC au M_1 maximal
(a,b)	coefficient de la droite $(a.q+b)$ qui définit l'axe principal des données dans l'EC
τ_1	seuil de convergence
τ_2	nombre de fausses alertes
V	vraisemblance des coefficients de la classe Cl_0 avec une loi du χ^2 centré
B_{eq}	bande équivalente
α	angle de rotation estimé pour l'ACP bidimensionnelle
α,β,γ	angles d'Euler estimés pour l'ACP tridimensionnelle
$\mathbf{R_2}(\alpha)$	matrice de rotation de dimension 2×2
$\mathbf{R_3}(\alpha,\beta,\gamma)$	matrice de rotation d'Euler de dimension 3×3
(r,θ,δ)	coordonnées sphériques (distance, azimut, élévation) d'une source sonore dans l'espace par rapport à l'intersection du plan horizontal et du plan médian séparant la gauche et la droite d'un auditeur.

ACRONYMES

AAC	*Advanced Audio Coding*
ACP	Analyse en Composante Principale
ALS	Atténuation des Lobes Secondaires
BC	Bande Critique
BCC	*Binaural Cue Coding*
BRIR	*Binaural Room Impulse Response* ou réponse impulsionnelle binaurale de salle
C	*Center channel* ou canal central
CCR	*Comparison Category Rating*
DM	Diffusion du Masquage
DSP	Densité Spectrale de Puissance
EC	Espace des Caractéristiques
ERB	*Equivalent Rectangular Bandwidth*
FFT	*Fast Fourier Transform*, algorithme rapide pour réaliser la TFD
HRTF	*Head-Related Transfer Function* ou fonction de transfert liée à la tête
IACC	*Interaural Cross Correlation*
ICC	*Inter-Channel Coherence*
ICLD	*Inter-Channel Level Difference*
ICPD	*Inter-Channel Phase Difference*
ICTD	*Inter-Channel Time Difference*
ILD	*Interaural Level Difference*
ITD	*Interaural Time Difference*
kbps	Kilo-bit par seconde
KLT	Karhunen-Loève *Transform*
L	*Left channel* ou canal gauche
LFE	*Low Frequency Effect* (.1) ou canal basses fréquences
Ls	*Left surround* ou canal arrière gauche
MDCT	*Modified Dicrete Cosinus Transform*
MUSHRA	*MUlti Stimuli with Hidden Reference and Anchors*
MPEG	*Moving Picture Expert Group*
OLA	OverLap-Add
OPD	*Overall Phase Difference*
OTT	*One-To-Two*
PQMF	*Pseudo-Quadrature Mirror Filter*
PS	*Parametric Stereo*
R	*Right channel* ou canal droit
$RCPA_{12}$	Rapport de puissance de la Composante Principale (D_1) à l'Ambiance (D_2)
$RCPA_{13}$	Rapport de puissance de la Composante Principale (D_1) à l'Ambiance (D_3)
Rs	*Right surround* ou canal arrière droit
RSB	Rapport Signal sur Bruit
RSDA	Rapport de puissance des Sources Directionnelles aux Ambiances
RSM	Rapport Signal à Masque
RTF	Représentation Temps-Fréquence
SBR	*Spectral Band Replication*
SPL	*Sound Pressure Level*
TF	Transformée de Fourier
TFCT	Transformée de Fourier à Court Terme
TFD	Transformée de Fourier Discrète
TTT	*Two-To-Three*
UIT	Union International des Télécommunications, –T (Télécommunications) et –R (Radiocommunications)

Introduction

Contexte et enjeux

L'homme dispose d'une sensibilité auditive qui le renseigne sur l'espace environnant. Alors que la vision est limitée aux directions frontales, l'ouïe permet la détection, l'identification et la localisation de sources sonores quelle que soit leur direction de provenance. Néanmoins, la reproduction vidéo a longtemps été privilégiée au détriment de la qualité sonore restituée par les systèmes audiovisuels. Cependant, aujourd'hui, avec l'évolution des techniques de prise de son et des systèmes de reproduction sonore, l'intérêt voué au rendu sonore de haute-qualité audio et « spatiale » est en plein essor. Les salles de cinéma, la télévision haute-définition, les ordinateurs ou encore les téléphones portables et autres objets communicants disposent désormais de systèmes de reproduction sonore performants et adaptés à la restitution de véritables « scènes sonores ». Autrement dit, l'auditeur peut désormais avoir la sensation d'être immergé dans un espace sonore virtuel.

Pour pouvoir re-créer ces environnements sonores, les contenus audio à haute-qualité spatiale sont constitués de plusieurs signaux acoustiques ou canaux. Ces canaux sont d'ailleurs, en général, issus d'une prise de son multi-microphonique (en un ou plusieurs points de l'espace) éventuellement suivie d'un *post*-traitement ou mixage réalisé par un ingénieur du son. En particulier, les composantes de ces signaux audio sont complexes puisqu'elles sont relatives à la fois à la technique de prise de son employée, à la nature des sources sonores ainsi qu'à leurs interactions avec l'environnement dans lequel elles évoluent.

La quantité d'information nécessaire à la représentation d'une scène sonore par un signal audio dit « multicanal », par exemple au format 5.1 ou 7.1, est directement liée au nombre de canaux, témoin de la fidélité de la reconstruction spatiale. Autrement dit, plus le nombre de canaux est important et plus la représentation spatiale de la scène sonore originale est fidèle. Finalement, pour pouvoir disposer d'un « son multicanal » sans pour autant être obligatoirement au cinéma, où les possibilités de stockage sont compatibles avec une reproduction spatiale de haute-qualité, des méthodes de compression audionumérique sont nécessaires pour faciliter la transmission ou le stockage de tels contenus. Ainsi, les communications mobiles ou les applications à débit contraint (sur internet notamment) pourront également bénéficier d'un son à haute-qualité spatiale à partir d'une représentation numérique compacte.

Problématique, orientations et plan de la thèse

Le sujet de cette thèse concerne le problème de la compression de signaux audio multicanaux (stéréo, 5.1, etc.) au moyen de la description, sous une forme concise et efficace, de l'ensemble des informations spatiales contenues dans ce type de signaux. Ces informations spatiales sont relatives à la fois à la position des sources sonores et aux caractéristiques de leur environnement (effet de salle). Le concept de compression implique non seulement l'idée de l'extraction des paramètres pertinents pour la représentation des informations spatiales, mais aussi l'idée de la minimisation de la quantité des informations nécessaires à cette représentation (compression de données). Finalement, l'objectif de la recherche a été de concevoir des algorithmes d'extraction et de codage des informations de spatialisation sonore contenues dans les signaux multicanaux.

Le travail de cette thèse a tout d'abord consisté à analyser l'état de l'art du domaine couvrant à la fois les techniques de prise et restitution du son spatialisé ainsi que les méthodes de codage audio multicanal présentées respectivement aux chapitres 1 et 2. En particulier, une étude approfondie des standards du codage stéréophonique et multicanal a été menée tout en attachant une attention particulière aux récents travaux du groupe de normalisation MPEG[2] audio. En effet, au cours de l'année 2004, les acteurs de ce groupe de travail international ont orienté leurs recherches sur cette même problématique en se basant sur un procédé de codage innovant, le *Binaural Cue Coding* [FAL04].

Dans ce contexte, la première orientation de nos travaux de recherche, décrite au chapitre 3, s'est tournée vers l'amélioration des procédés existants en utilisant des techniques de traitement du signal et de l'image pour définir une représentation adaptée à la compression multicanale. En particulier, nous nous sommes appuyés sur un processus de segmentation dans le plan temps-fréquence pour repérer les motifs spectraux porteurs des informations de spatialisation sonore. L'originalité de cette approche étant de permettre l'extraction de paramètres spatiaux à partir d'une résolution temps-fréquence non plus fixe mais adaptée à la nature du signal.

La seconde orientation de nos travaux, décrite au chapitre 4, a consisté à établir un modèle de mélange adapté aux signaux audio multicanaux. Basé sur cette représentation du son multicanal, nous avons mené une analyse en temps et en sous-bandes de fréquences de signaux suivant ce modèle. La comparaison de ces analyses et une limite à la séparation des composantes du mélange sont donnés. Nous proposons plus particulièrement une analyse bidimensionnelle et tridimensionnelle avec une approche paramétrique pour la décomposition de la covariance de tels signaux. Finalement, nous mettons en œuvre une Analyse en Composante Principale (ACP) au moyen de matrice(s) de rotation paramétrée(s) par des angles directement liés aux azimuts des sources sonores (restitution dans le plan horizontal). Les performances de l'ACP réalisée en temps ou en sous-bandes de fréquences sont discutées d'un point de vue énergétique en mesurant la concentration de l'énergie obtenue par ce matriçage adaptatif.

A partir de notre modélisation et de notre analyse, nous décrivons, au chapitre 5, une nouvelle méthode de codage paramétrique basée sur l'ACP. Nous justifions, au moyen des résultats délivrés par un test subjectif, le choix des paramètres utiles à la méthode de codage. Les détails de l'implémentation de notre méthode de codage stéréophonique sont donnés et le principe de l'extension au codage de signaux au format 5.0 est présenté. D'après notre évaluation subjective des performances de cette méthode de codage paramétrique, elle permet, en étant associée à un codeur audio monophonique, la reconstruction de signaux stéréophoniques de qualité équivalente à celle délivrée par le standard actuel opérant à un débit global équivalent. Par conséquent, une discussion est ensuite menée de manière à confronter les différentes approches de codage paramétrique. Enfin, les perspectives énumérées constituent les évolutions possibles à la méthode actuelle.

[2] *Moving Picture Expert Group* : groupe de normalisation rattaché à l'Organisme International de Normalisation (ISO).

1. Perception et reproduction du son spatialisé

Avant d'aborder les principes du codage audio multicanal, nous nous devons d'introduire les phénomènes liés à notre perception de la spatialisation sonore. Ce chapitre présente d'une part, au paragraphe 1.1, les paramètres qui régissent notre perception auditive et plus particulièrement la faculté de notre système auditif à localiser les sons (dans le plan horizontal) d'abord en champ libre [3] puis en présence de réflexions. D'autre part, le paragraphe 1.2 introduit plusieurs systèmes de reproduction audio spatialisée ainsi que les techniques de prises de son et de mixage artificiel qui leurs sont adaptées.

1.1 Les paramètres de la localisation auditive

Dans notre expérience quotidienne, nous percevons un environnement sonore en trois dimensions en analysant le son parvenant à nos deux oreilles. Cette perception spatiale des sons complète les informations récoltées par nos autres sens. Elle a un rôle informatif puisqu'elle nous renseigne sur la position des sources dans cet espace sonore. Finalement, nous sommes capables de trouver la position d'une source dans l'espace en estimant sa direction et la distance qui la sépare de nos oreilles : c'est la localisation auditive. On distingue, d'après [BLA97], trois types de mécanismes (*cf.* **Figure 1.1**) :

- la localisation dans le plan horizontal qui permet d'estimer l'azimut, θ, de la source,
- la localisation dans le plan médian qui permet d'estimer la position de la source en élévation (angle δ),
- la localisation en distance pour évaluer la distance qui sépare la source de l'auditeur (r).

Notre étude vise à analyser les signaux audio multicanaux habituellement restitués par des haut-parleurs disposés dans le plan horizontal et à hauteur des oreilles de l'auditeur (*cf.* paragraphe 1.2.2). Dans le cas d'une écoute au casque, nous faisons l'hypothèse que les sources sonores sont perçues au niveau de l'axe interaural[4] (*cf.* **Figure 1.1**). Par conséquent, nous nous limitons à présenter les caractéristiques de la localisation auditive dans le plan horizontal.

[3] Environnement dépourvu de parois (objets, obstacles, murs d'une salle, etc.) réfléchissantes, difficilement rencontré en milieu naturel mais approché en pratique avec les chambres « sourdes » ou anéchoïques (le son n'est pas réfléchi, pas de phénomène d'écho).

[4] En réalité, avec une écoute au casque, nous pouvons percevoir les sons à une certaine hauteur intracrânienne (au-dessus de l'axe interaural).

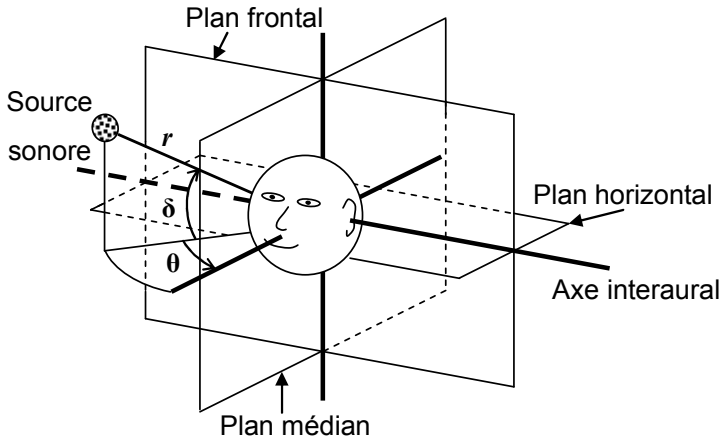

Figure 1.1 Localisation d'une source sonore dans l'espace à trois dimensions (intersection du plan horizontal, du plan frontal et du plan médian). La position de la source est repérée par ses coordonnées sphériques : son rayon r, son angle d'azimut θ et son angle d'élévation δ.

1.1.1 La localisation azimutale en champ libre

Cette section présente les mécanismes de la localisation auditive en champ libre en s'appuyant sur la *théorie duplex* introduite par Rayleigh dans [RAY07]. Nous nous intéressons, d'une part, à la localisation d'une source unique puis, d'autre part, à la localisation de multiples sources sonores en se basant sur deux phénomènes acoustiques : l'addition de la localisation et la superposition des évènements sonores.

1.1.1.1 Localisation d'une source

Pour localiser les sources sonores dans le plan horizontal, le système auditif utilise principalement les différences qu'il perçoit entre les signaux captés par les deux oreilles, c'est-à-dire les différences interaurales. Ces différences sont de deux types (*cf.* **Figure 1.2**) :

- la différence interaurale de temps[5] (*Interaural Time Difference* ITD) qui est engendrée par la différence de trajet acoustique des ondes qui vont exciter les tympans,
- la différence interaurale d'intensité (*Interaural Level Difference* ILD) qui provient essentiellement des phénomènes de diffraction et de masquage provoqués par la présence de la tête et qui n'est perceptible qu'aux hautes fréquences (à partir de 1,5kHz environ d'après [BLA97]).

D'après la théorie duplex [RAY07], les indices de localisation dans le plan horizontal dépendent de la fréquence : aux basses fréquences, l'azimut de la source est identifié sur la base des différences interaurales de temps tandis qu'aux hautes fréquences interviennent les différences interaurales d'intensité. En effet, les ondes basses fréquences dont la longueur d'onde est supérieure au diamètre de la tête ne présentent que peu d'ILD (pas d'effet de masquage par la tête) mais bien une ITD si la source sonore n'est pas située en face de l'auditeur. D'après les expérimentations de Blauert dans [BLA97], les indices interauraux estimés à partir de divers signaux réels prennent leurs valeurs dans l'intervalle [-1;1] ms pour l'ITD et [-30;30] dB pour l'ILD, valeurs à considérer comme des ordres de grandeur à titre indicatif.

[5] La différence de temps d'arrivée des ondes aux oreilles de l'auditeur peut également être considérée comme une différence de phase.

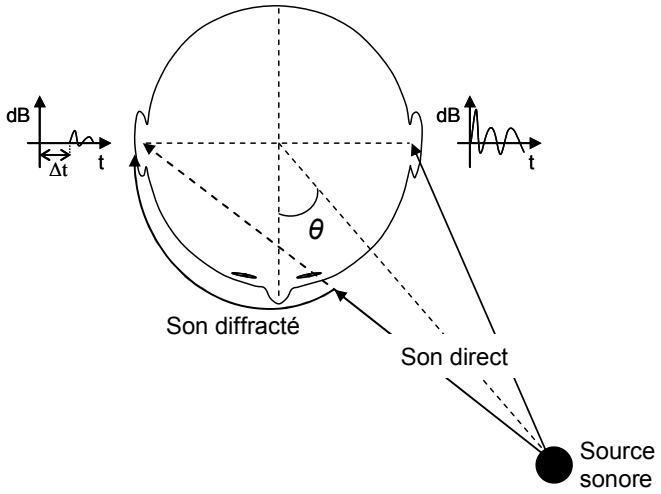

Figure 1.2 Les différences interaurales de temps et d'intensité résultent des différents trajets acoustiques des ondes qui atteignent les oreilles de l'auditeur.

Le déplacement d'une source sonore selon l'axe interaural dénoté *latéralisation* peut être paramétré par le couple d'indices ITD-ILD. La propagation des ondes sonores en champ libre est relative aux chemins acoustiques directs reliant la source à chaque oreille en omettant les trajets croisés (un seul signal atteint chaque oreille) et les réflexions du son émis. Par conséquent, une écoute au casque peut être considérée comme idéalement identique à une écoute en champ libre. En appliquant des différences de temps et d'intensité indépendamment à une même source sonore (signal acoustique) diffusée par les oreillettes du casque, la source peut être perçue à une position particulière sur l'axe interaural. Le schéma de principe d'une latéralisation, vers l'oreille droite, d'une source sonore s en champ libre est présenté à la **Figure 1.3** en considérant l'ITD et l'ILD définies par

$$\begin{cases} ITD = \Delta t & \text{ms} \\ ILD = 10 \times \log_{10}\left(\dfrac{g_2^2}{g_1^2}\right) & \text{dB} \end{cases}. \tag{1.1}$$

L'ITD est définie par le réel $\Delta t \in [0;1]$ ms et l'ILD par le logarithme du rapport d'énergie introduit entre les canaux gauche et droit par les gains réels définis tels que $g_1 < g_2$.

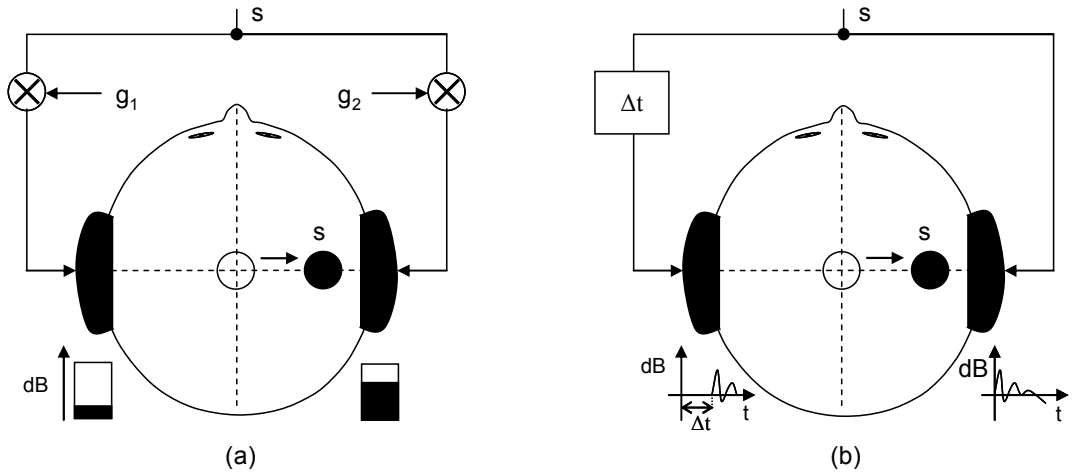

Figure 1.3 Latéralisation d'une source sonore s en contrôlant les différences interaurales **(a) -** d'intensité (avec $g_1<g_2$ tel que 0<ILD<30dB) et **(b) -** de temps (avec $\Delta t>0$ tel que 0<ITD<1ms).

Le couple (ITD=0, ILD=0) résulte en une source sonore perçue au milieu des oreilles de l'auditeur (θ=0°). Les expériences présentées par Blauert dans [BLA97] et plus tard complétées par celles de Cheng et Wakefield dans [CHE01] montrent que plus l'ILD est grand (et respectivement pour l'ITD) et plus la latéralisation est importante.

Il existe une théorie de la localisation auditive à la fois plus générale et plus complète que la théorie duplex basée sur les indices interauraux. Elle est basée sur des fonctions de transfert exprimant la propagation acoustique entre la source sonore située en un point donné de l'espace et les deux oreilles de l'auditeur. Il s'agit des fonctions de transfert liées à la tête (*Head Related Transfer Function*) HRTF [MOL92]. Ces fonctions de transfert modélisent l'ensemble des phénomènes qui vont affecter l'onde acoustique captée par le tympan et en particulier, elles rendent compte des phénomènes de diffraction par la tête, de réflexions sur le pavillon, le torse et les épaules de l'auditeur, ainsi que des temps d'arrivée de l'onde au niveau de chaque oreille. Les HRTF traduisent de manière exhaustive le codage acoustique de la position de la source sonore et contiennent donc toutes les informations dites « directionnelles » : les indices perceptifs ITD-ILD ainsi que les indices spectraux relatifs à notre perception dans le plan médian [BLA97].

Le paradigme de la localisation auditive est que le système auditif peut localiser des sons avec les informations provenant d'une seule oreille. Des expériences ont ainsi montré que les performances de localisation en élévation sont conservées même avec une seule oreille, et que des sujets sourds d'une oreille parviennent à localiser des sources en azimut dans certaines conditions. Cependant, une étude sur l'importance relative des indices interauraux et monauraux [JIN04] a montré que les réponses des sujets à un test de localisation sont principalement liées aux variations des indices interauraux et non aux variations des indices monauraux. Dans la suite du document, il ne sera considéré que l'importance perceptive des indices interauraux. Une approximation raisonnable consiste à affirmer que pour les directions frontales ($-90° \leq \theta \leq 90°$), la direction de la source repérée par l'angle θ (*cf.* **Figure 1.2**) détermine les valeurs des indices interauraux. Cette affirmation est valable sous l'hypothèse que le système auditif puisse pallier les confusions avant-arrière (par exemple par un mouvement de la tête de l'auditeur) : pour chaque direction frontale il existe une direction correspondante à l'arrière de l'auditeur qui résulte en un même couple d'ITD-ILD.

1.1.1.2 Localisation de plusieurs sources

En considérant une écoute stéréophonique sur haut-parleurs (*cf.* **Figure 1.4**), les signaux délivrés par les haut-parleurs peuvent être interprétés comme deux sources sonores réelles *a priori* distinctes. Cependant, un phénomène de sommation de la localisation auditive intervient lorsqu'un même signal acoustique (source sonore s) est délivré par les haut-parleurs : l'auditeur perçoit un évènement acoustique virtuel dont la position azimutale dépend de la position des deux sources réelles (haut-parleurs) et des indices interauraux relatifs à ces sources. Ces indices ont été définis au paragraphe 1.1.1.1 en considérant les signaux qui atteignent les oreilles de l'auditeur. Blauert, dans [BLA97], définit des indices inter-canaux relatifs aux signaux diffusés par les haut-parleurs. Ces indices inter-canaux sont considérés comme équivalents aux indices interauraux, relatifs aux signaux atteignant les oreilles, seulement dans le cas d'une écoute au casque. Littéralement, l'ITD est remplacée par la différence inter-canal de temps (*Inter Channel Time Difference*) ICTD et l'ILD par la différence inter-canal d'intensité (*Inter Channel Level Difference*) ICLD.

La **Figure 1.4** illustre la perception par sommation auditive de deux sources réelles *i.e.* haut-parleurs gauche et droit, relatives au même signal acoustique (source sonore s) avec différentes ICLD données par : $ICLD_i = 10 \times \log_{10}(g^2_{2i}/g^2_{1i})$ pour chaque situation d'indice i. La situation i=1 correspond à une $ICLD_1$ nulle ($g_{11}=g_{21}$) et la *source virtuelle* s_1 est perçue entre les deux sources réelles *i.e.* à l'intersection du plan horizontal et du plan médian en face de l'auditeur. La situation i=2 correspond à une $ICLD_2$ positive ($g_{12}<g_{22}$), l'intensité de la source réelle est plus grande à droite de l'auditeur et la source virtuelle s_2 se déplace dans cette direction. Enfin, la situation i=3 illustre le cas extrême où seule la source réelle gauche est active ($g_{23}=0$) et dont la position correspond alors à la position de la source virtuelle s_3

perçue. Finalement, ce phénomène de sommation de la localisation des sources réelles résulte en la localisation d'une source sonore virtuelle dont les indices interauraux (aux oreilles de l'auditeur) approximent les indices d'une source sonore qui serait physiquement présente entre les sources réelles *i.e.* haut-parleurs.

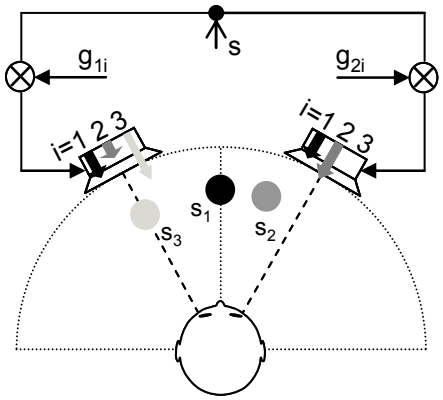

Figure 1.4 **Localisation d'une source virtuelle s_i entre deux sources réelles** (haut-parleurs), au moyen d'un même signal acoustique s[n], dont l'ICLD est contrôlée pour trois situations différentes indexées par i.

Si nous considérons maintenant que les signaux délivrés par les haut-parleurs *i.e.* les canaux, diffusent plusieurs sources sonores simultanément alors le même principe s'applique également. Nous considérons à la **Figure 1.5** des sources sonores indépendantes, par exemple, un signal de parole (s_1) et un instrument de musique (s_2).

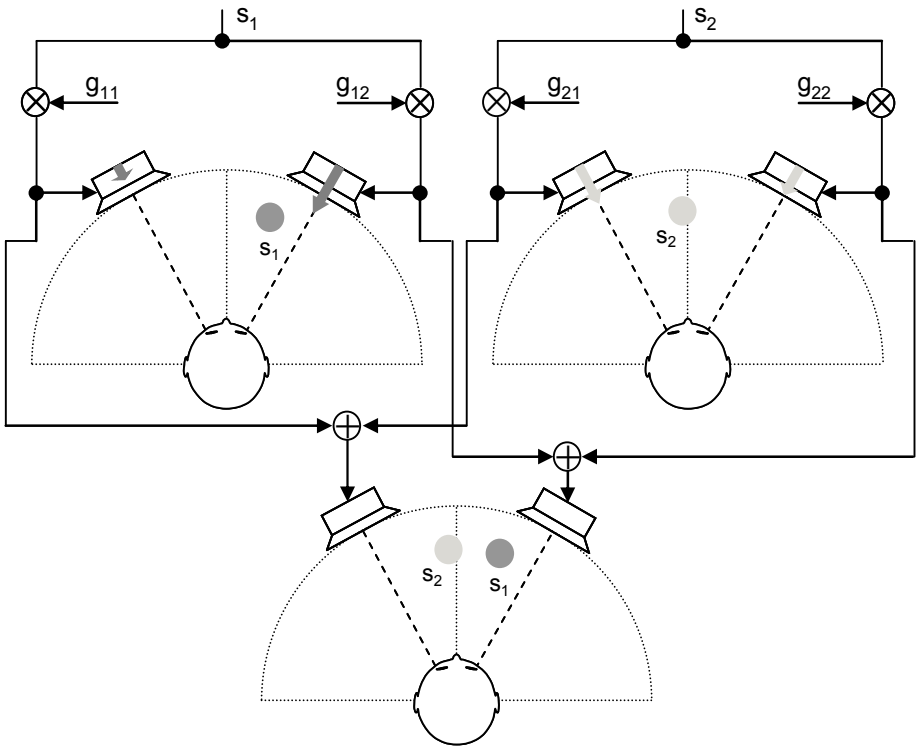

Figure 1.5 **Superposition de la localisation.** Un signal stéréo est synthétisé par la somme de deux signaux stéréo relatifs à un évènement sonore contrôlé par une ICLD fixée par un couple de gain (g_{i1}, g_{i2}). Deux sources virtuelles sont finalement localisées à des positions équivalentes à celles qui auraient été perçues à partir des signaux stéréo originaux.

Dans ce cas, le principe de superposition s'applique et le système auditif est capable de localiser les sources virtuelles relatives aux sources diffusées par les haut-parleurs. Bien que seule l'utilisation de l'ICLD soit présentée sur la **Figure 1.4** et la **Figure 1.5**, l'ICTD peut être utilisée de la même manière pour contrôler la position de la source virtuelle entre les deux sources. Cependant, il convient de limiter la valeur maximale de l'ICTD à 1ms pour éviter l'apparition d'un écho ou du phénomène d'antériorité décrit au paragraphe 1.1.2.2. L'utilisation conjointe des indices inter-canaux est également envisageable pour se rapprocher au maximum d'une situation réelle (écoute naturelle en champ libre); dans ce cas, la dépendance fréquentielle des indices interauraux doit être prise en compte par exemple, à la manière de la théorie duplex [RAY07].

Le même principe de superposition s'applique au cas de signaux diffusés au casque comme présenté par l'expérimentation de la **Figure 1.3**. Mis à part la différence de perception en termes de distance si l'on compare une écoute au casque à une écoute stéréo sur haut-parleurs, c'est grâce aux phénomènes d'addition de la localisation de signaux cohérents et de superposition des événements sonores que la compatibilité des signaux stéréo est rendue possible d'un casque à un système de haut-parleurs. C'est d'ailleurs à partir de ces phénomènes qu'ont été mises au point les techniques de mixage audio artificielles décrites au paragraphe 1.2.

1.1.2 La localisation azimutale en milieu réverbérant

Bien que les paramètres de localisation en champ libre donnent de solides bases à la compréhension des phénomènes acoustiques, cette configuration idéale est peu réaliste. De nombreux facteurs affectent une onde sonore avant qu'elle ne parvienne aux oreilles de l'auditeur. Les conditions climatiques peuvent avoir une influence sur la propagation des ondes en milieu naturel. De plus, les ondes sonores peuvent rencontrer sur leur chemin de propagation un certain nombre d'obstacles autres que le corps et les vêtements absorbants de l'auditeur. Lorsqu'une onde sonore rencontre un objet, l'objet lui-même peut absorber une partie de l'onde pendant que l'énergie résiduelle est réfléchie dans une autre direction. Rares sont les environnements naturels dits anéchoïques *i.e.* de type champ libre, exceptés de larges espaces ouverts tel le sommet d'une montagne.

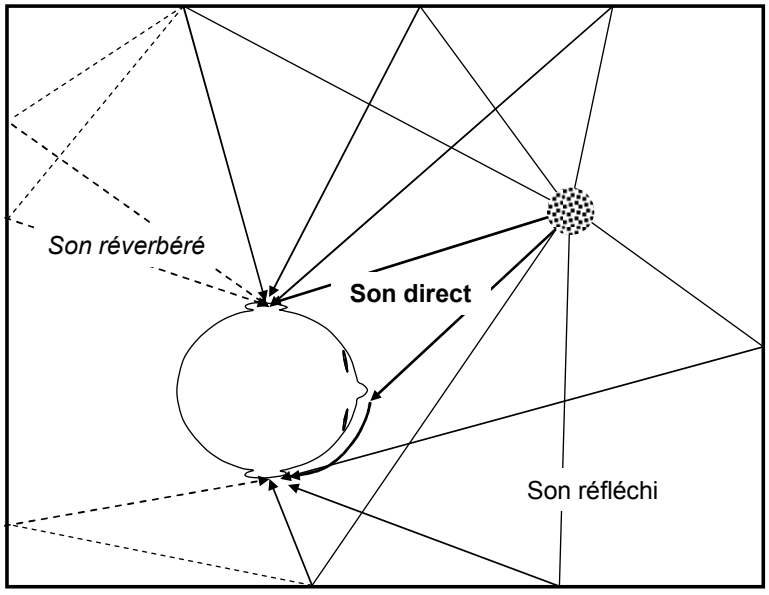

Figure 1.6 Son direct, réfléchi et réverbéré atteignant les oreilles de l'auditeur dans une salle. L'ordre d'arrivée des sons parvenant aux oreilles dépend du nombre de réflexions sur les parois de la salle. Le nombre d'ondes réfléchies peut atteindre plusieurs milliers.

Un environnement plus réaliste d'écoute est caractérisé par les parois d'une salle au revêtement plus ou moins absorbant, un sol réfléchissant, etc. Comme, le montre la **Figure 1.6**, l'auditeur perçoit d'abord le son direct, qui atteint les oreilles de l'auditeur sans obstruction, puis les ondes sonores qui se réfléchissent sur les parois de la salle et enfin le son réverbéré ou la partie diffuse du son qui s'est réfléchie plusieurs fois sur les murs de la salle avant d'atteindre les oreilles de l'auditeur.

La perception auditive dans une salle peut être caractérisée au moyen de la réponse impulsionnelle de la salle. Celle-ci dépend alors de la géométrie de la salle relativement à la position de l'auditeur ainsi que des matériaux constituant la salle qui vont déterminer le coefficient d'absorption de la salle et par suite, la réduction d'intensité du son réverbéré. Une réponse impulsionnelle de salle peut être décomposée en trois ou quatre sections temporelles (*cf.* **Figure 1.7**) [MOO79]-[JOT92]-[EME95]-[BAS03].

Le son direct ne dépend pas de la salle mais seulement des positions et caractéristiques de la source et de l'auditeur. Les réflexions primaires ou précoces correspondent à des versions retardées (20 à 50 ms) du signal émis, atteignant l'auditeur après une ou deux réflexions sur une paroi de la salle. Les réflexions secondaires correspondent à des versions du signal émis par la source retardé (70 à 100 ms) en ayant été réfléchies à au moins deux reprises par les parois de la salle. Les réflexions secondaires sont donc plus nombreuses que les réflexions primaires par unité de temps, elles sont difficilement distinguables et représentées par un bloc sur la **Figure 1.7**. Enfin, la réverbération tardive apparaît au moins 120 ms après le son direct et correspond à une accumulation de réflexions temporellement très denses, provenant de toutes les directions (champ diffus) et amorties par la salle (et l'atmosphère). Sa réponse est caractérisée par une décroissance exponentielle du niveau sonore (ou linéaire en dB) en fonction du temps.

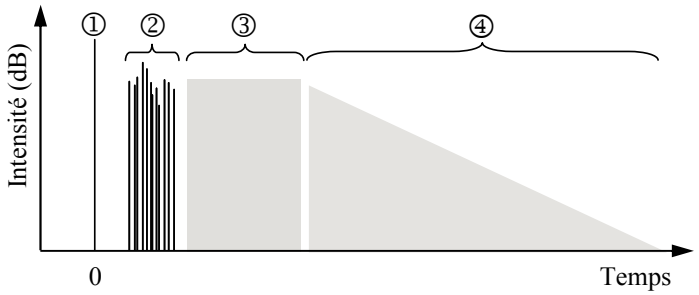

Figure 1.7 Réponse impulsionnelle de salle théorique. ① représente le son direct, ② représente les réflexions primaires, ③ représente les réflexions secondaires et ④ représente la réverbération tardive.

D'autres paramètres sont habituellement utilisés pour décrire les réponses impulsionnelles de salle comme le temps de réverbération, T_{60}, défini comme le temps à partir duquel le son est atténué de 60 dB. Ce paramètre peut être estimé à partir de connaissances géométriques propres à la salle comme proposé dans [GAR00]; plus la salle est réverbérante et plus le temps de réverbération est important. D'autre part, la distance de réverbération, définie au paragraphe 1.1.2.1, constitue un paramètre qui participe à la perception de la distance en milieu réverbérant.

La localisation des sources sonores devrait être rendue plus difficile dans les environnements réverbérants (champ diffus), cependant, le cerveau humain possède une capacité remarquable pour extraire l'information utile à la localisation même en présence de champ réfléchi. Le cerveau parvient à ne pas tenir compte de l'information de localisation auditive fournie par les réflexions des ondes sonores sur l'environnement. L'*effet d'antériorité*[6] est le phénomène lié à la capacité du système auditif à déterminer la direction

[6] L'effet d'antériorité est également connu sous le nom d'effet de précédence ou d'effet de Haas.

d'une source en présence de réflexions en favorisant la localisation du premier front d'onde par rapport aux réflexions qui lui succèdent [LIT99]. La direction perçue correspond alors à la direction du son direct éventuellement coloré par les premières réflexions *i.e.* le timbre du son direct s'en trouve altéré. L'effet d'antériorité n'apparaît toutefois que lorsque les signaux sont riches en transitoires. C'est le cas de toutes les percussions, des instruments à cordes et de la parole par exemple. D'autre part, l'effet d'antériorité se produit lorsque le décalage temporel entre les sources ne dépasse pas 5 ms pour les impulsions très brèves et 30 ou 40 ms pour des sons complexes ou des sons de parole [CAN00] (*cf.* **Figure 1.8**). Lorsque cette limite est dépassée, le signal retardé peut être perçu mais tout en localisant l'ensemble dans la direction du premier front d'onde *i.e.* la zone entre 30 et 40 ms contient le seuil d'apparition de l'écho. La séparation devient totale pour un retard de plus de 70 ms. La **Figure 1.8** illustre l'effet perçu lorsqu'un décalage temporel (Δt) de quelques millisecondes est introduit entre les signaux émis par deux haut-parleurs formant un système stéréophonique classique. Lorsque 0,6<Δt<40ms, le signal retardé peut provoquer un effet de coloration du timbre de la source virtuelle perçue au niveau des oreilles de l'auditeur. Cet effet dépend principalement de la nature de la source sonore et du nombre de réflexions.

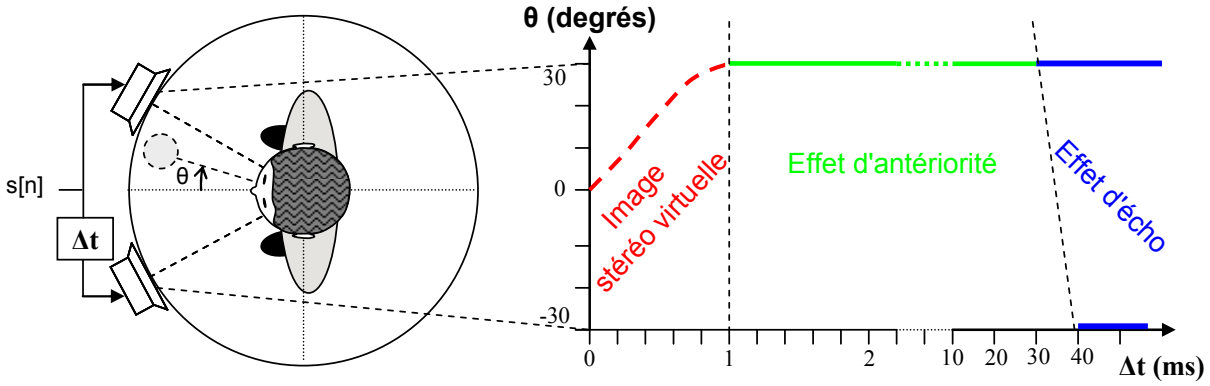

Figure 1.8 **L'effet d'antériorité -** tiré de [**THE01**]. Le décalage de temps introduit entre les signaux émis par les haut-parleurs (ICTD=Δt<1ms) procure à l'auditeur la perception d'un déplacement de la source virtuelle (4,4° pour 100 µs) repérée par l'angle θ qui évolue de 0° à 30°. Lorsque 1<Δt<40ms, l'effet d'antériorité est actif et la source sonore est perçue comme provenant du premier front d'onde *i.e.* à droite de l'auditeur (θ=30°). Lorsque Δt>40ms, l'effet d'écho permet à l'auditeur de distinguer deux sources sonores distinctes aux positions des deux haut-parleurs.

Outre la détection de la direction des sources dans le plan horizontal, notre système auditif parvient également à localiser les sources en distance sous certaines conditions et ceci d'autant mieux en présence de réflexions. Enfin, suivant la nature de l'environnement et donc de la cohérence des signaux parvenant à nos oreilles, la largeur apparente des sources devient variable et un « enveloppement sonore » peut être ressenti.

1.1.2.1 Localisation en distance

Il est difficile pour le système auditif d'identifier la distance d'une source sonore dans l'absolu. Les indices visuels ainsi que la connaissance du signal émis et de la source jouent un grand rôle. L'évaluation relative des distances est mieux maîtrisée. D'après Blauert [BLA97], trois indices interviennent principalement:

- le niveau sonore : plus il est fort, plus la source est perçue proche.

Les résultats issus de tests de localisation présentés dans [BRU99] démontrent la dépendance de l'ILD à la distance de la source. A l'écoute au casque, pour une source sonore perçue à un azimut θ=90° (au niveau de l'axe interaural), l'ILD augmente de l'ordre de 20 à 30 dB lorsque la distance de la source perçue diminue de 1 m à 12 cm.

- le rapport entre l'énergie du champ direct et l'énergie du champ réverbéré : un rapport faible donne une sensation d'éloignement.

L'effet de salle contribue à l'impression de profondeur/distance d'une source sonore et introduit la notion de plans sonores qui donnent du relief à la scène sonore. On peut considérer un auditeur comme situé à la distance de réverbération si l'intensité du son direct perçu est égale à l'intensité du son réverbéré. La distance de réverbération dépend finalement du volume de la salle et du temps de réverbération d'après [GAR00].

- le contenu spectral : puisque les hautes fréquences sont plus fortement atténuées par la propagation dans l'air, un son pourvu de composantes hautes fréquences renforce la sensation d'une source sonore proche.

On retiendra que compte tenu de sa difficulté d'évaluation en situation réelle la distance d'une source sonore est un paramètre délicat à simuler et à contrôler dans les champs sonores virtuels. En outre, il est fortement influencé par des indices non auditifs.

1.1.2.2 Largeur apparente de source et enveloppement sonore

Le phénomène subjectif qui correspond à la perception de la « largeur » apparente d'une source sonore a été notamment étudié pour la sonorisation des salles de concert [BER96]-[OKA98]. Les sons peuvent procurer une impression de volume au sens géométrique du terme.

Du point de vue de l'écoute binaurale *i.e.* au casque (*cf.* paragraphe 1.2.1), une autre caractéristique acoustique intervient dans l'impression spatiale : c'est la *cohérence interaurale* (*interaural cross-correlation* IACC) qui peut être définie, à partir des signaux atteignant les oreilles (les signaux binauraux) $o_g[n]$ et $o_d[n]$, comme le maximum de la valeur absolue de l'inter-corrélation normalisée telle que:

$$IACC = \max_{\Delta t} \frac{\left| \sum_{n=-\infty}^{n=\infty} o_g[n] \cdot o_d[n+\Delta t] \right|}{\sqrt{\sum_{n=-\infty}^{n=\infty} o_g^2[n] \cdot \sum_{n=-\infty}^{n=\infty} o_d^2[n+\Delta t]}}, \tag{1.2}$$

où la valeur de Δt, qui correspond au maximum de l'IACC, exprime l'ITD entre les signaux binauraux. L'ITF prend ses valeurs dans l'intervalle [-1;1] ms et l'IACC entre 0 et 1; IACC=1 signifie que les signaux sont cohérents (corrélés) à une ITD et une ILD près et IACC=0 signifie que les signaux sont indépendants. L'IACC peut également être définie sans considérer la valeur absolue de l'inter-corrélation, dans ce cas, IACC=-1 signifie que les signaux sont en opposition de phase.

La **Figure 1.9** présente les fonctions IACC estimées en sous-bandes de fréquences suivant l'échelle *Equivalent Rectangular Bandwidth* ERB, définie en Annexe B.1.2, à partir des signaux atteignant les oreilles de l'auditeur lorsque la source virtuelle est perçue à une position azimutale θ=30° d'après [PUL98]. On visualise sur la **Figure 1.9** que la valeur maximale de chaque fonction IACC en sous-bandes (adéquation avec l'équation (1.2) pour une estimation pleine bande) détermine une valeur significative de l'ITD ($\tau=\Delta t$ en ms) jusqu'à environ 1,5 kHz.

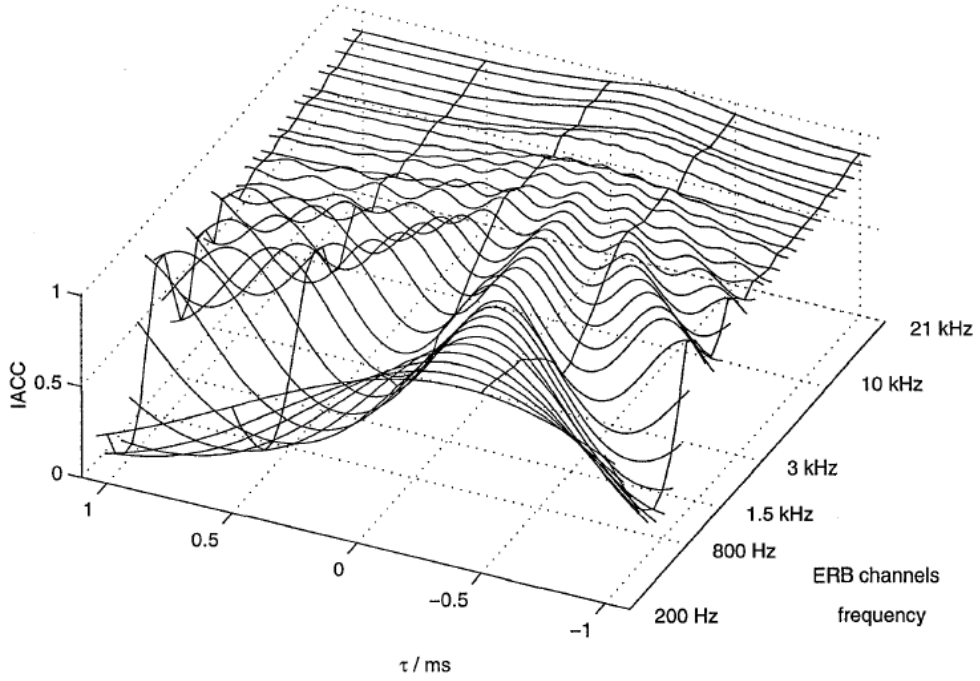

Figure 1.9 IACC estimées en sous-bandes de fréquences - tiré de [PUL98]. IACC estimée pour une source réelle à une position azimutale θ=30°.

La **Figure 1.10** présente certains résultats des expérimentations menées par Blauert dans [BLA97] pour mesurer la largeur apparente de la source perçue pouvant aller jusqu'à l'enveloppement de l'auditeur dans le cas d'une restitution multicanale. Lorsque deux signaux acoustiques (bruit blanc) identiques (IACC=1) sont diffusés par les oreillettes d'un casque ou par une paire de haut-parleurs (*cf.* **Figure 1.10-(a)-(b)**), la largeur apparente de la source est réduite et se confine le long de l'axe médian. Par contre lorsque la cohérence des signaux décroît, la largeur de la source apparente s'élargit pouvant aller jusqu'à la séparation en deux sources sonores réelles distinctes lorsque l'IACC=0 (cas non représenté sur la **Figure 1.10**). Le même principe s'applique au cas d'une restitution multicanale sur haut-parleurs (*cf.* **Figure 1.10-(c)**) : lorsque l'IACC entre les signaux émis par les haut-parleurs décroît alors l'enveloppement sonore autour de l'auditeur augmente.

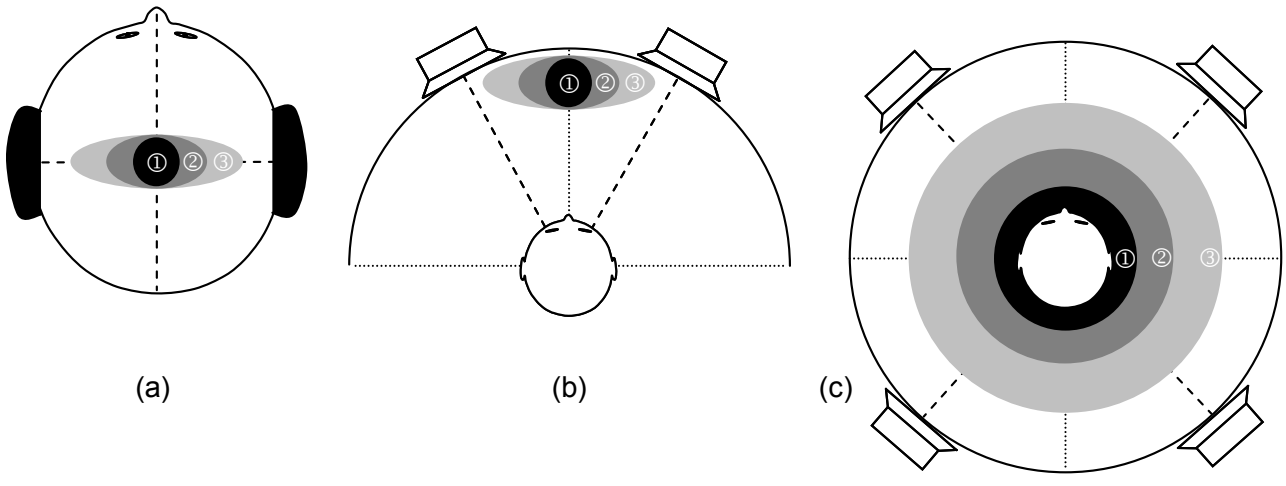

Figure 1.10 [Largeur apparente d'une source sonore ou enveloppement sonore avec une restitution : **(a)** - au casque, **(b)** - stéréophonique sur haut-parleurs, **(c)** – multi-haut-parleurs de type quadriphonie (haut-parleurs équi-répartis autour de l'auditeur)**]**. De la situation ① à ③, l'IACC entre les signaux diffusés voit sa valeur décroître.

Par souci de clarté nous avons considéré le même paramètre IACC pour les trois situations (**a**, **b** et **c**) de la **Figure 1.10**. Cependant, l'IACC ou cohérence interaurale est définie à partir des signaux atteignant les oreilles de l'auditeur (*cf.* équation (1.2)). Pour les signaux délivrés par les haut-parleurs (situations **b** et **c** de la **Figure 1.10**), nous parlerons désormais de *cohérence inter-canal* ICC.

En acoustique des salles, le même effet de volume se produit : l'impression d'espace plus ou moins large est assurée par le degré de cohérence entre le son direct et les réflexions provenant des parois. Ainsi, plus la salle est réverbérante et plus l'IACC a une valeur faible qui peut renseigner sur la largeur apparente de la source ou sur la sensation d'enveloppement de l'auditeur. Notons d'ailleurs que dans le contexte de la conception des salles assistée par ordinateur, il est possible de simuler l'écoute dans une salle (sensation de volume) dont la géométrie est connue (avant sa construction) en combinant les fonctions de transfert liées à la tête de l'auditeur, les HRTF, et la réponse impulsionnelle de la salle [EME95].

1.2 Reproduction sonore spatialisée dans le plan horizontal

Les techniques d'enregistrement et de reproduction audio spatialisées évoluent depuis plus d'un siècle et suscitent toujours autant d'intérêt. La reproduction d'un contenu audio naturel ou artificiel a pour vocation de reproduire à la fois le positionnement et les mouvements des sources sonores par rapport à l'auditeur mais aussi l'environnement acoustique qui l'entoure *i.e.* l'effet de salle.

Cette section présente différents systèmes de reproduction audio spatialisée dans le plan horizontal à hauteur de l'axe interaural. Les principes de la stéréophonie allant de la prise de son naturelle à la spatialisation artificielle sont abordés, au paragraphe 1.2.1, puis comparés à la prise et restitution binaurale. Le paragraphe 1.2.2 présente ensuite l'extension de ces techniques pour la reproduction audio multicanale.

1.2.1 De la stéréophonie au binaural

1.2.1.1 A partir d'un enregistrement naturel

Utilisant les indices acoustiques de localisation dans le plan horizontal que sont les différences de temps et de niveau entre les deux oreilles d'un auditeur (ITD et ILD), la stéréophonie constitue la première technique restituant des effets de spatialisation. Le concept de source virtuelle est né avec la stéréophonie (*cf.* paragraphe 1.1.1). L'effet droite-gauche ou de latéralisation (*cf.* paragraphe 1.1.1.1) est rendu de manière optimale si les positions des deux haut-parleurs et de l'auditeur forment un triangle équilatéral, soit un écartement de 60° entre les haut-parleurs. Les informations de retard et de gain peuvent être captées directement par un couple de microphones (*cf.* [MER93] pour les techniques de prise de son):

- *stéréophonie de temps* avec un couple de microphones omnidirectionnels non coïncidents *i.e.* écartés de quelques centimètres,
- *stéréophonie d'intensité* avec un couple de microphones coïncidents unidirectifs tels que : XY (*cf.* **Figure 1.11**), stéréosonic ou MS (moyennant un décodage),
- *stéréophonie mixte* avec un couple AB (microphones unidirectifs non coïncidents).

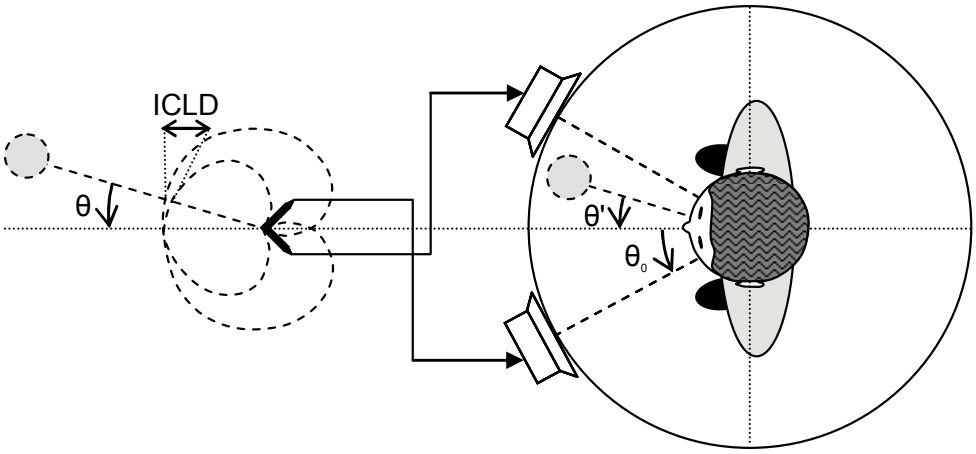

Figure 1.11 Principe d'une prise de son stéréophonique avec une paire de microphones cardioïdes coïncidents de type XY. A la restitution, la différence d'intensité captée par les microphones (ICLD) génère une source virtuelle placée entre les sources réelles *i.e.* les haut-parleurs espacés de $2\times\theta_0=60°$, à l'azimut $\theta'\sim\theta$.

Lorsqu'on évoque une restitution sonore spatialisée dans le plan horizontal, cette restitution n'est qu'incomplète étant donné que la source virtuelle ne peut se mouvoir qu'au sein de la portion d'espace comprise entre les deux haut-parleurs. En outre, même si une sensation de profondeur est présente et qu'il est possible d'identifier plusieurs plans sonores en fonction de la distance (notamment en présence de réverbération), il faut reconnaître que ce paramètre est mal contrôlé dans les systèmes stéréophoniques.

La perception associée à une restitution stéréophonique sur haut-parleurs dépend du phénomène de sommation de la localisation présentée au paragraphe 1.1.1.1. Les expériences de prises de son menées par Blumlein en 1931 [BLU31] ont démontré qu'en introduisant des différences d'intensité entre les signaux émis par les haut-parleurs (ICLD) *i.e.* prise de son coïncidente, des différences interaurales de phase (ITD) sont détectées au niveau des oreilles comme lors d'une écoute naturelle. Les techniques de prise de son non-coïncidentes se basent sur le même phénomène de sommation de la localisation à partir de signaux caractérisés par leur ICTD. Les techniques de la stéréophonie mixte permettent de contrôler l'azimut de la source virtuelle au moyen du couple ICLD-ICTD. Lorsque plusieurs sources sont enregistrées simultanément, le même principe s'applique et résulte en la restitution de plusieurs sources virtuelles (*cf.* principe de superposition présenté au paragraphe 1.1.1.2).

Les techniques de prise de son coïncidentes sont les plus communément employées pour la stéréophonie mais leur diversité est très importante et souvent caractéristique du savoir-faire de l'ingénieur du son. L'avantage d'une prise de son coïncidente est d'assurer une compatibilité avec les systèmes monophoniques par simple sommation de la paire de signaux sans provoquer de pertes d'information et d'amplification comme il peut arriver avec une paire de signaux déphasés ($ICTD \neq 0$) issus d'une prise de son non-coïncidente.

Les techniques binaurales, décrites notamment dans [MOL92], reposent sur une modélisation à la fois plus globale et plus rigoureuse que celle employée par la stéréophonie. Elles se proposent de reproduire le champ sonore induit au niveau des oreilles de l'auditeur. Ainsi, dans le champ restitué sont présents non seulement les effets de la propagation entre la source et l'auditeur (atténuation, retard, effet de salle, etc.) mais aussi l'ensemble des phénomènes engendrés par le corps de l'auditeur tels que la diffraction par la tête, les réflexions sur le haut du corps et le pavillon de l'oreille externe. En visant la reproduction fidèle du champ sonore excitant le tympan, les techniques binaurales restituent l'ensemble des indices qu'utilise l'appareil auditif pour interpréter le champ sonore : les différences interaurales pour la localisation horizontale et les indices spectraux (monoraux) pour la localisation verticale. Le champ sonore est alors spatialisé dans les trois dimensions de l'espace c'est à dire que les sources sonores virtuelles peuvent être perçues comme provenant de n'importe quelle direction.

A la différence de l'écoute stéréophonique au casque, la restitution binaurale a l'avantage d'externaliser le champ sonore perçu (à l'extérieur du crâne) comme le présente la **Figure 1.12**. L'écoute des signaux binauraux enregistrés auparavant dans ses propres oreilles permet de revivre l'expérience auditive passée et cela de façon très réaliste puisque les particularités environnementales et morphologiques propres à l'auditeur ont été prises en compte lors de l'enregistrement.

Pour des raisons pratiques, les enregistrements binauraux sont souvent réalisés au moyen d'une tête artificielle qui peut aller jusqu'à reproduire la moitié supérieure d'un corps humain. Cependant, une tête artificielle ne rend compte que d'une morphologie moyenne qui n'est exacte pour aucun individu. Les enregistrements sur tête artificielle sont susceptibles d'introduire dans le champ perçu des inversions avant/arrière *i.e.* une source censée être située devant l'auditeur est perçue derrière lui, ou des phénomènes de localisation intracrânienne.

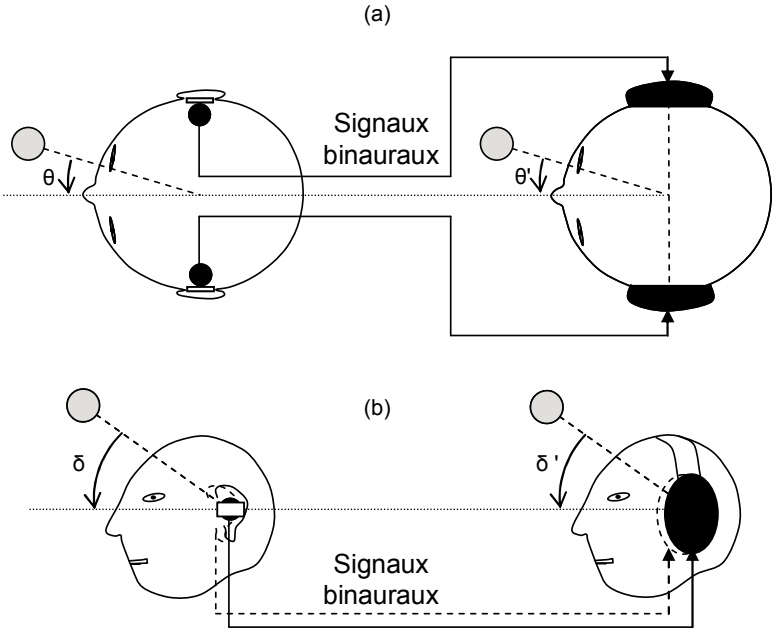

Figure 1.12 Prise et restitution binaurale (en champ libre). Prise de son avec une paire de microphones miniatures positionnés dans les conduits auditifs. La restitution est réalisée au casque **(a)** - Vue de dessus: la spatialisation en azimut est préservée $\theta'=\theta$. **(b)** - Vue de côté: la spatialisation en élévation est préservée $\delta'=\delta$.

Une des raisons principales qui freine le développement du « son binaural » sur le marché audiophile est liée à la dépendance du format à la morphologie de l'auditeur. Les avancées en matière d'individualisation des fonctions de transfert liées à la tête (HRTF), dans [PER03], ou des indices acoustiques de localisation, dans [BUS06], pourraient cependant permettre de réaliser une synthèse binaurale artificielle (*cf.* paragraphe 1.2.1.2).

D'autre part, la diffusion d'enregistrements binauraux sur un système de haut-parleurs stéréophonique ne permet pas une perception acoustique réaliste. Dans cette configuration, on remarque que chaque oreille perçoit à la fois le signal émis par le haut-parleur gauche et celui émis par le haut-parleur droit (trajets croisés). Il est pourtant possible de restituer un enregistrement binaural sur deux haut-parleurs en compensant la propagation des trajets croisés entre les haut-parleurs et les oreilles de l'auditeur [GAR97] de telle manière que les signaux atteignant les oreilles approximent les signaux binauraux. Cette compensation est réalisée en traitant les signaux binauraux par des filtres dont les réponses inversent les différentes fonctions de transfert reliant chaque haut-parleur à chaque oreille (voir [GAR97] pour plus de détails). La **Figure 1.13** décrit un tel dispositif avec son circuit de correction.

Une solution particulière est proposée par le système *stereo dipole* dont la particularité réside dans le positionnement très rapproché des haut-parleurs toujours face à l'auditeur suivant un axe parallèle à l'axe interaural. De tels systèmes de restitution audio spatialisée dépassent les possibilités offertes par la stéréophonie puisque les sources sonores peuvent être perçues en dehors du triangle formé par les haut-parleurs et l'auditeur [KIR97].

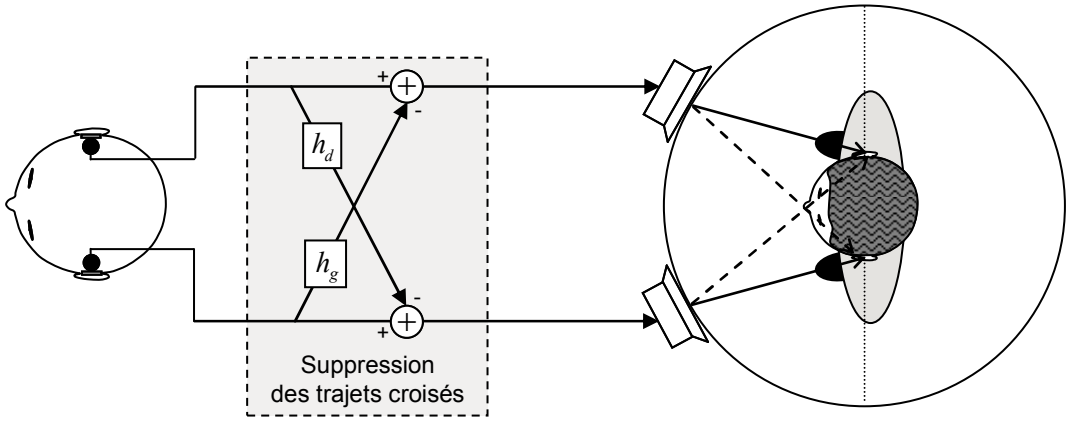

Figure 1.13 Dispositif de prise et restitution binaurale sur haut-parleurs. Les signaux binauraux sont traités (filtre gauche h_g et filtre droit h_d) de manière à compenser les trajets croisés entre les haut-parleurs et l'auditeur.

1.2.1.2 A partir d'un traitement artificiel

Nous avons présenté, au paragraphe 1.2.1.1, la génération des signaux stéréophoniques obtenus à partir d'une prise de son par un couple stéréophonique. L'effet de spatialisation est réalisé de façon acoustique et les différences de temps et/ou d'intensité sont introduites par le jeu du positionnement et des directivités des microphones.

Nous avons également introduit, au paragraphe 1.1.1.2, le principe de la génération artificielle d'un signal stéréo par mélange ou mixage de plusieurs sources sonores indépendantes. Ce principe, qui vise à introduire une différence d'intensité entre les signaux émis par les haut-parleurs (ICLD), est nommé *stéréophonie dirigée* ou plus communément *panoramique d'intensité* (*panpot*).

La **Figure 1.14** illustre ce procédé pour une source virtuelle évoluant du haut-parleur gauche au haut-parleur droit écartés de θ_0=30°. La relation entre l'ICLD et l'angle θ de la source virtuelle perçue est donnée.

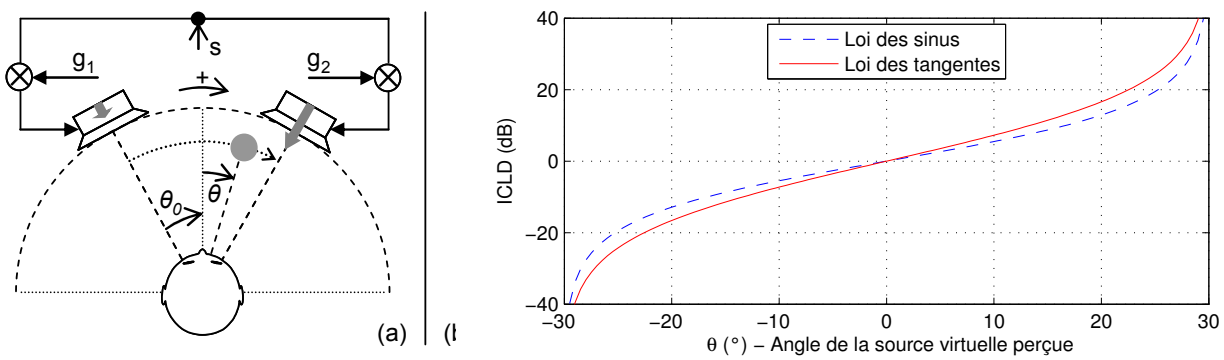

Figure 1.14 Panoramique d'intensité. (a) - Schéma de principe du procédé. **(b)** - Evolution de l'ICLD (dB) en fonction de l'angle de la source virtuelle perçue avec la loi des sinus et la loi des tangentes.

La direction de la source sonore virtuelle perçue entre les haut-parleurs peut être approximée par la loi des sinus, d'après les travaux de Blumlein [BLU31], telle que :

$$\frac{\sin\theta}{\sin\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{1.3}$$

où $0° \leq \theta_0 \leq 90°$ correspond à l'angle entre l'axe frontal (perpendiculaire à l'axe interaural) et l'un des haut-parleurs symétriques par rapport à cet axe, θ correspond à l'angle entre l'axe frontal et la direction de la source virtuelle perçue. g_1 et g_2 sont les gains qui fixent la valeur de l'ICLD entre les signaux définis par $ICLD = 10 \times \log_{10}(g^2_2/g^2_1)$. L'équation (1.3) permet de simuler approximativement une prise de son coïncidente (stéréophonie d'intensité) mais n'est valide que lorsque la tête de l'auditeur est orientée selon l'axe frontal. D'après Pulkki [PUL01a], lorsque l'orientation de la tête suit la position perçue de la source virtuelle, la loi des tangentes, définie par :

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{1.4}$$

constitue une meilleure approximation. Le comportement des lois sinus et tangente diffère d'autant plus que l'écartement entre les haut-parleurs augmente. Cependant, dans le cas d'un écartement des haut-parleurs de 60° cette différence peut être négligée [PUL97]. La loi des tangentes peut également être formulée sous forme vectorielle désignée sous le nom de VBAP (*Vector Base Amplitude Panning*) [PUL97]. Cette formulation vectorielle permet l'utilisation de la loi des tangentes pour n'importe quelle configuration de haut-parleurs en deux et trois dimensions. Pour cela, les gains de panoramique sont exprimés en termes de coordonnées de la source virtuelle dans la base établie par le triplet de haut-parleurs (les plus proches de la source virtuelle dans l'espace) formant un triangle du point de vue de l'auditeur. Suivant le même principe que la méthode VBAP mais en raisonnant en termes d'énergie, une autre approche a été dérivée pour les hautes fréquences sous le nom de VBIP (*Vector Base Intensity Panning*) [PER98]. Pour restituer l'ensemble du spectre, il conviendrait donc de coupler les deux approches VBAP et VBIP.

La stéréophonie artificielle obtenue par *panoramique de temps* (*cf.* **Figure 1.8**) pour simuler une prise de son stéréophonique non-coïncidente reste assez rare. La raison principale est la trop forte dépendance de cette technique de panoramique à la fréquence des signaux pour assurer la stabilité des sources virtuelles. Force est de constater que les différences de temps et/ou d'intensité restituées par la stéréophonie ne rendent compte de la perception spatiale d'une scène sonore que de façon incomplète. Mis à part le fait que la stéréophonie se limite à une restitution dans le plan horizontal, la captation ou la génération artificielle de l'effet de salle est indispensable pour restituer une perception réaliste qui ne se limite pas aux sources sonores. Pour combler cet écart à la réalité acoustique, l'ingénieur du son dispose de multiples techniques et post-traitements pour générer un effet de salle par exemple par un filtrage de type réverbération (*cf.* **Figure 1.7**).

Les techniques binaurales, elles aussi, ne se limitent pas à une prise et restitution du son décrites à la **Figure 1.12**; un traitement approprié peut permettre de créer artificiellement un champ sonore : c'est la synthèse binaurale. Pour cela, les HRTF ou les réponses impulsionnelles binaurales d'une salle (*Binaural Room Impulse Response* BRIR) sont utilisées pour filtrer chaque source sonore de façon à les positionner dans l'espace. Puisqu'un couple de HRTF ou BRIR (gauche et droite) doit être mesuré ou calculé pour chaque position de l'espace (avec un pas de quelques degrés en azimut et en élévation), il apparaît aussitôt que la synthèse binaurale en temps réel n'est pas une chose évidente à mettre au point. Cependant, nombres de recherches, notamment dans [BUS06], sont actuellement orientées vers une simplification des filtres (fonctions de transferts) HRTF ou BRIR pour limiter le nombre de positions spatiales et de coefficients spectraux perceptuellement significatifs pour finalement réduire la complexité des algorithmes de synthèse binaurale.

1.2.2 Reproduction multi-haut-parleurs

1.2.2.1 De la stéréophonie au 5.1

Avec l'apparition du DVD-vidéo (*Digital Versatile Disc*), la diffusion des contenus audio multicanaux s'est largement répandue. La stéréophonie jusqu'alors diffusée au moyen du CD audio trouve là un successeur qui se base sur les contenus audio habituellement réservés au cinéma.

Les signaux audio multicanaux sont diffusés au cinéma par l'intermédiaire d'un système multi-haut-parleurs qui entoure le public avec l'objectif de couvrir une large zone d'écoute et de donner aux auditeurs la sensation d'être plongé « dans la scène sonore » (*cf.* **Figure 1.15-(a)**). Au milieu des années 70, le laboratoire Dolby a introduit l'ajout de deux pistes audio supplémentaires sur les pellicules 35 mm. Grâce à un encodeur matriciel basé sur la phase des signaux (*cf.* paragraphe 2.3.1.1), quatre pistes audio peuvent alors être diffusées après décodage : le canal gauche, le canal central, le canal droit et un canal à l'arrière ou *surround* qui restitue une ambiance sonore au moyen de plusieurs haut-parleurs (*cf.* **Figure 1.15-(a)**). Précisons que suivant le type de décodeur employé, un ou deux canaux *surround* peuvent être générés à partir du matriçage *Dolby Stereo*. Cette technologie connu un large succès car les salles ont pu s'équiper à moindre coût, le système étant en outre compatible à un rendu monophonique, stéréophonique sur deux ou quatre canaux (*cf.* paragraphes 2.3.1.2 et 2.3.1.3). Aujourd'hui, les pellicules cinématographiques possèdent à la fois une piste audio numérique compatible avec le standard *Dolby Digital* ou AC-3 (*cf.* paragraphe 2.3.2.1) et deux pistes analogiques *Dolby Stereo* compatibles avec les décodeurs *Dolby Surround* et *Pro Logic* (*cf.* paragraphe 2.3.1).

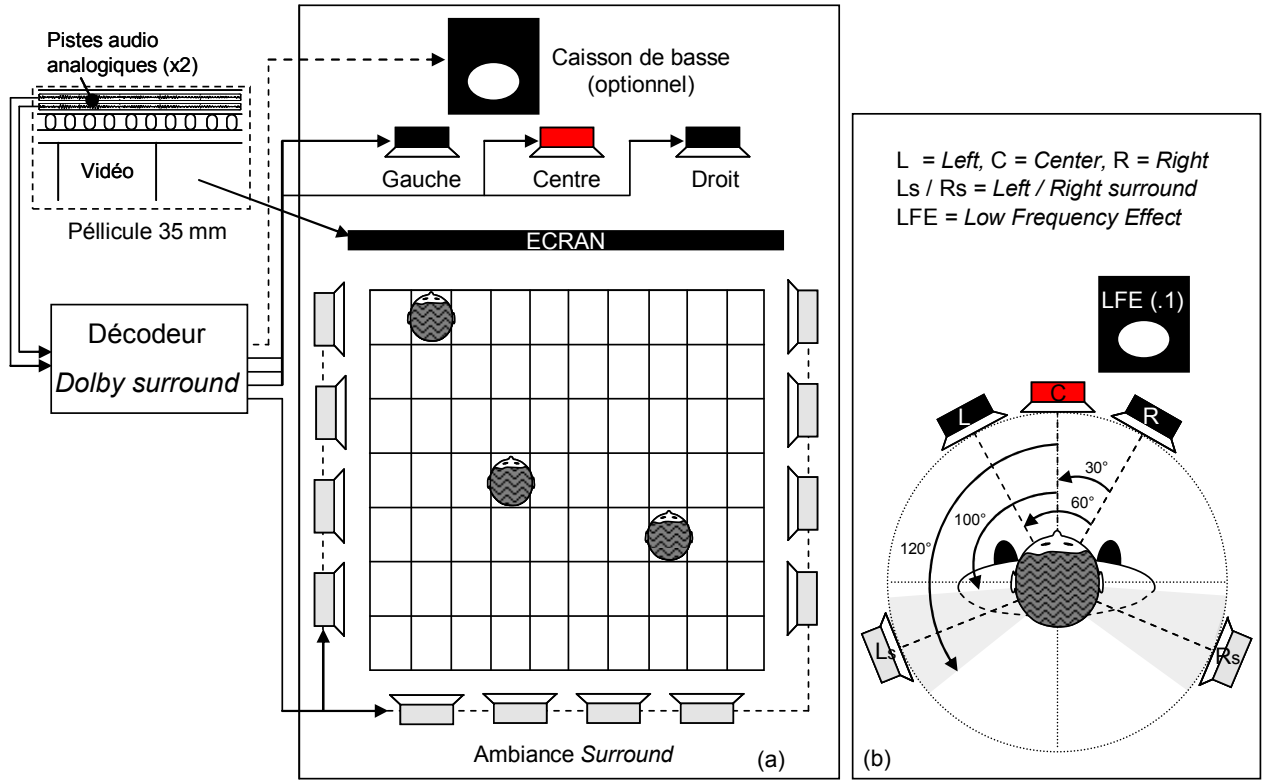

Figure 1.15 Dispositif de restitution sonore spatialisée dans le plan horizontal [(a) - au cinéma (4 canaux) le canal *surround* est diffusé par plusieurs haut-parleurs, **(b) -** de signaux au format 5.1 (6 canaux discrets)**]**.

L'adaptation des contenus audio cinématographiques pour le grand public a été rendue possible par l'intermédiaire des systèmes de restitution 5.1 ou *home theater* (*cf.* **Figure 1.15-(b)**). Le DVD-vidéo a été le premier support à permettre le stockage de 5 à 6 canaux discrets adapté pour la restitution sur un système 5.1. Le « .1 » ou canal basse fréquence est dédié à la diffusion des effets spéciaux, présents dans les contenus cinématographiques (explosion, tremblement de terre, etc.), qui ne peuvent être retransmis par les autres haut-parleurs; sa bande passante se limite généralement à 120 Hz alors que les autres canaux disposent d'une bande passante pouvant aller jusqu'à 24000 Hz. La **Figure 1.15-(b)** présente la disposition des haut-parleurs d'un système 5.1 suivant la recommandation de l'UIT [UIT775] également décrite dans [AES01]. Cette configuration des haut-parleurs assure une compatibilité avec les systèmes de restitution stéréophonique (*cf.* Annexe C.1 pour plus de détails sur les opérations de conversion audio) puisque les haut-parleurs gauche et droit, associés aux canaux *L* et *R*, occupent les positions traditionnelles de la stéréophonie. Le haut-parleur associé au canal central (*C*) est situé en face de l'auditeur et permet d'obtenir une scène sonore frontale plus stable (pour le dialogue d'un film notamment) lorsque celui-ci ne se trouve pas dans la zone d'écoute privilégiée (au centre du système). Enfin les haut-parleurs arrières, associés aux canaux *Ls* et *Rs*, permettent la diffusion de l'ambiance liée à la perception spatiale de la scène sonore (*cf.* chapitre 4 pour plus de précisions sur l'analyse d'une scène sonore multicanale).

La configuration du système 5.1 apparaît comme un compromis entre la fidélité de la reproduction spatiale (dans le plan horizontal) et le nombre de haut-parleurs disponibles. A la vue de l'écartement des haut-parleurs arrières (environ 140°, *cf.* **Figure 1.15**), il est difficile d'obtenir une image spatiale stable entre ces deux haut-parleurs. Cette remarque étant valable pour l'espace existant entre les haut-parleurs frontaux et les haut-parleurs arrières, il apparaît que le système 5.1 n'est pas disposé à offrir une image spatiale stable à 360° autour de l'auditeur. Comme le présente la section 4.1, ce système a été mis au point pour permettre la diffusion de contenus audio cinématographiques directement liés à une image (vidéo) présentée en face de l'auditeur.

Malgré les limitations intrinsèques au système 5.1, ce système est optimisé pour délivrer une image spatiale frontale, plus stable qu'avec la stéréophonie classique, qui est en outre, complétée par une image spatiale latérale (extrême gauche et droite de l'auditeur) et d'une autre image provenant de derrière l'auditeur. Même si l'image spatiale latérale est considérée comme instable et que l'image spatiale arrière est modérément satisfaisante (perception d'un « trou entre les haut-parleurs »), les canaux arrières supplémentaires permettent le positionnement de sources sonores ponctuelles derrière l'auditeur (aux positions des haut-parleurs arrières). Enfin, la zone d'écoute proposée par le système 5.1 s'est considérablement étendue comparée à celle offerte par la stéréophonie à deux canaux : bien que restant idéale au centre du système, un léger déplacement de l'auditeur n'implique pas une perception étroite orientée vers un haut-parleur *i.e.* la perception spatiale du son est conservée.

1.2.2.2 Génération des contenus audio au format 5.1

Même si les propriétés acoustiques attachées au système 5.1 sont encore basées sur les indices interauraux [MAR99], les techniques de prise de son multicanal sont actuellement en pleine évolution et offrent une large palette de systèmes microphoniques.

Les techniques de prises de son stéréophonique, comme la stéréophonie d'intensité, ont d'abord été étendues à la prise de son multicanal. Ces techniques peuvent être dissociées en deux familles : celle qui repose sur un arbre ou une antenne de microphones (système multi-microphonique) relativement proches les uns des autres, et celle qui propose de capter séparément l'image frontale et l'ambiance destinée aux canaux arrières du système de restitution. Signalons les travaux de Williams [WIL01] dans ce domaine, et plus généralement, un inventaire des techniques communément employées est donné dans [RUM01].

Les procédés de panoramique d'intensité pour la génération artificielle de signaux multicanaux tels que VBAP (cf. paragraphe 1.2.1.2) peuvent être remplacés par des lois de panoramiques liées à la théorie ambisonique. Ces lois sont basées sur l'optimisation de critères liés à la localisation auditive tels que les vecteurs vélocité (relatif à la direction pour les basses fréquences) et énergie (relatif à la puissance pour les hautes fréquences). Les définitions et interprétations relatives à ces paramètres caractéristiques d'un champ acoustique sont données par Daniel dans [DAN00]. Gerzon démontra que favoriser l'égalité de ces deux vecteurs (pour chaque position du plan horizontal) permet d'établir une loi de panoramique optimale pour de multiples systèmes multi-haut-parleurs [GERZ92a]. Nous retiendrons que les fonctions ambisoniques d'ordre O définissent les gains de panoramique $(g_m)_{1\leq m\leq M}$ associés aux M haut-parleurs (ou canaux) d'un dispotif de restitution dans le plan horizontal tels que :

$$g_m = \frac{1}{M}\left[1 + 2\cos\left(\Delta\theta_m\right) + 2\cos\left(2\times\Delta\theta_m\right) + \cdots + 2\cos\left(O\times\Delta\theta_m\right)\right], \tag{1.5}$$

où O correspond l'ordre de la décomposition spatiale et $\Delta\theta_m$ la différence angulaire entre l'azimut θ de la source sonore perçue et la position du haut-parleur d'indice m, θ_m'.

Au-delà de l'établissement de lois de panoramique, la théorie ambisonique, qui se base sur une décomposition en harmoniques sphériques du champ acoustique [DAN00], est utilisée pour enregistrer un champ sonore en un point de l'espace. Cette opération constitue un encodage du champ sonore en plusieurs composantes ambisoniques ordonnées suivant leur résolution spatiale. Autrement dit, plus le nombre de composantes extraites par encodage ambisonique est grand et plus la résolution spatiale du champ acoustique capté est fine [DAN04]. L'opération inverse, le décodage ambisonique, est définie à partir de la connaissance du système de restitution sonore (positionnement et nombre de canaux variables). Cependant, la reconstruction du champ sonore est acoustiquement imparfaite (critères directionnel et énergétique) lorsque la disposition des haut-parleurs n'est pas régulière. Bien qu'il existe des solutions de décodage optimisées pour le système 5.0 [DAN00], le positionnement non-régulier des haut-parleurs (écartement trop important notamment) résulte en une moins bonne reconstruction sur les côtés et à l'arrière de l'auditeur (la scène frontale pouvant être rendue plus stable). Dans cette perspective, les récents travaux menés par Laborie et al. dans [LAB04] ont permis de mettre au point un système multi-microphonique adapté au système 5.0. Bien que le système multi-microphonique (huit microphones) soit de type non-coïncident, il est associé à une unité de traitement (décrite dans [LAB04]) qui transforme le champ sonore capté comme s'il avait été capté par cinq microphones coïncidents tels que les directivités de ces microphones « coïncident » parfaitement avec l'écartement des haut-parleurs du système 5.0.

1.3 Conclusion

Les caractéristiques de notre perception auditive « spatiale » (dans le plan horizontal) ont été brièvement présentées dans ce chapitre. Les paramètres liés à certains phénomènes acoustiques et perceptuels (latéralisation, sommation et superposition de la localisation, loi du premier front d'onde) ont été introduits tels que l'ITD, l'ILD (ICTD, ICLD) et l'IACC (ICC). Nous avons vu comment ces paramètres peuvent être distingués les uns des autres en tenant compte de notre faculté à localiser et à percevoir les sons. Les différences interaurales de temps et d'intensité ITD-ILD peuvent être qualifiées d'indices directionnels que se soit pour une reproduction sonore en champ libre ou en présence de réflexions. Ces indices de la localisation auditive sont principalement liés aux positions et au contenu fréquentiel des sources sonores qui entourent l'auditeur. En outre, la cohérence interaurale IACC peut être mise en correspondance avec notre perception de l'espace sonore environnant. En effet, le degré de réverbération, la largeur d'une scène sonore ou encore l'enveloppement de l'auditeur peuvent être en partie caractérisés par la valeur de l'IACC.

Nous avons ensuite présenté plusieurs systèmes de reproduction sonore adaptés à la restitution du son spatialisé. Nous nous sommes focalisés sur les principes de base liés à la prise de son naturelle et à la génération artificielle de contenus audio spatialisés. Enfin, nous avons indiqué les possibilités offertes, en termes de perception spatiale, et certaines limites caractéristiques des systèmes de reproduction stéréo, binaural et multicanal (5.1 canaux).

2. Etat de l'art des procédés de codage stéréo et multicanal

Le codage audio vise à établir une représentation du signal adaptée à une transmission efficace ou un stockage à dimension réduite des échantillons audionumériques. Bien que les réseaux de télécommunications et les systèmes de stockage voient leurs capacités s'accroître, les procédés de codage audio suscitent toujours autant d'intérêt. En effet, les applications à débit contraint (communication mobile par exemple) ou encore les applications temps-réel (*streaming* audio/vidéo sur internet) nécessitent une transmission du signal à bas débit. De plus, la quantité d'information des représentations audionumériques est en constante augmentation avec des résolutions « haute-définition » pour la conversion analogique-numérique (fréquence d'échantillonnage, allocation binaire) et surtout de par l'augmentation du nombre de canaux avec les formats audio dits multicanaux (5.1, 7.1, 10.2, etc.).

L'objectif commun des procédés de codage audio est d'obtenir une représentation compacte de l'information tout en maximisant la qualité audio qui sera restituée. Une représentation compacte est relative à la minimisation de la quantité d'information (compression) et finalement au débit du procédé de codage. La fidélité correspond à la capacité d'un procédé de codage à restituer un signal audio objectivement, ou du moins subjectivement, indissociable du signal audio original. La conception d'un codec (codeur-décodeur) audio repose sur un compromis entre la compacité de la représentation, la fidélité de la reconstruction et la complexité de l'algorithme. Selon l'application visée, d'autres contraintes peuvent apparaître comme par exemple le délai de codage.

Alors que les modèles de production de la parole [BOI87] permettent un codage de la parole adapté à de très bas débits, aucun modèle équivalent n'existe pour la caractérisation des signaux audio (musique et parole). Plutôt que coder indépendamment chaque signal ou canal original, les procédés de codage audio multicanal combinent les principes d'exploitation de la redondance et de réduction de l'information perceptuellement non-significative tout en cherchant à s'adapter au mieux à la nature du signal.

Ce chapitre vise à introduire la modélisation des signaux audio multicanaux basée sur les paramètres de la localisation auditive, présentés au chapitre 1, et associée aux principes du codage audio dit monocanal ou monophonique décrit au paragraphe 2.1.

Le paragraphe 2.2 présente en détail les méthodes employées pour le codage des signaux stéréophoniques. L'évolution des méthodes est présentée au travers de l'utilisation conjointe de la réduction des redondances et de l'extraction des indices de la localisation auditive. Les principes et performances des méthodes de codage stéréo paramétrique sont présentés.

Enfin, les caractéristiques des standards du codage audio multicanal sont données, au paragraphe 2.3, à partir des principes évoqués dans ce chapitre. Nous présentons

particulièrement comment les techniques de codage stéréophonique sont étendues au codage audio multicanal. Nous décrivons, dans ce contexte, l'émergence d'un nouveau standard qui utilise l'ensemble des principes utiles au codage et à la représentation paramétrique de notre perception spatiale.

2.1 Du codage monocanal vers le codage multicanal

Dans cette première partie du chapitre, nous introduisons différentes méthodes de codage audio appliquées à la compression d'un signal monophonique. La section 2.1.1 présente la représentation audionumérique utilisée par les supports audio actuels. La section 2.1.2 rappelle les principes du codage audio perceptuel. Les méthodes de codage basées sur une modélisation hybride du signal sont brièvement présentées au paragraphe 2.1.3. Enfin, la section 2.1.4 introduit la représentation paramétrique associée aux principes employés pour la compression des données audio afin de coder efficacement un signal audio multicanal.

2.1.1 Codage audio sans perte

Une première distinction entre les procédés de codage audio est liée à la qualité objective de la reconstruction du signal issu d'une compression avec ou sans perte. Le format audio PCM (*Pulse Code Modulation*) permet la représentation d'un signal analogique sous forme numérique par simple échantillonnage du signal acoustique et quantification uniforme des coefficients. La qualité de cette représentation est fonction de la fréquence d'échantillonnage et du nombre de bits associé à chaque échantillon lors de l'étape de quantification [GER92] qui approxime les coefficients avec une précision finie. Par exemple, la « qualité CD » (Disque Compact) est obtenue avec 16 bits par échantillon et une fréquence d'échantillonnage égale à 44100 Hz. Le débit de codage PCM d'un signal stéréophonique à qualité CD est donc: $2\times16\times44100=1{,}4$ Mbps.

Pour réduire ce débit de référence, les codeurs audio sans perte visent à reconstruire parfaitement le signal original (format PCM) en tirant profit du principe de réduction des redondances. Pour cela, les techniques de prédiction linéaire exploitent l'inter-dépendance des échantillons (un signal peut être prédit à partir des échantillons passés [HAN01]). Basé sur ce principe, le standard de la norme MPEG-4 (décrite au paragraphe 2.3.2.2) incorpore un codeur audio sans-perte (MPEG-4 Audio Lossless Coding [LIE04]) qui repose à la fois sur la prédiction linéaire du signal à coder et sur le codage entropique (codage de Huffman ou codage arithmétique [GER92]) de l'erreur de prédiction.

Les procédés de réduction de la redondance intra-canal tels que les techniques de prédiction linéaire peuvent être également considérées pour la réduction des redondances inter-canal. Notamment, la prédiction linéaire stéréo permet de regrouper les informations communes à coder avec l'information non-redondante (erreurs de prédiction) extraite du signal stéréo [LIE02]. Ce principe est également étendu pour le codage sans perte des signaux audio multicanaux à la manière du procédé *Meridian Lossless Packing* MLP, décrit dans [GER99], qui est utilisé pour le stockage de divers formats audio sur les supports DVD-vidéo et DVD-audio. Pour fixer les idées, un DVD-audio d'une capacité de 4,7 Go (10^6 octets) peut stocker jusqu'à 4 heures de musique en stéréo ou 100 minutes de musique au format 5.1 en considérant ces deux représentations audionumériques échantillonnées à 96 kHz, quantifiées sur 24 bits puis encodées au format MLP. Sans cette compression audio sans perte, ce même DVD-audio pourrait stocker seulement 45 minutes de cette représentation audio multicanale au format 5.1 puisque le taux de compression MLP est environ d'un rapport 2:1 par rapport au format PCM.

2.1.2 Codage audio perceptuel

Le second principe utilisé pour la réduction du débit de codage repose sur l'exploitation des informations non-significatives pour l'oreille. Autrement dit, l'information qui ne sera pas transmise est considérée comme inaudible *i.e.* cette information n'a pas d'influence sur notre perception du signal audio reconstruit. Ce type de codage appartient à la famille des codeurs audio dits perceptuels. La réduction du débit par ces codeurs, qui exploitent les propriétés du système auditif, peut alors être conséquente : le masquage (temporel et fréquentiel) ainsi que la limite absolue fixée par le seuil de l'audition (*cf.* Annexe B.1.1). L'idée principale, sur laquelle repose les procédés de codage perceptuel, est de s'assurer que le bruit généré par le procédé de quantification soit perceptuellement inaudible, plus précisément, en dessous du seuil de masquage calculé par un modèle psychoacoustique [PAI00] associé au codeur (*cf.* Annexe B.2 et la **Figure 2.1**).

Les codeurs audio perceptuels opèrent dans le domaine des sous-bandes ou un domaine dit transformé. Le signal peut être décomposé en sous-bandes grâce à un banc de filtres dont la conception [PRI86] relève d'un compromis entre la résolution temporelle et fréquentielle. Le découpage temporel vise à s'adapter à la nature du signal (stationnaire ou transitoire) pour éviter une surestimation du bruit de quantification lorsqu'une composante transitoire est codée (*cf.* Annexe B.1.3.2). Le découpage en fréquence se base habituellement sur la résolution fréquentielle de l'oreille *i.e.* les bandes critiques (*cf.* Annexe B.1.2). Par exemple, le codeur audio MPEG-1 couches I et II, décrit dans [BRA94], utilise un banc de filtres particulier dénommé *Pseudo Quadrature Mirror Filters* (PQMF) [PRI95] qui décompose le signal en 32 sous-bandes.

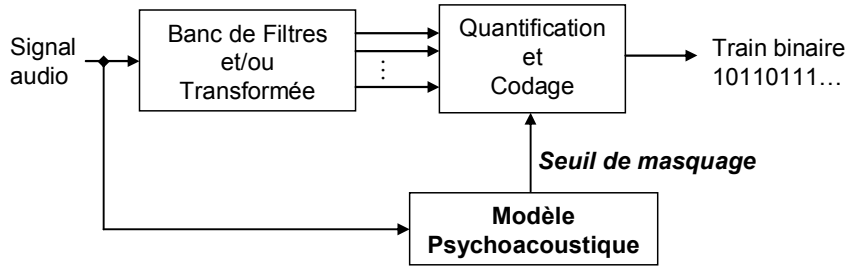

Figure 2.1 Schéma de codage générique d'un codeur audio perceptuel. Le signal d'entrée est décomposé en sous-bandes de fréquences par un banc de filtres associé ou non à une transformation temps-fréquence. Ces signaux (en sous-bandes) sont quantifiés et codés. Le modèle psychoacoustique contrôle l'erreur de quantification de manière à ce qu'elle soit en-dessous du seuil de masquage.

En considérant le fait qu'un signal (en sous-bande) peut être représenté différemment que par une suite d'échantillons à amplitude variable au cours du temps, le découpage en fréquence réalisé par un banc de filtres peut être suivi d'une transformation visant à rendre plus compactes les informations. Les représentations qui présentent un intérêt particulier pour la réduction des redondances sont : la Transformée de Fourier Discrète (TFD) [RAB78]-[DUH90], la Transformée en Cosinus Discret Modifiée (*Modified Discrete Cosine Transform* MDCT) [RAO90] et la Transformée en Ondelettes Discrètes (*Discrete Wavelet Transform* DWT) [VET95]. Ce principe est utilisé par le codeur MPEG-1 couche III ou MP3, décrit dans [BOS02], qui, malgré une augmentation de la complexité pour le calcul des coefficients MDCT, délivre une qualité audio qualifiée de transparente à 128 kbps (192 kbps constaté) pour un codage stéréophonique (taux de compression MP3 d'un rapport 10:1 par rapport au format PCM). Notons d'ailleurs, que le codec MP3 utilise les techniques de la stéréo jointe, décrites au paragraphe 2.2.1, pour réduire les redondances et parvenir à un tel taux de compression.

Le schéma de codage perceptuel, présenté à la **Figure 2.1**, est largement utilisé par les standards pour le codage des signaux audio (ISO/MPEG [PAN95]-[ISO11172]-[ISO13818]) et à moindre mesure pour le codage de la parole avec un modèle psychoacoustique largement simplifié [ZEL77] (codeur de l'UIT G722.1 [UIT]). Nous verrons au paragraphe 2.3.2.2 comment ce schéma de codage est associé aux techniques de la stéréo jointe pour le codage des signaux audio multicanaux par les standards MPEG-2 et MPEG-4 audio.

2.1.3 Codage audio hybride

Bien que les modèles sinusoïdaux aient été utilisés pour le codage de la parole et la synthèse sonore, les applications au codage audio sont apparues plus tardivement. Levine, Verma, dans [LEV98]-[VER99], et Goodwin, dans [GOO97], se sont appuyés sur un modèle sinusoïdal - qui permet de représenter un signal stationnaire au moyen de trois paramètres que sont l'amplitude, la fréquence et la phase du signal – pour la compression des signaux audio. Le procédé de codage de référence est dénommé *Analysis/Synthesis Audio Codec* ASAC [EDL95] et vise à extraire, quantifier et coder les paramètres de la synthèse sinusoïdale au regard de critères perceptuels.

Le procédé de codage audio perceptuel ASAC a été amélioré par le codeur *Harmonic and Individual Lines Plus Noise Coder* HILN [PUR97] basé sur une modélisation hybride de type : harmoniques, fréquences pures et bruit. Suivant la même philosophie, Daudet, dans [DAU00], propose un procédé de codage hybride qui décompose un signal en trois couches : une partie tonale (sinusoïdes) codée avec des coefficients en cosinus discrets, des transitoires pour lesquels les coefficients en ondelettes sont quantifiés et enfin une partie stochastique (bruit) codée par les techniques de modélisation de l'enveloppe spectrale.

La modélisation « sinusoïdes + transitoires + bruit » a cependant le désavantage de générer des artefacts à l'écoute qui sont liés en majorité à la mauvaise reconstruction des composantes transitoires. Boyer, dans [BOY02]-[BOY03], présente plusieurs extensions du modèle sinusoïdal classique (modèle sinusoïdal amorti et retardé) pour permettre une meilleure reconstruction des transitoires.

Ces méthodes de codage audio hybride peuvent être également qualifiées de paramétriques dans la mesure où les paramètres relatifs à chaque « objet » du modèle sont extraits, codés et transmis pour reconstruire le signal audio à de faibles débits [BRI02]. Cependant, les procédés de codage audio multicanal se basent sur une autre modélisation paramétrique liée à notre perception spatiale du son.

2.1.4 Codage paramétrique des signaux audio multicanaux

Nous considérons dorénavant non plus un signal audio *i.e.* un seul canal, mais un signal audio multicanal *i.e.* signal stéréo, 5.1, etc. qui, en pratique, alimente un système de restitution sonore approprié. Appliquer les méthodes de codage audio sans perte et perceptuel à chacun des canaux, indépendamment les uns des autres, résulte en une augmentation linéaire du débit total. Pour réduire la quantité d'informations à transmettre, les méthodes actuelles en matière de codage audio multicanal (*cf.* **Figure 2.2**) se basent sur les mêmes principes de compression classiquement utilisés pour le codage « monocanal » mais en exploitant, en outre, les redondances inter-canal et une représentation paramétrique des canaux. Comme l'indique la **Figure 2.2**, le codage audio multicanal dit *paramétrique* modélise la scène sonore originale par

- un signal somme (monophonique ou stéréophonique), dérivé des canaux d'entrée par réduction des redondances, qui contient l'information audio basique de la scène originale
- un flux de paramètres dérivés de notre perception spatiale des sons.

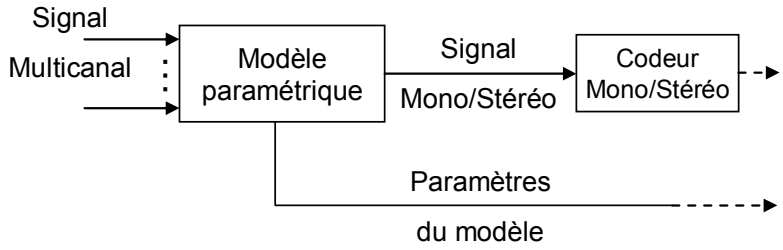

Figure 2.2 Schéma de codage générique d'un codeur audio multicanal dit paramétrique.

Ainsi, en associant cette représentation aux techniques classiques de compression (codage perceptuel, quantification, etc.), un codeur audio multicanal paramétrique permet une réduction drastique de la quantité d'informations à transmettre. En effet, le débit des paramètres du modèle est considéré comme faible de façon à approximer le débit total comme équivalent à celui d'un codage monophonique ou stéréophonique. D'abord appliquée au codage des signaux stéréophoniques (*cf.* paragraphe 2.2.3) avec un débit légèrement supérieur à celui d'un codeur monophonique, cette modélisation a été ensuite appliquée au codage des signaux audio multicanaux. Puisque l'erreur de modélisation de ce procédé est perceptuellement significative, le standard dérivé dit MPEG *surround*, décrit au paragraphe 2.3.3, prévoit le codage de ce signal d'erreur pour pouvoir reconstruire un signal multicanal subjectivement indissociable du signal original.

2.2 Codage audio stéréophonique

Cette section du chapitre introduit les procédés utilisés pour le codage stéréophonique. Les paragraphes 2.2.1 et 2.2.2 présentent l'association des techniques de matriçage aux méthodes de codage audio classiques (quantification perceptuelle dans le domaine transformé notamment). Le paragraphe 2.2.3 présente comment les paramètres liés à la localisation auditive, présentés au paragraphe 1.1, sont utilisés pour définir une représentation paramétrique des signaux stéréophoniques dans un contexte de codage.

2.2.1 Les techniques de la stéréo jointe

Les techniques de la stéréo jointe, dénommées codage somme-différence (*Mid-Side coding*) et codage d'intensité (*Intensity coding*), exploitent les redondances inter-canal pour réduire la quantité d'informations à transmettre.

2.2.1.1 Codage stéréo par matriçage somme-différence

Plutôt que de coder indépendamment chaque canal par un codeur monophonique perceptuel, Johnston et Ferreira, dans [**Erreur ! Source du renvoi introuvable.**], ont proposé de coder les signaux somme et différence définis, pour un signal stéréophonique ($c_1[n]$,$c_2[n]$), tels que

$$\begin{cases} c_m[n] = \frac{1}{\sqrt{2}}\left(c_1[n] + c_2[n]\right), \\ c_s[n] = \frac{1}{\sqrt{2}}\left(c_1[n] - c_2[n]\right). \end{cases} \quad (2.1)$$

Les composantes similaires s'additionnent dans le canal somme ou *middle* $c_m[n]$ et s'annulent dans le canal différence ou *side* $c_s[n]$. Sous l'hypothèse que les canaux originaux sont

corrélés, l'entropie du canal différence est réduite. Ce matriçage est inversible, ce qui facilite d'autant plus l'opération de décodage (mode de matriçage transparent en l'absence de quantification/codage). Comme l'indique la **Figure 2.3**, le codage stéréo par matriçage somme-différence repose sur un double codage monophonique des canaux matricés sous l'hypothèse que le modèle psychoacoustique atteste que le bruit de codage est masqué. Dans le cas de signaux stéréophoniques, le seuil de masquage ne peut plus être calculé à partir d'une modélisation monophonique du son atteignant nos oreilles. Le calcul du seuil de masquage est établi en tenant compte des différences interaurales (au niveau des oreilles de l'auditeur) engendrées par la diffusion de plusieurs signaux simultanément. Cette dépendance est décrite par le terme *binaural masking level difference* (BMLD) décrit dans [BLA97] et [MOO03]. Finalement, le codage stéréo par matriçage somme-différence est alors remplacé par un double codage monophonique des canaux originaux si les niveaux des seuils de masquages binauraux sont inférieurs au niveau du bruit de codage (voir [**Erreur ! Source du renvoi introuvable.**], [HER92] et [KAT92] pour plus de détails).

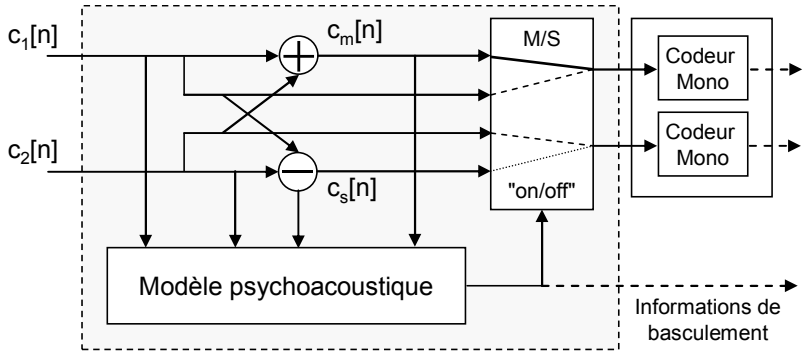

Figure 2.3 Codage stéréophonique par matriçage somme-différence (*Mid-Side* M/S).

Finalement, le codage par somme-différence exploite les redondances en assurant le masquage « spatial » du bruit de quantification par matriçage, fonction de la valeur des seuils de masquage et des différences binaurales.

2.2.1.2 Codage d'intensité

Le codage d'intensité, décrit dans [HER94], écarte la composante différence $c_s[n]$ dite résiduelle et exploite le fait que la perception des composantes hautes fréquences est principalement liée aux enveloppes temporelles (énergétiques) du signal. Une forte réduction du débit de codage est obtenue en transmettant uniquement le signal moyen $c_m[n]$ accompagné de facteurs d'échelle définis dans le domaine des sous-bandes entre les canaux originaux. Ces facteurs d'échelle peuvent être considérés comme des ICLD extraites par sous-bandes de fréquences (*cf.* paragraphe 1.1.1.2). Le signal stéréophonique peut alors être reconstruit à partir du signal c_m' perceptuellement décodé et des facteurs d'échelle. Bien que cette technique de codage puisse délivrer la reconstruction parfaite de signaux stéréophoniques issus d'un panoramique d'intensité (*cf.* paragraphe 1.2.1.2), ce n'est plus le cas lorsque les canaux présentent des composantes décorrélées.

En pratique, les techniques de la stéréo jointe sont combinées : le codage somme-différence est utilisé pour coder les composantes basses fréquences et le codage d'intensité pour les composantes hautes fréquences. Autrement dit, une meilleure reconstruction, en termes de qualité audio, est obtenue lorsque le signal différence est transmis (en complément du signal somme) au moins pour les composantes basses fréquences *i.e.* les facteurs d'échelle ou ICLD constituent, en hautes fréquences, une approximation perceptuellement acceptable. D'abord utilisées pour le codage stéréophonique, ces techniques se sont alors étendues au codage des signaux audio multicanaux notamment au sein du codec MPEG-2/4 AAC [BOS97] (*cf.* paragraphe 2.3.2).

2.2.2 Codage stéréo basé sur un matriçage adaptatif des canaux

Le matriçage somme-différence peut être vu comme un cas particulier du matriçage adaptatif, basé sur la covariance des signaux, utilisé par van der Waal et Veldhuis dans [WAA91]. La méthode vise à exploiter la corrélation des canaux dans le domaine des sous-bandes. Chaque canal est décomposé en 32 sous-bandes uniformes par un banc de filtres décrit dans [WAA91]. Les signaux temporels en sous-bandes sont alors projetés sur la base des vecteurs propres (*cf.* **Figure 2.4**).

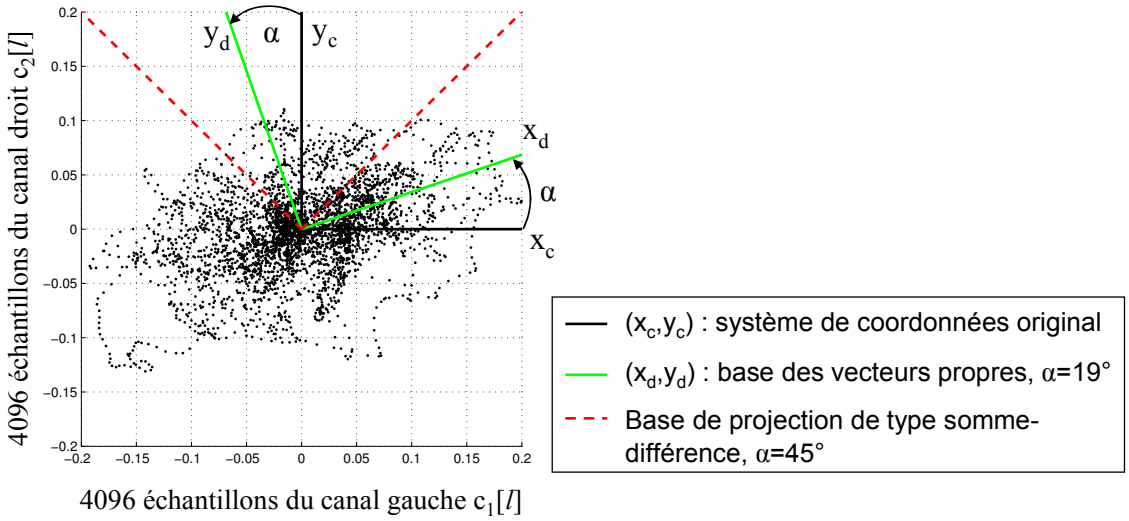

Figure 2.4 **Bases de projection des vecteurs propres et de type somme/différence.** Les valeurs du signal stéréo sont normalisées et extraites d'une portion de signal (4096 échantillons à 48000 Hz de fréquence d'échantillonnage) tirée de « Layla » par E. Clapton enregistré en concert.

Cette transformation est connue sous le nom d'Analyse en Composante Principale (ACP) ou Transformation de Karhunen-Loève (KLT) dont les fondements théoriques et les performances de séparation sont présentés au chapitre 4. La matrice de projection est orthogonale et considérée comme une matrice de rotation qui définit la transformation des portions temporelles d'indice l des canaux (C_1, C_2) telle que

$$\begin{pmatrix} d_1[l] \\ d_2[l] \end{pmatrix} = \begin{pmatrix} \cos(\alpha[l]) & \sin(\alpha[l]) \\ -\sin(\alpha[l]) & \cos(\alpha[l]) \end{pmatrix} \cdot \begin{pmatrix} c_1[l] \\ c_2[l] \end{pmatrix} \tag{2.2}$$

où la dépendance à l'indice des sous-bandes est omise pour raison de clarté. L'angle de rotation est estimé à partir des éléments de la matrice de covariance donnée par

$$\mathbf{\Gamma}_{\mathbf{c}_2} = (r_{ij}),\ r_{ij} = \mathrm{E}\left[(c_i - \mathrm{E}[c_i]) \cdot (c_j - \mathrm{E}[c_j])^T \right],\ \forall i, j \in [1;2] \tag{2.3}$$

$$\text{tel que :}\ \tan(2\alpha[l]) = \frac{2 \times r_{12}[l]}{r_{11}[l] - r_{22}[l]}. \tag{2.4}$$

^{T} désigne l'opération de transposition et E[.] l'espérance mathématique définie comme la moyenne arithmétique des échantillons d'une portion de signal à N échantillons.

Cet angle de rotation peut être physiquement interprété comme la direction de l'image stéréophonique perçue. Sous l'hypothèse que le mélange d'une source, $s[n]$, dans les canaux est réalisé au moyen d'un panoramique d'intensité suivant la loi des tangentes (*cf.* paragraphe 1.2.1.2), alors $c_1[n] = \sin(\theta) \times s[n]$ et $c_2[n] = \cos(\theta) \times s[n]$, d'après [PUL97], et la source virtuelle est perçue à l'azimut θ entre les haut-parleurs écartés de $2 \times \theta_0 = 90°$. Notons que dans ce cas particulier, la correspondance entre θ et α est possible en utilisant l'équation (2.4). Dans le cas d'un signal stéréo issu d'un enregistrement naturel (*cf.* paragraphe 4.1.2.1 pour la caractérisation des composantes d'un tel signal), l'angle α peut être mis en correspondance avec l'azimut de la source « dominante » contenue dans le signal stéréo (voir paragraphe 4.3.3.1). Cette approche, basée sur l'ACP, est également utilisée dans [IRW02] pour réaliser l'upmix d'un signal stéréo en un signal audio multicanal (voir l'Annexe C.2 pour plus de détails).

La situation où $\alpha=0$ correspond au cas où les signaux sont décorrélés et se traduit alors par un double codage audio monophonique puisque les redondances ne peuvent être réduites. La situation où $\alpha=\pi/4$ correspond au cas où les signaux sont fortement corrélés (à des niveaux d'énergie proches) et se traduit par un encodage de type somme-différence. Dans tous les autres cas, les rotations s'adaptent à la corrélation des canaux en sous-bandes et le signal dominant d_1 est une meilleure approximation du signal stéréophonique (au sens de l'énergie) que le signal moyen ou somme introduit au paragraphe 2.2.1.1.

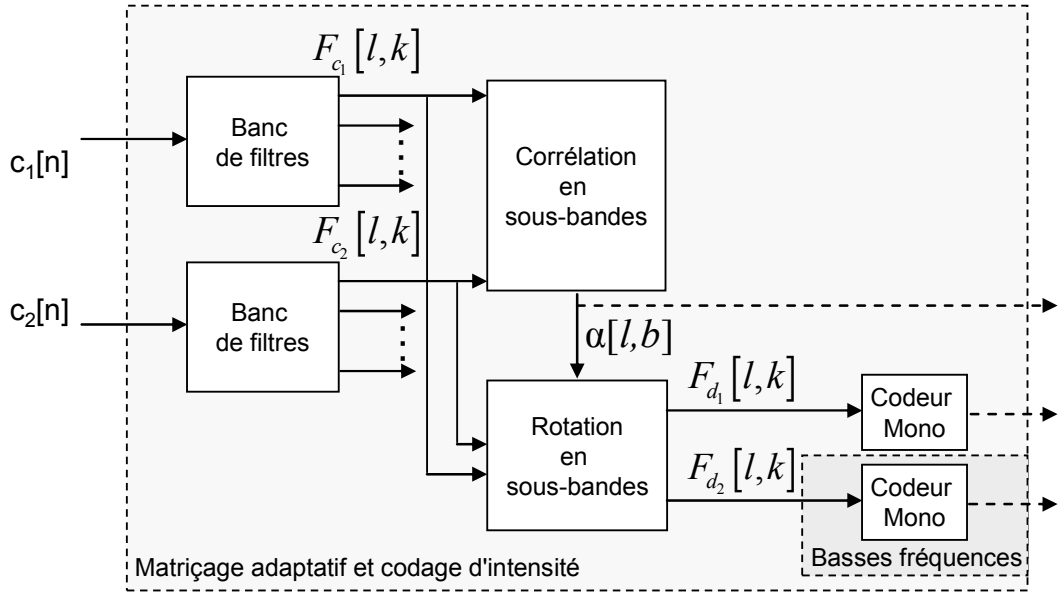

Figure 2.5 **Matriçage adaptatif associé au codage d'intensité**. Le signal résiduel est codé seulement pour les sous-bandes en basse fréquence (< 10 kHz).

D'après [WAA91], en considérant un schéma de codage dit « stéréo adaptatif » où à la fois les signaux D_1, D_2 et les angles de rotation sont quantifiés et codés, la concentration de l'énergie dans le signal dominant D_1 ne suffit pas à générer un gain de codage (en nombre de bits lors de l'allocation binaire) suffisant pour permettre la transmission supplémentaire des angles de rotation. En d'autres termes, la quantité d'informations nécessaire à la transmission des angles de rotation compense le gain de codage introduit par le matriçage adaptatif. Cependant, un autre schéma de codage, proche de la stéréo jointe (*cf.* paragraphe 2.2.1), est envisagé dans [WAA91] et présenté à la **Figure 2.5**. Le procédé consiste à associer le codage stéréo adaptatif à un codage d'intensité où seuls le signal dominant et les angles de rotation sont transmis. Les résultats présentés dans [WAA91] démontrent que les performances d'un tel schéma de codage dépassent celles de la stéréo jointe avec une fréquence de transition entre les modes de codage autour de 10 kHz (codage stéréo adaptatif en basse fréquence et codage d'intensité en hautes fréquences présenté à la **Figure 2.5**).

Récemment, l'utilisation de la méthode de codage stéréo adaptative a été reconduite, dans [BAR05], avec une décomposition des canaux par un banc de filtres complexe. Les coefficients sont regroupés en sous-bandes selon l'échelle des Barks (*cf.* Annexe B.1.2). La concentration de l'énergie est significativement supérieure comparée à celle obtenue avec le matriçage fixe de type somme-différence (de l'ordre de 4 dB d'après [BAR05]).

2.2.3 Les méthodes de codage stéréo paramétrique

L'évolution des techniques de la stéréo jointe pour le codage stéréophonique s'est orientée vers une modélisation paramétrique du signal. Outre le contenu audio « basique » des canaux, la spatialisation présente dans le signal est représentée au moyen de paramètres perceptifs basés sur les indices de la localisation auditive (*cf.* paragraphe 1.1). Le codeur vise d'une part à générer un signal somme par *downmix* des canaux originaux et d'autre part à extraire les paramètres qui caractérisent l'image spatiale stéréophonique (*cf.* **Figure 2.6**). Au décodage, la compatibilité avec les systèmes monophoniques est assurée et la transmission des paramètres spatiaux permet la synthèse d'un signal stéréophonique dont les indices perceptifs inter-canaux approximent ceux présents dans le signal original.

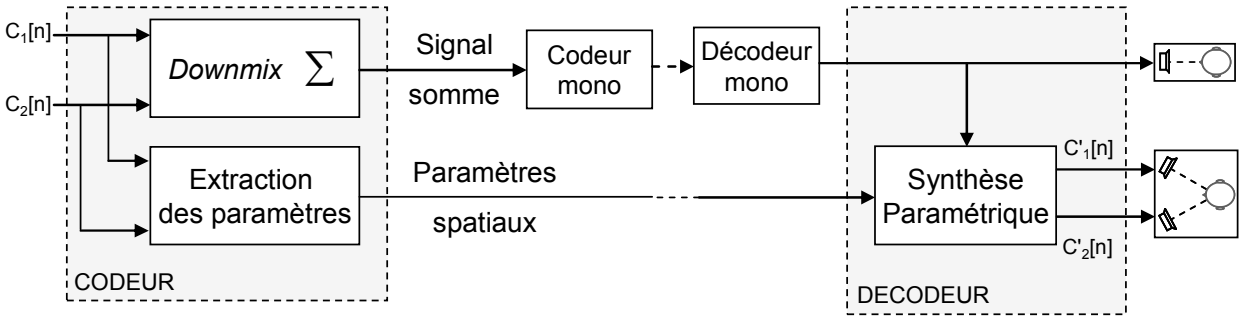

Figure 2.6 Schéma de principe d'un système de codage/décodage stéréo paramétrique.

Le procédé *Binaural Cue Coding* (BCC), initié par Faller [FAL02] et Baumgarte [BAU02a], est basé sur une représentation des canaux en sous-bandes obtenues par un banc de filtres dont la résolution approxime la résolution fréquentielle de l'oreille. Le procédé BCC peut être vu comme une extension du codage stéréo d'intensité en termes de bande passante puisque les différences d'intensité (ICLD) sont extraites pour chaque sous-bande de fréquence. D'autre part, le procédé BCC se base sur une modélisation paramétrique et perceptuelle des signaux audio spatialisés à coder plus complète que celle utilisée par le codage stéréo d'intensité en considérant en outre les différences de temps introduites entre les canaux (ICTD) et la cohérence inter-canal (ICC). Alors que la théorie duplex (*cf.* paragraphe 1.1.1), dérivée de mesures réalisées aux niveaux des oreilles à partir de sources sonores réelles, formule une prépondérance de l'ITD pour les basses fréquences (f<1500 Hz) et une prépondérance de l'ILD pour les hautes fréquences (f>1500 Hz), le procédé BCC considère des paramètres définis pour chaque fréquence. En effet, la validité de la théorie duplex, non plus pour des indices interauraux mais, pour les indices inter-canaux extraits de signaux stéréo issus d'une prise de son naturelle ou générés artificiellement n'est absolument pas garantie. Rien ne permet d'affirmer que les différences d'intensité captées par une paire de microphones coïncidents seraient prépondérantes seulement pour les fréquences supérieures à 1500 Hz. De même, les panoramiques d'intensité ne sont pas forcément réalisés avec une dépendance fréquentielle respectant cette théorie.

Outre le fait que le procédé *Parametric Stereo* (PS) [BRE05a] utilise un banc de filtres hybride complexe dont les principes et avantages (faible complexité par exemple) sont présentés dans [SCH04], la différence majeure qui le distingue du procédé BCC réside en l'utilisation d'un filtre de décorrélation et d'un matriçage basé sur l'ACP au niveau du

décodeur. Comme le procédé BCC, la méthode de codage PS a tout d'abord été implémentée en se basant sur la TFD des canaux [FAL02]-[BRE05a]. Les spectres sont alors décomposés en sous-bandes par regroupement des coefficients spectraux selon une échelle perceptuelle (*cf.* paragraphe B.1.2). Par la suite, nous ne ferons pas de distinction catégorique entre les deux approches et nous nous focaliserons sur une implémentation dans le domaine fréquentiel par TFD [BRE05a] (ou par sa version à faible complexité Transformée de Fourier Rapide - *Fast Fourier Transform* FFT) pour sa simplicité de mise en œuvre.

2.2.3.1 Opérations affectées au codeur

Extraction de paramètres spatiaux basés sur la localisation auditive

Les paramètres spatiaux extraits par les procédés de codage stéréo paramétrique correspondent aux indices de la localisation auditive dans le plan horizontal considérés aussi bien pour une diffusion sur haut-parleurs qu'au casque. Par conséquent, les différences inter-canal de temps et d'intensité (ICTD-ICLD) sont extraites à partir des canaux du signal stéréophonique de façon à pouvoir repositionner les sources. L'ICTD peut aussi être considérée comme une différence inter-canal de phase entre les canaux (*Inter Channel Phase Difference* ICPD) comme proposé dans [BRE05a]. Cependant ces paramètres ne donnent qu'une information relative à la direction des sources dans l'image sonore. Pour combler cette insuffisance à la caractérisation de la scène sonore, la cohérence inter-canal (ICC) est extraite pour pouvoir régénérer la même impression spatiale (largeur de l'image stéréo notamment *cf.* paragraphe 1.1.2.2).

Les paramètres spatiaux sont qualifiés de paramètres perceptuellement significatifs dans la mesure où ils sont extraits dans le domaine spectral selon une échelle perceptuelle dérivée des bandes critiques (*cf.* Annexe B.1.2). En notant $F_{C1}[l,k]$ et $F_{C2}[l,k]$ les Transformées de Fourier à Court Terme (TFCT définie dans l'Annexe A.1.1) des canaux $c_1[n]$ et $c_2[n]$, les paramètres spatiaux sont extraits par sous-bandes de fréquences, pour chaque portion glissante de signal d'indice l, à partir de l'indice fréquentiel k_b à l'indice k_{b+1} tels que :

$$\mathrm{ICLD}[l,b] = 10 \times \log_{10}\left(\frac{\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_1}^{*}[l,k]}{\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_2}[l,k]\cdot F_{c_2}^{*}[l,k]}\right) \ \mathrm{dB} \qquad (2.5)$$

où l'indice des sous-bandes b = [1,…K_b]. L'implémentation, présentée dans [BRE05a], préconise une décomposition spectrale allant de K_b = 20 à 34 sous-bandes suivant l'échelle ERB, échelle qui définit les valeurs de k_b et k_{b+1} pour chaque sous-bande b (*cf.* Annexe B.1.2). La fenêtre d'analyse est de taille variable puisqu'elle s'adapte au contenu des signaux. En effet, si une impulsion ou « attaque » est détectée alors un basculement vers une fenêtre de taille plus petite est réalisé (*cf.* Annexe B.1.3.2). Ainsi, les effets de pré-écho et d'antériorité sont diminués. Les différences de phases inter-canal ou ICPD sont données par :

$$\mathrm{ICPD}[l,b] = \angle\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_2}^{*}[l,k]\right) \qquad (2.6)$$

et s'expriment en radians. L'ICC prend ses valeurs entre 0 et 1 lorsqu'elle est estimée indépendamment de la phase des canaux (module de l'inter-spectre au numérateur) telle que :

$$\mathrm{ICC}[l,b] = \frac{\left|\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_2}^{*}[l,k]\right|}{\sqrt{\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_1}^{*}[l,k]\right)\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_2}[l,k]\cdot F_{c_2}^{*}[l,k]\right)}}. \qquad (2.7)$$

Notons que dans ce cas, l'ICC et l'ICLD sont extraites après avoir aligné les signaux en phase au regard de la valeur de l'ICPD extraite (la procédure est décrite au paragraphe suivant). A

l'inverse, lorsque l'information de phase n'est pas transmise par l'encodeur, comme c'est le cas pour le procédé PS décrit dans [BRE05a], l'ICC dépend de la différence de phase entre les canaux telle que :

$$\mathrm{ICC}_2[l,b] = \frac{\Re\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_2}^*[l,k]\right)}{\sqrt{\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_1}^*[l,k]\right)\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_2}[l,k]\cdot F_{c_2}^*[l,k]\right)}} \tag{2.8}$$

ainsi, ICC_2 prend ses valeurs entre -1 (signaux en opposition de phase) et 1 (signaux en phase).

Le procédé BCC définit l'ICTD, équivalente à l'ICPD, en termes de décalage temporel entre les canaux (nombre d'échantillons) comme l'indice temporel du maximum de la fonction de cohérence entre les signaux temporels en sous-bandes (issus du banc de filtres décrit dans [BAU02a]).

Downmix **basé sur les paramètres pour assurer la conservation de l'énergie**

L'opération de *downmix* d'un signal stéréophonique ($c_1[n]$, $c_2[n]$) en un signal monophonique $c_m[n]$ consiste à définir la combinaison linéaire des canaux originaux telle que l'énergie totale soit conservée. Typiquement, la simple sommation des canaux, même pondérée par les coefficients donnés par la recommandation de l'UIT-R (*cf.* Annexe C.1), peut résulter en une amplification ou une atténuation des canaux originaux dans le signal somme.

Faller, dans [FAL04], propose d'égaliser le signal somme tel que sa puissance approxime la puissance correspondante des canaux d'entrée. L'addition des canaux peut alors être corrigée par un facteur tel que :

$$F_{c_m}[l,k] = e[l,k]\times\left(F_{c_1}[l,k] + F_{c_2}[l,k]\right) \tag{2.9}$$

avec le facteur de correction $e[l,k]$ défini par :

$$e[l,k] = \sqrt{\frac{\left|F_{c_1}[l,k]\right|^2 + \left|F_{c_2}[l,k]\right|^2}{\left|F_{c_\Sigma}[l,k]\right|^2}} \tag{2.10}$$

où $F_{c_\Sigma}[l,k]$ correspond à la TFCT du signal somme $c_\Sigma[n] = c_1[n]+c_2[n]$.

Cependant, des différences de phase mesurées entre les canaux peuvent également détériorer les composantes du signal somme par un effet de filtrage en peigne décrit en Annexe C.2.3. Par conséquent, un recalage temporel ou une modification de la phase des canaux doit être appliquée à chaque canal stéréophonique avant de les additionner. Le procédé PS préconise l'estimation de la différence de phase totale (*overall phase difference* OPD) entre le signal somme C_Σ et un canal d'entrée, par exemple C_1 telle que :

$$\mathrm{OPD}[l,b] = \angle\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{c_1}[l,k]\cdot F_{c_\Sigma}^*[l,k]\right) \tag{2.11}$$

L'OPD n'est pas forcément égale à la moitié de l'ICPD puisque ce paramètre dépend à la fois de l'ICLD et de l'ICC comme l'ont démontré Lapierre et *al.* dans [LAP06]. Par conséquent, avant de réaliser la somme des canaux, indiquée par l'équation (2.9), il convient de modifier la phase des canaux en ajoutant la valeur (OPD) à la phase de C_1 et en ajoutant la valeur (ICPD-OPD) à la phase de C_2, comme indiqué par la **Figure 2.7**.

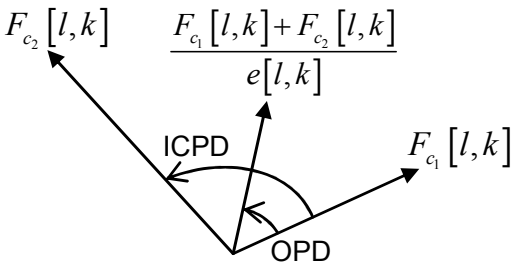

Figure 2.7 Différences de phases inter-canal (ICPD) et totale (OPD).

Ajoutons également que l'OPD peut être estimée à partir des valeurs de l'ICLD, de l'ICC et de l'ICPD, d'après [LAP06], tel que :

$$\mathrm{OPD}[l,b] = \arctan_{2\pi}\left(10^{\mathrm{ICLD}[l,b]/20} + \mathrm{ICC}[l,b]e^{j\mathrm{ICPD}[l,b]}\right) \tag{2.12}$$

où la fonction arc-tangente est définie pour les quatre quadrants. Par conséquent, l'OPD peut ne pas être transmise explicitement au décodeur si les erreurs introduites par le processus de quantification (Q *cf.* **Figure 2.8**) ne sont pas perceptuellement significatives (*cf.* [LAP06] pour les détails de l'expérimentation). Dans le cas où le signal somme a été généré sans précaution, les mêmes auteurs ont proposé une méthode permettant de compenser l'amplification ou l'atténuation de composantes dans le signal somme, au décodage et seulement à partir des paramètres ICLD, ICPD et ICC.

La **Figure 2.8** présente le schéma de codage utilisé par le procédé PS basé sur la transformée de Fourier des canaux. Le codeur génère un signal monophonique à partir de la transformée de Fourier inverse (FFT^{-1}) du signal somme dont les portions temporelles sont additionnées avec la méthode *overlap-add* (OLA).

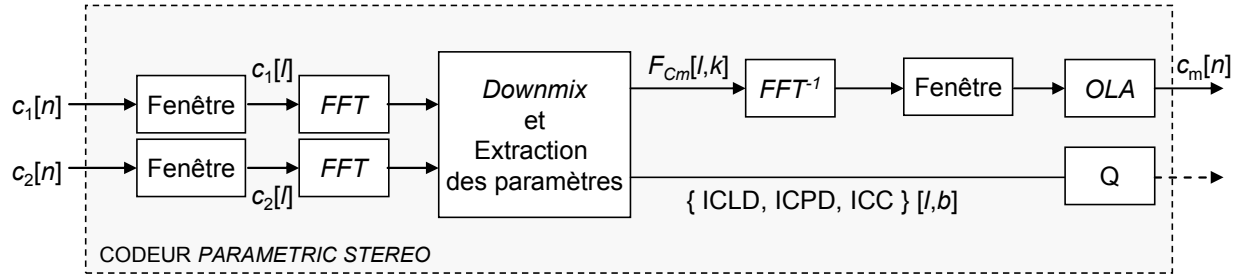

Figure 2.8 Principe du codeur PS : extraction des paramètres puis downmix des canaux dans le domaine spectral en sous-bandes.

En se basant sur des critères perceptuels, le processus de quantification (Q) est de type non-uniforme pour chaque paramètre. Par exemple, plus la valeur de l'ICLD à quantifier est grande et plus la sensibilité auditive aux variations de ce paramètre est faible. Les pas de quantification et les techniques de codage différentiel utilisées sont donnés dans [BRE05a]. Le débit moyen du flux de paramètres spatiaux extraits par l'encodeur PS avec une décomposition en 34 sous-bandes et un taux de rafraichissement des paramètres de 23 ms, a été estimé sur une large base d'apprentissage autour de 8 kbps [BRE04]. Le débit de l'ensemble {ICLD, ICC} atteint environ 6 kbps alors que celui de l'ensemble {ICPD, OPD} est de l'ordre de 2 kbps car les différences de phases sont seulement transmises pour les basses fréquences inférieures à 2 kHz sous l'hypothèse que le système auditif est peu sensible aux différences de phase pour les composantes hautes fréquences. D'après [BRE05a], le débit des paramètres peut être réduit jusqu'à 1,5 kbps avec un nombre de sous-bandes limité à 20 et un taux de rafraichissement des paramètres (sans ICPD-OPD) de 46 ms.

2.2.3.2 Opérations affectées au décodeur

A partir du signal somme transmis, les décodeurs stéréo paramétrique de type BCC ou PS synthétisent un signal stéréophonique tel que ses différences inter-canal approximent celles du signal original. D'un point de vue de la perception spatiale, une synthèse paramétrique limitée aux paramètres ICTD/ICPD-ICLD délivre les positions des sources (dans le plan horizontal) et les effets de coloration dus aux réflexions primaires mais peut souffrir d'une réduction de la largeur de la scène sonore. Comme nous l'avons introduit au paragraphe 1.1.2.2, l'IACC ou ICC tient une place importante dans notre perception de l'espace sonore environnant (largeur, profondeur, etc.). Par conséquent, les décodeurs stéréo paramétriques se basent sur le signal monophonique, les paramètres spatiaux et une version décorrélée du signal monophonique d'entrée (cf. **Figure 2.9**). La synthèse paramétrique vise d'une part à reproduire les positions des sources à partir du signal monophonique et du couple de paramètres {ICTD/ICPD, ICLD} et d'autre part à synthétiser un signal stéréophonique dont la cohérence est définie par le mélange du signal monophonique et de sa version décorrélée au regard du paramètre ICC.

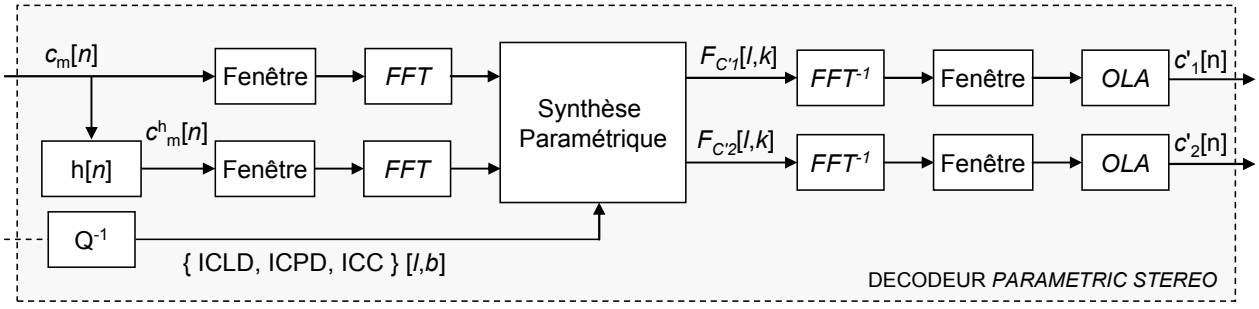

Figure 2.9 **Principe du décodeur PS**: synthèse stéréophonique en sous-bandes de fréquences à partir du signal monophonique $c_m[n]$, de sa version décorrélée $c^h_m[n]$ et des paramètres spatiaux.

Filtre de décorrélation synthétique ou de réverbération tardive

Comme nous le présentons en Annexe C.2.3, l'introduction d'un simple décalage temporel constant permet de décorréler les signaux lorsque le décalage temporel est compris entre 10 et 40 ms (pas ou peu d'effet d'antériorité ni d'écho, *cf.* paragraphe 1.1.2). Cependant, les composantes hautes fréquences d'un signal évoluent plus rapidement dans le temps que celles des basses fréquences. Les composantes hautes fréquences sont donc plus sensibles aux décalages temporels; le délai introduit doit donc être réduit pour les hautes fréquences. Le décodeur PS utilise un filtre passe-tout décorrélateur $h[n]$ dont le retard (modification de phase) introduit est dépendant de la fréquence. La réponse impulsionnelle d'un tel filtre est donnée par l'expression suivante :

$$h[n] = \sum_{k=0}^{N_h/2} \frac{2}{N_h} \cos\left(\frac{2\pi kn}{N_h} + \frac{2\pi k(k-1)n}{N_h} \right), \quad 0 \le n \le N_h, \tag{2.13}$$

où N_h = 640 échantillons (de l'ordre de 15 ms) d'après [BRE05a]. La réponse impulsionnelle $h[n]$ est présentée à la **Figure 2.10-(a)**. La modulation du cosinus évolue de manière non-linéaire avec le terme en $k(k\text{-}1)$ de l'équation (2.13). Les réponses en module et en phase de la transformée de Fourier de $h[n]$ sont présentées aux **Figure 2.10-(b)-(c)**. La réponse en module n'est pas tout à fait plate du fait de la modulation importante de phase (non-linéaire). La convolution du signal monophonique $c_m[n]$ par le filtre $h[n]$ délivre un signal $c^h_m[n]$ dont l'énergie s'étale suivant l'axe temporel comme l'illustre la **Figure 2.11**.

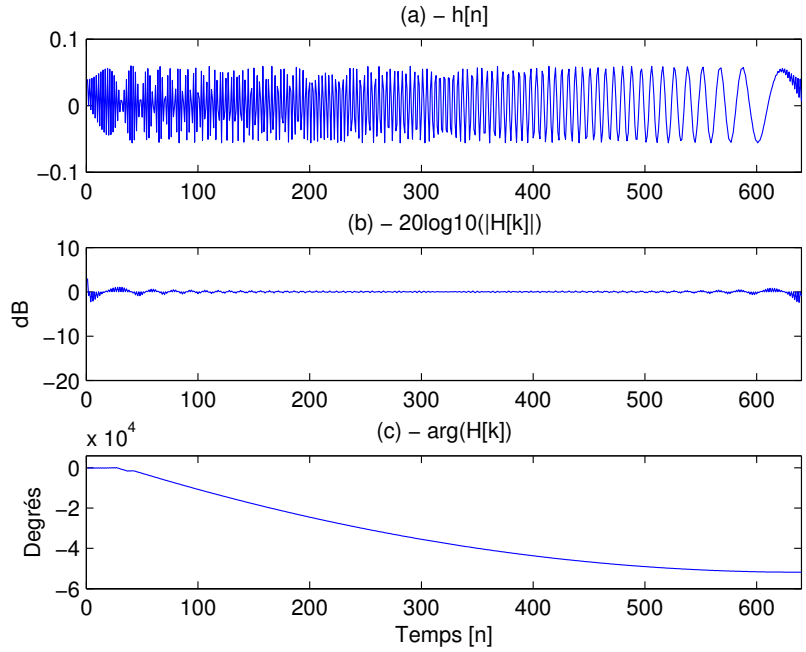

Figure 2.10 Filtre passe-tout décorrélateur à phase non-liénaire. (a) - Réponse impulsionnelle, [**(b) -** Module, **(c) -** Phase] de la transformée de Fourier de h[n].

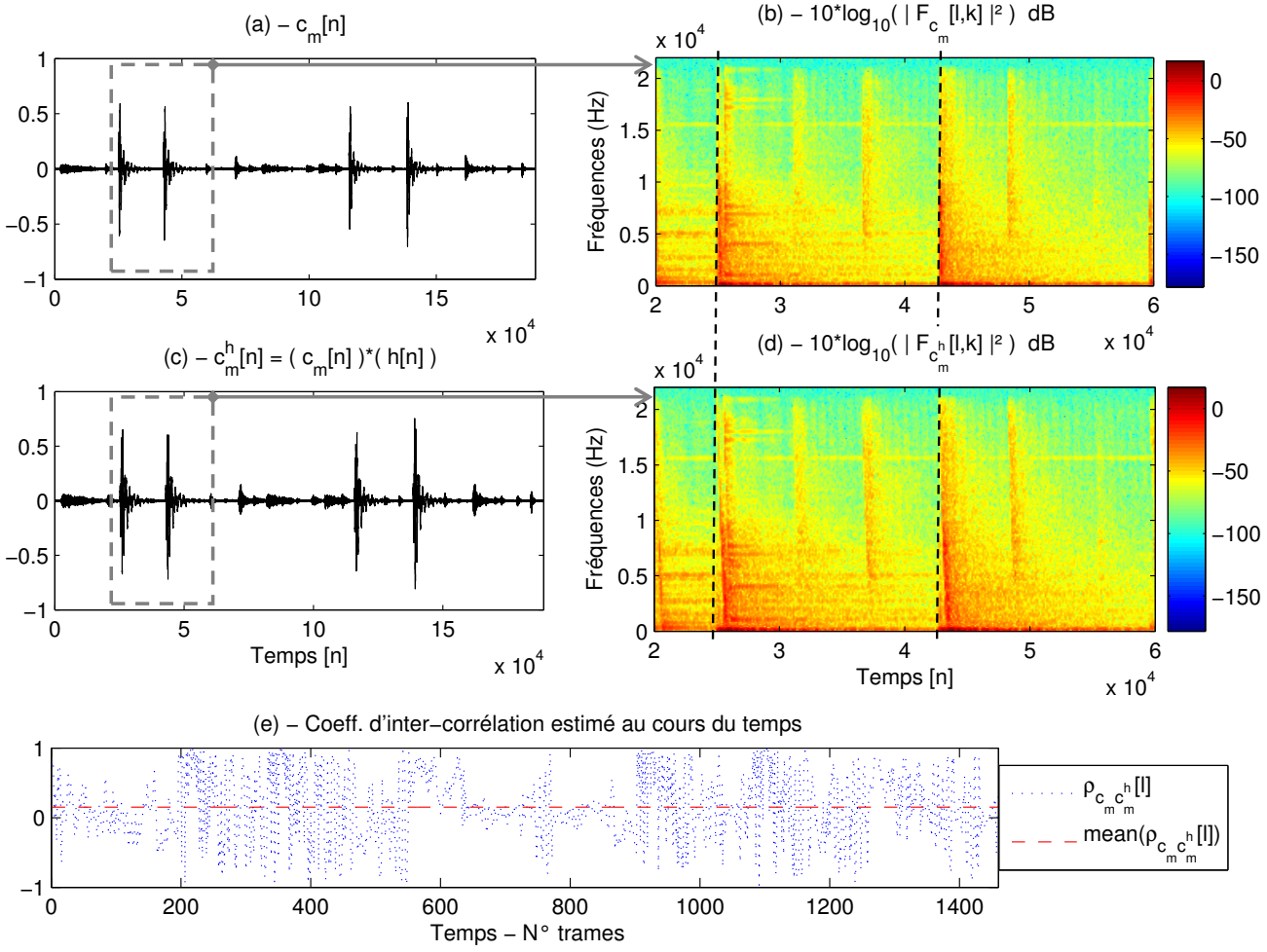

Figure 2.11 Effet du filtrage passe-tout décorrélateur à phase non-liénaire. (a) - Signal original percussif extrait d'une ligne de batterie, **(b) -** Agrandissement du spectrogramme du signal original, **(c) -** Signal convolué par le filtre h[n] **(d) -** Agrandissement du spectrogramme du signal filtré, **(e) -** Indice de corrélation estimé au cours du temps et sa valeur moyenne.

Les spectrogrammes du signal original et du signal filtré sont calculés à partir de la TFCT des signaux avec une fenêtre de Hanning de longueur N=256, Z=256 et un recouvrement de 50% (*cf.* Annexe A.1.1). La comparaison, à la **Figure 2.11-(b)** et à la **Figure 2.11-(d)**, des agrandissements des spectrogrammes du signal original et filtré illustre le décalage temporel introduit de façon non-linéaire en fonction de la fréquence. L'indice d'inter-corrélation entre le signal original et le signal filtré est calculé à partir de l'équation (C.4) en considérant des trames de longueur N=128 points. L'évolution de l'inter-corrélation au cours du temps et sa valeur moyenne sont présentées à la **Figure 2.11-(e)**. La valeur moyenne de l'inter-corrélation au cours du temps est égale à 0,15; ce qui témoigne bien du pouvoir de décorrélation d'un tel filtre. Cependant, cette inter-corrélation reste ponctuellement forte notamment pour les composantes transitoires du signal à forte énergie.

D'un point de vue subjectif, l'écoute stéréophonique ($c_m[n]$ à gauche et $c^h{}_m[n]$ à droite ou l'inverse) procure une sensation d'élargissement de la scène sonore : le signal percussif de batterie, dans notre cas, n'est pas perçu comme provenant d'une direction particulière. Par contre, une coloration haute fréquence se fait sentir sur les attaques *i.e.* les composantes hautes fréquences deviennent prédominantes d'après la loi du premier front d'onde.

Le filtre de décorrélation utilisé par le procédé BCC, présenté dans [FAL06a], est basé sur la fin d'une réponse impulsionnelle de salle (*cf.* **Figure 1.7**) : la réverbération tardive. La réponse impulsionnelle du filtre de décorrélation $h_r[n]$ est dérivée d'un bruit blanc gaussien $b_g[n]$ d'amplitude exponentiellement décroissante tel que :

$$h_r[n] = b_g[n]\left(1-\frac{1}{f_s T_h}\right)^n, \quad 0 \le n \le N_h \tag{2.14}$$

où la longueur du filtre N_h est de l'ordre de 14000 échantillons, soit 0,3 seconde, et la valeur de T_h est de l'ordre de 0,4 seconde équivalent à un temps de réverbération de 2,8 secondes, d'après [FAL06a].

Notons que l'utilisation d'un filtre passe-tout décorrélateur, équation (2.13), comparé à un filtre de réverbération tardive, équation (2.14), à l'avantage de ne pas produire de réverbération si le signal à coder en est dépourvu. Par contre, un filtre de réverbération tardive a l'avantage de moins colorer le signal (par filtrage en peigne décrit en Annexe C.2.3) et ainsi de moins dénaturer le signal s'il contient de l'effet de salle.

Synthèse paramétrique dans le domaine spectral ou des sous-bandes

Une fois le signal monophonique filtré, le décodeur réalise une synthèse paramétrique à partir du mélange des signaux $c_m[n]$ et $c^h{}_m[n]$ et des paramètres spatiaux. Le découpage des signaux est réalisé de la même manière qu'à l'encodeur : fenêtrage (indice l), TFCT (indice des fréquences k), traitement en sous-bandes (indice b) dont les limites ($k_b \le k \le k_{b+1}$) sont données par l'échelle ERB. Le matriçage dynamique, d'après [BRE05a], est défini par l'expression suivante :

$$\begin{bmatrix} F^b_{c_1'}[l,k] \\ F^b_{c_2'}[l,k] \end{bmatrix} = \sqrt{2}\mathbf{R}_1[l,b]\mathbf{R}_2[l,b]\mathbf{R}_3[l,b]\begin{bmatrix} F^b_{c_m}[l,k] \\ F^b_{c^h_m}[l,k] \end{bmatrix}, \tag{2.15}$$

$$\text{où : } \mathbf{R}_1[l,b] = \begin{bmatrix} e^{j\mathrm{OPD}[l,b]} & 0 \\ 0 & e^{j(\mathrm{OPD}[l,b]-\mathrm{ICPD}[l,b])} \end{bmatrix} \tag{2.16}$$

est la matrice qui exprime la modification de phase des signaux monophoniques telle que

- la différence de phase moyenne entre les signaux synthétisés $c'_1[n]$ et $c'_2[n]$ correspond à l'ICPD transmis,
- la différence de phase moyenne entre les signaux $c_m[n]$ et $c'_1[n]$ correspond à l'OPD transmise.

Les expressions de $\mathbf{R_2}$ et $\mathbf{R_3}$ sont données dans [BRE04] sous la réserve de vérifier les contraintes suivantes:

- la différence d'intensité entre les signaux synthétisés doit approximer l'ICLD transmise,
- la cohérence des signaux synthétisés doit correspondre à l'ICC transmise,
- l'énergie moyenne des signaux de sortie doit être égale à l'énergie du signal monophonique à l'entrée du décodeur, et
- la proportion de signal provenant de $c_m[n]$ doit être maximale dans les signaux de sortie.

La matrice $\mathbf{R_2}$ correspond à la transposition de la matrice de rotation décrite par l'équation (2.2) qui définit la projection du signal stéréophonique sur la base des vecteurs propres de la covariance (du signal d'entrée). La rotation définie par $\mathbf{R_2}$ correspond à la rotation inverse en considérant les signaux $c_m[n]$ et $c^h{}_m[n]$ comme étant les signaux projetés sur la base des vecteurs propres. Par conséquent cette opération se justifie seulement lorsque les canaux $c_m[n]$ et $c^h{}_m[n]$ sont synchronisés (par $\mathbf{R_1}$) et d'énergie équivalente à celle des signaux (D_1 et D_2) qui auraient été obtenus par la rotation de l'équation (2.2). La matrice $\mathbf{R_2}$ qui permet de repositionner les sources est donc donnée par :

$$\mathbf{R}_2[l,b] = \begin{bmatrix} \cos(\alpha[l,b]) & -\sin(\alpha[l,b]) \\ \sin(\alpha[l,b]) & \cos(\alpha[l,b]) \end{bmatrix} \tag{2.17}$$

où l'angle de la rotation opérant sur les coefficients spectraux de $c_m[n]$ et $c^h{}_m[n]$ est exprimé à partir des paramètres ICC et ICLD, tel que :

$$\alpha[l,b] = \begin{cases} \dfrac{\pi}{4}, & \text{si } \left(\mathrm{ICC}[l,b], \mathrm{ICLD}[l,b]\right) = (0,0) \\ \mathrm{mod}\left(\dfrac{1}{2}\arctan\left(\dfrac{2\times \mathrm{ICC}[l,b]\times 10^{\mathrm{ICLD}[l,b]/20}}{10^{\mathrm{ICLD}[l,b]/10}-1}\right), \dfrac{\pi}{2}\right), & \text{sinon} \end{cases}. \tag{2.18}$$

La valeur de l'angle de rotation est fixée à π/4 (matriçage somme-différence, *cf.* paragraphe 2.2.1) lorsque la corrélation et la différence d'intensité entre les canaux originaux sont nulles. En effet, dans ce cas, aucune direction prédominante n'aurait pu être estimée à l'encodeur à partir de canaux (originaux) décorrélés et de même énergie. Dans le cas contraire, la direction de la source dominante (dans la sous-bande considérée) est estimée sous l'hypothèse que l'ICC est positive puisque les signaux sont synchronisés en temps au moyen de la matrice $\mathbf{R_1}$. Cependant, un modulo π/2 est utilisé pour s'assurer que l'angle estimé appartienne au premier quadrant *i.e.* $\alpha \in [0; \pi/2]$. Nous verrons, au chapitre 4, qu'à partir des spectres en sous-bandes ou des portions glissantes des signaux originaux, l'équation (2.4), équivalente à l'équation (2.18), permet l'estimation de l'azimut de la source dominante (entre les haut-parleurs du système de reproduction) en apportant une correction à cette estimation. Enfin, l'opération est complètement définie en exprimant la matrice qui ajuste l'intensité des signaux $c_m[n]$ et $c^h{}_m[n]$ dans les canaux synthétisés (en maximisant l'énergie de $c_m[n]$ par rapport à celle de $c^h{}_m[n]$ dans les canaux $c'_1[n]$ et $c'_2[n]$) telle que :

$$\mathbf{R}_3[l,b] = \begin{bmatrix} \cos(\nu[l,b]) & 0 \\ 0 & \sin(\nu[l,b]) \end{bmatrix}, \tag{2.19}$$

$$\text{où : } \nu[l,b] = \arctan\left(\sqrt{\frac{1-\sqrt{\mu[l,b]}}{1+\sqrt{\mu[l,b]}}}\right) \tag{2.20}$$

$$\text{avec : } \mu[l,b] = 1 + \frac{4\left(\text{ICC}^2[l,b]-1\right)}{\left(10^{\text{ICLD}[l,b]/20} + 1/10^{\text{ICLD}[l,b]/20}\right)^2}. \tag{2.21}$$

Enfin, comme l'illustre la **Figure 2.9**, le codeur PS génère un signal stéréophonique à partir de la FFT inverse (FFT^{-1}) des canaux synthétisés par reconstruction selon la méthode *overlap-add* (OLA).

L'approche proposée dans [FAL06a] offre une solution équivalente à la synthèse stéréophonique (PS) dans le domaine spectral à partir des paramètres spatiaux et d'un signal monophonique décomposé en sous-bandes par un banc de filtres d'analyse. La **Figure 2.12** présente le traitement appliqué au signal temporel $c_m[n,b]$ issu du filtrage par le $b^{\text{ième}}$ filtre passe-bande.

L'indice temporel lié au fenêtrage est ignoré ici pour plus de clarté. Le signal de sortie en sous-bandes s'exprime de la manière suivante :

$$\begin{cases} c_1'[n,b] = g_{11} \times c_m[n-\Delta t_1, b] + g_{12} \times c_m^{h1}[n,b] \\ c_1'[n,b] = g_{21} \times c_m[n-\Delta t_2, b] + g_{22} \times c_m^{h2}[n,b] \end{cases} \tag{2.22}$$

où les signaux $c^{hi}{}_m[n]$ (i=1,2) résultent de la convolution du signal $c_m[n]$ avec deux filtres de réverbération tardive $h_{r1}[n]$ et $h_{r2}[n]$, obtenus à partir de l'équation (2.14) et deux bruits blancs indépendant, tels que :

$$c_m^{hi}[n] = h_{ri}[n] * c_m[n], \ \forall i \in [1;2]. \tag{2.23}$$

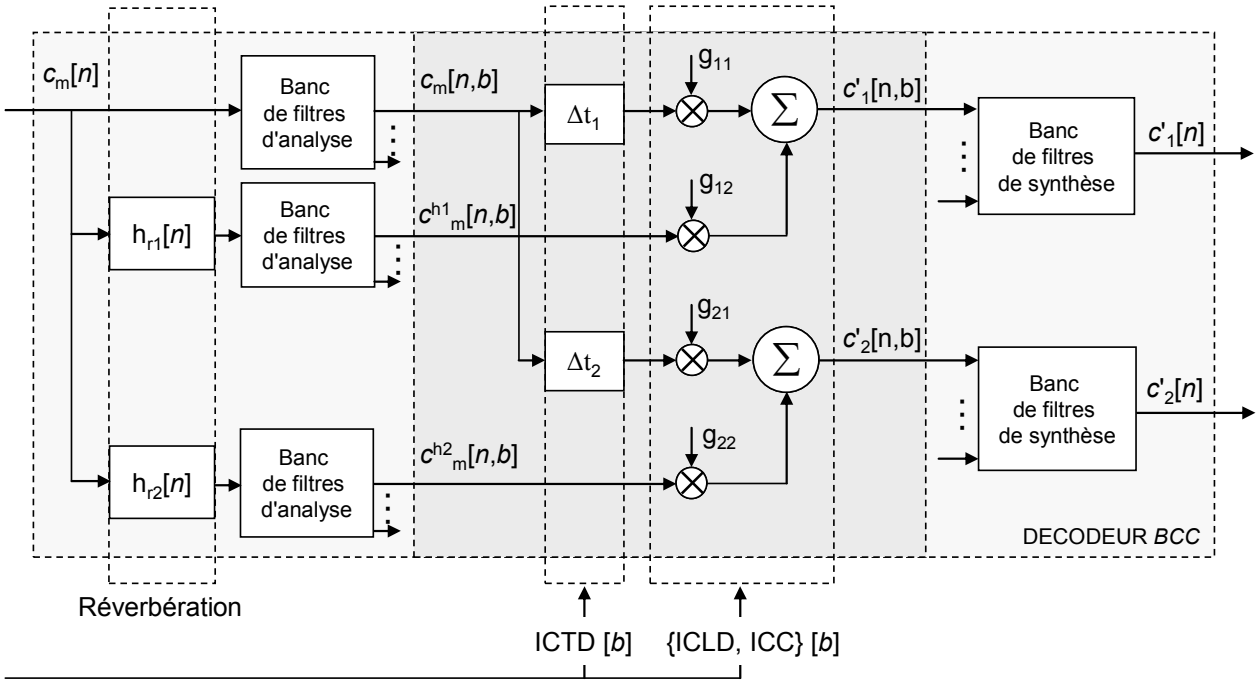

Figure 2.12 Principe du décodeur stéréo BCC (traitement d'une sous-bande) : la synthèse s'appuie sur les paramètres spatiaux et deux filtres de réverbération tardive dont les réponses impulsionnelles sont décorrélées.

A partir des signaux $c_m[n,b]$, $c^{h1}{}_m[n,b]$ et $c^{h2}{}_m[n,b]$ issus du banc de filtres, la synthèse paramétrique définit les retards (Δt_1, Δt_2) et les gains (g_{11}, g_{12}, g_{21} et g_{22}) de façon à ce que les critères énoncés pour la synthèse du décodeur PS soient respectés et que, en outre, la proportion de réverbération tardive soit équivalente dans les canaux synthétisés. Ainsi, la diffusion de l'effet de salle s'établie de manière homogène entre les canaux. La résolution du système (2.22) menée dans [FAL06a] détermine les valeurs suivantes :

$$\begin{cases} \Delta t_1 = \max\left\{-\mathrm{ICTD}[b], 0\right\} \\ \Delta t_2 = \max\left\{\mathrm{ICTD}[b], 0\right\} \\ g_{i1} = \sqrt{\dfrac{\mathrm{ICC}[b]\times 10^{\mathrm{ICLD}[b]/20} + (-1)^{i-1}\times\left(10^{\mathrm{ICLD}[b]/10}-1\right)}{2\times\left(10^{\mathrm{ICLD}[b]/10}+1\right)}} \\ g_{i2} = \sqrt{\dfrac{\left(10^{\mathrm{ICLD}[b]/10} - \mathrm{ICC}[b]\times 10^{\mathrm{ICLD}[b]/20}+1\right)\times P_{c_m}[n,b]}{2\times\left(10^{\mathrm{ICLD}[b]/10}+1\right)\times P_{c_m^{hi}}[n,b]}} \end{cases}, \; i\in[1;2] \qquad (2.24)$$

où $P_{cm}[n,b]$, $P_c{}^{h(1,2)}{}_m[n,b]$ correspondent aux puissances à court-terme des signaux en sous-bandes $c_m[n,b]$ et $c^{h(1,2)}{}_m[n,b]$ respectivement.

2.2.3.3 Performances des méthodes de codage stéréo paramétriques

Nous présentons ici une vue générale de la qualité audio et spatiale, tiré de [FAL04] et [BRE05a], obtenue par les procédés de codage stéréo paramétriques (BCC et PS).

En premier lieu, le procédé BCC, d'abord mis en œuvre avec une synthèse paramétrique sans filtre de réverbération, renforce les codeurs audio monophoniques (perceptuels) opérant à des débits inférieurs à 70 kbps (*cf.* **Figure 2.13**). La qualité audio et spatiale est alors jugée comme intermédiaire pour des débits d'informations spatiales (ICTD-ICLD-ICC) autour de 2 kbps, d'après [FAL06a]. Notons que dans ce cas, la corrélation entre les canaux synthétisés est obtenue non plus à partir de signaux décorrélés mais à partir de fonctions non-linéaires ou aléatoires visant à faire varier les paramètres spatiaux (ICTD-ICLD) différemment d'un signal à un autre (*cf.* [FAL04] pour plus de détails).

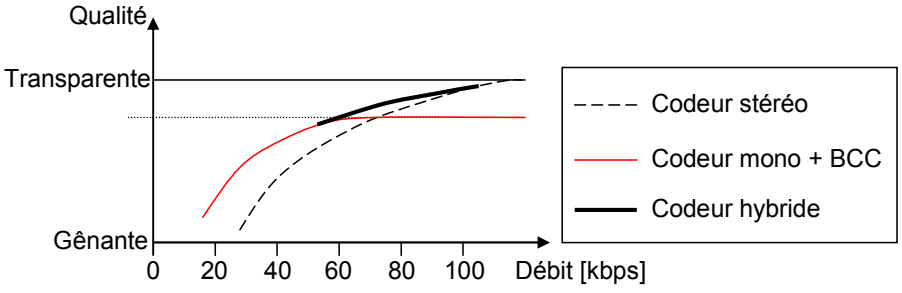

Figure 2.13 Qualité audio obtenue par le procédé de codage stéréo paramétrique BCC, d'après [FAL04].

Pour établir une transition douce entre les codeurs stéréophoniques opérant à débits élevés (haute-qualité) et les codeurs mono renforcés par BCC, un codage hybride a été mis en œuvre [FAL06a]. La méthode propose la transmission d'un signal stéréophonique pour les basses fréquences et la transmission d'un signal monophonique accompagné de paramètres spatiaux pour les hautes fréquences. La fréquence de coupure a été choisie de manière à ce que le codeur hybride opère dans la zone de débits où les performances des codeurs

stéréophoniques et monophoniques ne sont pas optimales. Notons que ce codeur hybride peut être considéré comme proche de la combinaison des techniques de la stéréo jointe (*Mid-Side* + *Intensity Coding*) comme le propose le codeur AAC.

Les résultats de tests subjectifs donnés dans [FAL06a], démontrent la nette amélioration apportée par la synthèse de la cohérence avec les filtres de réverbération tardive pour un débit d'informations spatiales identique au cas précédent. On peut donc imaginer surélever la courbe (rouge) de la **Figure 2.13** relative à un codeur mono associé au BCC. Il a été notamment démontré dans [BRE05a] que le procédé PS est capable de générer un signal stéréo, à partir d'un signal mono (non codé) et de 5 à 8 kbps d'informations spatiales, dont la qualité est équivalente à celle d'un signal stéréo codé par le codeur MPEG-1 couche III (MP3) à 128 kbps.

2.3 Codage des signaux audio multicanaux

En exploitant les principes du codage audio perceptuel et ceux des techniques de codage stéréophonique, les standards du codage audio multicanal sont destinés à la transmission efficace de signaux au format 5.1 (soit 6 canaux discrets *cf.* paragraphe 1.2.2) en particulier. Nous présentons dans cette section l'évolution des procédés et des standards qui ont vu leurs performances s'accroître (en termes de débit pour plusieurs niveaux de qualité audio et spatiale) durant ces dernières années.

Cette section introduit d'abord, au paragraphe 2.3.1, les premiers systèmes dit de « matriçage » qui ont permis la diffusion du son multicanal par réduction du nombre de canaux utiles. Au paragraphe 2.3.2, nous présentons les standards du codage audio multicanal qui exploitent les principes de la compression audio. Enfin, le paragraphe 2.3.3 aborde les principes de modélisation et de compression utilisés par une technologie émergente basée sur un codage audio « spatial » du son multicanal.

2.3.1 Réduction des redondances par matriçage

2.3.1.1 L'encodeur Dolby *Stereo*

L'UIT-R préconise un jeu de coefficients fixes au court du temps pour générer un signal stéréophonique le plus réaliste possible (*cf.* Annexe C.1) en comparaison au signal multicanal d'origine. Dans un contexte de compression audio multicanal, d'autres coefficients ont été définis pour réduire le nombre de canaux à transmettre tout en favorisant l'approximation du signal multicanal original.

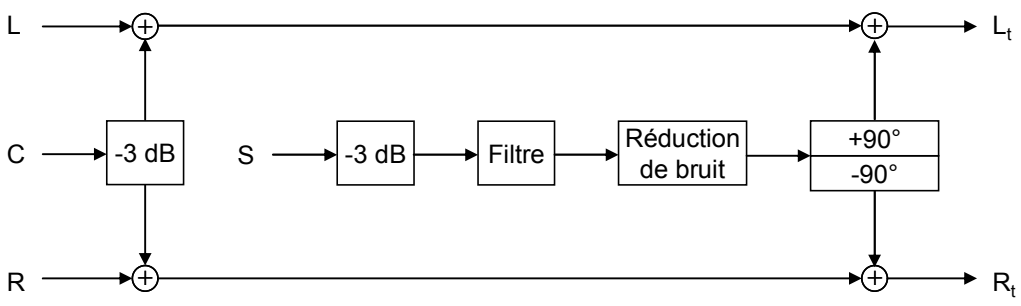

Figure 2.14 Encodage (4 canaux vers 2 canaux) de type Dolby *Stereo* - tiré de [DRE00]

Le premier système d'encodage du son multicanal dit Dolby *Stereo* réalise un downmix assurant la compatibilité stéréophonique en ne considérant que quatre canaux en entrée : le canal gauche (*L*), le canal central (*C*), le canal droit (*R*) et un seul canal *surround* (*S*). Le schéma de principe de l'encodeur *Dolby Stereo* est présenté à la **Figure 2.14**.

Les canaux L_t et R_t (« *t* » pour total) constituent alors l'information utile à transmettre relative à la fois au contenu audio (sources sonores) mais également à l'environnement sonore diffusé par les haut-parleurs arrières des systèmes de restitution présentés à la **Figure 1.15**. Le matriçage *Dolby Stereo* (*cf.* **Figure 2.14**) s'écrit de manière simplifié :

$$\begin{cases} L_t = L + \frac{1}{\sqrt{2}}\left(C + \Delta\varphi \cdot S\right) \\ R_t = R + \frac{1}{\sqrt{2}}\left(C - \Delta\varphi \cdot S\right) \end{cases}, \tag{2.25}$$

où le facteur $\Delta\varphi$ désigne un déphasage de 90°. Alors que le canal central est réparti également dans les canaux L_t et R_t (avec une réduction de gain de -3 dB), un traitement particulier est appliqué au signal surround (*S*) dont les étapes successives sont :

- une atténuation de gain à -3 dB,
- une réduction de la bande passante à la plage des fréquences [100 ; 7000] Hz,
- une réduction du niveau de bruit potentiellement présent par un algorithme propriétaire (*Dolby B-type noise reduction* [DRE00]),
- un déphasage total de 180° entre les composantes de *S* introduites dans L_t et R_t.

Ainsi, la mise en quadrature des signaux présents à la fois dans *S* et dans *L* ou *R* permet d'éviter de les annuler mutuellement et en préserve l'énergie totale. L'intérêt et les limitations de ce matriçage sont présentés à partir de l'analyse de l'opération de décodage passif Dolby *Surround*, au paragraphe 2.3.1.2, et actif Dolby *Pro Logic*, au paragraphe 2.3.1.3. A la différence des procédés d'*upmix* en aveugle décrits en Annexe C.2, les décodages Dolby *Surround* et *Pro Logic* visent à reconstruire le signal multicanal d'origine à partir d'un signal stéréo matricé.

2.3.1.2 Le décodage passif Dolby *Surround*

Les décodeurs de « première génération » nommés *Dolby Surround* consistent en un simple matriçage passif qui est décrit par le système d'équation suivant :

$$\begin{pmatrix} L' \\ R' \\ C' \\ S' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} L_t \\ R_t \end{pmatrix} \tag{2.26}$$

Des modules de traitement supplémentaires n'apparaissent pas sur le schéma de principe simplifié de la **Figure 2.15**. Notamment, un délai ajustable (entre 20 et 30 ms) peut être introduit entre les canaux frontaux et le canal *surround*. Ainsi, la localisation de sources provenant de l'image frontale peut être améliorée d'après la loi du premier front d'onde décrite au paragraphe 1.1.2.

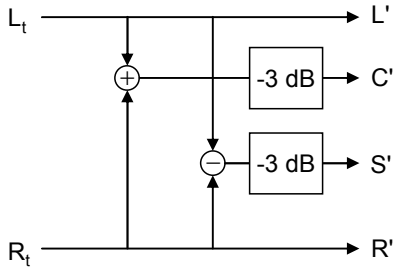

Figure 2.15 Décodage simplifié (2 canaux vers 4 canaux) de type *Dolby Surround*.

Ce système purement passif, décrit dans [DRE00], est capable d'assurer une séparation correcte (pas de mélange d'informations) entre les canaux « diamétralement opposés » (entre *C'* et *S'* ou *L'* et *R'*). Cependant, il offre une piètre séparation de 3 dB (effet de diaphonie) entre les canaux adjacents (entre *L'* ou *R'* et *C'* ou *S'*), qui se traduit par une dégradation de la qualité spatiale en termes de localisation et d'impression spatiale (largeur de l'image frontale réduite et effet de salle mal reproduit). En outre, une forte limitation de cette technique de matriçage provient de l'association implicite de ce procédé à une organisation particulière de l'image sonore : les sources sonores sont délivrées par l'image frontale et l'effet de salle par l'image *surround* (signal multicanal de type I défini au paragraphe 4.1.2.2). Sous cette hypothèse, la somme des canaux L_t et R_t peut en effet générer un canal central *C*' composé principalement des sources primaires alors que leur différence, *S*', approxime l'ambiance du signal multicanal original.

2.3.1.3 Le décodage actif Dolby *Pro Logic*

Le décodeur de « seconde génération » *Dolby Surround Pro Logic*, décrit dans [DRE00], vise à améliorer la séparation entre les canaux adjacents et ainsi renforcer les effets directionnels. Il est donc nécessaire de modifier les coefficients de la matrice (2.26) en fonction du contenu les signaux L_t et R_t, ou plus précisément en fonction de l'information directionnelle suggérée par les signaux issus du décodage passif (*cf.* **Figure 2.15**). Il s'agit donc d'un décodage actif mettant en jeu une matrice adaptative (*steering logic*), principe que l'on retrouve avec quelques variantes notamment dans le système *Logic 7* [GRI96].

Le renforcement directionnel se base sur la détection d'une direction prédominante à tout instant, définie d'après les rapports des niveaux sonores (en dB) entre L' et R' et entre C' et S' *i.e.* entre l'image gauche/droite et l'image avant/arrière, tels que

$$\begin{cases} \text{DominanceL/R}=20\times\log_{10}\left(\dfrac{L_t}{R_t}\right) \\ \text{DominanceC/S}=20\times\log_{10}\left(\dfrac{L_t+R_t}{L_t-R_t}\right) \end{cases}. \qquad (2.27)$$

Un angle relatif à la direction du son dominant et son amplitude peuvent être dérivés du système d'équations (2.27) pour fournir une représentation vectorielle de la dominance présentée à la **Figure 2.16**.

Dit simplement, le système *Pro Logic* vise à mettre en avant l'information dominante et à atténuer l'information restante dans les canaux décodés. Dans le cas d'une détection frontale par exemple, le signal *C*' est retranché aux signaux *L*' et *R*' afin qu'il soit diffusé uniquement par le haut-parleur central (« procédé d'annulation » [DRE00]) et les autres canaux sont atténuées. Le danger associé au décodage actif est la fluctuation intempestive des qualités sonores et spatiales : la dégradation et l'instabilité des « arrière-plans sonores » au profit d'une source jugée prépondérante à un instant donné, les variations du niveau sonore (effets « de pompe »)... autant de manifestations qui se révèlent rapidement gênantes et fatigantes en

dépit d'un effet de focalisation espéré sur l'événement prépondérant. Pour définir une action plus pertinente, les décodeurs doivent tenir compte de paramètres dynamiques supplémentaires – degré de prédominance, niveau sonore et rapidité de leur variation – en fonction desquels différents modes d'action sont déclenchés [DRE00]. Cela ne suffit pas en général à assurer une restitution fiable et cohérente par rapport au signal multicanal original. Il semble capital, particulièrement avec le *Dolby Pro Logic*, que le travail de production (mixage) soit réalisé en fonction de la chaîne de codage-décodage.

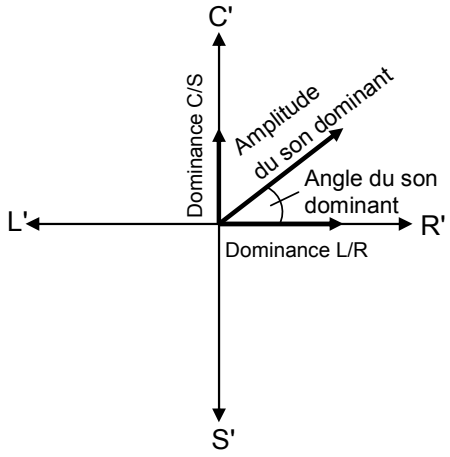

Figure 2.16 Vecteur dominance défini pour le décodage actif (2 canaux vers 4 canaux).

Dans un contexte de diffusion musicale, aux exigences esthétiques souvent plus contraignantes que pour le cinéma, Griesinger [GRI96] énonce quelques critères essentiels pour un système *surround* performant. L'équilibre énergétique des sources (niveaux sonores relatifs) doit être préservé, l'effet de localisation original doit être conservé au mieux, enfin, la diffusion spatiale du champ sonore d'arrière-plan exige une bonne décorrélation latérale des canaux surround pour être maximale.

Dans un contexte d'*upmix* des signaux stéréophoniques, nous présentons en Annexe C.2, comment un matriçage fixe de type somme-différence peut être judicieusement remplacé par un matriçage adaptatif des canaux basé sur l'ACP (équivalent au décodage par rotations employé par le procédé décrit au paragraphe 2.2.2) différent de celui proposé par le décodage Dolby *Pro Logic*.

2.3.1.4 Performances des procédés de codage basé sur un matriçage des canaux

Les performances de plusieurs procédés ont été évaluées au moyen de tests subjectifs menés par Rumsey [RUM01]. L'expérimentation vise à comparer les performances de quatre algorithmes (présentés de façon anonyme) lorsque ils sont utilisés pour convertir un signal stéréophonique original (non-issu d'un matriçage particulier) en un signal multicanal au format standard 5.0[7]. Les résultats démontrent que l'ensemble des sujets expriment percevoir une qualité réduite de l'image frontale (largeur et profondeur de l'image réduite). Par contre, l'impression spatiale globale est sensiblement améliorée et préférée par les sujets notamment grâce à l'enveloppement sonore procuré. Ces résultats n'étant pas particulièrement sensibles à la méthode de matriçage utilisée [RUM01].

[7] Cette conversion qui opère en aveugle est également dénommée *upmix*, nous présentons en Annexe C.2 deux méthodes particulières qui ne sont pas directement évaluées par Rumsey dans [RUM01].

2.3.2 Les standards du codage audio multicanal

A partir du début des années 90, les activités du groupe de normalisation MPEG ont progressivement évolué du codage audio mono et stéréo vers le codage audio multicanal. MPEG-2 audio a été l'un des premiers standards à prendre en considération le codage des signaux au format 5.1. En premier lieu, la norme MPEG-2 audio a proposé un codage audio multicanal compatible avec la norme MPEG-1 audio et les systèmes monophoniques et stéréophoniques : il s'agit du codec MPEG-2 BC (*Backward Compatible*), décrit dans [BOS02]. Le cœur du codec est semblable au cœur des codecs de la norme MPEG-1 audio avec, de plus, l'utilisation du procédé de codage d'intensité (dénommé *dynamic crosstalk*) et de filtres de prédiction adaptatifs pour extraire l'information non-redondante du canal central (*C*) et des canaux surround (*Ls* et *Rs*) [FUC93]. Pour permettre à un décodeur MPEG-1 de décoder le flux au format MPEG-2 BC, l'encodeur MPEG-2 BC utilise un matriçage équivalent à ceux définis en Annexe C.1 pour opérer le *downmix* du signal multicanal en un signal stéréo. Finalement, le flux binaire au format MPEG-2 BC contient à la fois le flux du signal matricé codé au format MPEG-1 et le flux de l'extension multicanale relative au codage des canaux *C*, *Ls* et *Rs*. D'après [BOS02], les tests menés suivant les spécifications de la recommandation de l'UIT-R BS.1116 [UIT1116] ont montré qu'une bonne qualité peut être obtenue avec un débit moyen de 640 kbps pour le codage MPEG-2 BC (couche II) des signaux au format 5.1. A partir de 1994, les acteurs en normalisation MPEG ont cherché à définir une méthode de codage audio multicanal plus performante en relâchant la contrainte de compatibilité avec les systèmes stéréophoniques. A la même période, les laboratoires Dolby développent un concurrent direct aux codecs MPEG avec le codec AC-3 également connu sous son appellation commerciale *Dolby Digital*.

Dans cette section, nous introduisons les principes utilisés par le codec AC-3 et sa flexibilité vis-à-vis de l'utilisateur. En second lieu, nous présentons l'évolution du codec AAC, initialement défini dans la norme MPEG-2 audio et ensuite intégré au cœur de la norme MPEG-4. Nous verrons comment les techniques de la stéréo jointe ont été remplacées, avec succès, par un matriçage des canaux basé sur la transformation de Karhunen-Loève.

2.3.2.1 Principes du codec AC-3 (*Dolby Digital*)

Le codec AC-3, décrit dans [DAV95], s'est d'abord imposé comme le standard pour le son multicanal au cinéma (*Batman Returns* en 1991). Avec l'apparition des systèmes de restitution multicanal « grand public », l'AC-3 s'est ensuite introduit chez les particuliers par le biais de décodeurs spécifiques utiles au visionnage des DVD-Vidéo. De plus, le codec AC-3 constitue le standard en matière de télévision numérique haute-définition [ATS01] aux Etats-Unis.

Codage multicanal

Outre l'utilisation d'un banc de filtres particuliers (*Time-Domain Alisiasing Cancellation* TDAC [PRI86]) dont la résolution temps-fréquence dépend de la détection ou non d'une composante transitoire, et d'un modèle psychoacoustique qui diffère sensiblement de celui des codeurs normalisés MPEG, la principale caractéristique du codec AC-3 réside dans l'exploitation des redondances inter-canal avec l'utilisation du codage somme-différence dénommé *Rematrixing* et du codage d'intensité dénommé *Channel Coupling* [[FIE96] (*cf.* paragraphe 2.2.1 pour les détails sur les techniques de la stéréo jointe). En réalité, ces deux techniques sont combinées seulement dans le cas d'un codage stéréophonique. Lorsque le signal à coder présente plus de deux canaux, seul le procédé *Channel Coupling* est utilisé par l'AC-3. Cet outil de codage opère dans le domaine spectral (sur les coefficients MDCT) et repose sur l'hypothèse que les indices de la localisation auditive dépendent principalement de l'énergie de l'enveloppe spectrale des signaux et non sur leur structure temporelle. Cette hypothèse étant valable pour les hautes fréquences, le module de *Channel Coupling* est

appliqué à tous les coefficients spectraux des canaux au-dessus d'une certaine fréquence (*coupling frequency*) qui peut évoluer d'une trame à une autre.

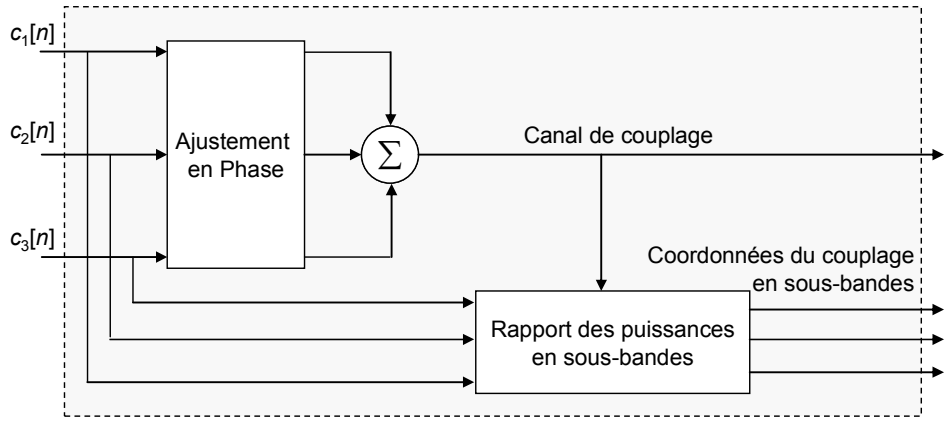

Figure 2.17 Exemple du couplage de 3 canaux par le codeur AC-3, d'après [**Erreur ! Source du renvoi introuvable.**].

Comme indique la **Figure 2.17**, le couplage des canaux consiste d'une part à générer un signal somme (canal de couplage) après un ajustement en phase des canaux (fonctionnalité optionnelle) pour assurer la conservation de l'énergie totale. D'autre part, le signal de couplage est transmis au décodeur avec des informations complémentaires dénommées coordonnées de couplage qui correspondent aux rapports d'énergie des spectres en sous-bandes entre chaque canal d'entrée et le canal de couplage [**Erreur ! Source du renvoi introuvable.**]. Le nombre de sous-bandes spectrales peut varier entre 1 et 18 mais typiquement, 14 sous-bandes sont utilisées. Finalement, le décodeur AC-3 réalise l'opération inverse (découplage) à partir du canal de couplage et des coordonnées de couplages quantifiées et transmises par l'encodeur pour reconstruire les canaux originaux, du moins leur enveloppe spectrale au-delà de la fréquence de couplage. L'AC-3 utilise, en outre, une stratégie de codage des enveloppes spectrales des canaux de couplage et des canaux codés individuellement qui repose sur une représentation à virgule flottante (exposant-mantisse) dont les détails sont présentés dans [**Erreur ! Source du renvoi introuvable.**].

Ainsi, à partir d'un signal multicanal 5.1 au format PCM (sur 16 bits et échantillonné à 48 kHz), le débit d'origine 5×16×48000 + 1×16×240 = 3,84 Mbps (5 canaux large bande et un canal basses fréquences inférieures à 120 Hz) est réduit d'un facteur 10 pour atteindre le débit de prédilection du codec AC-3, soit 384 kbps (typiquement 192 kbps pour la stéréophonie). Les résultats de test d'évaluation de la qualité délivrée par les codeurs AC-3 et MPEG-2 BC et AAC sont présentés dans [SOU98] pour un codage stéréophonique. Les résultats montrent d'une part que le codec AC-3 délivre une qualité supérieure à celle du codec MPEG-2 (couche II) pour des débits équivalents (128, 160 et 192 kbps). D'autre part, l'AC-3 à 192 kbps délivre une qualité équivalente à celle du MPEG-2 AAC (*cf.* paragraphe 2.3.2.2) opérant au débit de 160 kbps.

Flexibilité pour l'utilisateur

La flexibilité du codec AC-3 se mesure d'une part par la possibilité donnée à l'utilisateur de décoder le flux binaire de multiples manières à partir d'un décodeur *Dolby Digital*. Comme l'illustre la **Figure 2.18**, le flux binaire est rendu compatible avec les systèmes de restitution mono/stéréo par *downmix* du signal multicanal décodé (décodeurs « C » et « D »). De plus, la compatibilité avec le système *Dolby Surround* (*cf.* paragraphe 2.3.1.2) est possible à partir du signal stéréo issu du *downmix* (*Dolby Stereo* décrit au paragraphe 2.3.1) de la scène sonore multicanale (décodeur « B »). Enfin, si l'utilisateur dispose d'un système de restitution 5.1, comme celui présenté à la **Figure 1.15**, le décodeur *Dolby Digital* réalise le décodage (« A ») inverse de l'opération qui a généré le flux AC-3.

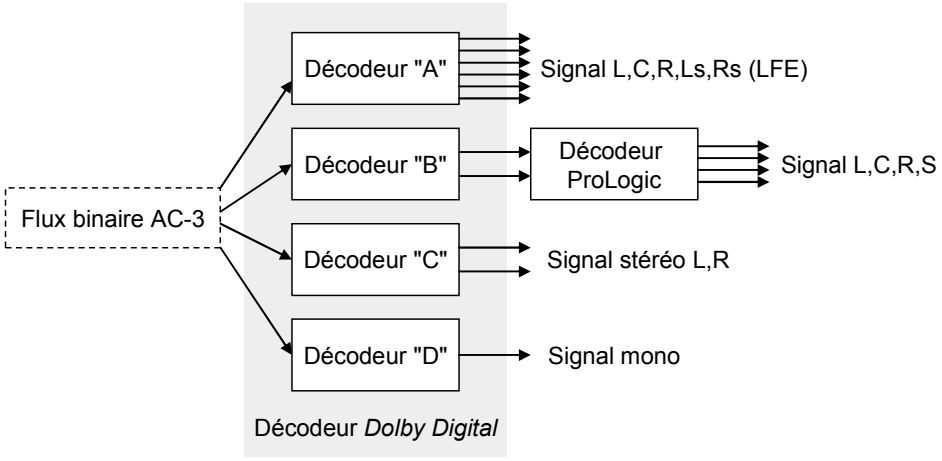

Figure 2.18 Décodages multiples du flux binaire AC-3 par un décodeur *Dolby Digital*.

En outre, le flux AC-3 contient des informations auxiliaires liées aux options de multi-langages, de protection de contenu (*copyright*), de normalisation du dialogue et plus particulièrement des paramètres de contrôle de la dynamique des signaux. L'outil *Dynamic Range Control* définit des paramètres de contrôle sur le niveau sonore des canaux décodés (précision de 0,25 dB toutes les 5,3 ms) [BOS02]. Ainsi, les sons quasi-inaudibles, détectés par l'encodeur, peuvent être accentués, au décodeur via les paramètres de contrôle, pour être perçus à un niveau équivalent à celui du dialogue notamment pour les contenus audio cinématographiques.

2.3.2.2 Principes du codec MPEG-2/4 AAC

Avec le relâchement de la contrainte de compatibilité avec MPEG-1 audio, le codec MPEG-2 AAC [ISO13818] initialement nommé *Non Backward Compatible* (MPEG-2 NBC) a une efficacité supérieure à celle du MPEG-2 BC (meilleure qualité pour un débit divisé par 2). La norme est constituée de trois profils : un profil principal (MP, *Main Profile*), un profil à complexité réduite (LC, *Low Complexity*) et un profil permettant de générer un train binaire hiérarchique (SSR, *Scalable Sampling Rate*). Plusieurs outils sont définis, certains étant optionnels suivant la configuration (profil) utilisée.

La commande de gain (*Gain Control*) permet la séparation du signal d'entrée en quatre sous-bandes avec possibilité d'abandonner des bandes en cours de transmission. La transformation du signal d'entrée ou du signal issu des sous-bandes de commande de gain (seulement pour le profil SSR) est obtenue par MDCT sur 2048 ou 256 échantillons (commutation de fenêtres en fonction de la stationnarité du signal). La nouvelle fonctionnalité introduite dans la norme permet la mise en forme temporelle du bruit de quantification (TNS, *Temporal Noise Shaping*) pour réduire les effets de démasquage lors du décodage des composantes transitoires. L'idée repose sur le fait que la prédiction d'un signal dans le domaine temporel (respectivement spectral) permet la représentation de son enveloppe fréquentielle (respectivement temporelle). Basé sur cette dualité, les signaux en sous-bandes sont filtrés par prédiction linéaire des composantes fréquentielles [BOS97]. Ainsi l'enveloppe temporelle du bruit de quantification est adaptée à l'enveloppe temporelle du signal à coder même s'il est non-stationnaire. En outre, le codec MPEG-2 AAC utilise un modèle psychoacoustique un peu plus évolué que celui utilisé par les codecs de la norme MPEG-1 audio (*cf.* Annexe B.2) à savoir le *MPEG Psychoacoustic Model* 2, décrit dans [BOS02], utile à la quantification des coefficients spectraux. Pour réduire les redondances, le procédé de codage utilise les techniques de la stéréo jointe explicitées au paragraphe 2.2.1. Le codage somme-différence (*cf.* paragraphe 2.2.1) est appliqué aux paires de canaux définies à partir des positions diamétralement opposées par rapport à l'axe séparant les haut-parleurs frontaux des haut-parleurs arrières (*L-Ls*, *R-Rs* notamment). Le codage d'intensité

est étendu, comme dans le cas du codec AC-3 (*cf.* paragraphe 2.3.2.1), mais cette fois-ci en partageant les facteurs d'échelles (rapport d'énergie du couplage) entre plusieurs paires de canaux [BOS97] *i.e.* les informations redondantes sont mutualisées au maximum.

Le codec MPEG-2 AAC supporte jusqu'à 48 canaux en entrée et délivre un train binaire au débit maximal allant de 48 à 576 kbps par canal (typiquement 64 kbps) pour des fréquences d'échantillonnage allant de 8 à 96 kHz. Les tests d'évaluation de la qualité délivrée par le codec MPEG-2 AAC pour coder de la musique au format 5.0 (5 canaux avec la bande de fréquence complète) [BOS97] ont montré que ce système de codage multicanal est transparent pour un débit de 320 kbps. De plus, le codec MPEG-2 AAC opérant à 320 kbps offre une qualité supérieure à celle du MPEG-2 BC au débit double de 640 kbps.

De MPEG-2 audio vers MPEG-4 audio

Les normes ISO/IEC MPEG-4 audio version 1 et 2 ont été formulées respectivement en 1998 et 1999. Cette norme a été définie pour permettre la composition et la manipulation de divers objets sonores au sein d'une même application [ISO14496-1]. Pour cette raison, l'approche générique utilisée pour les normes MPEG-1 audio et MPEG-2 audio n'a pu être poursuivie. MPEG-4 audio [ISO14496-3] propose l'utilisation d'un codage adapté en fonction du contenu (parole, musique) et des fonctionnalités souhaitées (compression, scalabilité, résistance aux erreurs, spatialisation, manipulation etc.). Dans ce contexte, le codec AAC, initialement défini par la norme MPEG-2, a été choisi pour la compression des signaux audio en raison de ses excellentes performances. La notion de scalabilité correspond au fait qu'un décodeur MPEG-4 audio peut générer un signal audio intelligible à partir d'une partie (et non l'intégralité) du train binaire de manière à répondre efficacement aux besoins des applications à débit contraint (téléphonie mobile par exemple). Pour y parvenir le codage dit scalable d'un signal audio est réalisé de manière hiérarchique : le cœur du codeur délivre un flux bas débit contenant l'information principale à transmettre puis les informations manquantes sont codées et transmises en complément et cela de manière récursive. Cette imbrication des codeurs résulte en une augmentation du débit total par palier (notion de granularité du débit). Ainsi, suivant la capacité de la chaîne de transmission, le décodeur peut traiter une portion ou l'intégralité du train binaire.

La norme MPEG-4 audio [ISO14496-3] propose de multiples schémas de codage pour le codage de la parole avec le codeur paramétrique HVXC (*Harmonic Vector eXcitation Coding*) opérant à bas débit entre 2 et 4 kbps et les codeurs de parole multi-débits CELP (*Code-Excited Linear Prediction*) bande étroite et bande élargie opérant à des débits couvrant la gamme 3,85 à 24 kbps. En outre, la norme MPEG-4 permet la transmission de signaux de parole à très bas débit sous forme textuelle avec l'interface *Text-To-Speech* et la synthèse sonore algorithmique ou par table d'ondes (d'instruments de musique) sous l'appellation SA (*Structured Audio*). Enfin, le codage AAC cohabite avec la Twin-VQ (*Transform Weighted INterleave-Vector Quantization*) qui est une technique de codage par transformé (codage prédictif) associée à la quantification vectorielle des coefficients normés et entrelacés [NAO95]. La Twin-VQ a montré de meilleures performances que l'AAC pour les faibles débits inférieurs à 16 kbps (débit minimal 6 kbps) d'après [BOS02]. Pour la même gamme de débit, le codec HILN [PUR97] a également été incorporé à la norme pour son approche paramétrique utile à la transmission de signaux musicaux à de faibles débits avec une qualité intermédiaire. La hiérarchisation des codeurs pour la scalabilité propose d'utiliser un codeur cœur de type CELP ou Twin-VQ puis des modules d'AAC d'amélioration de la qualité par augmentation de la bande transmise lorsque le débit disponible augmente. Pour l'interopérabilité des plates-formes (mobile, PC, etc.), MPEG-4 audio définit plusieurs profils basés sur les principales applications : « *Main - Scalable - Speech - Synthetic Audio - High Quality Audio - Low-Delay Audio - Natural Audio - Mobile Audio Internetworking - Error Protection* ». Enfin, signalons que des outils de composition, décrits dans la partie système de la norme [ISO14496-1], permettent la composition et la spatialisation de divers objets sonores dans une scène sonore complexe au format MPEG-4. La **Figure 2.19** présente un décodeur complet suivant la norme MPEG-4. Le décodeur inclut à la fois les décodeurs des objets audio définis dans la partie audio de la norme [ISO14496-3] et les outils de

démultiplexage et de composition de scène sonore décrits dans la partie système du standard [ISO14496-1] et dans [VAA04].

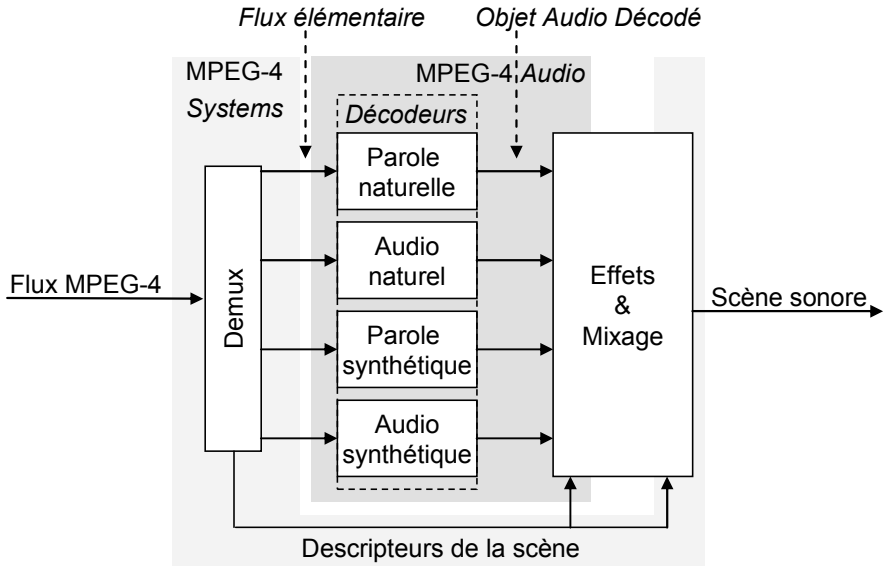

Figure 2.19 Schéma de principe d'un décodeur MPEG-4, d'après [WOL03].

Au début des années 2000, des améliorations ont été apportées au schéma de codage AAC. La première amélioration est inspirée du codage de la parole qui utilise couramment un codage paramétrique des composantes de bruit. Ce principe a été appliqué au codage audio avec succès sous l'appellation PNS, *Perceptual Noise Substitution*. Les composantes de bruit détectées ne sont pas considérées par l'étape de quantification/codage, seule leur intensité est transmise au décodeur qui peut alors injecter le niveau de bruit voulu pour les composantes fréquentielles considérées. En parallèle, les outils de prédiction ont été améliorés pour la prédiction à long terme des signaux stationnaires (LTP, *Long Term Prediction*) mais l'amélioration qui apporte la plus grande réduction du débit est basée sur la réduction de la bande effective à coder par le codeur MPEG-4 AAC associé à un outil d'extension de bande : SBR pour *Spectral Band Replication*.

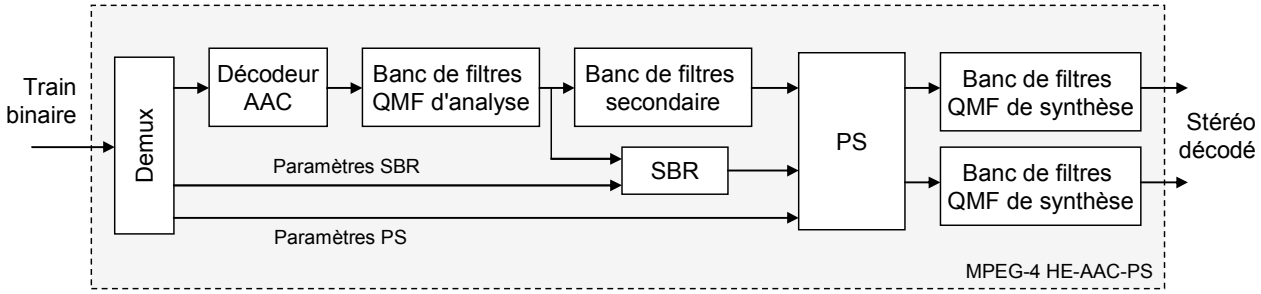

Figure 2.20 Structure du décodeur stéréophonique MPEG-4 HE-AAC-PS, d'après [BRE05a].

L'association du procédé SBR au codeur MPEG-2/4 AAC s'appelle *High Efficiency Advanced Audio Coding* ou HE-AAC [WOL03]. L'outil SBR a été mis en œuvre pour répondre à un besoin exprimé en partant de l'observation que le bruit de codage dépasse le seuil de masquage fréquentiel (*cf.* Annexe B.1.3.1) en basse fréquence lorsque le codeur MPEG-4 AAC opère à très bas débit (caractéristique transposable à d'autres codeurs). La perception du bruit peut être évitée (niveau du bruit sous le seuil de masquage) en limitant la bande de fréquence du signal original à coder (typiquement entre 5 et 6 kHz). Le procédé

SBR, décrit dans [DIE02], permet d'étendre la bande du signal décodé en suivant un principe de réplication de la « bande basse » vers la « bande haute » en utilisant un filtre blanchisseur, un générateur de bruit et de fréquences pures qui sont paramétrés par les informations auxiliaires (niveau du bruit, des fréquences pures, etc.) extraites et transmises par l'encodeur. Les tests subjectifs, décrits dans [WOL03], ont démontré que le codec MPEG-4 HE-AAC stéréo opérant à 48 kbps (2,5 kbps pour les informations auxiliaires) est capable de générer un signal de qualité légèrement supérieure à celle obtenue par un codage MPEG-4 AAC stéréo (pleine bande) opérant à 60 kbps.

Enfin, l'outil de codage *Parametric Stereo* (PS), introduit par Breebaart et *al.* dans [BRE05a] et présenté au paragraphe 2.2.3, a été associé au codeur MPEG-4 HE-AAC pour former le codec très bas débit HE-AAC-PS ou HE-AACv2 selon le contexte d'utilisation qui lui est attribué. Comme l'illustre la **Figure 2.20**, la structure du décodeur MPEG-4 HE-AAC-PS utilise un premier banc de filtres d'analyse pour assurer la compatibilité avec l'extension de bande du procédé SBR qui opère dans le domaine transformé obtenu par le banc de filtres QMF. La synthèse stéréo paramétrique est alors réalisée dans le domaine transformé obtenu par un banc de filtres hybride équivalent à la mise en cascade des deux bancs de filtres (voir [BRE05a] pour les détails d'implémentation). Les paramètres spatiaux offrent la possibilité de réduire le débit de codage global pour la transmission d'un signal stéréo avec une fidélité équivalente. Les résultats de tests subjectifs, dans [BRE05a], montrent que le codeur HE-AAC à 32 kbps bénéficie d'une réduction de débit de l'ordre de 8 kbps sans introduire de dégradation perceptible avec l'utilisation de l'outil PS *i.e.* l'HE-AAC à 32 kbps délivre une qualité équivalente à celle de l'HE-AAC-PS à 24 kbps.

Matriçage optimisé pour la réduction des redondances

Alors qu'un matriçage particulier a été défini pour assurer la compatibilité mono/stéréo d'un flux MPEG-2 AAC, dans [SZC03], à la manière du MPEG-2 BC, Yang et *al.* dans [YAN04] ont introduit l'utilisation d'un matriçage multicanal adaptatif basé sur la transformation de Karhunen-Loève (KLT) en lieu et place du matriçage somme-différence dans le cœur du codec MPEG-2/4 AAC. Dans [YAN03]-[YAN04], les auteurs définissent un codec AAC modifié (*Modified* AAC *with Karhunen-Loève Transform* MAACKLT) puisque les coefficients MDCT issus du banc de filtres sont matricés par KLT de façon à réduire les redondances inter-canal et à compacter l'énergie des canaux d'entrée.

Le matriçage adaptatif utilisé pour le codage stéréophonique, présenté au paragraphe 2.2.2, réalise la projection des données stéréophoniques sur la base des vecteurs propres de la matrice de covariance du signal d'entrée. Cette transformation correspond à la transformation KL [LOE48] originellement inspirée par la formulation de l'ACP par Hotelling dans [HOT33]. Comme l'indique la **Figure 2.21**, appliquer ce matriçage à un signal multicanal $(C_1,...,C_M)^T$ offre la possibilité de réduire le nombre de canaux à transmettre au travers du signal multicanal $(D_1,\ldots,D_P)^T$ tel que $P \leq M$. En effet, d'après [KRA56], les coefficients du matriçage qui minimisent l'erreur quadratique moyenne (EQM) entre les canaux originaux $(c_1,...,c_M)^T$ et les canaux reconstruits $(C'_1,...,C'_M)^T$, à partir du signal multicanal $(D_1,\ldots,D_P)^T$ et du matriçage inverse, sont définis par les vecteurs propres de la matrice de covariance. De plus, ce matriçage qui s'adapte à la nature des signaux, est linéaire et, par suite, permet la reconstruction du signal multicanal original à partir de la transposition de la matrice utilisée à l'encodage puisqu'elle est orthogonale (*cf.* paragraphe 4.3 pour plus de détails). En d'autres termes, cette reconstruction minimise d'autant plus l'EQM que le nombre de canaux utilisés (transmis) pour le matriçage inverse est proche du nombre de canaux originaux ($P=M$ implique une reconstruction parfaite). Ce principe est d'ailleurs utilisé par les auteurs de [YAN04] puisqu'ils proposent une méthode de codage scalable en termes de débit et de qualité dans le sens où les canaux issus du matriçage sont transmis suivant un ordre d'importance (critère énergétique). A la manière du codage MPEG-4 scalable décrit au paragraphe 2.3.2.2, les applications à débit contraint peuvent ainsi décoder la partie du flux binaire permettant une reconstruction partielle avec une qualité intermédiaire. Dans le cas où la contrainte de débit n'est pas problématique, le train binaire est complètement décodé et la qualité restituée atteint la transparence.

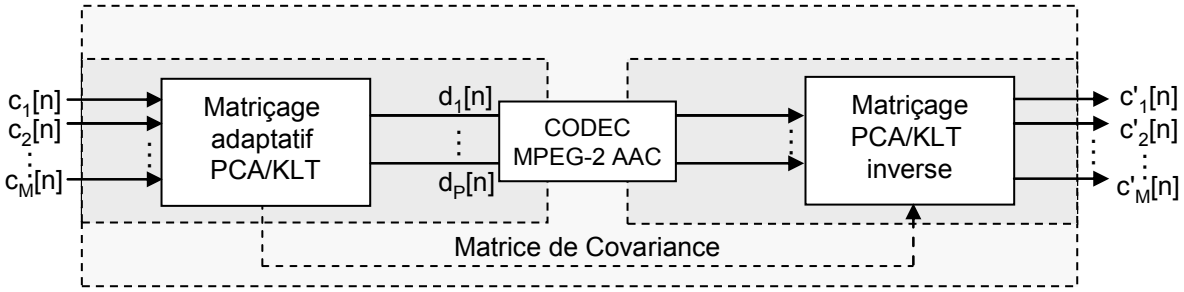

Figure 2.21 Principe du codec MAACKLT. Le codec MPEG-2 AAC est modifié pour assurer le codage des canaux matricés $(d_1,\ldots,d_P)^T$ et par la désactivation du module de matriçage somme-différence. Le matriçage adaptatif est appliqué aux coefficients MDCT des canaux d'entrée.

Les expérimentations menées par Yang et *al.* dans [YAN03] montrent que la concentration de l'énergie dans les canaux matricés augmente d'autant plus que le nombre de canaux du signal d'entrée (M) est grand. Ainsi, le gain de codage avec ce type de matriçage, nécessitant la transmission d'informations auxiliaires (covariance, vecteurs propres ou matrice de transformation), parfois réduit dans le cas de signaux stéréophoniques (*cf.* paragraphe 2.2.2), s'accroît nettement dans le cas de signaux aux formats 5.1, 7.1, etc. En outre, les auteurs montrent, à partir d'un signal multicanal provenant d'une prise de son naturelle, que l'inter-corrélation des canaux matricés est plus faible avec un matriçage appliqué aux coefficients MDCT, issus du banc de filtres du codeur AAC, qu'avec un matriçage réalisé sur les canaux d'entrée dans le domaine temporel [YAN03].

Finalement, le codec MACCKLT procède par matriçage des canaux d'entrée dans le domaine transformé à une période de l'ordre de 10 secondes qui est considérée comme le compromis optimal entre le suivi temporel du contenu des signaux (pour obtenir une décorrélation convenable) et la quantité d'informations à transmettre pour réaliser l'opération inverse (*cf.* **Figure 2.22**). D'après [YAN04], puisque la matrice de covariance est réelle et symétrique, seuls $M\times(M+1)/2$ éléments de la matrice de covariance doivent être quantifiés et transmis au décodeur pour que ce dernier puisse calculer les vecteurs propres. Typiquement, 240 bits sont nécessaires pour la transmission de la matrice de covariance, relative à une transformation, d'un signal multicanal au format 5.1.

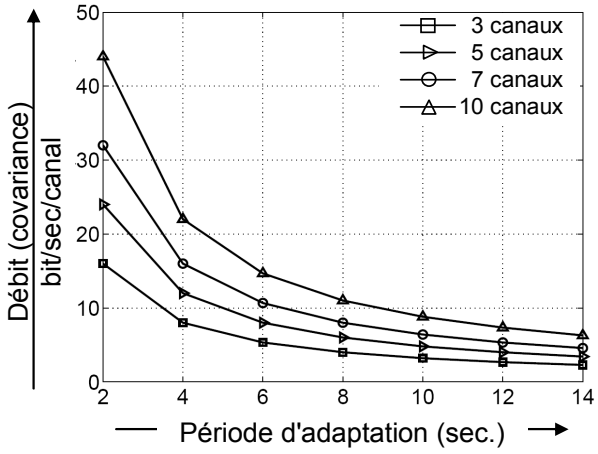

Figure 2.22 Débit des informations auxiliaires (éléments de la covariance quantifiés sur 16 bits) en fonction de la période d'adaptation utilisée pour la transformation KL. Tiré de [YAN03].

Les tests subjectifs, menés dans [YAN04], montrent que le codec MACCKLT délivre une qualité légèrement supérieure au codec MPEG-2 AAC avec un débit de 64 kbps pour le codage de signaux aux formats 5.1 (qualité globale intermédiaire ou faible puisque le MPEG-2 AAC est transparent à 64 kbps pour un codage monophonique). L'intérêt majeur de cette

méthode de codage réside en la hiérarchisation du flux binaire qui permet une transmission scalable de l'information par ordre d'importance qui résulte en la scalabilité de la qualité audio restituée.

2.3.3 Emergence de la technologie MPEG *surround*

Basé sur les avancées du codage stéréo paramétrique (BCC et PS) étendu au codage audio multicanal, le groupe de standardisation MPEG audio a démarré un nouveau sujet d'étude en mars 2004 pour enrichir la norme MPEG-4 audio d'un module de codage multicanal paramétrique dit « spatial » (*Spatial Audio Coding*, SAC) sous l'appellation MPEG *surround*.

A la différence des techniques *downmix/upmix*, décrites en Annexe C.2, qui permettent de conserver en partie (*downmix*) ou d'étendre arbitrairement (*upmix* aveugle) l'information spatiale initiale, le codage audio spatialisé utilise une description paramétrique de la scène sonore pour parvenir à la reconstruction du signal multicanal original.

2.3.3.1 Le codage audio multicanal paramétrique

Les procédés de codage stéréo paramétrique à bas débit (BCC et PS), décrits au paragraphe 2.2.3, exploitent les propriétés de notre perception spatiale du son avec l'extraction de paramètres spatiaux liés à la localisation auditive (*cf.* paragraphe 1.1) tout en réduisant les redondances avec la transmission d'un signal mono obtenu par *downmix* des canaux d'entrée. L'extension de ces procédés de codage stéréo au codage des signaux audio multicanaux propose une approche différente de celle des codeurs MPEG-4 AAC et Dolby AC-3, présentés au paragraphe 2.3.2, dans la mesure où la compatibilité avec les systèmes à moindre capacité est, par nature, assurée (*cf.* **Figure 2.23**). De plus, le codage multicanal paramétrique propose un codage unifié des canaux qui était seulement initié avec les techniques de la stéréo jointe (*channel coupling*) pour les standards du codage multicanal.

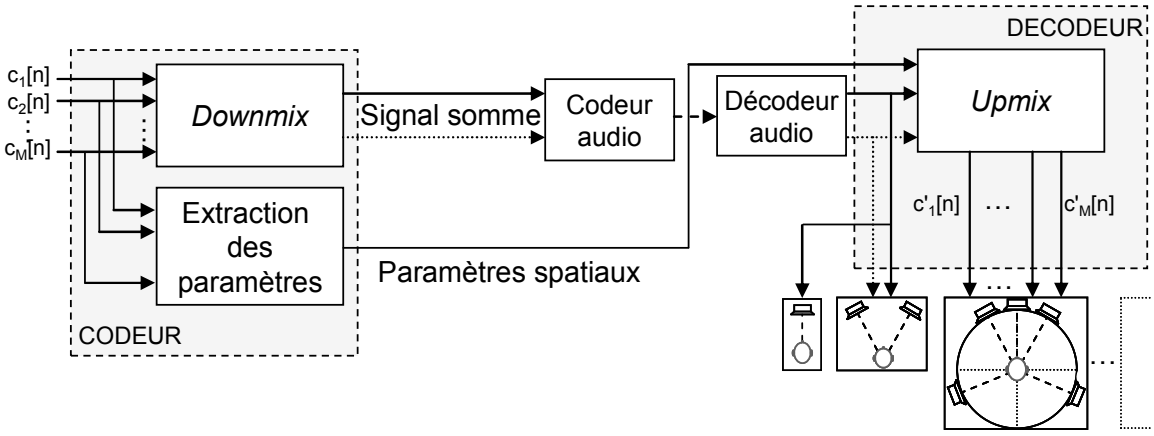

Figure 2.23 Structure d'un codec audio multicanal dit spatialisé, basé sur le procédé BCC étendu pour la norme MPEG *surround*.

Le procédé BCC, décrit dans [FAL04], propose le codage des signaux audio multicanaux de manière générique et flexible. De manière générale, un signal multicanal à M canaux peut être compressé en un signal à P canaux ($P<M$) accompagné d'un flux de paramètres spatiaux. Comme pour le codage stéréo paramétrique employant le procédé BCC, le signal multicanal est matricé en un signal somme mono ou stéréo ou à P canaux en conservant l'énergie initiale (*cf.* paragraphe 2.2.3). D'autre part, les paramètres spatiaux sont extraits non plus à partir d'une paire de canaux mais à partir de tous les canaux. Dans le cas d'un *downmix* mono, comme l'indique la **Figure 2.24**, les paramètres (ICTD, ICLD et ICC) sont extraits (dans le

domaine des sous-bandes) entre un canal de référence et tous les autres canaux. *M*-1 paramètres directionnels (ICTD, ICLD) sont extraits de cette manière alors que le débit du paramètre de corrélation, ICC, est réduit de manière pertinente en ne considérant que les deux canaux ayant le maximum d'énergie (parmi l'ensemble des sous-bandes).

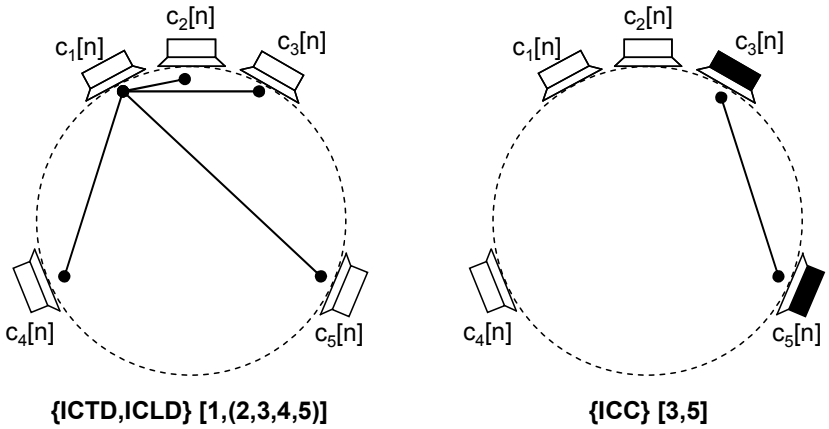

Figure 2.24 Extraction des paramètres spatiaux par le procédé BCC appliqué à un signal au format 5.1 dans le cas d'un *downmix* mono, d'après [FAL04] les indices temporels et des sous-bandes sont omis pour raison de clarté.

L'opération de décodage, définie en détail dans [FAL06a], préconise l'utilisation de *M*-1 filtres de réverbération tardive de manière à générer *M* signaux décorrélés entre eux dans le cas d'un *downmix* monophonique. Ensuite, la synthèse spatiale utilise les paramètres spatiaux transmis pour reconstruire un signal multicanal dont la qualité sera fonction du débit de codage du codec audio utilisé pour la transmission du signal somme et du débit des paramètres spatiaux.

Les travaux menés par le groupe de standardisation MPEG, pour la définition de la norme MPEG *surround*, ont consisté à définir une implémentation du codec qui autorise un large choix de débits correspondant à une large gamme de qualités subjectives. Plusieurs expérimentations ont été menées de manière à comparer les performances d'une implémentation à faible complexité (*Low Power* MPEG *surround*) et d'une implémentation haute qualité (*High Quality* MPEG *surround*) pour lesquelles le débit des informations spatiales varie entre 0 et 192 kbps voire au-delà [ISON8324].

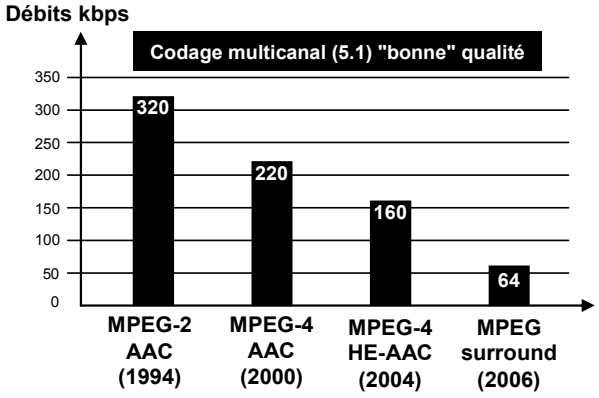

Figure 2.25 Evolution des débits des standards du codage audio multicanal relatifs à une « bonne » qualité de reconstruction des signaux au format 5.1 – tiré de [HER04].

L'objectif principal étant d'obtenir une « bonne » qualité avec un débit global (compression audio associée au débit du flux de paramètres) de l'ordre de 64 kbps pour la compression d'un signal au format 5.1. Comme l'indique la **Figure 2.25** , le débit requis pour reconstruire un signal audio au format 5.1 avec une qualité acceptable a été fortement réduit depuis le codage MPEG-2 AAC (non compatible avec MPEG-1) et sa version améliorée intégrée à MPEG-4 opérant à 220 kbps [HER04]. L'utilisation de la méthode SBR, décrite au paragraphe 2.3.2.2, au contexte du codage multicanal permet d'obtenir une qualité équivalente à un débit moindre compris entre 128 et 160 kbps [HER04]. Ainsi, l'association du codeur MPEG-4 HE-AAC avec les techniques de codage paramétrique telles que BCC et PS a servi de point de départ aux travaux de normalisation pour tenir les objectifs fixés.

2.3.3.2 Architecture de la technologie MPEG surround

Configurations de l'encodeur MPEG *surround*

L'architecture initiée par les acteurs de MPEG *surround* est présentée dans [HER05] et [BRE05b]. De manière à être compatible avec la structure du codec MPEG-4 HE-AAC, présentée à la **Figure 2.20**, l'encodeur MPEG surround utilise un banc de filtres complexe QMF appartenant à une structure hybride décrite dans [SCH04]. D'une manière générale, pour convertir un signal à M canaux en un signal à P canaux ($P<M$), l'encodeur MPEG surround réalise un encodage hiérarchique (en arbre) au moyen de modules de conversion nommés *One-To-Two element* (OTT) et *Two-To-Three element* (TTT).

Le nom des modules correspond respectivement au nombre de canaux en entrée et en sortie du point de vue de la synthèse multicanale au décodeur. Comme le montre la **Figure 2.26**, la mise en cascade de ces modules délivre une approche hiérarchique de compression/synthèse multicanale. Plusieurs configurations sont supportées par l'architecture MPEG *surround* de façon à assurer une compatibilité stéréo (configuration 525) ou mono avec une séparation avant-arrière (configuration 515-1) ou gauche-droite (configuration 515-2) des canaux d'un signal 5.1. Notons d'ailleurs que ce principe s'étend pour des signaux au format 7.1, etc. (voir le texte de la norme [ISON8324]).

Chaque module OTT extrait des paramètres spatiaux de type ICLD et ICC en considérant que les différences de temps ou de phase (ICTD-ICPD) ne sont pas significatives dans le sens où leur utilisation n'améliore pas la qualité du signal multicanal synthétisé. Un module TTT extrait les paramètres de corrélation ICC avec des paramètres énergétiques ICLD ou de prédiction CPC (*Channel Prediction Coefficient*) qui constituent un moyen idéal pour extraire un canal central à partir d'un downmix stéréo lors de la synthèse spatiale au niveau du décodeur. La quantité d'informations spatiales transmise est flexible de manière à proposer un débit variable notamment pour le paramètre ICC qui peut être transmis pour chaque sous-bande ou réduit à une seule valeur pour l'ensemble du spectre. De plus, le taux de rafraichissement des paramètres variable (résolution temporelle) et la résolution fréquentielle allant de 5 à 40 sous-bandes procurent une large palette de débits pour les paramètres spatiaux allant de 0 kbps et 32 kbps [ISON8324].

En parallèle, les modules OTT et TTT réalisent le *downmix* (matriçage) de deux ou trois canaux en un signal mono ($c_m[n]$) ou stéréo ($c^1_{st}[n]$, $c^2_{st}[n]$) dominant accompagné d'un canal résiduel ($c_s[n]$).

$$\begin{bmatrix} c_m[n] \\ c_s[n] \end{bmatrix} = \mathbf{R}_{\mathrm{OTT}} \begin{bmatrix} c_1[n] \\ c_2[n] \end{bmatrix} \tag{2.28}$$

$$\begin{bmatrix} c_{st}^{1}[n] \\ c_{st}^{2}[n] \\ c_{s}[n] \end{bmatrix} = \mathbf{R}_{\mathrm{TTT}} \begin{bmatrix} c_{1}[n] \\ c_{2}[n] \\ c_{3}[n] \end{bmatrix} \tag{2.29}$$

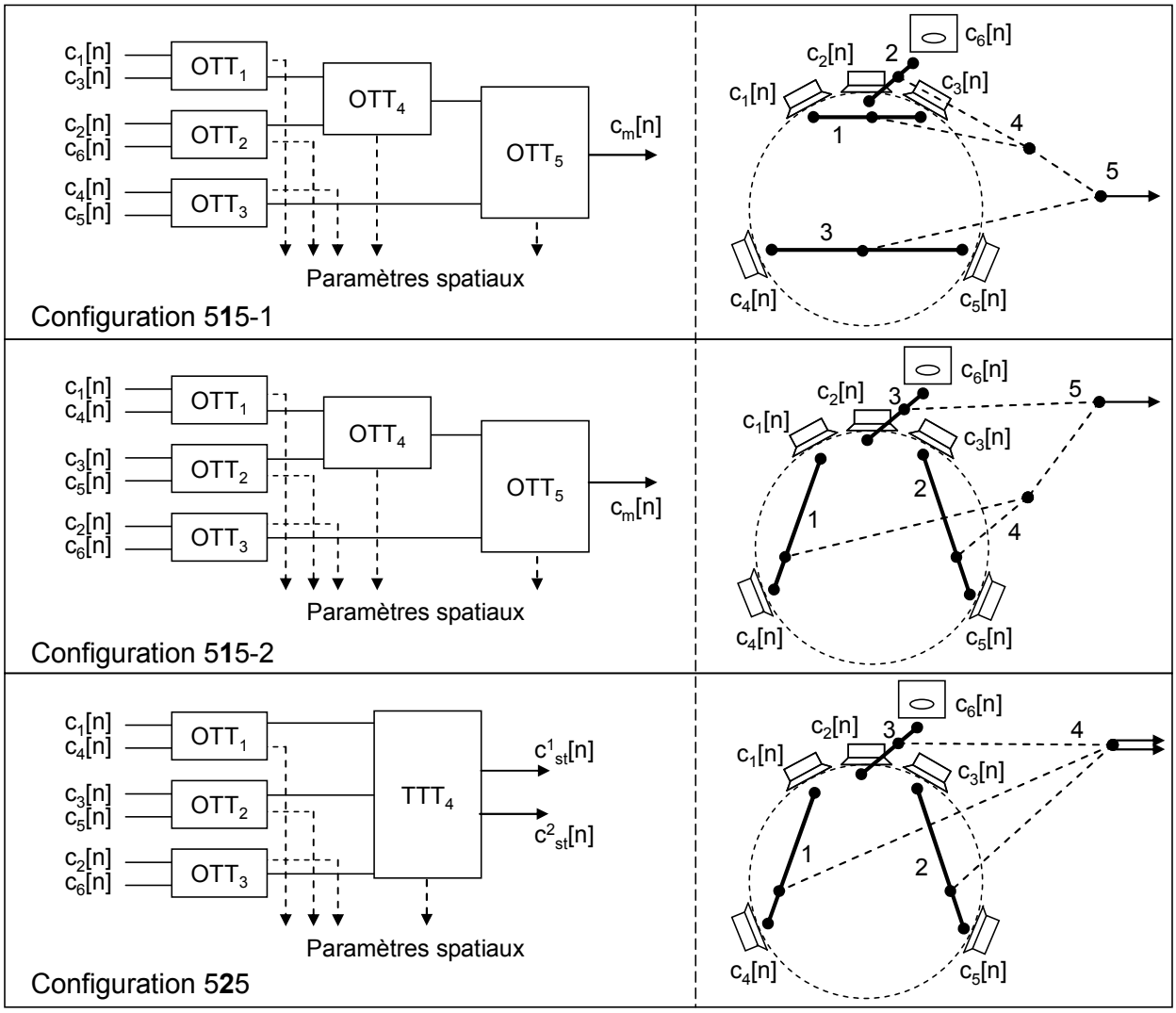

Figure 2.26 Configuration 515-1, 515-2 et 525 de l'encodeur MPEG *surround*.

Les matrices $\mathbf{R}_{\mathrm{OTT}}$ et $\mathbf{R}_{\mathrm{TTT}}$ sont définies telles que l'énergie du signal résiduel, qui peut être considéré comme l'erreur de modélisation, soit minimale étant donné les performances du modèle établies à partir des paramètres spatiaux. D'après [HER05], le signal résiduel est calculé comme étant la différence entre les signaux originaux et les signaux synthétisés à partir du signal dominant mono/stéréo, généré par *downmix*, et des paramètres extraits (décodeur local). Par exemple, dans le cas d'un matriçage TTT, le signal résiduel correspond à l'erreur de prédiction obtenue à partir des paramètres CPC et du signal stéréo issu du *downmix*. Le niveau de corrélation entre les canaux d'entrée (ICC) reflète finalement la perte d'information inhérente à la prédiction.

Etant donné que la modélisation paramétrique d'un signal multicanal ne permet pas d'obtenir une qualité audio transparente, la technologie MPEG *surround* définit un codeur basse complexité (simplification des paramètres et utilisation d'un banc de filtre complexe en basses fréquences et réel pour les hautes fréquences) et un codeur haute qualité hybride qui associe la transmission des paramètres spatiaux, du signal matricé dominant et de signaux

résiduels (relatifs à chaque module utilisé). Ces signaux résiduels sont perceptuellement encodés et la qualité globale est contrôlée par un compromis entre la largeur de la bande de fréquence des signaux résiduels et le débit associé à cette transmission d'informations supplémentaires. Typiquement, seules les composantes basses fréquences des signaux résiduels sont transmises pour assurer un décodage haute-qualité pour un débit variable pouvant atteindre 192 kbps et au-delà [ISON8324].

Précisons également qu'un post-traitement [ISON8324] peut être appliqué aux signaux matricés à partir des paramètres spatiaux de manière à rendre le système compatible avec les codeurs Dolby *Surround, Pro Logic*, *Circle Surround*, etc. (*cf.* paragraphe 2.3.1).

Synthèse spatiale par le décodeur MPEG *surround*

A la manière de l'encodeur, le décodeur MPEG *surround* transforme les canaux transmis par l'encodeur avec un banc de filtres d'analyse (QMF) appliqué à chaque canal. Ensuite, une synthèse spatiale est réalisée dans le domaine transformé à partir des paramètres spatiaux et de filtres de décorrélation. Enfin, les signaux sont synthétisés à partir d'un banc de filtres de synthèse (QMF) appliqué à chaque canal.

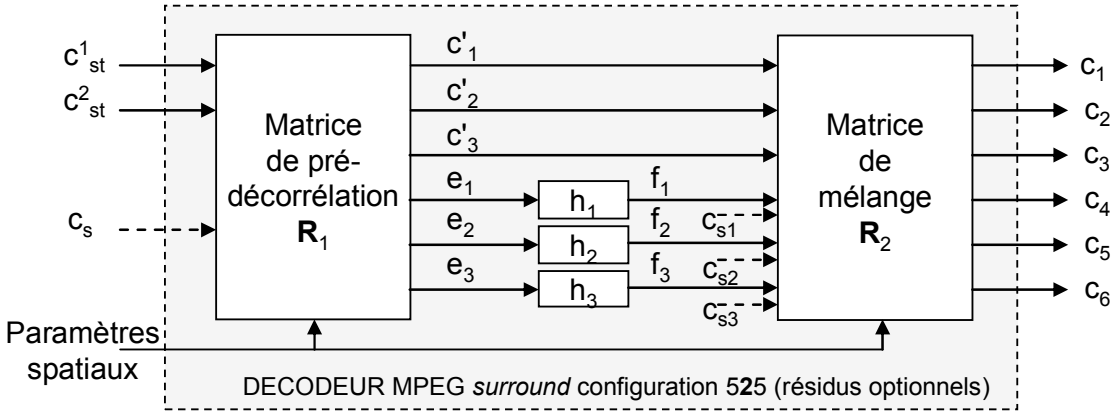

Figure 2.27 Structure du décodeur MPEG surround basé sur un signal stéréo et un flux de paramètres spatiaux.

La **Figure 2.27** présente la structure générale du décodeur MPEG *surround* dans la configuration 525. Les canaux stéréo ($c^1_{st}[n]$, $c^2_{st}[n]$) et le signal résiduel ($c_s[n]$ optionnel) en sous-bandes sont d'abord matricés par la matrice $\mathbf{R}_1$ qui correspond en partie à la matrice inverse de la matrice $\mathbf{R}_{TTT}$ utilisée à l'encodeur. Basé sur les paramètres de prédiction CPC, un canal c'_3 est généré à partir des canaux c^1_{st} et c^2_{st}. Les filtres passe-tout décorrélateurs (*cf.* paragraphe 2.2.3.2) sont intégrés aux modules OTT et TTT et opèrent dans le domaine des sous-bandes. D'après [BRE05a], ces filtres sont définis tels que le signal de sortie soit très faiblement corrélé avec le signal d'entrée sans que les enveloppes temporelles et spectrales du signal soient modifiées. Deux outils spécifiques sont utilisés pour s'assurer de ce critère de conservation des enveloppes du signal (*Guided envelope shaping* - GES et *Subband domain Temporal Processing* - STP) par mesure de l'énergie des signaux en sous-bandes [ISON8324]. De manière à réduire la corrélation de ces signaux, un étage de décorrélation est utilisé par l'intermédiaire des filtres h_1, h_2 et h_3 (réponses impulsionnelles différentes) pour être capable de synthétiser une scène sonore à ambiance diffuse (*cf.* paragraphe C.2.3). Finalement, les signaux obtenus sont matricés par la matrice de mélange $\mathbf{R}_2$ qui peut être considérée comme équivalente à la combinaison des matrices $\mathbf{R}_2$ et $\mathbf{R}_3$ utilisées par le décodeur PS présenté au paragraphe 2.2.3.2 (matrice de gains et de rotation en sous-bandes). Dans le cas où un ou plusieurs signaux résiduels (c_{s1}, c_{s2} et c_{s3}) sont transmis au décodeur, un décodage hybride haute-qualité utilise les composantes basses fréquences des signaux résiduels c_{s1}, c_{s2} et c_{s3} en lieu et place des composantes basses fréquences des signaux f_1, f_2 et f_3 (voir [HER05]-[BRE05a] pour les détails de la mise en œuvre).

Précisons également que le décodeur MPEG *surround* est capable d'opérer en aveugle dans la mesure où un décodage uniquement basé sur les signaux transmis est supporté. Le mode de fonctionnement *Enhanced Matrixed Mode* réalise finalement une synthèse spatiale à partir de paramètres spatiaux extraits du signal stéréophonique transmis par l'encodeur [ISON8324].

Enfin, le codec MPEG *surround* qui repose sur la génération automatique d'un *downmix* mono/stéréo peut également supporter un signal extérieur au processus c'est-à-dire un *downmix* dit « artistique » *i.e.* provenant du mixage réalisé par un ingénieur du son. Les traitements spécifiques pour assurer la compatibilité entre le *downmix* généré automatiquement par les modules OTT-TTT cascadés et le *downmix* artistique sont présentés dans [VIL06] et [ISON8324].

Performances du codage spatialisé MPEG surround

La technologie MPEG surround permet une large scalabilité compatible avec les applications très bas débit et pouvant aller jusqu'à délivrer un signal multicanal d'une qualité transparente par rapport à l'original et ceci pour la majorité des scènes sonores même complexes.

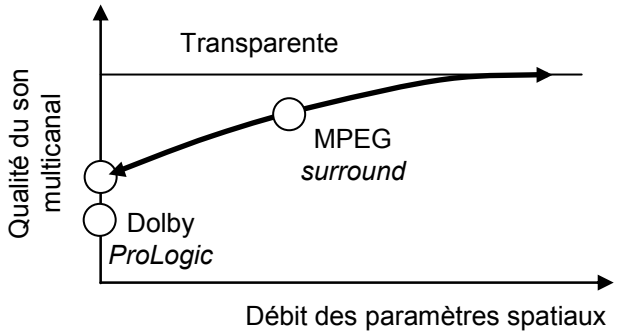

Figure 2.28 Scalabilité en termes de débit/qualité du codec MPEG *surround*: de la qualité obtenue par le mode de fonctionnement en aveugle jusqu'à la transparence.

La qualité délivrée par ce procédé de codage multicanal dépasse largement celle obtenue avec le système *Pro Logic* II (successeur du procédé décrit au paragraphe 2.3.1.3) même dans le cas d'un *upmix* aveugle (*Non-Guided*). Les résultats de tests subjectifs sont présentés dans [ISON7138] et illustrés à la **Figure 2.28**. Les résultats de tests publiés dans [BRE05b] démontrent que la technologie MPEG *surround* à 160 kbps (codeur cœur MPEG-4 AAC opérant à 128 kbps pour la compression du *downmix* stéréo) délivre une qualité supérieure à celle du MPEG-4 AAC opérant à 192 kbps.

Extension : synthèse spatiale pour une écoute multicanale au casque

Récemment, une extension a été incorporée au codec MPEG *surround* pour permettre l'écoute multicanale non seulement sur un système de reproduction multi-haut-parleurs mais également au casque en utilisant les technologies binaurales (*cf.* paragraphe 1.2.1). Deux approches sont présentées dans [VIL06] et [ISON7138]. La première est illustrée par la **Figure 2.29** où une synthèse binaurale est réalisée, au décodeur, à partir du signal stéréo décodé (configuration 525) et de la combinaison des paramètres spatiaux avec une base de filtres HRTF (*cf.* paragraphe 1.1.1.1). Notons que la synthèse binaurale est réalisée dans le domaine QMF obtenu par l'analyse en sous-bandes des canaux. De plus, deux modes de fonctionnement sont proposés (basse et haute complexité), ils reposent sur le mode d'expression des HRTF (ou BRIR). Les HRTF peuvent être exprimées de façon « paramétrique » comme un égaliseur de niveau associé à une valeur de retard ou bien d'une manière plus générale comme des filtres en sous-bandes [ISON7138]. En outre, le choix de la base de HRTF ou de BRIR (anéchoïques ou échoïques *i.e.* présence de réverbération) modélisées, influence également la qualité de la spatialisation et de l'effet de salle restitué.

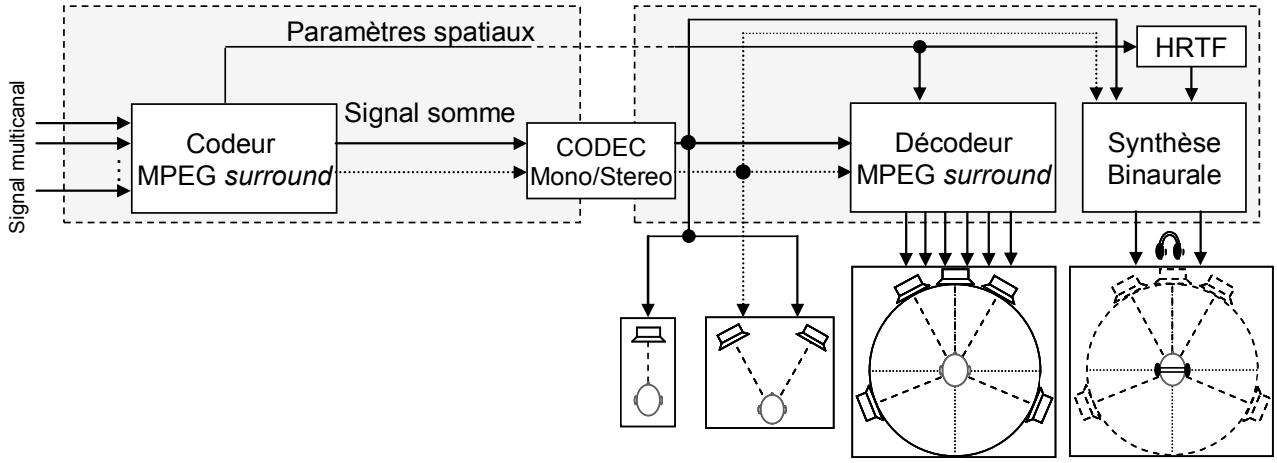

Figure 2.29 Synthèse binaurale intégrée au décodeur MPEG *surround*.

La seconde approche, décrite dans [ISON7138], repose sur une synthèse binaurale réalisée à l'encodeur. Ainsi, le signal binaural transmis et décodé permet une écoute multicanale au casque d'une manière directe. En outre, le flux MPEG *surround* reste compatible avec la stéréo classique (les HRTF paramétriques sont transmises au décodeur pour réaliser l'opération de filtrage inverse) et également avec une écoute multicanale sur haut-parleurs via l'opération de décodage complète (avec les paramètres spatiaux) décrite dans cette section.

2.4 Conclusion

Nous avons présenté dans ce chapitre les différents procédés de compression audio (mono) ainsi que l'évolution des techniques de codage stéréo. Les principes utilisés par les méthodes de la stéréo jointe ont été étendus. D'une part, la réduction des redondances par matriçage a été généralisée par la méthode des rotations en sous-bandes qui assure une concentration maximale de l'énergie (supérieure au matriçage somme-différence). D'autre part, l'extraction de paramètres directement liés aux paramètres de la localisation auditive permet la reconstruction d'un signal stéréo dont les indices inter-canaux approximent les indices originaux. Finalement, les codeurs stéréo paramétriques offrent une qualité (audio et spatiale) supérieure à celle d'un codeur stéréo classique à un débit équivalent. De plus, à qualité équivalente à celle délivrée par un codeur stéréo classique, le débit total est réduit par les procédés de codage stéréo paramétrique. Même si le codage stéréo paramétrique était initialement destiné aux applications à débits contraints comme le *streaming* audio par exemple (très bas débit pour une qualité intermédiaire), l'avantage principal donné par l'utilisation d'une représentation paramétrique est d'avoir la possibilité de manipuler la scène sonore au moyen des paramètres spatiaux [FAL06b]. En effet, ces paramètres à la fois directionnels (ICTD-ICLD) et, dans une moindre mesure, liés à l'effet de salle (ICC) permettent de contrôler la position des sources contenues dans le signal original (selon la limite posée par le recouvrement des supports fréquentiels des sources). On imagine alors la possibilité de contrôler le champ acoustique, auquel appartiennent les sources, à partir de paramètres liés aux réponses impulsionnelles de salles (limites temporelles qui séparent les différentes sections de la réponse, temps de réverbération, etc.). Cette possibilité est d'ailleurs prise en compte dans le cadre de la description de scène sonore multicanale menée pour la norme MPEG-4 BIFS (*BInary Format for Scenes*) [VAA04].

A partir des principes de la compression audio mono et stéréo, nous avons successivement présenté les différentes solutions existantes pour coder efficacement les signaux audio multicanaux. Les solutions initiales données par le matriçage Dolby *Stereo* et le codage AC-3, flexible et optimisé pour réduire les redondances, ont vu évoluer en parallèle le standard

MPEG-1 vers MPEG-2 puis MPEG-4 audio. Nous avons présenté les améliorations qui ont été apportées au codeur MPEG-2/4 AAC qui se présente désormais comme le procédé de codage le plus sophistiqué. L'utilisation conjointe de ce procédé et d'une représentation paramétrique du signal, déjà intégrée avec succès au codec MPEG-4 HE-AAC pour le codage stéréo, a été mise en oeuvre pour le codage multicanal par le codec MPEG *surround*. L'apport des paramètres liés à la localisation auditive est désormais indéniable en termes de compression et procure une qualité audio spatiale de reconstruction qui est directement liée au débit de ces informations spatiales. Finalement, les acteurs du groupe de normalisation MPEG ont amélioré les procédés de codage pour faire chuter le débit de 320 (débit représentatif d'une bonne qualité) à 64 kbps (débit moyen pour une qualité intermédiaire) de façon à faciliter la transmission du son multicanal (5.1, 7.1, etc.). La solution MPEG *surround* offre désormais un rapport de compression d'ordre 60:1 par rapport au son non-compressé (cas d'un signal 5.1) avec une reconstruction spatiale acceptable. Cependant, le modèle paramétrique utilisé par le codec MPEG *surround* ne permet pas toujours une reconstruction audio subjectivement acceptable pour tous les signaux notamment pour certains signaux critiques tels que les applaudissements d'une foule. Dans ce cas, le codec MPEG *surround* est contraint de transmettre les composantes du signal qui ne sont pas correctement modélisées (signaux résiduels) pour atteindre la transparence (à un débit globalement beaucoup plus élevé).

3. Segmentation Temps-Fréquence des signaux audio multicanaux

Comme nous l'avons vu au chapitre 2, le standard en matière de codage audio multicanal qui offre actuellement une haute qualité de reconstruction (*cf.* paragraphe 2.3.3) se base sur l'extraction de paramètres à dépendance temporelle et fréquentielle. Même si certaines implémentations peu complexes reposent sur l'utilisation de bancs de filtres hybrides (*cf.* [SCH04] pour plus de détails), nous considérons que les signaux en sous-bandes sont obtenus par Transformée de Fourier à Court Terme (TFCT, *cf.* Annexe A.1.1) dont les coefficients spectraux sont groupés de façon non uniforme selon une échelle perceptuelle (*cf.* Annexe B.1.2). En règle générale, la résolution temps-fréquence employée par les procédés de codage paramétrique est fixe *i.e.* analyse en sous-bandes qui ne s'adaptent pas au contenu fréquentiel des signaux. Par conséquent, nous avons cherché à extraire les paramètres intercanaux liés à notre perception spatiale (*cf.* paragraphe 1.1) à partir d'une analyse temps-fréquence automatique. L'idée est finalement d'obtenir un ensemble de paramètres perceptuellement pertinents définis sur des supports temps-fréquence adaptés à la nature des signaux à coder. La Représentation Temps-Fréquence (RTF) utilisée par le processus de segmentation proposé dans [HOR00] étant le module carré de la TFCT, l'application de ce processus au codage paramétrique des signaux multicanaux a pu être considérée.

Les travaux de Hory, présentés dans [HOR00], ont permis la mise en œuvre d'un processus d'interprétation automatique du signal en exploitant le contenu de sa RTF. C'est à partir de la description statistique du plan temps-fréquence qu'une interprétation sur le contenu des signaux non-stationnaires est menée. La compréhension du comportement aléatoire des coefficients temps-fréquence pour un modèle de signal temporel très général laisse la possibilité d'appliquer cette méthode d'interprétation à une large gamme de signaux. Les travaux de Saxod dans [SAX03] se sont alors naturellement orientés vers l'identification et la classification de signaux non-stationnaires. De par la nature de l'application visée, le processus de segmentation automatique a été appliqué aux RTF de signaux d'avalanche de façon à caractériser ces signaux.

D'une part, l'objet de notre étude vise à extraire des différences inter-canal à partir des RTF (de chaque canal) segmentées par le processus automatique. D'autre part, étant donné que la représentation paramétrique utilisée par les procédés de codage paramétrique résulte en une erreur de modélisation importante (*cf.* paragraphe 2.3.3), nous avons également cherché à identifier des motifs spectraux caractéristiques d'une dégradation apportée par un procédé de codage audio paramétrique.

3.1 Procédé de segmentation dans le plan Temps-Fréquence

De manière générale, l'objectif fixé par le processus de segmentation consiste à localiser les structures de la RTF analysée, appelées motifs spectraux, par la caractérisation de leur contenu énergétique. Un motif spectral peut par exemple être défini comme un ensemble de coefficients spectraux connexes d'énergie en moyenne supérieure à l'énergie moyenne de tous les coefficients de la RTF. L'idée de base du processus de segmentation est de représenter un motif par un ensemble de paramètres descriptifs aussi informatifs que possible. Partant de l'hypothèse que le signal analysé contient une partie aléatoire, la RTF est considérée comme un ensemble de réalisations de variables aléatoires décrites par les paramètres de leur loi. Il a été démontré dans [HOR00] que le comportement statistique des coefficients temps-fréquence est régi par un mélange de loi du χ^2 centrée et décentrée. Le processus de segmentation vise à estimer les paramètres d'un mélange de lois et à classifier ces paramètres pour finalement délivrer une RTF segmentée en classes de motifs spectraux.

3.1.1 Modèle et mélange de lois du χ^2

L'hypothèse de départ est que le signal analysé x correspond à la somme d'un signal déterministe d_t et d'une composante aléatoire blanche et gaussienne b_g, stationnaire à l'ordre deux, de moyenne nulle et de variance σ^2 telle que:

$$x[n] = d_t[n] + b_g[n]. \tag{3.1}$$

Ce modèle de signal temporel très général est différent des modèles de mélange utilisés en séparation aveugle de sources notamment (*cf.* paragraphe 4.5.1). Cependant, rien ne s'oppose à la validité de l'équation (3.1) pour modéliser un signal audio puisqu'il peut être considéré comme constitué de silences suivant l'axe du temps et à la fois suivant l'axe des fréquences en considérant un signal pourvu de fréquences muettes. De plus, ce modèle est suffisamment général pour n'introduire aucune connaissance *a priori* sur la composante déterministe du signal. Ceci permet d'englober les signaux tels que les composantes bande-étroite ou large-bande, les fréquences pures ou signaux harmoniques, etc. Cette approche s'apparente bien à l'analyse de signaux de musique ou de parole dont nos connaissances *a priori* sont relativement faibles (nature et nombre de sources, réverbération ou effet de salle, spatialisation, etc.).

Suivant cette hypothèse, les non-stationnarités sont portées par la composante déterministe d_t. Le signal temporel x est un ensemble de N_T variables aléatoires gaussiennes statistiquement indépendantes, d'espérance mathématique $E[x[n]] = d_t[n]$ et de variance σ^2 :

$$x[n] \sim \mathsf{N}\left(d_t[n], \sigma^2\right). \tag{3.2}$$

$F_x[l,k]$, la TFCT de $x[n]$ (définie en Annexe A.1.1), peut être interprété comme le produit scalaire de x par un élément d'une base de fonctions analysantes c'est-à-dire comme une combinaison linéaire des $x[n]$. Si x suit le modèle de l'équation (3.1), $F_x[l,k]$ est donc une variable aléatoire gaussienne :

$$F_x[l,k] \sim \mathsf{N}\left(D_t[l,k], \sigma^2\right), \tag{3.3}$$

dont la moyenne $D_t[l,k]$ est le coefficient de la TFCT de la partie déterministe dt et la variance est la variance du bruit additif σ^2. De plus, la RTF du signal x (spectrogramme défini en Annexe A.1.1), s'écrit :

$$S_{F_x}[l,k] = \left(\Re\left(F_x[l,k]\right)\right)^2 + \left(\Im\left(F_x[l,k]\right)\right)^2. \tag{3.4}$$

En supposant que la taille de la fenêtre analysante N est infinie, C. Hory a montré que les parties réelle ($\Re$) et imaginaire ($\Im$) d'un coefficient de la TFCT sont des variables gaussiennes indépendantes telles que :

$$\Re\left(F_x[l,k]\right) \sim \mathsf{N}\left(\Re\left(D_t[l,k]\right), \frac{\sigma^2}{2}\right), \tag{3.5}$$

$$\Im\left(F_x[l,k]\right) \sim \mathsf{N}\left(\Im\left(D_t[l,k]\right), \frac{\sigma^2}{2}\right). \tag{3.6}$$

Par conséquent, le coefficient $S_{F_x}[l,k]$ est la somme du carré de deux variables aléatoires gaussiennes de même variance autrement dit une variable aléatoire du $\chi^2(p_1,p_2,p_3)$ de coefficient de proportionnalité $p_1 = \frac{\sigma^2}{2}$, à $p_2 = 2$ degrés de liberté[8], et de paramètre de décentrage :

$$p_3 = \frac{\left(\Re\left(D_t[l,k]\right)\right)^2 + \left(\Im\left(D_t[l,k]\right)\right)^2}{2} = S_{F_d}[l,k], \tag{3.7}$$

soit le coefficient de la RTF de la composante déterministe *dt*. Finalement, les coefficients de la RTF suivent une loi du χ^2 définie telle que :

$$S_{F_x}[l,k] \sim \chi^2\left(\frac{\sigma^2}{2}, 2, S_{F_d}[l,k]\right). \tag{3.8}$$

Si $S_{F_x}[l,k]$ ne porte pas d'énergie (ne contient que du bruit), alors le paramètre de décentrage est nul et $S_{F_x}[l,k]$ est une variable du χ^2 centré :

$$S_{F_x}[l,k] \sim \chi^2\left(\frac{\sigma^2}{2}, 2\right). \tag{3.9}$$

La RTF du signal analysé est donc composée d'un ensemble de variables aléatoires $S_{F_x}[l,k]$ suivant les lois du χ^2 centré (bruit) et décentré (signal).

La méthode proposée par Hory dans [HOR00] consiste à caractériser chaque coefficient temps-fréquence par le contenu statistique de son voisinage en introduisant des cellules d'analyse $\Xi_{l,k}$ centrées en (l,k) de N_C points (*cf.* **Figure 3.1**).

La loi de la variable parente $S_{F_x}[l,k]$ de la cellule $\Xi_{l,k}$ *i.e.* la loi du mélange, correspond à la somme de N_C-P_d lois du χ^2 centré et de P_d lois du χ^2 décentré telle que :

$$f_{S_{F_x}[l,k]}(x) = (1-p) \times f_{\chi^2\left(\sigma^2/2,2\right)}(x) + p \times f_{\chi^2\left(\sigma^2/2,2,\bar{S}_{Fd}[l,k]\right)}(x), \tag{3.10}$$

[8] δ=2 car on suppose l'indépendance des parties réelle et imaginaire.

avec un paramètre de décentrage commun $\overline{S}_{F_d}[l,k]$ défini par :

$$\overline{S}_{F_d}[l,k] = \frac{1}{P_d} \sum_{(l',k') \in \Xi_{l,k}} S_{F_d}[l',k']. \tag{3.11}$$

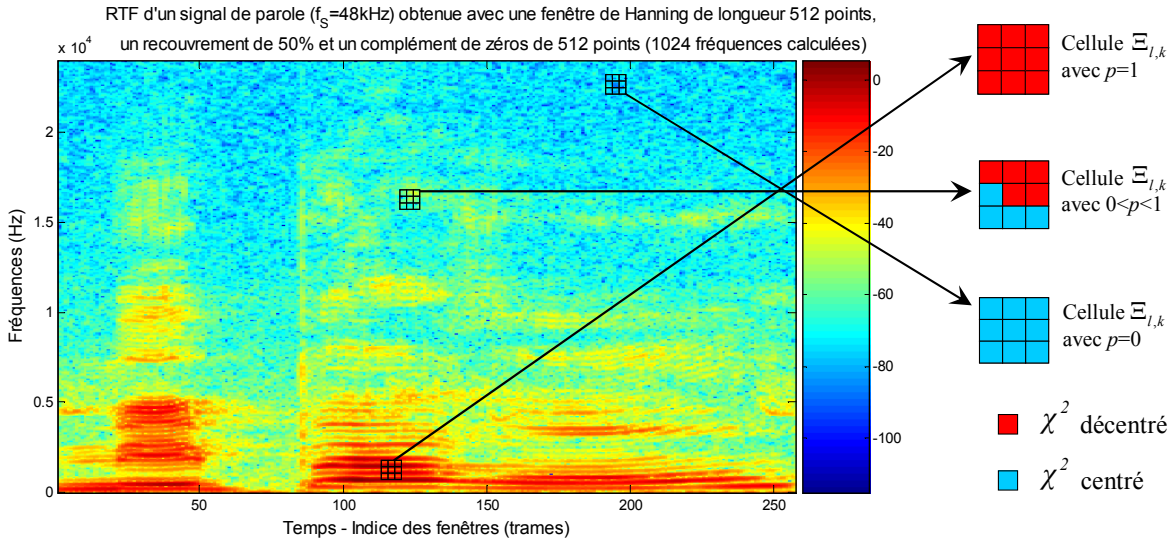

Figure 3.1 **Parcours de la RTF par la cellule $\Xi_{l,k}$.** à N_C=9 points (3x3), centrée en $[l,k]$. Elle est composée de P_d points portant l'énergie d'une composante déterministe, soit une proportion $p = P_d/N_C$ de coefficients suivant une loi du χ^2 décentré. Les N_C-P_d autres coefficients de la cellule suivent une loi du χ^2 centré.

En effet, le lissage engendré par la fenêtre d'analyse utilisée pour construire la RTF permet de considérer les variations d'énergie du signal déterministe faibles en temps et en fréquence sur la (petite) surface de la cellule. Il est donc possible d'approcher les P_d paramètres de décentrage par leur moyenne $\overline{S}_{F_d}[l,k]$ sur la cellule. A partir de cette moyenne énergétique, on peut définir le rapport signal à bruit local (*RSB*) sur la cellule noté r :

$$r[l,k] = \frac{\overline{S}_{F_d}[l,k]}{\sigma^2}. \tag{3.12}$$

D'un point de vue statistique, la segmentation consiste à discriminer les coefficients de la RTF suivant une loi du χ^2 centré (bruit) de ceux suivant une loi du χ^2 décentré (signal + bruit). Le processus d'interprétation se ramène alors à un problème d'estimation des paramètres de la loi du mélange (caractérisation statistique énergétique) puis à une classification des coefficients temps-fréquence (segmentation). Pour faire cela, les coefficients de la RTF sont projetés dans un espace des caractéristiques dont les axes sont les estimations des moments d'ordre un et deux de $S_{F_x}[l,k]$ (moments estimés sur les coefficients de la cellule).

3.1.2 Projection dans l'espace des caractéristiques

Un Espace des Caractéristique (EC) est construit à partir de l'estimation des moments d'ordre un et deux (espérance mathématique et variance) des coefficients d'une cellule $\Xi_{l,k}$. Deux approches ont été mise en œuvre par Hory dans [HOR00]; à savoir le calcul des moments

centrés et non-centrés. La seconde approche a été privilégiée car elle permet de fournir une estimation des paramètres *p* et *r* du mélange local (équations des espérances mathématiques des moments centrés non inversibles). Les axes de l'EC sont donc obtenus en calculant sur la cellule :

- la moyenne empirique telle que :

$$M_1[l,k] = \frac{1}{N_C} \sum_{(l',k') \in \Xi_{l,k}} S_{F_x}[l',k'], \tag{3.13}$$

- le moment non-centré d'ordre deux tel que :

$$M_2[l,k] = \frac{1}{N_C} \sum_{(l',k') \in \Xi_{l,k}} \left(S_{F_x}[l',k'] \right)^2 . \tag{3.14}$$

Pour établir l'EC, le calcul de M_1 et M_2 est réalisé sur les coefficients de la cellule et ceci pour chaque variable parente de toutes les cellules. De plus, il y a un phénomène de recouvrement entre les cellules lors du parcours de la RTF qui procure, au final, autant de couples (M_1,M_2) que de coefficients temps-fréquence. Cette projection des coefficients temps-fréquence dans l'EC constitue alors les données d'entrée au processus de segmentation.

Lorsque la cellule contient des coefficients ne portant que l'énergie du bruit ($p=r=0$) alors $E[M_1] = \sigma^2$ d'après [SAX03]. Ainsi, la moyenne des caractéristiques M_1 prises sur des cellules ne portant que sur l'énergie du bruit est un estimateur efficace de la puissance du bruit σ^2. Par contre l'estimation locale sur une cellule n'est plus efficace du fait de la grande variance de l'estimateur sur cette petite quantité de données.

3.1.3 Segmentation de la RTF en classes

3.1.3.1 Confinement du bruit

La représentation établie avec l'EC permet la discrimination des coefficients temps-fréquence en fonction des propriétés statistiques du signal. Les cellules constituées de coefficients ne portant que l'énergie du bruit sont représentées dans l'EC par un nuage de points confiné vers l'origine des axes appelé région de confinement du bruit R_b (*cf.* **Figure 3.2**). L'orientation de ce nuage est donnée par son axe principal dont l'équation est estimée en diagonalisant la matrice de variance-covariance reliant M_1 et M_2. Une approximation est réalisée via une ACP effectuée sur les données d'après [HOR00]. Grâce à l'équation de cette droite on définit la région de confinement du bruit telle que :

$$R_b = \left\{ \left(M_1, M_2 \right) / M_1 \le q \ et \ M_2 \le \tilde{a} q + \tilde{b} \right\} \tag{3.15}$$

$$\text{où } q \text{ tel que : } Prob\left\{ M_1 \ge q \right\} = p_{fa} . \tag{3.16}$$

p_{fa} est la probabilité de fausse alarme, c'est-à-dire la probabilité pour qu'un point soit attribué à une classe[9] contenant de l'énergie du signal alors que le point correspond à du bruit. La détermination de R_b nécessite la connaissance de la puissance du bruit σ^2 de manière à estimer la loi de M_1. Cette estimation des paramètres du bruit va redéfinir à chaque itération (de la boucle de segmentation) la nouvelle région de confinement du bruit.

[9] Une classe correspond à un ensemble de points portant la même étiquette ou numéro (n°) de classe (*cf.* paragraphe 3.1.3.2).

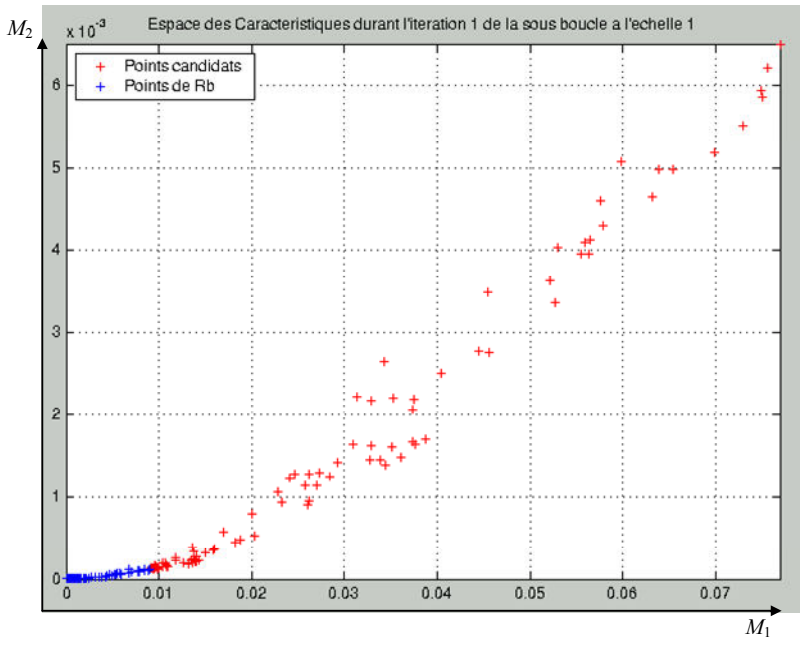

Figure 3.2 **Espace des caractéristiques à l'état initial du processus de segmentation.** Le calcul de R_b permet de discriminer les coefficients de la RTF du signal de parole (cf. Figure 3.1) ne portant que l'énergie du bruit (en bleu) de ceux portant l'énergie du signal (en rouge) « candidats » à la segmentation.

Le modèle de mélange, qui décrit localement la RTF, permet le calcul des M_1 et M_2. Ces moments d'ordre 1 et 2 sont utilisés pour identifier quelques points caractéristiques d'une structure spectrale par leur position dans l'EC (moment d'ordre 1) en leur associant une mesure d'incertitude (moment d'ordre 2). La contrainte posée pour permettre la segmentation des différents motifs spectraux est la connexité des points dans le plan temps-fréquence. La méthode s'est alors tournée vers un algorithme de segmentation de type croissance de région.

3.1.3.2 Croissance des régions

La segmentation des points de l'EC a pour objectif de séparer les coefficients temps-fréquence en classes Cl_0, Cl_1, Cl_2, …. La classe Cl_0 étant réservée aux coefficients temps-fréquence ne portant que l'énergie du bruit. A l'origine du processus, la classe Cl_0 contient tous les points de l'EC. Le processus extrait alors successivement les autres classes Cl_i jusqu'à la limite de segmentation atteinte lorsque tous les points « candidats » (points de la classe Cl_0 n'appartenant pas à R_b) ont été segmentés.

Pour extraire ces classes Cl_i (avec $i \neq 0$) des points de l'EC, le processus procède par extraction de « germes » concentrés dans un cercle de centre le point de l'EC ayant M_1 maximal. Son rayon R_y (*cf.* **Figure 3.3**) est égal à l'écart-type des M_1 pour r (défini à équation (3.12)) maximal tel que p soit également maximal.

Chaque germe de la classe Cl_i contamine parmi ses huit voisins de la RTF ceux appartenant aux candidats qui seront alors affectés de l'étiquette Cl_i. Ces nouveaux points de Cl_i, considérés alors comme germes de Cl_i, peuvent à leur tour propager la contamination en contaminant parmi leurs huit plus proches voisins de la RTF ceux considérés comme des candidats. Les points ayant été contaminés, durant la propagation d'une classe, ne peuvent plus être candidats lors de la propagation d'une autre classe effectuée durant la même itération. Ces points sont d'ailleurs définitivement attachés à cette classe si sa propagation est validée. Ces germes extraits correspondent aux maxima d'énergie de la RTF qui ne sont d'ailleurs pas forcément connexe dans le plan temps-fréquence. C'est pour cette raison qu'à une classe correspondent souvent plusieurs motifs situés à des endroits plus ou moins proches dans le plan temps-fréquence.

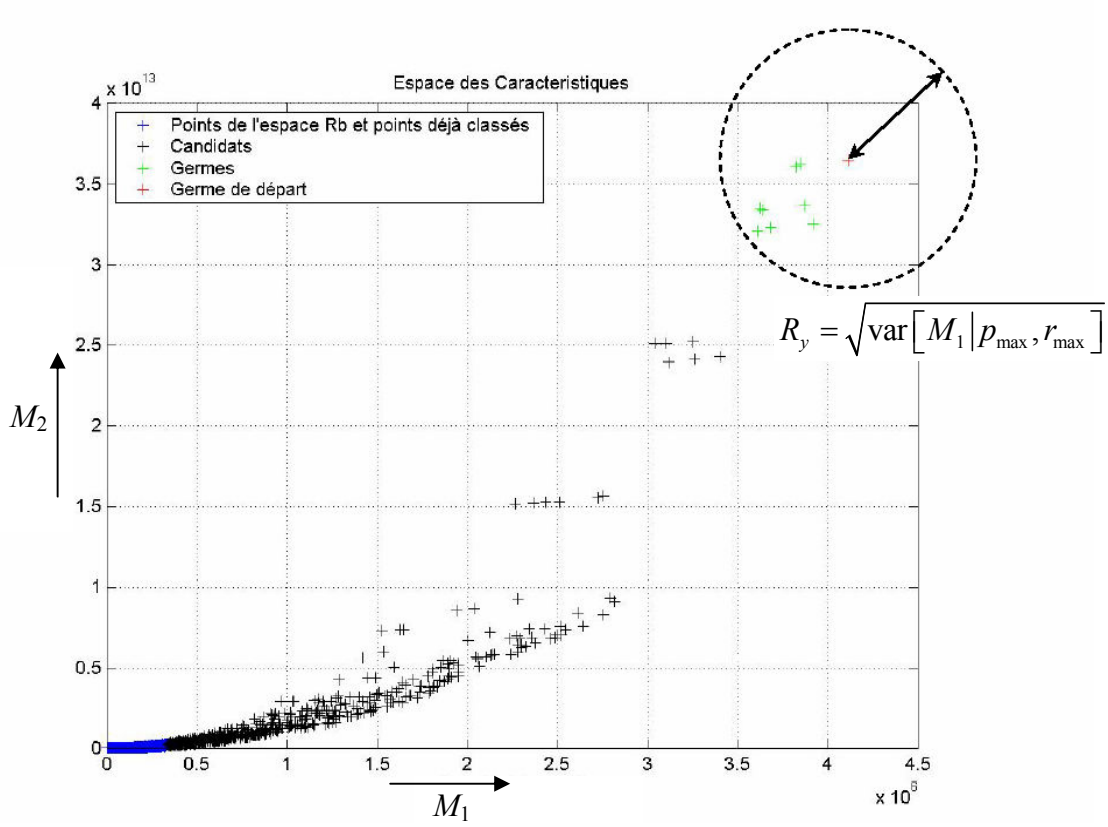

Figure 3.3 **Extraction des germes dans l'Espace des Caractéristiques.** Le point rouge correspond au germe originel, le cercle permet d'extraire les autres germes de la classe. Ces germes permettent l'extraction de la classe Cl_i (avec $i \neq 0$) par contamination des candidats suivant le critère de connexité dans le plan temps-fréquence.

La validation de la propagation d'une classe est rendue possible en testant l'adéquation des données restantes à une loi du χ^2 centré (loi des points ne portant que l'énergie du bruit). Cette adéquation est mesurée par la distance maximale entre la fonction de répartition empirique et celle estimée sur les données (distance de Kolmogorov-Smirnov utilisée dans [HOR00]). Si la propagation n'est pas validée, Saxod dans [SAX03] propose d'attribuer une étiquette demi-entière uniquement aux points issus de la contamination ayant entraîné la non-validation. Ces points à demi-étiquette ne seront alors plus des candidats à la segmentation jusqu'à la prochaine itération. En effet, une étiquette demi-entière attribuée à l'ensemble des points constituant la classe Cl_i inclus les points de cette classe ayant été extraits lors d'itérations précédentes. Ainsi, lorsque la contamination d'une classe n'était pas validée, tous ses points étaient remis en jeu lors de l'itération suivante. Ce qui va à l'encontre du principe selon lequel les classes extraites le plus tôt au cours de la segmentation étaient celles qui, d'un point de vue statistique, étaient les plus éloignées du bruit. Cette gestion de la non-validation d'une contamination avait parfois pour conséquence d'empêcher la convergence de l'algorithme.

3.1.3.3 Limites de segmentation

L'algorithme inclut deux tests d'arrêt: un pour la sous-boucle de propagation, un autre pour la boucle incluant toutes les opérations de la segmentation proprement dite.

Test d'arrêt de la sous-boucle de propagation

La probabilité de fausse alarme p_{fa} fixe la limite de la région de confinement du bruit et peut-être considérée comme la probabilité d'associer un point de la classe du bruit à une classe de la composante déterministe du signal. Le nombre de coefficients de l'ensemble des

candidats susceptibles d'être mal classés à l'issue de la propagation de l'ensemble des classes déjà extraites est défini tel que:

$$\tau_2 = Card\left\{[l,k] \in Cl_0\right\} \times p_{fa} \tag{3.17}$$

Afin d'accélérer le processus de segmentation, il a été choisi de définir une nouvelle classe uniquement si le nombre de candidats est supérieur au nombre de fausses alertes τ_2.

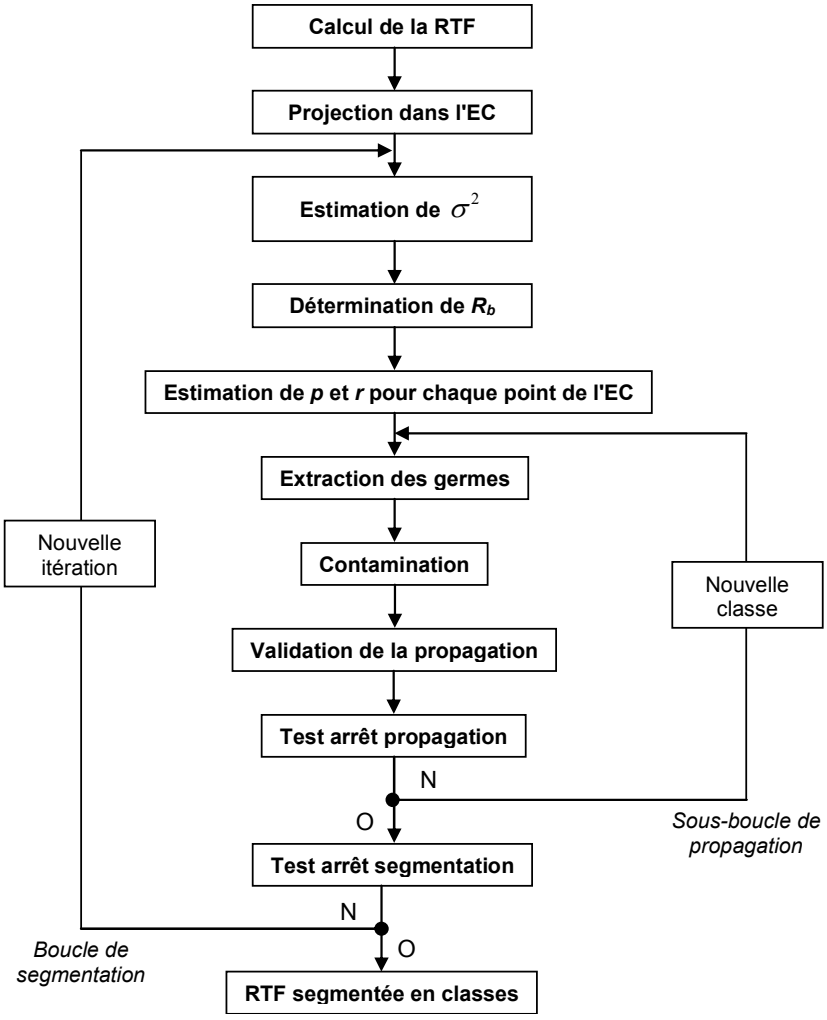

Figure 3.4 Etapes de l'algorithme de segmentation temps-fréquence. Le processus repose sur une boucle principale de segmentation qui englobe une sous-boucle de propagation.

Test d'arrêt de la boucle de segmentation

Ce test repose sur la convergence de la vraisemblance des caractéristiques M_1 de la classe Cl_0 du bruit, autrement dit ce test vérifie l'adéquation des points de Cl_0 avec une loi du χ^2 centré. Cependant, afin de s'affranchir de la variation du nombre de points constituant la classe Cl_0 d'une itération à l'autre, la vraisemblance à l'itération *it*, soit $V[it]$, a été normalisée par le nombre de données d'après [HOR00]. La segmentation s'arrête lorsque cette vraisemblance normalisée n'évolue plus (à un seuil τ_1 près) entre l'itération *it* et l'itération *it*+1, c'est-à-dire lorsque :

$$\frac{V^{[it+1]} - V^{[it]}}{V^{[it]}} \leq \tau_1, \tag{3.18}$$

τ_1 est appelé le « seuil de convergence » et est laissé au choix de l'utilisateur (en pratique, τ_1=0.001). D'un point de vue statistique, le processus peut être vu comme la classification des points temps-fréquence projetés dans l'EC contribuant le moins à l'adéquation de la loi des données à une loi du χ^2 centré.

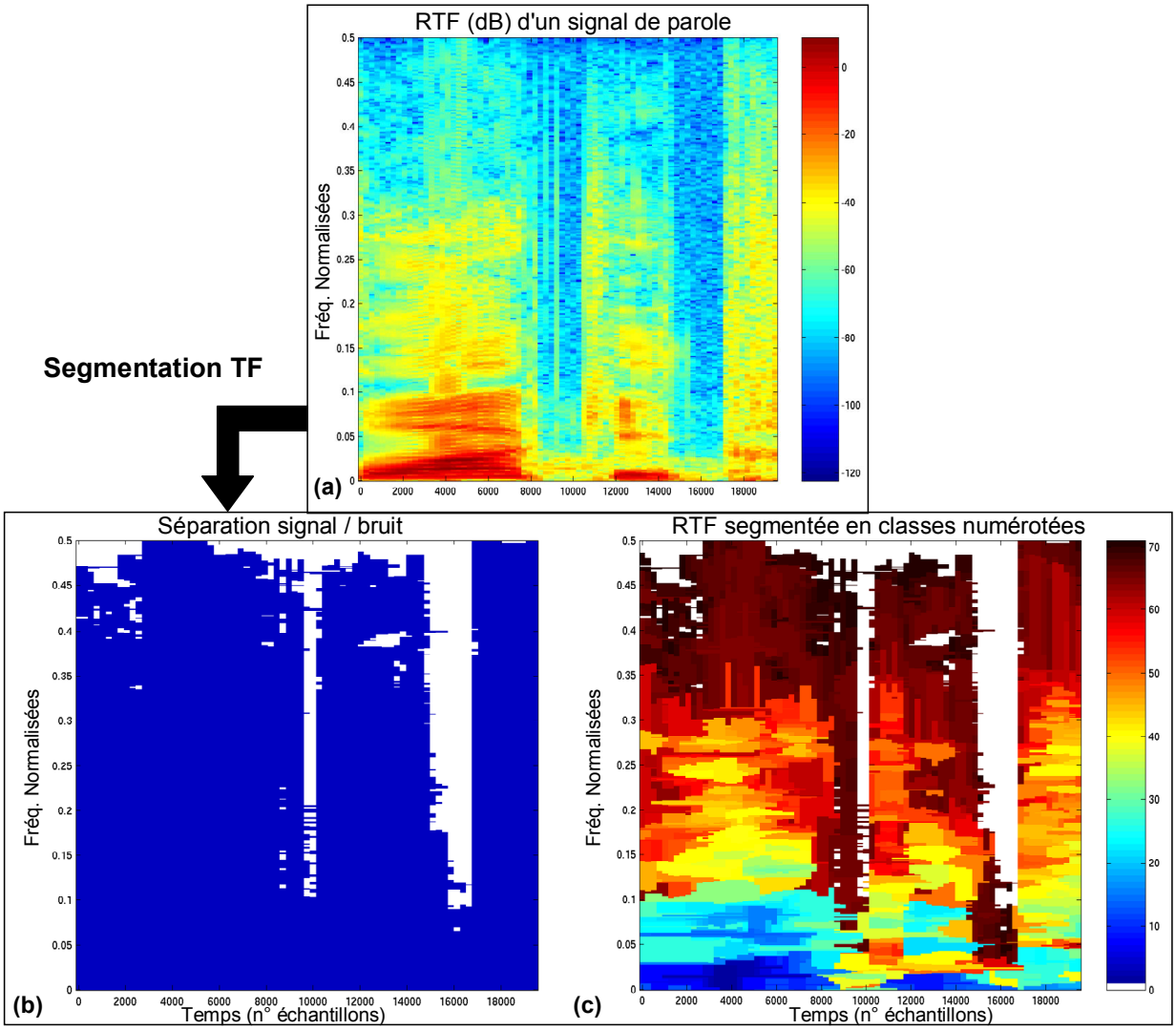

Figure 3.5 **Algorithme de segmentation temps-fréquence appliqué à (a) -** la RTF (spectrogramme) d'un signal de parole. **(b) -** Séparation des points de signal segmentés en bleu et du bruit (points non segmentés de la classe Cl_0) en blanc. **(c) -** RTF segmentée en 70 classes de signal numérotées (classe du bruit Cl_0 en blanc).

A l'issu du processus de segmentation, on obtient une RTF segmentée en classes elles-mêmes constituées d'un ou plusieurs motifs spectraux (*cf.* **Figure 3.5**). L'étiquette ou numéro de chaque classe correspond à son ordre d'arrivée (création) au cours du processus. Les premières classes extraites par le processus de segmentation correspondent aux motifs spectraux les plus énergétiques qui sont réparties sur toute la RTF (*cf.* **Figure 3.5-(c)**). Les dernières classes extraites correspondent aux motifs spectraux proches du bruit et sont, par conséquent, les plus étendues. La fusion des classes signal (Cl_i avec i>0) permet alors d'obtenir un masque temps-fréquence qui sépare le signal du bruit (*cf.* **Figure 3.5-(b)**).

3.2 Segmentation temps-fréquence pour le codage audio multicanal

3.2.1 Le procédé BCC dans le plan temps-fréquence

Les contraintes liées au contexte du codage audio multicanal définissent un système de codage qui doit assurer la compatibilité avec les systèmes de diffusion mono et stéréo. D'après les paragraphes 2.2.3 et 2.3.3, seule l'approche du codage audio paramétrique est capable de restituer une qualité subjective satisfaisante (non transparente) à de très bas débits. L'extraction des paramètres de spatialisation, sur lesquels se base le procédé *Binaural Cue Coding* (BCC décrit au paragraphe 2.2.3), est réalisée à partir de portions de signal analysées et cela pour chaque sous-bande de fréquence (établies au regard d'une échelle perceptuelle équivalente aux bandes critiques), *cf.* Annexe B.1.2.

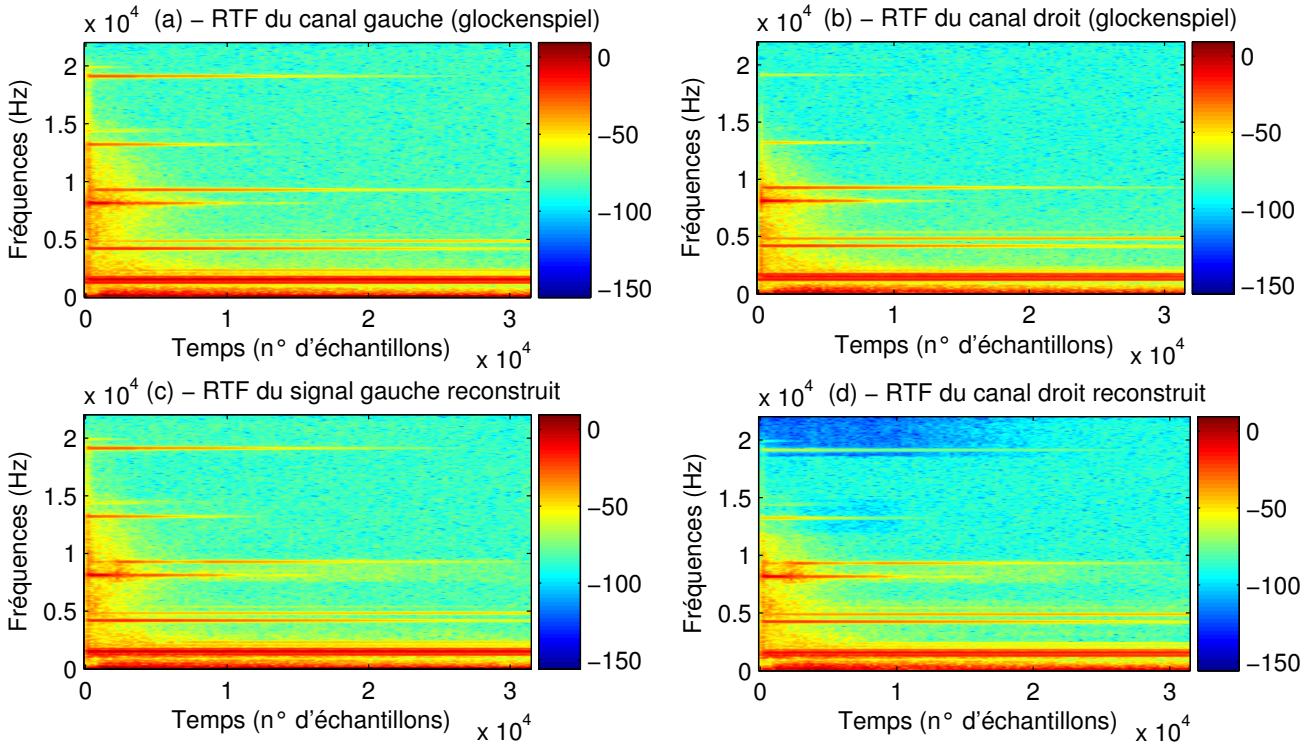

Figure 3.6 RTF obtenues avec une fenêtre de Hanning à N=512 points (Z=512 soit 1024 coefficients spectraux calculés et un recouvrement de 50% entre les fenêtres glissantes). Le signal stéréo original est un mélange de sons de glockenspiel et de grosse caisse dont les amplitudes diffèrent du canal gauche au canal droit (pas de différence de temps entre les canaux). **(a) -** RTF du canal gauche original. **(b) -** RTF du canal droit original. **(c) -** RTF du canal gauche reconstruit par le procédé BCC. **(d) -** RTF du canal droit reconstruit par le procédé BCC.

A titre d'illustration, nous présentons le cas particulier d'un signal stéréo (au contenu fréquentiel harmonique) encodé et reconstruit par le procédé BCC. La **Figure 3.6-(a)** et la **Figure 3.6-(b)** présentent les RTF des canaux originaux. La **Figure 3.6-(c)** et la **Figure 3.6-(d)** présentent les RTF des canaux reconstruits par le procédé BCC.

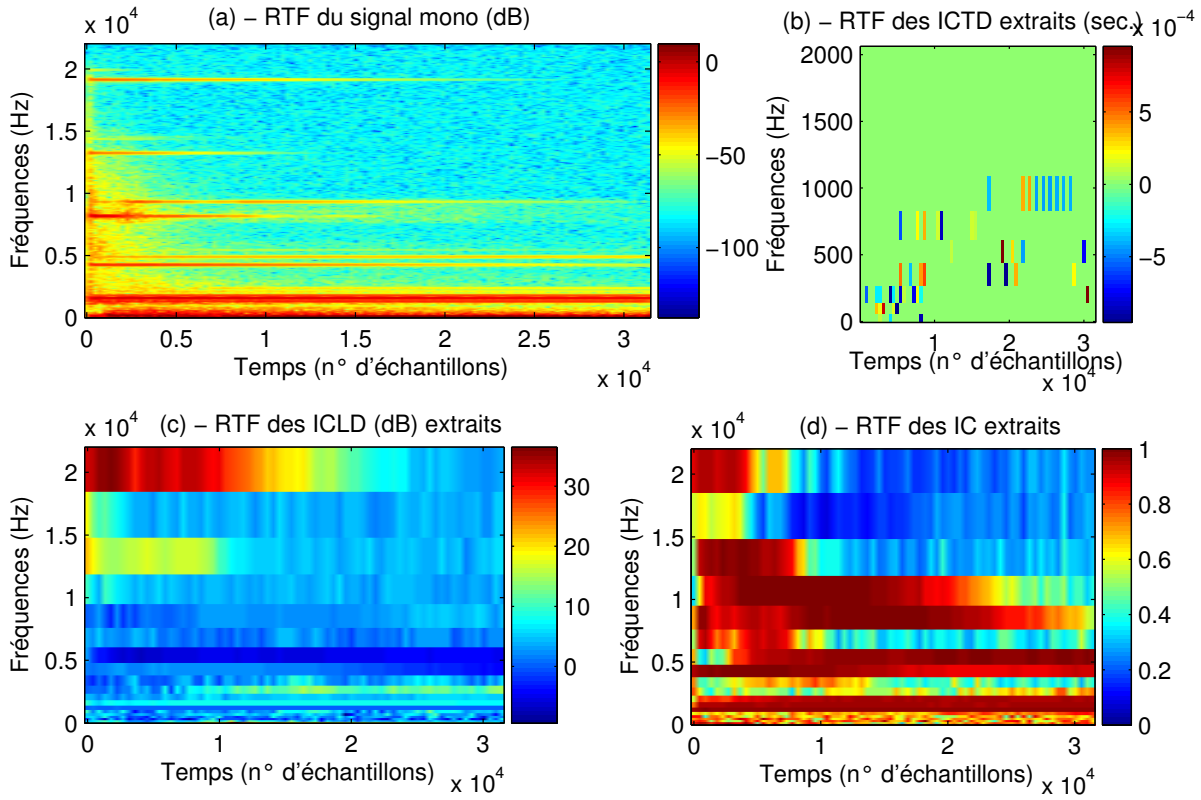

Figure 3.7 RTF **[(a) -** du signal mono obtenu avec une fenêtre de Hanning à N=512 points (1024 coefficients spectraux calculés et un recouvrement de 50% entre les fenêtres glissantes). **(b) -** des ICTD (en seconde). **(c) - d**es ICLD (en dB). **(d) -** des ICC**]**.

La **Figure 3.7** présente les RTF du signal monophonique et des paramètres spatiaux à partir desquels le décodeur BCC reconstruit approximativement le signal stéréo. Le signal monophonique a été généré par la simple somme des canaux d'origine corrigée par un facteur d'égalisation tel que la puissance du signal mono approxime celle des canaux d'entrée (*cf.* paragraphe 2.2.3.1). La RTF du signal ainsi généré est présentée à la **Figure 3.7-(a).**

Les paramètres spatiaux ont été extraits avec une analyse par fenêtres glissantes (type sinus) de N=1024 échantillons (50 % de recouvrement) et un découpage des (N/2+1) coefficients des TFCT (prise en compte de la symétrie hermitienne) en K_b=20 sous-bandes de fréquences selon l'échelle ERB (*cf.* Annexe B.1.2). La **Figure 3.7-(b)** montre que les différences de temps ou ICTD extraites (jusqu'à 2 kHz) entre les canaux ne sont pas significatives à la vue de leur faible nombre et de leur variation au cours du temps *i.e.* pas de continuité entre les valeurs extraites d'une portion glissante à une autre. La **Figure 3.7-(c)** et **Figure 3.7-(d)** présentent respectivement les RTF des ICLD et des ICC extraites (pour chaque portion glissante et chaque sous-bande). Les ICC extraites traduisent la corrélation des harmoniques du signal de glockenspiel et la faible corrélation du bruit (silence). Les ICLD extraites sont relativement faibles (entre 0 et 10 dB) excepté pour certaines harmoniques des sous-bandes b=1, 18, 19 et 20 notamment. La forte différente d'intensité entre les harmoniques hautes fréquences présentes dans chacun des canaux originaux est identifiée par l'encodeur/décodeur BCC au travers de l'ICLD extraite à la 20ème sous-bande. Cependant, la résolution fréquentielle suivant l'échelle ERB ne permet de conserver une séparation suffisante de ces hautes fréquences. Il en résulte une reconstruction approximative des signaux originaux comme le montre la **Figure 3.6-(d)**.

L'exemple choisi peut être critiquable du fait que l'harmonique haute fréquence approximativement reconstruite par le procédé BCC ne sera pas perceptible puisqu'elle apparaît aux alentours de 19 kHz (voir les caractéristiques de l'oreille en Annexe B.1). Cependant, cet exemple illustre parfaitement comment le procédé BCC peut synthétiser ou

plutôt approximer un signal stéréo (ou multicanal) en le caractérisant par ses différences d'intensités entre les canaux (ICLD) définies par la résolution temps-fréquence employée. A partir de l'analyse de ce cas particuliers, nous soulignons une faiblesse caractéristique de cette approche de codage qui utilise une résolution temps-fréquence fixe (analyse en sous-bandes) indépendante de la nature des signaux. Par conséquent, l'application de la segmentation temps-fréquence au codage des signaux audio multicanaux s'est naturellement orientée vers l'extraction de paramètres temps-fréquence définis à partir des caractéristiques spectrales des signaux. De façon à réduire les redondances entre les canaux, nous avons cherché à extraire les différences inter-canal à partir des motifs spectraux identifiés par le processus de segmentation.

3.2.2 Extraction de différences inter-canal

Les procédés de codage audio paramétrique (de type BCC) utilisent des paramètres en sous-bandes relatifs aux différences inter-canal pour générer une approximation du signal multicanal à partir du signal somme. Par conséquent, notre première approche s'est attachée à appliquer le procédé de segmentation aux signaux audio multicanaux de façon à identifier des motifs spectraux propres à chaque canal.

3.2.2.1 Principe

Dans un premier temps, la comparaison des RTF segmentées est conduite de manière à extraire les coordonnées temps-fréquences des motifs spectraux communs à chaque canal. Ensuite, l'idée est de pouvoir en dériver une carte ou grille temps-fréquence relative aux classes extraites mais qui diffèrent d'un canal à un autre. Ainsi, cette grille temps-fréquence caractéristique des différences inter-classes pourrait être utilisée au sein d'un codeur paramétrique de type BCC comme l'indique la Figure 3.8.

L'analyseur temps-fréquence utilisé pour extraire les paramètres spatiaux pourrait ainsi s'adapter au contenu fréquentiel des canaux. Les paramètres directifs liés aux sources dominantes {ICTD-ICLD} seraient estimés pour les zones temps-fréquences considérés comme identiques (composantes corrélées) à l'issue de la comparaison des RTF segmentées. A l'inverse, les paramètres de corrélation ICC pourraient être estimés uniquement pour les zones temps-fréquence définies par les coordonnées des différences inter-classes.

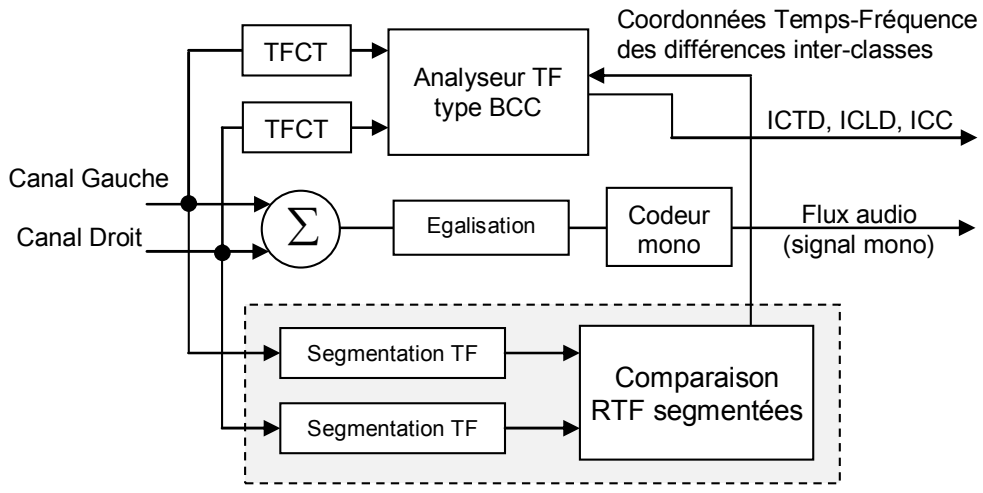

Figure 3.8 Schéma de principe d'un encodeur stéréo paramétrique (type BCC) qui extrait des paramètres spatiaux pertinents à partir des différences inter-classes extraites des RTF segmentées.

3.2.2.2 Expérimentation

Le processus de segmentation dans le plan temps-fréquence peut alors être vu comme un outil à l'extraction des différences spectrales inter-canal qui n'ont pas pu être extraites à partir du pavage temps-fréquence fixe utilisé par la technique BCC.

La **Figure 3.9** présente les RTF et les RTF segmentées obtenues à l'issue du processus de segmentation appliqué aux RTF gauche et droite du signal stéréo mélange de sons de glockenspiel et grosse caisse. La même probabilité de fausse alarme et le même seuil de vraisemblance ont été utilisés pour chacun des processus de segmentation (p_{fa}=0.001 et τ_1=0.001).

Les premières classes de signal extraites (Cl_1, Cl_2 et Cl_3) par le processus de segmentation sont relatives aux composantes les plus énergétiques ici localisées en basses fréquences (0-500 Hz) avec le signal de grosse-caisse. La différence d'énergie de ce contenu basse fréquence entre le canal gauche et droit (*cf.* **Figure 3.9 (a)-(b)**) est observée à partir des trois premières classes extraites qui diffèrent entre les deux RTF (des canaux gauche et droit) segmentées (*cf.* **Figure 3.9 (c)-(d)**). La classe Cl_3 obtenue avec la segmentation de la RTF du canal gauche (en bleu foncé sur la **Figure 3.9 (c)**) s'étant sur l'ensemble de la portion de signal temporel analysé pour les fréquences comprises entre 0 et 500 Hz. Pour cette même région temps-fréquence, la classe Cl_3 extraite par le processus à partir de la RTF du canal droit (en bleu foncé sur la **Figure 3.9-(d)**) est divisée en deux motifs à l'image de l'énergie aux basses fréquences (0-500 Hz) qui diminue entre les échantillons 10000 et 15000 sur la RTF originale du canal droit (*cf.* **Figure 3.9 (b)**).

On peut également remarquer qu'il n'y a pas de correspondance entre les numéros de classe qui correspondent seulement à l'ordre de création des classes de signal. En effet, les classes Cl_1 et Cl_2 caractérisent les mêmes contenus basses fréquences (à gauche et à droite) seulement si on inverse les numéros attribués à ces classes pour l'une où l'autre des RTF segmentées (*cf.* **Figure 3.9 (c)-(d)**).

Les harmoniques du signal de glockenspiel présentes dans chacun des canaux sont extraites en plusieurs classes de signal dont l'interprétation reste délicate. Les **Figure 3.9 (c)-(d)** montrent que certaines harmoniques sont regroupées dans la même classe à gauche et identiquement à droite mais pour des numéros de classe différents et parfois même pour des harmoniques différentes. De plus, la résolution temps-fréquence établie par les RTF originales n'est pas conservée à l'issue du processus puisque les classes extraites ont tendance à englober une partie de l'énergie du bruit pour finalement réduire la précision d'analyse selon l'axe fréquentiel.

Même si ce signal stéréo est très corrélé (*cf.* la RTF des indices de corrélation à la **Figure 3.7-(d)**) le processus de segmentation génère des RTF segmentées avec des limites différentes entre les classes signal fusionnées et la classe de bruit (*cf.* **Figure 3.9 (e)-(f)**). En effet, le processus de segmentation appliqué à la RTF du canal gauche a délivré 14 classes de signal en 25 itérations alors qu'en l'appliquant à la RTF du canal droit (moins énergétique), ce dernier a délivré 16 classes en 19 itérations. La **Figure 3.9 (d)** illustre le cas où une partie du bruit a été considéré par l'algorithme comme du signal en formant notamment les classes Cl_{12} et Cl_{15}.

La **Figure 3.10** présente les RTF des canaux gauche et droit masquées par les classes n° 9 (canal gauche) et n°14 (canal droit) qui correspondent aux classes extraites pour une même zone temps-fréquence *i.e.* celle qui contient l'harmonique haute fréquence du glockenspiel. Les classes extraites pour une même région temps-fréquence, au numéro pas nécessairement identique, ne caractérisent pas exactement les mêmes contenus spectraux.

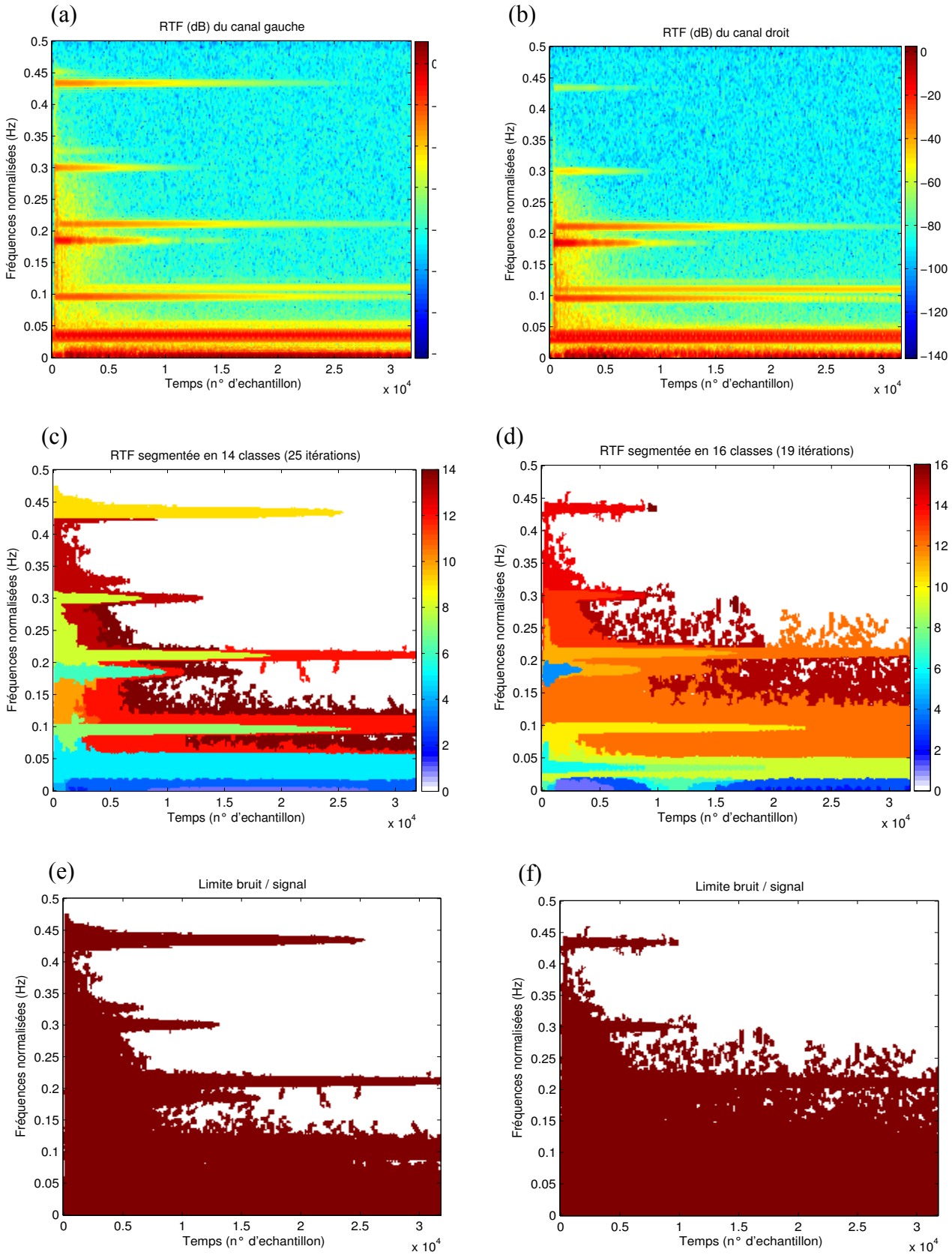

Figure 3.9 **(a)-(b) -** RTF (en dB) des canaux gauche et droit obtenues avec une fenêtre de Hanning à N=512 points (1024 coefficients spectraux calculés et un recouvrement de 50% entre les fenêtres glissantes). **(c)-(d) -** RTF segmentées avec une p_{fa}=0.001. **(e)-(f) -** Séparation des classes de signal fusionnées et de la classe de bruit (Cl_0).

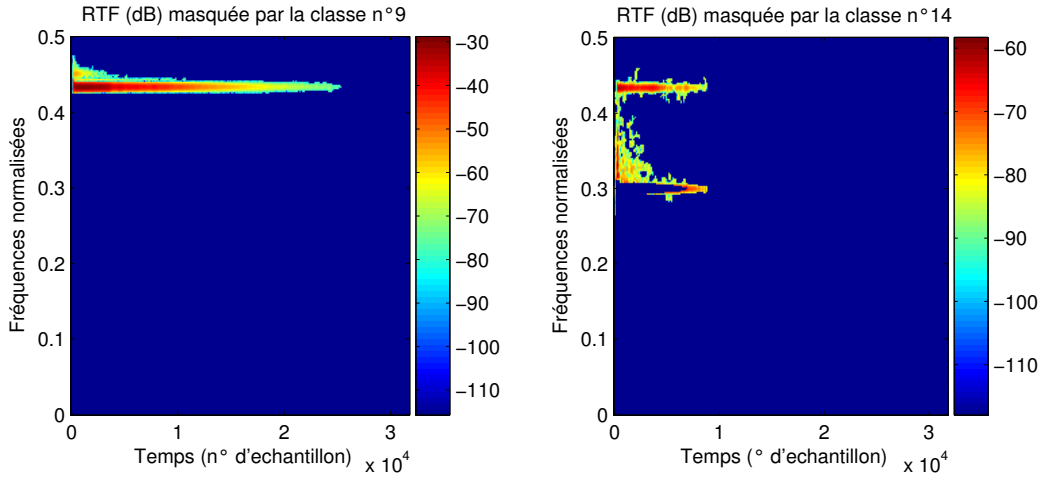

Figure 3.10 RTF (en dB) des canaux gauche et droit masquées par les classes n°9 (pour la RTF du canal gauche) et n°14 (pour la RTF du canal droit) qui ont été extraites par le processus de segmentation pour la même région temps-fréquence.

En conclusion, même si la durée du signal analysé par le processus de segmentation est très courte (moins d'une seconde pour le signal présenté dans ce paragraphe), la richesse spectrale des signaux audio ne permet pas une comparaison immédiate des classes extraites par le processus de segmentation.

Au regard des expérimentations menées sur une multitude de signaux au contenu fréquentiel variés (*cf.* Annexe A.2), l'état actuel du processus de segmentation délivre en règle général un nombre trop important de classes pour pouvoir réaliser une comparaison directe. De plus, le processus délivre trop souvent des classes constituées de trop nombreux motifs spectraux étalés sur l'ensemble du plan temps-fréquence. Une fusion des classes suivant un critère de corrélation et/ou de proximité temps-fréquence devrait cependant permettre d'améliorer la comparaison des RTF pour dégager une grille temps-fréquence propres aux signaux analysés.

3.2.3 Segmentation de l'erreur de reconstruction audible générée par un procédé de codage audio paramétrique

La deuxième approche que nous avons abordée vise à utiliser la segmentation dans le plan temps-fréquence pour identifier les motifs spectraux caractéristiques d'une dégradation apportée par un procédé de codage audio paramétrique. Les techniques de codage paramétrique de type BCC génèrent une approximation du signal multicanal à partir du signal somme des canaux et de paramètres spatiaux. Comme nous l'avons vu au paragraphe 2.3.3, la technologie MPEG *surround* prévoit la transmission de signaux résiduels caractéristiques de l'erreur de modélisation pour reconstruire un signal multicanal de haute qualité.

Nous avons donc cherché à appliquer le processus de segmentation à l'erreur de reconstruction, issue d'un tel procédé de codage, qui correspond au signal somme des différences entre les canaux originaux et les canaux approximés ou reconstruits. L'objectif étant de pouvoir identifier les zones de la RTF qui sont particulièrement critiques dans le sens où elles nécessitent une attention particulière par le procédé de codage. Il pourrait alors être envisagé d'augmenter la résolution temps-fréquence pour l'extraction des paramètres de spatialisation pour certaines zones temps-fréquence considérées comme critiques.

Bien que nous ne soyons pas parvenu à extraire des informations caractéristiques à partir des différences entre les RTF des canaux segmentées par le processus automatique de

segmentation (*cf.* paragraphe 3.2.2), nous considérons ici un seul canal relatif à l'erreur de modélisation totale issue d'un codage paramétrique multicanal. De plus, nous faisons l'hypothèse que ce signal d'erreur a une énergie faible et un contenu spectral appauvri facilitant l'analyse dans le plan temps-fréquence. Sous cette hypothèse, le processus de segmentation appliqué au signal d'erreur de reconstruction serait utilisé comme un outil de classification des motifs spectraux dégradés par un procédé de codage. En outre, cette approche permettrait d'ajouter un critère objectif aux tests subjectifs habituellement utilisés pour évaluer les performances des procédés de codage dont l'erreur de modélisation est perceptuellement significative.

3.2.3.1 Principe

La première étape consiste à établir un signal résiduel d'erreur entre chaque canal original et chaque canal reconstruit par codage/décodage paramétrique. Nous proposons d'associer ces signaux d'erreurs à un critère perceptuel basé sur le masquage fréquentiel (*cf.* Annexe B.1.3.1). En effet, si des défauts de codage apparaissent sur les signaux reconstruits autrement dit, si l'erreur de reconstruction est non nulle mais qu'elle est inaudible alors le procédé de codage reste subjectivement valide. La seconde étape consiste alors à calculer un seuil de masquage propre à chaque signal reconstruit puis à l'appliquer à l'erreur de reconstruction correspondante. De cette manière, le signal d'erreur obtenu caractérise l'erreur de reconstruction « audible » ou non-masquée engendrée par l'opération de codage/décodage. Finalement, la somme des erreurs de reconstruction audibles constitue le signal d'entrée au processus de segmentation.

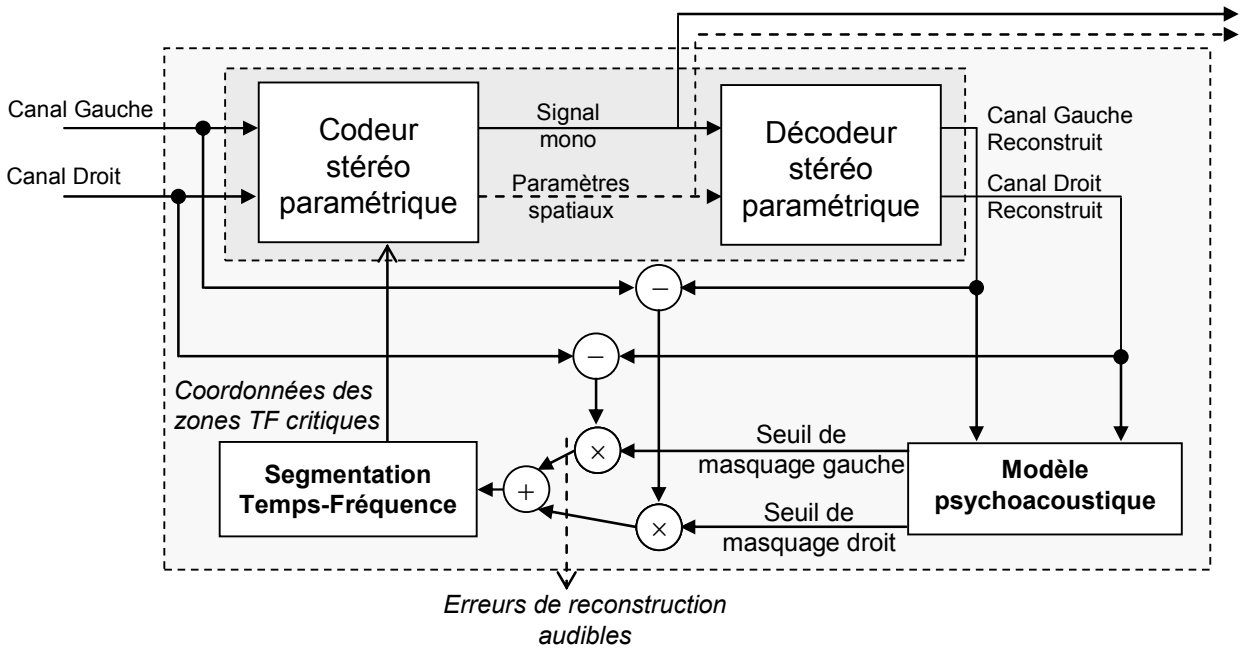

Figure 3.11 Calcul des erreurs de reconstruction audibles et segmentées pour contrôler la résolution temps-fréquence de l'analyseur temps-fréquence du codeur BCC. Un modèle psychoacoustique est utilisé pour calculer les seuils de masquage relatifs aux signaux reconstruits. Les erreurs de reconstruction audibles sont segmentées dans le plan temps-fréquence pour extraire les coordonnées temps-fréquence des zones critiques pour lesquelles les paramètres seront redéfinis.

L'étape finale consiste alors à utiliser le processus de segmentation pour déceler les zones temps-fréquence où l'erreur de reconstruction audible totale témoigne d'une mauvaise reconstruction du signal multicanal. En transmettant les coordonnées temps-fréquence des

zones critiques, plusieurs stratégies permettant de conserver cette information perdue ou dégradée par le procédé de codage peuvent être imaginées :

- extraction des paramètres spatiaux avec une haute résolution fréquentielle pour les zones critiques
- transmission auxiliaire d'une partie ou de la totalité du signal résiduel d'erreur, cette transmission peut être hiérarchisée puisque certaines composantes sont particulièrement critiques d'après l'analyse temps-fréquence menée,
- modification de la somme des canaux (opération de *downmix* décrite au paragraphe C.1) en favorisant la conservation des composantes critiques.

La **Figure 3.11** présente un schéma de principe alliant l'utilisation d'un modèle psychoacoustique et du processus de segmentation temps-fréquence pour calculer les coordonnées des régions temps-fréquence perceptuellement mal reconstruites par un codec stéréo paramétrique de type BCC. Nous pouvons également considérer ce schéma de principe comme l'architecture d'un codeur, certes complexe puisqu'il incorpore un décodeur de façon à reconstruire localement les signaux décodés. Comme l'indique la **Figure 3.11**, nous avons considéré un traitement indépendant entre les canaux puisque nous disposons d'un modèle psychoacoustique monophonique décrit en Annexe B.2.

3.2.3.2 Expérimentation

L'expérimentation, basée sur le schéma de principe de la **Figure 3.11**, a été mise en œuvre à partir de signaux audio aux formats 5.1 encodés et décodés par le codec relatif à la norme MPEG *surround* (codage paramétrique associé au codec MPEG-4 AAC, *cf.* paragraphe 2.3.3). De manière à calculer l'erreur de reconstruction engendrée par ce procédé de codage, nous avons considéré chaque canal indépendamment les uns des autres.

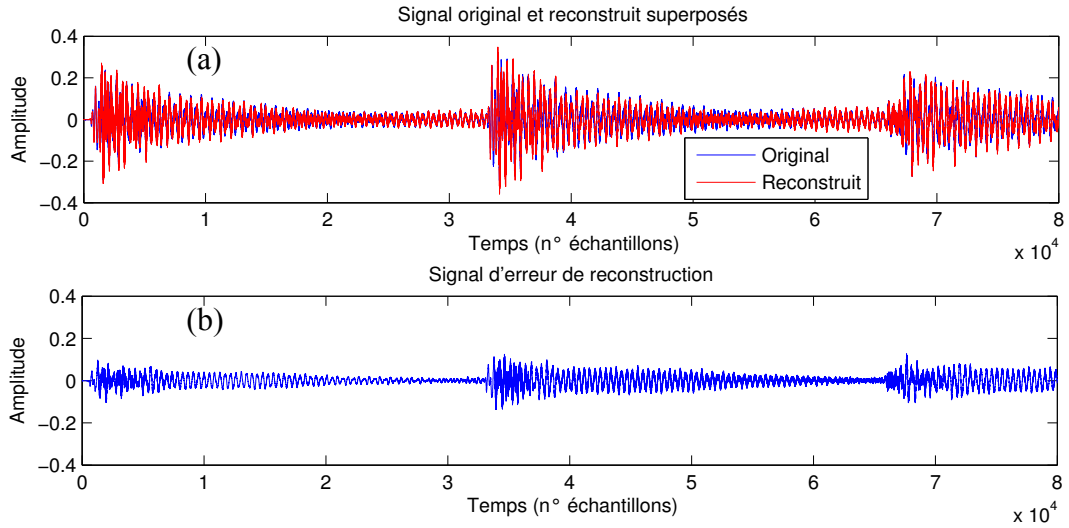

Figure 3.12 (a) - Signal original et reconstruit par le codec MPEG *surround*. **(b) -** Erreur de reconstruction établie par différence du signal original et du signal reconstruit.

La **Figure 3.12** présente une portion du signal original (canal avant gauche du signal 5.1 original mélange de sons de glockenspiel et de grosse caisse) et du signal reconstruit par le procédé de codage, ainsi que leur différence *i.e.* l'erreur de reconstruction. L'erreur de reconstruction issue du codage/décodage MPEG *surround* est très énergétique et d'autant plus lors des attaques du signal de glockenspiel (*cf.* **Figure 3.12-(b)**). Cela s'explique en partie par le fait que le décodeur est pourvu de filtres de décorrélation (*cf.* paragraphe 2.3.3.2) qui déphases les composantes fréquentielles des canaux. Dit simplement, le codec MPEG *surround* approxime la phase des signaux puisqu'aucune ICTD ou ICPD n'est

tranmise. En effet, les ICTD extraites entre le signal original et le signal reconstruit, présentés à la **Figure 3.13-(b)**, présentent une continuité dans le temps pour certaines sous-bandes notamment aux alentous de 2 kHz et pour les premières sous-bandes (les plus énergétiques). Par contre, les ICLD extraites, présentées à la **Figure 3.13-(a)**, traduisent bien le fait que les énergies en sous-bandes du canal original et du canal reconstruit sont globalement équivalentes excepté pour la dernière sous-bande où le codage audio (MPEG-4 AAC) du signal somme a réduit la bande passante du signal. Enfin, les ICC extraites, présentés à la **Figure 3.13-(c)**, témoignent de la forte corrélation des signaux.

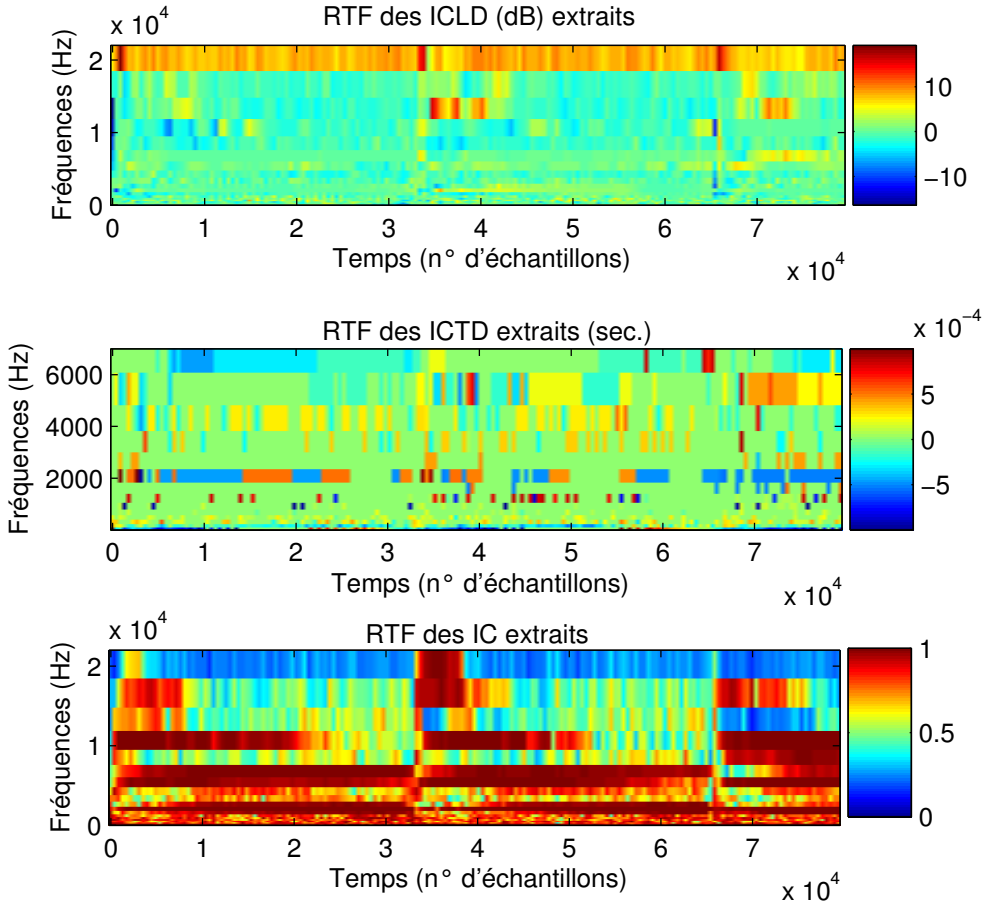

Figure 3.13 RTF des paramètres spatiaux extraits entre le signal original (canal avant gauche) et le signal reconstruit après codage/décodage MPEG *surround*. (a) – ICLD, en dB, estimées pour les 20 sous-bandes. **(b) –** ICTD, en seconde, estimées pour les sous-bandes couvrant les fréquences 0-7 kHz. **(c) –** ICC estimées pour les 20 sous-bandes.

L'étape suivante vise à prendre en compte les seuils de masquages relatifs aux canaux reconstruits. Pour cela, le modèle psychoacoustique 1 du standard MPEG-1 a été mis en œuvre (*cf.* Annexe B.2) pour calculer un seuil de masquage minimum. Ce seuil de masquage minimum correspond à la valeur minimale du seuil de masquage global pour chaque sous-bande suivant l'échelle des Barks (*cf.* **Figure 3.14**).

Plutôt que de calculer l'erreur de reconstruction pour chaque canal codé/décodé puis de les sommer entre elles pour obtenir une erreur totale, nous avons, dans un premier temps, cherché à calculer l'erreur de reconstruction audible propre à chaque canal.

Ainsi, l'erreur de reconstruction non-masquée ou audible est obtenue en conservant les coefficients spectraux de l'erreur de reconstruction qui sont au-dessus du seuil de masquage (les coefficients spectraux en-dessous du seuil sont pondérés par un coefficient de façon à être complètement inaudibles) calculé pour chaque portion glissante du canal reconstruit.

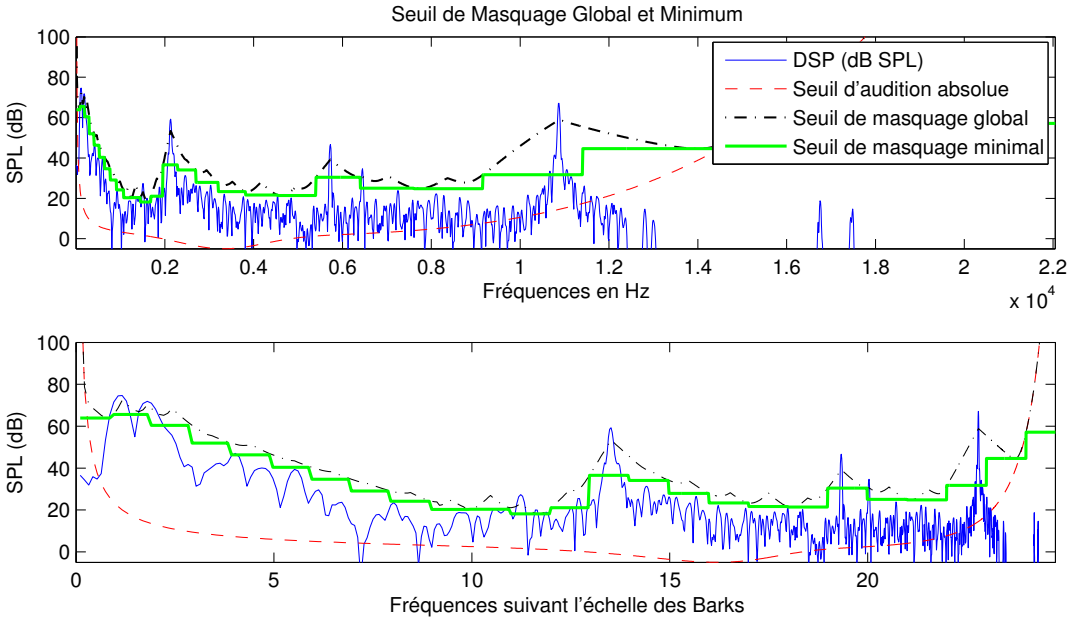

Figure 3.14 **Seuil de masquage global et minimum** calculés à partir de la Densité Spectrale de Puissance (DSP de 4096 points en dB SPL) estimée à partir de la troisième portion (d'une longueur de 2048 échantillons) glissante du signal reconstruit à l'issu de l'encodage/décodage MPEG *surround*.

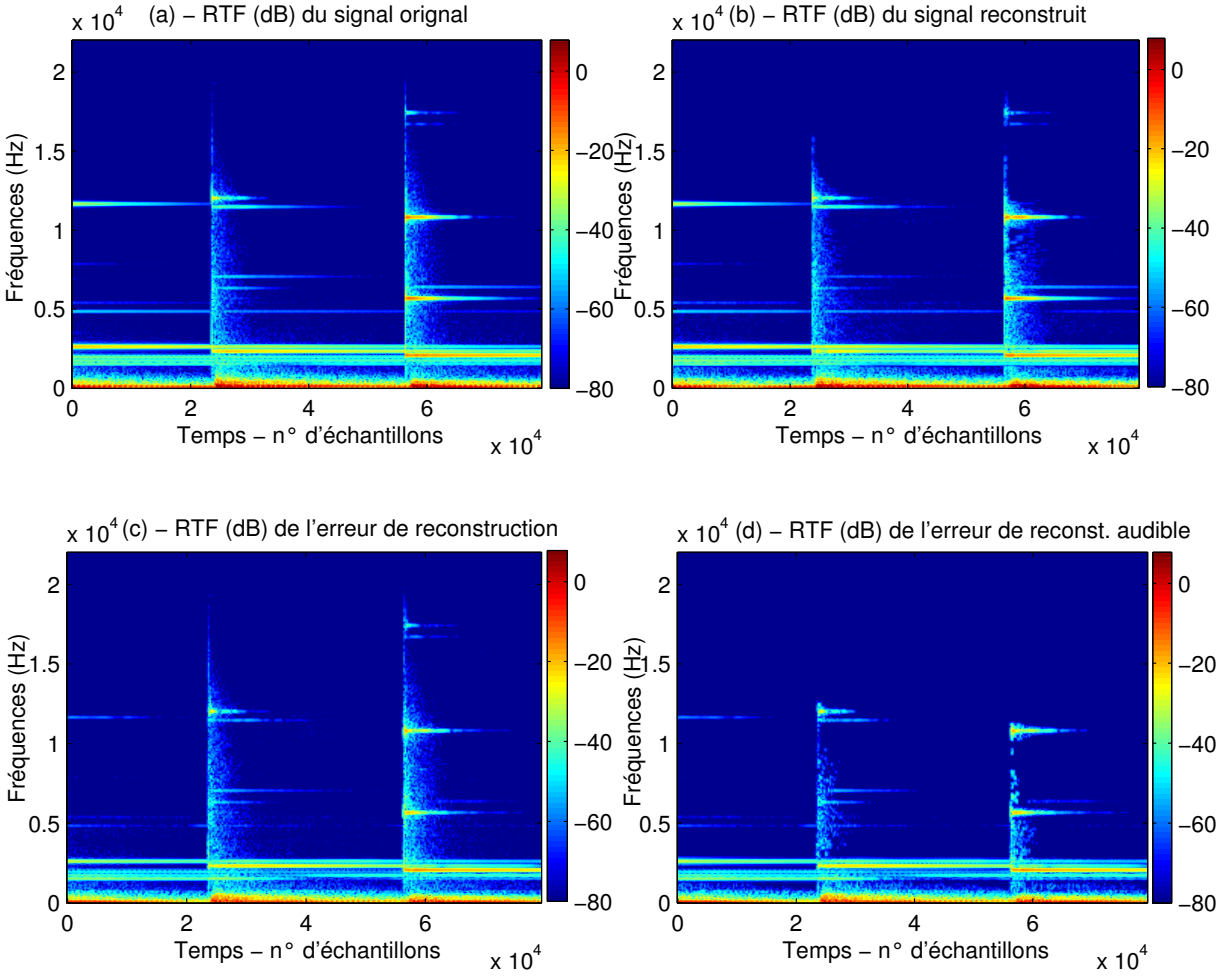

Figure 3.15 **(a) -** RTF du signal original. **(b) -** RTF du signal reconstruit à l'issu du codage/décodage MPEG *surrround*. **(c) –** RTF de l'erreur de reconstruction. **(d) -** RTF de l'erreur de reconstruction audible ou non-masquée. RTF (en dB) obtenues avec une fenêtre de Hanning à N=512 points (1024 coefficients spectraux calculés, un recouvrement de 50% entre les fenêtres glissantes) et un seuil énergétique minimum fixé à 80 dB.

Finalement, les RTF : du signal original, du signal reconstruit suite à l'encodage/décodage MPEG *surround*, de l'erreur de reconstruction et de l'erreur de reconstruction audible sont présentées à la **Figure 3.15**. La RTF de l'erreur de reconstruction audible présentée à la **Figure 3.15-(d)** a un niveau d'énergie non négligeable et surtout une forte répartition dans le plan temps-fréquence. Les composantes harmoniques les plus énergétiques de la RTF originale sont toujours présentes dans la RTF de l'erreur de reconstruction audible alors que ces composantes sont censées être les mieux reproduites par le procédé de codage. En effet, cette erreur a été calculée à partir d'un signal codé/décodé avec une qualité audio perçue comme bonne en considérant une écoute du signal multicanal décodé sur un système de diffusion 5.1 (*cf.* résultats des tests subjectifs menés, dans [HER05], pour l'évaluation du codec MPEG *surround* dans la configuration de test 1 *i.e.* haute qualité). Toutefois, nous considérons ici l'erreur de reconstruction relative à un des canaux du signal multicanal traité. Dans ce cas, les différences perçues à l'écoute au casque, plus critique, entre le canal original et le canal reconstruit sont significatives pour l'oreille. Ces différences perçues ne justifient pas pour autant le niveau d'énergie de l'erreur de reconstruction et sa répartition étendue dans l'ensemble du plan temps-fréquence « audible », qui apparaîssent comme disproportionnés par rapport aux dégradations perçues.

Sans prise en compte de la phase des signaux, l'erreur de reconstruction non-masquée est trop énergétique et ne correspond pas à une bonne indication des dégradations perçues par notre oreille. Finalement, elle ne nous permet pas de déceler des zones temps-fréquence précises où le signal est particulièrement dégradé. C'est pour cette raison que nous n'avons pas abordé la classification des motifs spectraux de l'erreur au moyen du processus de segmentation automatique.

3.3 Conclusion

L'algorithme de segmentation temps-fréquence constitue un processus d'analyse et d'interprétation qui a d'abord été destiné à l'identification et la classification automatique de différents types de signaux (signaux sismiques et d'avalanches). L'application du processus de segmentation au codage des signaux audio multicanaux a été initiée dans ce chapitre.

L'objectif premier du travail accompli ici a été d'appliquer le processus de segmentation aux signaux audio (parole et musique) pour classifier les motifs spectraux de leur RTF. Pour cela, la résolution des RTF a été ajustée (*cf.* Annexe A.1) ainsi que les différents paramètres qui conditionnent le processus automatique de segmentation. Le premier constat tiré des résultats obtenus prouve la capacité du processus de segmentation à séparer le signal du « bruit » (fréquences muettes, silences) dans le plan temps-fréquence. Le second a été de constater la multiplicité des informations obtenues à l'issus du processus de segmentation et ceci même sur un signal de très courte durée (inférieure à la seconde). Ces informations sont relatives aux classes extraites et aux trop nombreux motifs présents sur les RTF segmentées. Ce nombre important de classes et motifs extraits a rendu les interprétations délicates.

En tenant compte de ces premières observations, l'objectif principal de cette étude a été d'associer la segmentation temps-fréquence à un procédé de codage audio paramétrique de type BCC. Cette technique de codage est basée sur l'extraction de paramètres spatiaux dans le plan temps-fréquence avec une résolution temporelle adaptative (commutation de fenêtre courte et longue, *cf.* Annexe B.1.3.2) mais avec une résolution fréquentielle fixe basée sur une échelle perceptuelle. L'interrogation a porté principalement sur la possibilité d'extraire des informations spectrales pertinentes à partir d'une résolution fréquentielle adaptée à la nature du signal. Pour cela, la comparaison des résultats obtenus par le processus de segmentation appliqué indépendamment à différents canaux a été menée. Il en est ressorti que l'état actuel du processus de segmentation ne permet pas la caractérisation d'indices fréquentiels ou de régions temps-fréquence pour lesquelles la spatialisation sonore intervient dans le cadre des signaux audio multicanaux (stéréo et 5.1). En effet, les classes obtenues par segmentation ne concordent pas forcément d'un canal à un autre même si les signaux

analysés sont très corrélés. En outre, la limite entre les classes de signal et la classe de bruit n'est pas nécessairement identique à l'issue du processus appliqué indépendamment à des canaux très corrélés. Finalement, la comparaison entre les RTF segmentées de signaux même très corrélés n'est pas directe et nécessite au préalable une fusion des motifs spectraux associés à une classe dite de signal, selon des critères de corrélation par exemple.

La seconde approche que nous avons abordé a été d'utiliser le processus de segmentation pour localiser les zones temps-fréquence où l'erreur de reconstruction générée par le codage/décodage paramétrique de type BCC n'est pas négligeable (audible) à l'écoute des signaux reconstruits. Pour cela, le calcul du seuil de masquage fréquentiel des signaux reconstruits a été mis en œuvre. Par contre, l'erreur de reconstruction construite par différence entre les signaux originaux et les signaux reconstruits ne permet pas de justifier les dégradations perçues à l'oreille. En effet, la RTF d'une erreur de reconstruction ainsi établie ne reflète pas d'une manière fiable les dégradations qui peuvent être perçues à l'oreille. Dans ces conditions, la segmentation de la RTF de l'erreur de reconstruction audible ne semble pas être adaptée au codage audio spatialisé qui s'appuie sur la perception auditive et non sur une représentation visuelle des signaux.

3.4 Perspectives

Malgré la teneur des résultats obtenus avec les différentes tentatives menées, des perspectives ou orientations de recherche peuvent être envisagées au regard de nos premières expérimentations.

Toutes nos tentatives se sont basées sur le processus de segmentation temps-fréquence classifiant les motifs spectraux de la RTF d'un seul canal. Ensuite, nous avons effectué des comparaisons et analyses à partir de plusieurs RTF segmentées indépendamment les unes des autres. Par conséquent, il serait intéressant de pouvoir segmenter simultanément les RTF d'un signal multicanal. Dans ce cas, l'espace des caractéristiques (EC) contiendrait alors les couples de moments non-centrés d'ordre un et deux (M_1,M_2) de tous les signaux analysés. Pour réduire le nombre de points dans l'EC et surtout pour réaliser une segmentation progressive en fréquence, le procédé pourrait être répété pour chaque sous-bande de fréquence approximant les bandes critiques. Les points de l'EC aux (M_1,M_2) identiques (pour une sous-bande particulière) pourraient alors représenter la classe commune aux signaux considérés. Il conviendrait alors de segmenter les points de l'EC restants (aux caractéristiques différentes) en respectant la contrainte de connexité dans le plan temps-fréquence et cela pour chaque signal à classifier. Cette approche pourrait alors faire évoluer le processus de segmentation jusqu'ici monocanal vers un processus de segmentation multicanal (éventuellement en sous-bandes de fréquences).

De plus, le processus de segmentation jusqu'alors a été appliqué au module carré de la TFCT (spectrogramme) c'est-à-dire sans prendre en compte la phase des signaux si importante sur notre perception des sons (*cf.* la description de l'ITD au paragraphe 1.1). Par conséquent, on peut imaginer mettre en place un processus de segmentation du phasogramme monocanal ou éventuellement multicanal.

L'approche actuellement traitée par Millioz et *al.* dans [MIL06] a été de mettre en place la segmentation de la TFCT (*cf.* Annexe A.1) et par suite de comparer les résultats à ceux obtenus avec la segmentation du spectrogramme. Le principe repose alors sur une caractérisation de la partie réelle et imaginaire des coefficients temps-fréquence de la TFCT. L'estimation de la variance du bruit à partir de la partie réelle [MIL06] permet notamment de réduire le nombre de fausses alarmes du processus de segmentation comparé à celui décrit dans [HOR02]. Le fait d'obtenir la TFCT du signal segmentée autorise une marge de manœuvre plus grande en offrant la possibilité de reconstruire les signaux directement par TFCT inverse. Ces RTF segmentées pourraient alors être utilisées comme des masques temps-fréquence (éventuellement en sous-bandes de fréquences) utiles à la reconstruction de signaux temporels aux caractéristiques fréquentielles classifiées.

4. Modélisation et analyse des signaux audio multicanaux

Avant d'établir un modèle général pour les signaux audio multicanaux, il convient de rappeler la provenance de ces signaux pour pouvoir caractériser le contenu d'une « scène sonore » reproduite par un signal audio multicanal. Ensuite, la modélisation des signaux audio multicanaux constitue l'étape essentielle avant l'analyse des canaux puis l'interprétation qui nous a orienté vers un traitement adapté au contexte de compression audio.

Avant tout formalisme, nous rappelons brièvement, au paragraphe 4.1, comment sont générés les signaux audio multicanaux et ceci indépendamment de leur format. Nous définissons alors les composantes ou objets sonores associés à un signal audio multicanal qui permet la reproduction d'une scène sonore à partir d'un système de restitution approprié. Nous proposons ensuite, au paragraphe 4.2, un modèle de mélange basé sur cette caractérisation du contenu des signaux multicanaux. Enfin, l'analyse des signaux suivant cette modélisation est menée au paragraphe 4.3, les performances et les limites de cette analyse sont données vis-à-vis de notre application à la compression.

4.1 De la scène sonore au signal audio multicanal perçu

4.1.1 Enregistrement et synthèse des signaux audio multicanaux

Une « scène sonore » est par nature liée à des évènements sonores eux-mêmes attachés à des sources ou objets sonores positionnés dans l'espace où se déroule la scène. Outre les sons émis par les sources, l'interaction des sources avec l'espace environnant ajoute à la scène un environnement sonore qualifié dans la littérature d'effet de salle. Une personne située au cœur de la scène sonore identifie les sources en leur associant des attributs spatiaux tels qu'une direction et une distance relative à chaque source sonore. De plus, l'effet de salle perçu par l'auditeur le renseigne sur l'environnement qui l'entoure. La perception de l'effet de salle ne s'attache pas à une direction prédominante mais se réfère à la nature et à la dimension de l'espace environnant. La perception de l'environnement sonore dans une petite pièce est nettement différente de celle obtenue dans un stade par exemple.

Pour qu'un auditeur puisse écouter une scène sonore, soit il est présent au moment où se déroule la scène soit il écoute en différé un enregistrement de la scène considérée. L'enregistrement audio d'une scène sonore consiste à capter simultanément en un ou plusieurs points de l'espace les sons attachés aux sources et l'environnement sonore qui leurs

est associé. Les techniques d'enregistrement audio sont actuellement en pleine effervescence et disposent de multiples techniques de prise de son et de microphones (*cf.* paragraphes 1.2.1.1 et 1.2.2.2). Malgré la diversité de ces techniques, elles ont la caractéristique commune de capter les sources avec un niveau d'énergie et un temps d'arrivée propre à la position de chaque source dans l'espace et par suite, propre à chaque signal capté. De plus, selon l'environnement considéré lors de l'enregistrement (salle de concert, studio, etc.), le son direct peut se mélanger avec ses propres réflexions sur les éléments de l'espace environnant pour donner aux signaux, captés par les microphones, divers degrés de réverbération (*cf.* paragraphe 1.1.2).

Les signaux audio multicanaux peuvent provenir d'un enregistrement réalisé en milieu naturel éventuellement suivi d'un traitement artificiel pour ajuster la dynamique des signaux par exemple. Cependant, la majeure partie des enregistrements audio sont réalisés en studio et suivis d'un mixage dit artificiel. L'enregistrement en studio consiste à enregistrer individuellement les sources sonores (instruments) et à les traiter c'est-à-dire à réaliser un mixage artificiel. Le traitement des sources sonores enregistrées consiste à appliquer des gains ou fonctions de pondération dépendantes du temps qui attribuent aux sources une position dans l'espace sonore de restitution (*cf.* paragraphe 1.2.1.2). Ces sources sonores pondérées sont habituellement réverbérées par des filtres de réverbération artificielle pour simuler la perception d'un environnement naturel [WIN04]-[RUM01]. De façon à accroître la perception spatiale, des réponses impulsionnelles de réverbération (*cf.* **Figure 1.7**) faiblement corrélées sont utilisées pour synthétiser chaque canal.

Les avancées scientifiques et technologies en matière de synthèse sonore permettent aujourd'hui de synthétiser des sons d'instruments, des voix chantées qui apparaissent encore trop synthétiques (typiquement robotique pour la voix chantée) pour reproduire des sons réels. Cependant, les études menées en matière de synthèse sonore environnementale [DOB03] laissent entrevoir la possibilité de synthétiser des scènes sonores complexes capables de reproduire une réalité sonore virtuelle.

4.1.2 Perception spatiale à partir d'un signal audio multicanal

La diffusion d'un signal audio multicanal avec un système de restitution adapté vise à reproduire au mieux la perception dite « spatiale » de la scène sonore originale qui a été enregistrée et éventuellement mixée. Pour cela, les systèmes de restitution sonore multi-haut-parleurs (systèmes stéréo, 5.1 présenté au paragraphe 1.2.2.1, 6.1, 7.1, etc.) enveloppent l'auditeur et diffusent des évènements sonores qui proviennent d'une multitude de directions. Plus le nombre de haut-parleurs est élevé (*cf.* **Figure 4.1**) et plus la scène sonore originale pourra être fidèlement reproduite *i.e.* plus la perception spatiale sera riche (enveloppement, immersion, profondeur, localisation, etc.).

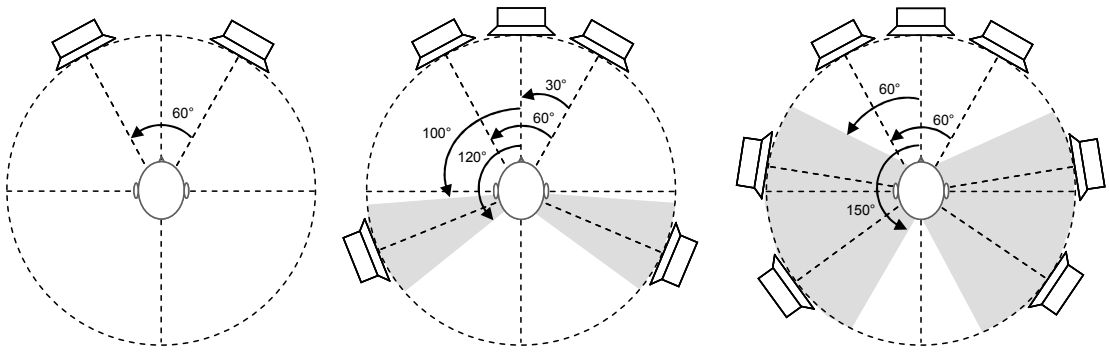

Figure 4.1 Systèmes de diffusion stéréo, 5.1 et 7.1 (de gauche à droite) selon la recommandation de l'UIT-R BS.775-1 [UIT775]. Le canal basse fréquence (Low Frequency Effect LFE ou .1) n'est pas représenté. Les parties grisées représentent les écartements qui peuvent être affectés aux haut-parleurs arrières.

Les dispositions de haut-parleurs recommandés par l'UIT [UIT775] ont été proposées, à l'origine, pour rendre compatible les systèmes audio utilisés au cinéma avec ceux utilisés pour la télévision numérique haute-définition tout en assurant la compatibilité avec les systèmes de diffusion stéréo/mono dits classiques (*cf.* Annexe C.1).

4.1.2.1 Définition des composantes d'un signal audio multicanal

La perception spatiale d'une scène sonore issue de la diffusion d'un signal audio multicanal s'articule autour de la perception des sources sonores d'intérêt dans l'espace de restitution. Ces sources d'intérêt (instrument de musique, parole, etc.) sont considérées comme intelligibles et aisément localisables par l'auditeur. Nous qualifions ces objets sonores de « sources directionnelles », en référence au son direct, dans la mesure où leur perception s'attache à une direction et une distance relatives à la position de l'auditeur. Chaque source directionnelle appartient à au moins un des canaux du signal multicanal. Si la direction attachée à une source ne coïncide pas avec la position d'un haut-parleur alors l'énergie de cette source est répartie dans au moins deux canaux diffusés par les haut-parleurs dont la position est la plus proche de la source *i.e.* opération de mixage ou mélange. Par conséquent, la corrélation inter-canal de ces canaux, qui contiennent chacun une partie de l'énergie de la source considérée, est élevée.

La sensation de réalisme par immersion dans la scène sonore est rendue possible grâce à la diffusion d'un environnement sonore notamment lorsque la zone d'écoute est étendue au moyen d'un système multi-haut-parleurs (*cf.* **Figure 4.1**). L'environnement sonore restitué est souvent qualifié dans la littérature d' « ambiance » sonore. Outre les sources sonores directionnelles, l'ambiance sonore est caractéristique des signaux audio multicanaux et constitue une information supplémentaire à la fois riche et complexe. L'ambiance sonore est située perceptuellement en arrière plan et informe l'auditeur du lieu ou contexte dans le lequel se déroule la scène sonore. L'ambiance sonore peut alors être définie comme « le son du lieu où les sources se propagent ». Un tel contenu audio se réfère alors à l'effet de salle (réflexions et réverbération des sources directionnelles) et à l'accumulation de sources sonores qualifiées de secondaires autrement dit les sources sonores qui n'interviennent pas pour rendre intelligible la scène sonore principale. L'ambiance sonore peut être considérée comme partie intégrante de chaque canal avec un contenu plus ou moins perceptible. En effet, l'ambiance sonore relative à la réverbération des sources est généralement moins perceptible qu'un environnement sonore dissociable des sources (applaudissements par exemple). Une ambiance relative à un environnement sonore peut donc être constituée d'informations dissociables des sources directionnelles *i.e.* des sources secondaires, et à la fois d'informations relatives aux sources directionnelles *i.e.* réflexions et réverbération des sources. Selon les choix liés aux techniques de prise de son et de mixage, les contenus des informations d'ambiances peuvent être mélangés ou attribués à des canaux en particulier. Dans le cas d'une prise de son multicanal dite classique, chaque signal d'ambiance est capté à des points différents de l'espace. Si la distance entre ces points de l'espace est suffisamment grande, ces signaux d'ambiance peuvent être considérés comme décorrélés entre eux et décorrélés des sources directionnelles. Cette hypothèse de faible corrélation entre les composantes d'ambiance d'un signal multicanal est par ailleurs valide dans le cas d'un mixage artificiel puisque les sources sonores sont habituellement filtrées avec des réponses impulsionnelles de réverbération décorrélées d'un canal à un autre (par déphasage comme indiqué au paragraphe 2.2.3.2 ou en Annexe C.2.3 par exemple).

4.1.2.2 Différentes catégories de signaux audio multicanaux

Bien qu'à l'origine les signaux audio multicanaux faisaient partie intégrante des contenus audio-visuels (cinématographiques, télévisuels), les créations musicales et radiophoniques utilisent de plus en plus une représentation (prise et restitution) multicanale du son. Cette évolution des applications nous permet de distinguer principalement deux catégories de signaux [RUM01].

La première catégorie de signaux audio multicanaux est principalement dédiée aux contenus audio-visuels dans la mesure où une analogie entre le son et l'image est utilisée. Les évènements sonores attachés au premier plan de l'image (personnages d'un film par exemple) sont reproduits par les canaux frontaux (dialogue sur le canal central notamment). A l'inverse, les canaux arrières diffusent principalement les informations sonores liées aux évènements visuels qui apparaissent en arrière-plan. Cette séparation avant-arrière des canaux diffusés par les haut-parleurs correspond à la représentation la plus communément employée. Ce type de signal audio multicanal (type I) retranscrit une scène sonore où l'information principale des sources directionnelles provient de l'image frontale issue des haut-parleurs frontaux (en face de l'auditeur). A l'inverse, les haut-parleurs arrières diffusent une ambiance sonore (applaudissements, pluie, musique de fond, etc.) qui renseigne l'auditeur sur l'environnement de la scène originale.

La seconde catégorie (type II) de signaux utilisée pour la musique ou la radio haute-définition se veut plus générale. En effet, les signaux multicanaux de musique, notamment sur les DVD-audio, ont parfois l'objectif de procurer à l'auditeur la sensation d'être plongé au milieu du groupe c'est-à-dire entouré par les musiciens en percevant les instruments à des positions différentes et qui ne se limitent pas aux positions frontales. Dans cette configuration, la séparation avant-arrière des canaux (type I) reste un cas particulier restrictif d'un point de vue de l'utilisation de l'espace sonore.

A titre d'illustration, la **Figure 4.2** présente deux signaux audio multicanaux aux formats 5.0 (le canal basse fréquence n'est pas représenté). La **Figure 4.2-(a)** présente un signal multicanal de type I où les sources directionnelles S_i (i=1,…,4) sont toutes diffusées par les haut-parleurs frontaux. La **Figure 4.2-(b)** illustre l'occupation complète de l'espace sonore par les sources directionnelles d'un signal multicanal de type II. Pour ces deux catégories, chaque canal C_i (i=1,…,5) est constitué d'un signal d'ambiance A_i (i=1,…,5). Ces ambiances sont relatives à des informations propres aux sources directionnelles (réflexions, réverbération) et éventuellement à des sources secondaires dissociables des sources directionnelles. Par conséquent, il n'y a pas de correspondance directe entre les signaux d'ambiance du signal de type I de la **Figure 4.2-(a)** et ceux du signal de type II de la **Figure 4.2-(b)**.

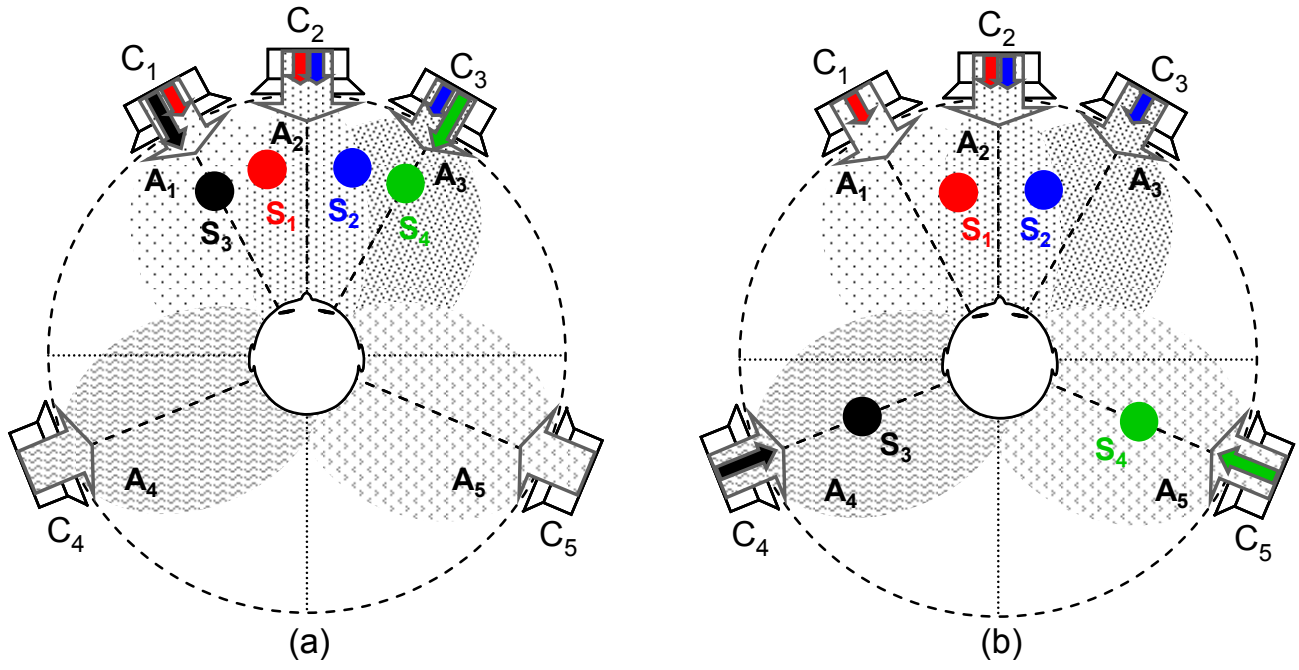

Figure 4.2 **Exemples de signaux audio multicanaux (C_1,…,C_5) au format 5.0 [(a) - de type I, (b) - de type II]**. Ces signaux sont constitués des sources directionnelles S_i (i=1,…,4) et des ambiances sonores (A_1,…,A_5).

4.2 Mélange instantané de sources directionnelles et d'ambiances

Un signal audio multicanal, issu d'un enregistrement en milieu naturel et/ou d'un mixage artificiel, peut être caractérisé par la notion de mélange de plusieurs sources directionnelles et d'ambiances sonores.

4.2.1 Définitions et hypothèses

D'après la caractérisation d'un signal multicanal établie au paragraphe 4.1, nous définissons un modèle de mélange instantané pour les signaux multicanaux de type I et II. Le mélange fait intervenir à la fois les sources directionnelles et l'ambiance sonore qui regroupe toutes les autres sources que l'on qualifie de secondaires ainsi que l'effet de salle. Les canaux sont tous définis comme la somme pondérée de sources directionnelles et d'ambiances sonores. Par nature, un mélange peut être considéré comme sur-déterminé (plus de canaux que de sources) ou déterminé (autant de canaux que de sources) mais en général si l'on considère un mélange multicanal « réel » (cinématographique, musical, etc.) par opposition à synthétique, le mélange est sous-déterminé (moins de canaux que de sources).

Une source directionnelle ou une ambiance sonore est un signal audio numérique qui est considéré comme la réalisation (résultat d'une expérience) d'un processus stochastique à temps discret uniformément espacé selon une fréquence d'échantillonnage (celle-ci étant choisie deux fois supérieure à la plus grande composante fréquentielle du processus) [HAY01]. Une source directionnelle est un processus stochastique fonction du temps et caractérisée par une position dans l'espace. Une ambiance est également un processus stochastique fonction du temps et de l'espace tel que ses réalisations en deux points de l'espace sont faiblement corrélées. Les réalisations d'un processus stochastique d'ambiance sont décorrélées si la distance entre les points de l'espace est suffisamment grande et, par nature, décorrélées de la réalisation d'une source directionnelle.

Une source directionnelle S_d, $\forall d \in [1,\ldots,D]$ étant le numéro de la source, peut être considérée comme une variable aléatoire qui prend ses valeurs $s_d[n]$ aux instants discrets n=[1,…,N_T]. De même, les réalisations d'une ambiance A s'écrivent $a[n]$ avec n=[1,…,N_T]. Ainsi, un canal C défini comme le mélange instantané d'un vecteur aléatoire de D sources directionnelles $\mathbf{S_D} = \left(S_1,\cdots,S_d,\cdots,S_D\right)^T$ et d'une ambiance sonore A s'écrit :

$$C = \mathrm{G}_\mathrm{D}^T \cdot \mathrm{S}_\mathrm{D} + A, \tag{4.1}$$

où ^{T} dénote la transposition matricielle. Précisément, le vecteur aléatoire de sources directionnelles est pondéré par une matrice de gains réels $\mathbf{G_D} = \left(G_1,\cdots,G_d,\cdots,G_D\right)^T$ constituée des fonctions de pondération G_d qui prennent les valeurs $g_d[n]$ aux instants n=[1,…, N_T] et ceci $\forall d \in [1,\ldots,D]$.

Ce mélange de variables aléatoires défini par l'équation (4.1) peut également être considéré comme une variable aléatoire à part entière. Par conséquent, une réalisation particulière aux instants discrets n=[1,…, N_T] d'un canal C suivant le mélange décrit par l'équation (4.1) s'exprime :

$$c[n] = \sum_{d=1}^{D}\left(g_d[n] \times s_d[n]\right) + a[n]. \tag{4.2}$$

Ainsi, chaque canal est défini comme la somme de sources directionnelles, pondérées par des gains qui traduisent leurs positions dans l'espace sonore de restitution, avec un signal

d'ambiance. Par conséquent, un signal multicanal $\mathbf{C_M} = \left(C_1, \cdots, C_m, \cdots, C_M\right)^T$ à M canaux, est constitué du mélange de D sources directionnelles $\mathbf{S_D} = \left(S_1, \cdots, S_d, \cdots, S_D\right)^T$ pondérées par les gains réels $\left(G_{md}\right)_{\substack{1 \le m \le M \\ 1 \le d \le D}}$ dont l'expression matricielle s'écrit :

$$\mathbf{G_{MD}} = \begin{pmatrix} G_{11} & \cdots & G_{1D} \\ \vdots & \ddots & \vdots \\ G_{M1} & \cdots & G_{MD} \end{pmatrix}. \tag{4.3}$$

Nous considérons donc qu'un signal multicanal à M canaux contient M signaux d'ambiance $\mathbf{A_M} = \left(A_1, \cdots, A_m, \cdots, A_M\right)^T$, c'est-à-dire un par canal. Finalement, l'expression du signal audio multicanal $\mathbf{C_M}$ est donnée par :

$$\mathbf{C_M} = \mathbf{G_{MD}} \cdot \mathbf{S_D} + \mathbf{A_M}. \tag{4.4}$$

Finalement, une réalisation particulière du canal d'indice $m \in [1,\ldots,M]$, $c_m[n]$, s'écrit :

$$c_m[n] = \sum_{d=1}^{D} \left[g_{md}[n] \cdot s_d[n] \right] + a_m[n] \tag{4.5}$$

avec G_{md} les gains de pondération, aux instants n=[1,…N_T], $g_{md}[n]$ appliqués respectivement à une réalisation particulière de chaque source directionnelle S_d en un signal à temps discret $s_d[n]$ $\forall d \in \left[1, \cdots, D\right]$. De même, les sources directionnelles pondérées sont sommées à une réalisation particulière de A_m en un signal $a_m[n]$ avec n=[1,…, N_T].

4.2.2 Décomposition en valeurs propres de la covariance

De façon à étudier les relations entre les canaux et à analyser les informations spatiales d'un signal multicanal, nous nous intéressons maintenant à la matrice de covariance de tels signaux. La covariance d'un signal multicanal suivant les hypothèses du modèle qui est défini par l'équation (4.4) s'écrit :

$$\begin{aligned} \mathbf{\Gamma_{C_M}} &= \mathrm{E}\left[\overline{\mathbf{C}}_\mathbf{M} \cdot \overline{\mathbf{C}}_\mathbf{M}^{\,T} \right] \\ \mathbf{\Gamma_{C_M}} &= \mathbf{G_{MD}} \cdot \mathrm{E}\left[\overline{\mathbf{S}}_\mathbf{D} \cdot \overline{\mathbf{S}}_\mathbf{D}^{\,T} \right] \cdot \mathbf{G_{MD}}^T + \mathrm{E}\left[\overline{\mathbf{A}}_\mathbf{M} \cdot \overline{\mathbf{A}}_\mathbf{M}^{\,T} \right] \\ \mathbf{\Gamma_{C_M}} &= \mathbf{G_{MD}} \cdot \mathbf{\Gamma_{S_D}} \cdot \mathbf{G_{MD}}^T + \mathbf{\Gamma_{A_M}} \end{aligned} \tag{4.6}$$

L'estimation de la matrice de covariance $\mathbf{\Gamma_{C_M}}$ repose sur le signal multicanal centré $\overline{\mathbf{C}}_\mathbf{M}$ défini tel que chaque canal centré d'indice m soit obtenu de la manière suivante :

$$\overline{C}_m = C_m - \mathrm{E}\left[C_m\right], \forall m \in \left[1, \cdots, M\right], \tag{4.7}$$

avec E[.] définie comme l'espérance mathématique. Par conséquent, la matrice de covariance $\mathbf{\Gamma_{C_M}}$ est fonction de la matrice de covariance des sources directionnelles $\mathbf{\Gamma_{S_D}}$ et de la matrice de covariance des ambiances sonores $\mathbf{\Gamma_{A_M}}$.

Les sources directionnelles sont considérées comme étant décorrélées à l'ordre 2, par conséquent, la matrice de covariance $\mathbf{\Gamma}_{\mathbf{S_D}}$ est diagonale. En définissant la variance de la source directionnelle ou variable aléatoire S_d comme :

$$\sigma_{S_d}^2 = \mathrm{E}\left[\left(S_d - \mathrm{E}\left[S_d\right]\right)\left(S_d - \mathrm{E}\left[S_d\right]\right)^T\right], \tag{4.8}$$

la matrice de covariance des sources directionnelles s'écrit :

$$\mathbf{\Gamma}_{\mathbf{S_D}} = \begin{pmatrix} \sigma_{S_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{S_2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{S_D}^2 \end{pmatrix}. \tag{4.9}$$

De la même manière, en définissant la variance d'un signal d'ambiance comme :

$$\sigma_{A_i}^2 = \mathrm{E}\left[\left(A_i - \mathrm{E}\left[A_i\right]\right)\left(A_i - \mathrm{E}\left[A_i\right]\right)^T\right],\ \forall i \in \left[1; M\right] \tag{4.10}$$

et l'indice de corrélation des signaux d'ambiances A_i et A_j comme :

$$\rho_{A_iA_j} = \frac{E\left[\left(A_i - \mathrm{E}\left[A_i\right]\right)\cdot\left(A_j - \mathrm{E}\left[A_j\right]\right)^T\right]}{\sqrt{\sigma_{A_i}^2\sigma_{A_j}^2}},\ \forall i, j \in \left[1; M\right] \tag{4.11}$$

la matrice de covariance des ambiances sonores $\mathbf{\Gamma}_{\mathbf{A_M}}$ s'écrit alors :

$$\mathbf{\Gamma}_{\mathbf{A_M}} = \begin{pmatrix} \sigma_{A_1}^2 & \rho_{A_1A_2}\sqrt{\sigma_{A_1}^2\sigma_{A_2}^2} & \cdots & \rho_{A_1A_M}\sqrt{\sigma_{A_1}^2\sigma_{A_M}^2} \\ \rho_{A_2A_1}\sqrt{\sigma_{A_2}^2\sigma_{A_1}^2} & \sigma_{A_2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{A_{M-1}A_M}\sqrt{\sigma_{A_{M-1}}^2\sigma_{A_M}^2} \\ \rho_{A_MA_1}\sqrt{\sigma_{A_M}^2\sigma_{A_1}^2} & \cdots & \rho_{A_MA_{M-1}}\sqrt{\sigma_{A_M}^2\sigma_{A_{M-1}}^2} & \sigma_{A_M}^2 \end{pmatrix} \tag{4.12}$$

D'après l'équation (4.6) et en combinant les équations (4.3), (4.9) et (4.12), les éléments d'auto-covariance notés $\left(r_{ii}\right)_{1\le i\le M}$, c'est-à-dire les termes de la diagonale de $\mathbf{\Gamma}_{\mathbf{C_M}}$, s'écrivent :

$$r_{ii} = \sum_{d=1}^{D} g_{id}^2\sigma_{S_i}^2 + \sigma_{A_i}^2 . \tag{4.13}$$

De la même manière, les éléments d'inter-covariance notés $\left(r_{ij}\right)_{1\le i,j\le M}$ de $\mathbf{\Gamma}_{\mathbf{C_M}}$ s'écrivent :

$$r_{ij} = \sum_{d=1}^{D}\left(g_{id}\cdot g_{jd}\right)\sigma_{S_d}^2 + \rho_{A_iA_j}\sqrt{\sigma_{A_i}^2\sigma_{A_j}^2} . \tag{4.14}$$

La matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_M}}$ est symétrique et par conséquent diagonalisable. La diagonalisation de la matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_M}}$ est établie par une décomposition en valeurs propres décrite par :

$$\mathbf{V_M} \cdot \mathbf{\Gamma}_{\mathbf{C_M}} \cdot \mathbf{V_M^{-1}} = \mathbf{\Lambda_M}, \text{ avec: } \mathbf{\Lambda_M} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_M \end{pmatrix}. \tag{4.15}$$

L'équation (4.15) fait intervenir la matrice $\mathbf{V_M}$ des vecteurs propres de la matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_M}}$ donnée par :

$$\mathbf{V_M} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1M} \\ v_{21} & v_{22} & \cdots & v_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{M1} & v_{M2} & \cdots & v_{MM} \end{pmatrix} = \begin{pmatrix} V_1 & V_2 & \cdots & V_M \end{pmatrix}^T. \tag{4.16}$$

Par nature, $\mathbf{V_M}$ est une matrice orthogonale de dimension $M \times M$, ainsi les vecteurs $V_i = \left(v_{i1}, \cdots, v_{iM}\right)$ et $V_j = \left(v_{j1}, \cdots, v_{jM}\right)$ sont orthogonaux $\forall i \neq j$ et $i, j \in \left[1, \cdots, M\right]$. Par conséquent, la diagonalisation de la matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_M}}$ revient à calculer la matrice des $\mathbf{V_M}$ et son inverse par simple transposition matricielle puisque $\mathbf{V_M^{-1}} = \mathbf{V_M^T}$, pour toute matrice orthogonale [GOL96]. Pour cela, il convient en premier lieu de calculer les valeurs propres $\lambda_m, \forall m \in \left[1, \cdots, M\right]$ racines du polynôme caractéristique noté :

$$P_M = \det\left(\mathbf{R}_{\mathbf{C_M}} - \lambda \mathbf{I_M}\right). \tag{4.17}$$

où $\mathbf{I_M}$ est la matrice identité. Les M vecteurs propres peuvent ensuite être calculés comme indiqué par l'équation (4.18).

$$\left(\mathbf{R}_{\mathbf{C_M}} - \lambda_i \mathbf{I_M}\right) \cdot V_i = 0, \ \forall i \in \left[1, \cdots, M\right] \tag{4.18}$$

La décomposition en valeurs propres d'un signal à M canaux permet d'extraire M sous-espaces propres orthogonaux V_{i}, $i \in \left[1, \cdots, M\right]$ relatifs aux M valeurs propres définies telles que : $\lambda_1 > \lambda_2 > \cdots > \lambda_M$.

La distribution des valeurs propres de la covariance délivre des informations sur la répartition énergétique des composantes du signal multicanal. Des analyses dans les domaines temporel et fréquentiel sont menées à partir de la covariance de dimension deux (signaux stéréophoniques) aux paragraphes suivants.

4.2.2.1 Analyse temporelle de dimension deux

Principe

L'analyse des caractéristiques d'un signal stéréophonique (M=2), exprimé dans le domaine temporel, suivant le modèle présenté dans le paragraphe 4.2.1 est réalisée au moyen de la matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_2}}$, obtenue à partir de l'équation (4.6) en posant M=2, dont les termes sont définis par les équations (4.13) et (4.14) tels que :

$$\begin{cases} r_{11} = \sum_{d=1}^{D} g_{1d}^2 \sigma_{S_d}^2 + \sigma_{A_1}^2, \\ r_{12} = r_{21} = \sum_{d=1}^{D} \left(g_{1d} \cdot g_{2d} \right) \sigma_{S_d}^2 + \rho_{A_1 A_2} \sqrt{\sigma_{A_1}^2 \sigma_{A_2}^2}, \\ r_{22} = \sum_{d=1}^{D} g_{2d}^2 \sigma_{S_d}^2 + \sigma_{A_2}^2. \end{cases} \tag{4.19}$$

Par conséquent, les racines du polynôme caractéristique de l'équation (4.17) de dimension deux correspondent aux valeurs propres estimées à partir de la covariance $\mathbf{\Gamma}_{\mathbf{C_2}}$ d'un signal stéréophonique telles que :

$$\lambda_{1,2} = \frac{1}{2}\left(r_{11} + r_{22} \pm \sqrt{\left(r_{11} - r_{22} \right)^2 + \left(2 \times r_{12} \right)^2} \right). \tag{4.20}$$

Finalement, l'expression des valeurs propres dépend des gains appliqués aux sources, contenus dans la matrice $\mathbf{G_{2D}}$, des variances des sources directionnelles $\sigma_{S_1}^2, \cdots, \sigma_{S_D}^2$ et des ambiances $\sigma_{A_1}^2$ et $\sigma_{A_2}^2$ (*cf.* équations (4.8) et (4.10)) ainsi que du coefficient de corrélation des ambiances sonores $\rho_{A_1 A_2}$ (*cf.* équation (4.11)). λ_1 a été choisie arbitrairement comme la plus grande valeur propre en considérant un signe positif dans l'équation (4.20).

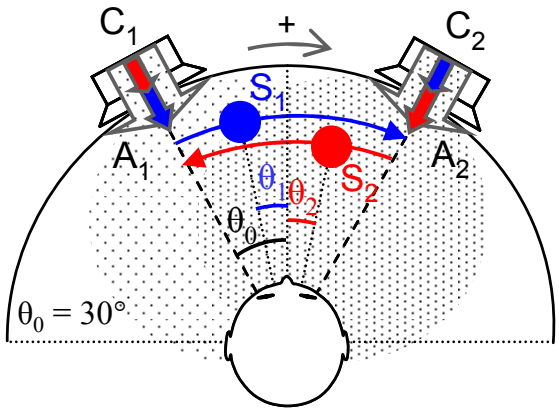

Figure 4.3 **Mélange synthétique par mixage artificiel de signaux enregistrés en milieu naturel.** Le signal stéréophonique (C_1,C_2) est généré par la somme de D=2 sources directionnelles (S_1 en bleu et S_2 en rouge) et d'un signal stéréophonique d'ambiances sonores (A_1, A_2 représentées par des textures). Les positions des sources directionnelles perçues, repérées par les angles θ_1 et θ_2, dépendent des gains fixés par le panoramique d'intensité.

Un signal stéréophonique synthétique (C_1,C_2), présenté à la **Figure 4.3**, est généré d'après l'équation (4.5) avec M=2, par la somme de :

- D=2 sources directionnelles S_1 et S_2 pondérées par la matrice $\mathbf{G_{22}}$
- d'un signal stéréophonique d'ambiances sonores (A_1, A_2).

Les sources directionnelles sont pondérées par la matrice de gains $\mathbf{G_{2D}}$ qui peut-être à dépendance temporelle de façon à faire évoluer l'azimut des sources au cours du temps par un panoramique d'intensité (*cf.* paragraphe 1.2.1.2).

Nous avons négligé l'effet de salle sur les sources directionnelles puisqu'elles ne sont pas réverbérées. En d'autres termes, les composantes réverbérées (réflexions) des sources directionnelles sont considérées comme négligeables d'un point de vue énergétique comparé aux ambiances sonores déjà utilisées. C'est d'ailleurs de ce point de vue que nous définissons le rapport de puissance des sources directionnelles aux ambiances (RSDA) :

$$\mathrm{RSDA} = 10 \times \log_{10} \left(\frac{\frac{1}{D} \sum_{d=1}^{D} \sigma_{S_d}^2}{\frac{1}{M} \sum_{m=1}^{M} \sigma_{A_m}^2} \right) \mathrm{dB} \tag{4.21}$$

comme le rapport (en dB) de la moyenne des puissances des sources directionnelles sur la moyenne des puissances des ambiances. Il est alors possible de synthétiser un signal stéréo dont la puissance moyenne des ambiances est relative à la puissance moyenne des sources directionnelles.

Expérimentation

A partir de l'expression théorique des valeurs propres de la covariance d'un signal stéréophonique (équation (4.20)) suivant le modèle proposé, un cas particulier basé sur des signaux audio tirés d'enregistrements naturels est présenté de façon à en déduire la distribution des valeurs propres d'un signal au contenu totalement maîtrisé.

Nous avons fixé les gains $\left(g_{md}\right)_{1 \le m,d \le 2}$ de la matrice $\mathbf{G_{22}}$ (D=2) en utilisant la loi des sinus, définie au paragraphe 1.2.1.2, pour réaliser le panoramique d'intensité. La position de la source S_d, repérée par l'angle θ_d, est obtenue en la pondérant par le couple de gains ($g_{1d}[n]$, $g_{2d}[n]$) donné par :

$$\frac{g_{1d}[n]}{g_{2d}[n]} = \frac{\sin\theta_0 + \sin\theta_d[n]}{\sin\theta_0 - \sin\theta_d[n]}, \ \forall d \in [1;2], \tag{4.22}$$

où θ_0 correspond à l'écartement entre les haut-parleurs. Nous traitons le cas où les sources directionnelles ont des positions initiales et des trajectoires opposées : $\theta_2[n]$=-$\theta_1[n]$, n=[1,…,N_T].

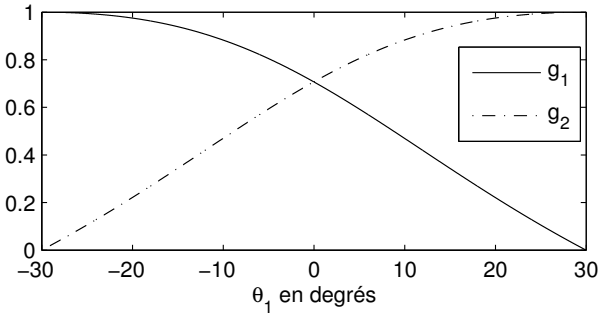

Figure 4.4 Gains g_1 et g_2 utilisés pour le panoramique d'intensité selon la loi des sinus d'une source (S_1 repérée par l'angle θ_1) se déplaçant du haut-parleur gauche (-30°) au haut-parleur droit (+30°).

De cette manière, les sources directionnelles sont perçues comme se rapprochant l'une de l'autre jusqu'à coïncider au milieu de l'image sonore stéréo avant de s'écarter jusqu'à des positions finales opposées à leur position initiale (*cf.* **Figure 4.3)**. Ces trajectoires opposées simplifient la matrice de gains $\mathbf{G_{22}}$ telle que :

$$\begin{cases} g_{11}[n] = g_{22}[n] = g_1[n] \\ g_{12}[n] = g_{21}[n] = g_2[n] \end{cases} \tag{4.23}$$

La **Figure 4.4** présente le couple de gains ($g_1[n]$,$g_2[n]$), établi d'après l'équation (4.22), appliqué à la source S_1 dont la position azimutale évolue de $\theta_l[1].$=-30° à $\theta_l[N_T].$=30°.

Le signal stéréo (C_1,C_2) a été généré à partir de la source directionnelle S_1 relative à un signal de parole chantée et de la source S_2 relative à un signal harmonique de glockenspiel (train d'impulsions). Le signal stéréophonique d'ambiances sonores (A_1, A_2) provient d'un enregistrement naturel réalisé dans le hall d'un aéroport (bruits de pas, de chariots, etc.). La **Figure 4.6** présente les RTF (*cf.* Annexe A.1) des signaux originaux (S_1, S_2, A_1 et A_2) ainsi que celles des canaux synthétisés (C_1 et C_2) d'après les équations (4.5), (4.22) et (4.23) avec un RSDA de +15 dB. Ces RTF sont toutes obtenues avec une fenêtre de *Hanning* à N=512 points, 1024 coefficients spectraux sont calculés (Z=512) et un recouvrement de 50% entre les fenêtres glissantes est utilisé.

A partir de l'équation (4.11), les indices d'inter-corrélation des signaux originaux et synthétisés sont estimés au cours du temps avec une analyse par fenêtre glissante de taille N=2048 points qui se recouvrent à 50%. D'après la **Figure 4.5**, malgré la faible corrélation des signaux d'ambiance, les canaux synthétisés sont corrélés et cela d'autant plus que les gains g_1 et g_2 convergent vers la même valeur.

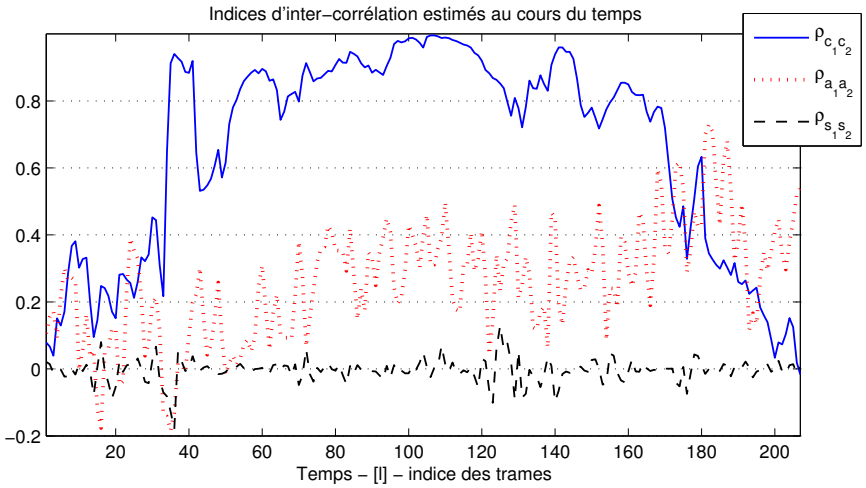

Figure 4.5 **Indices d'inter-corrélation :** des canaux C_1 et C_2 (en trait plein bleu), des signaux d'ambiance A_1 et A_2 (en pointillés rouge) et des sources directionnelles S_1 et S_2 (en trait discontinu noir).

D'un point de vue perceptif, plus les canaux synthétisés sont corrélés et plus les sources directionnelles sont perçues au milieu de l'image sonore restituée qui est réduite (centrale). A l'inverse, lorsque les canaux du signal stéréo synthétisé sont faiblement corrélés, nous pouvons percevoir les sources à des azimuts différents *i.e.* l'image sonore stéréo est élargie.

Plutôt que de faire évoluer la puissance des ambiances au cours du temps, nous avons synthétisé quatre signaux stéréo dont la puissance des ambiances est pondérée par un coefficient déduit du RSDA, défini à l'équation (4.21), allant de +5 à +20 dB avec un pas de +5 dB.

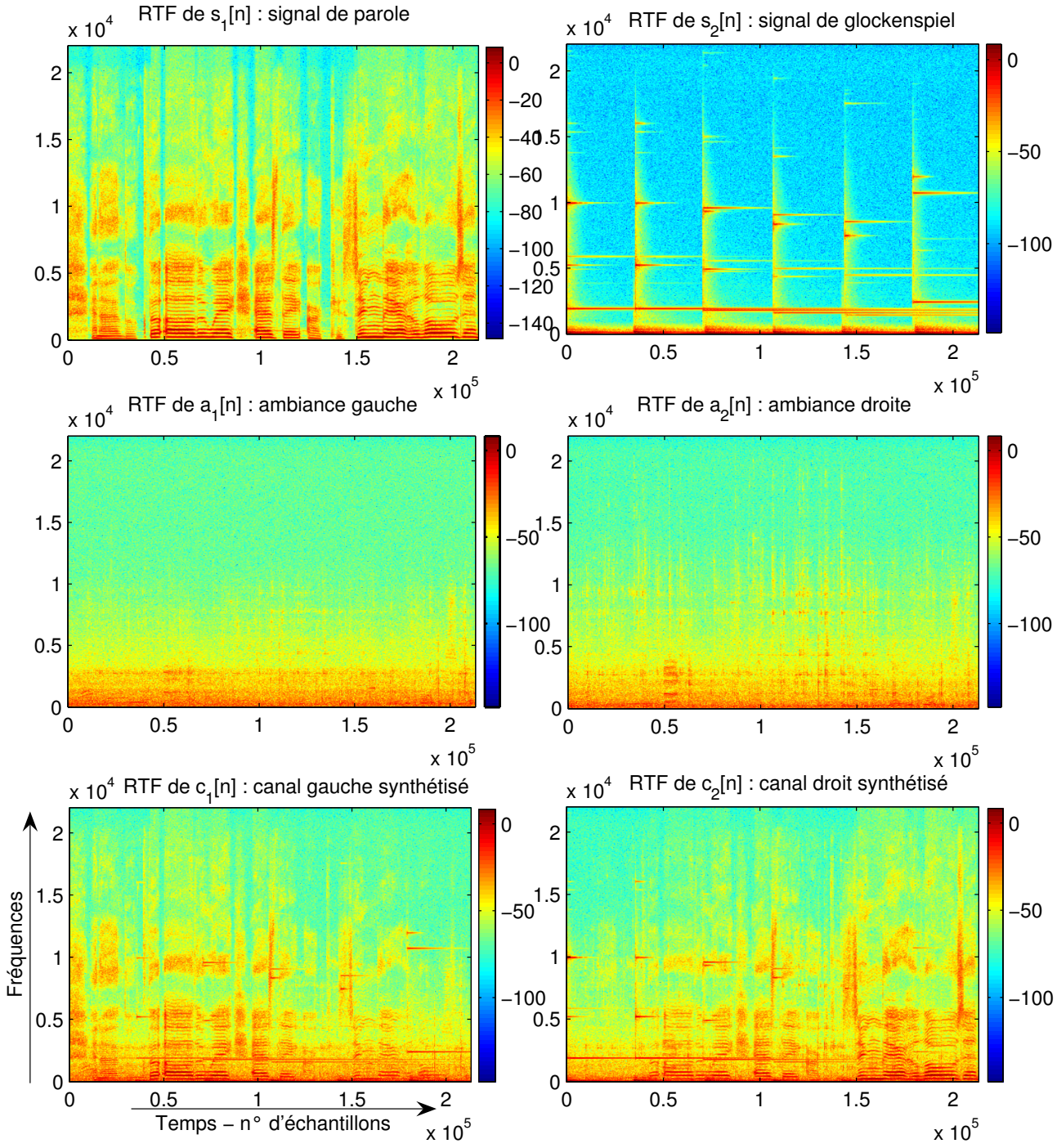

Figure 4.6 **RTF en dB de (a) –** S_1 : signal de parole, **(b) –** S_2 : signal de glockenspiel, **(c) –** A_1 : signal d'ambiance gauche, **(d) –** A_2 : signal d'ambiance droit, **(e) –** C_1 : canal gauche synthétisé et **(f) –** C_2 : canal droit synthétisé. Le RDSA vaut 15 dB.

Finalement, les covariances de ces signaux stéréophoniques sont estimées au cours du temps au regard du système d'équations (4.19). Les puissances des signaux et les indices de corrélation des ambiances sont estimés en temps sur des portions glissantes de signal définies par les paramètres suivants :

- fenêtrage des signaux par une fenêtre sinus (indice de trame l)
- fenêtre de longueur N=1024 échantillons
- recouvrement des fenêtres à 50%.

Les valeurs propres sont estimées au cours du temps à partir des équations (4.19) et (4.20). La **Figure 4.7** présente une comparaison entre :

- la plus grande valeur propre λ_1 et la puissance de la source directionnelle dominante, définie comme :

$$\sigma^2_{S_{\max}} = 10 \times \log_{10}\left(\max\left(\sigma^2_{S_1}, \sigma^2_{S_2}\right)\right) \tag{4.24}$$

- la plus petite valeur propre λ_2 et la puissance moyenne des ambiances :

$$\sigma^2_{A_{mean}} = 10 \times \log_{10}\left(\frac{\sigma^2_{A_1} + \sigma^2_{A_2}}{2}\right) \tag{4.25}$$

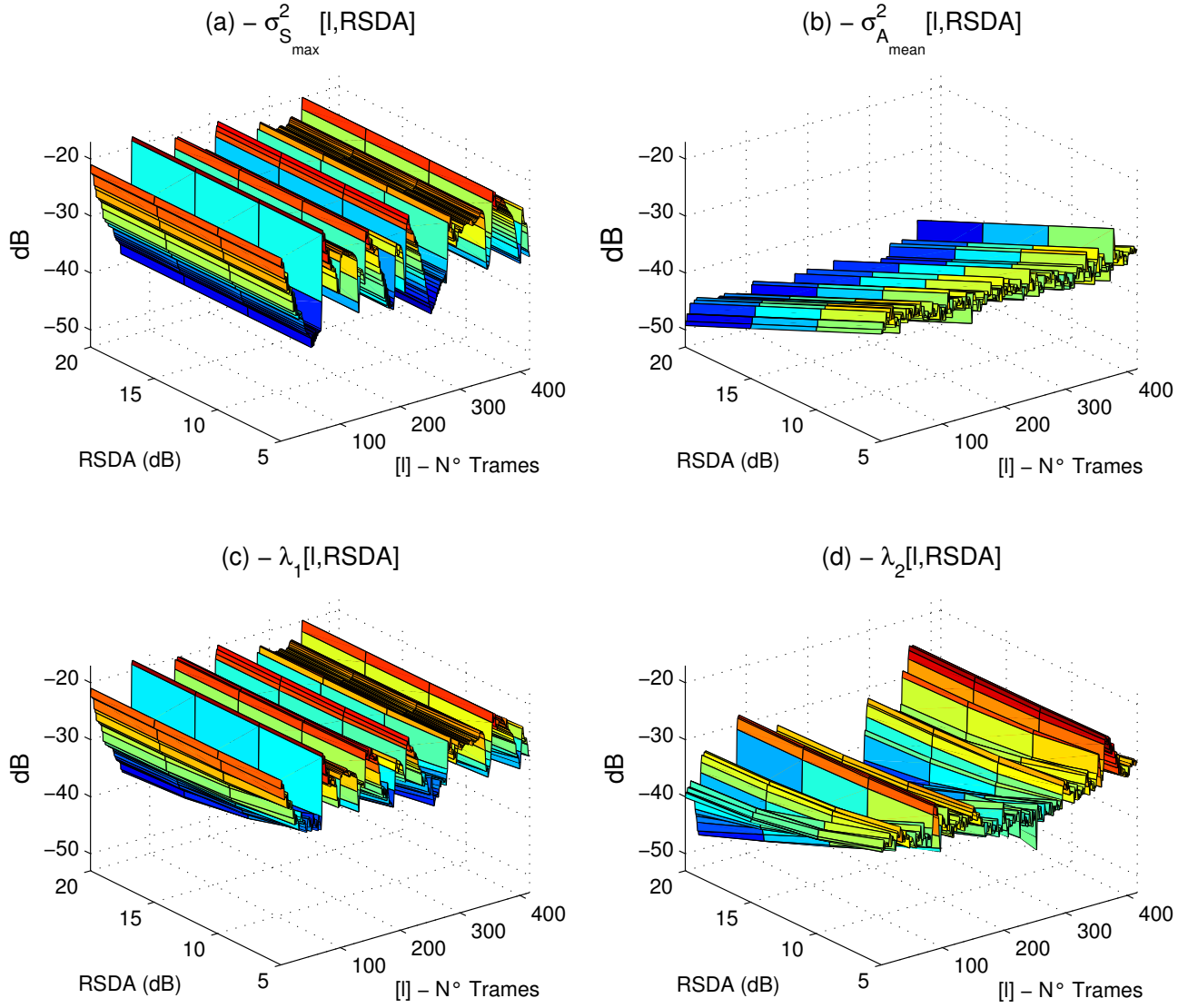

Figure 4.7 **Comparaison des puissances de la source dominante (a) et de l'ambiance moyenne (b) aux valeurs propres estimées (c) - (d)** avec une analyse par fenêtre glissante à partir de 4 signaux stéréo aux RSDA différents.

D'après la **Figure 4.7-(a)**, la puissance de la source directionnelle dominante varie au cours du temps mais ne dépend pas du RSDA. La puissance moyenne des ambiances sonores évolue au cours du temps et également en fonction du RSDA d'après la **Figure 4.7-(b)**. La **Figure 4.7-(c)** montre que la plus grande valeur propre correspond à la puissance de la source dominante sommée avec une partie de la puissance des ambiances lorsque le RSDA diminue. La plus petite valeur propre correspond à la puissance moyenne des ambiances

sommée à la puissance des sources secondaires, d'après la **Figure 4.7-(d)**. Nous considérons qu'à tout instant, les sources directionnelles ont des niveaux d'énergie différents et que la source directionnelle d'énergie la plus faible est la source secondaire par opposition à dominante.

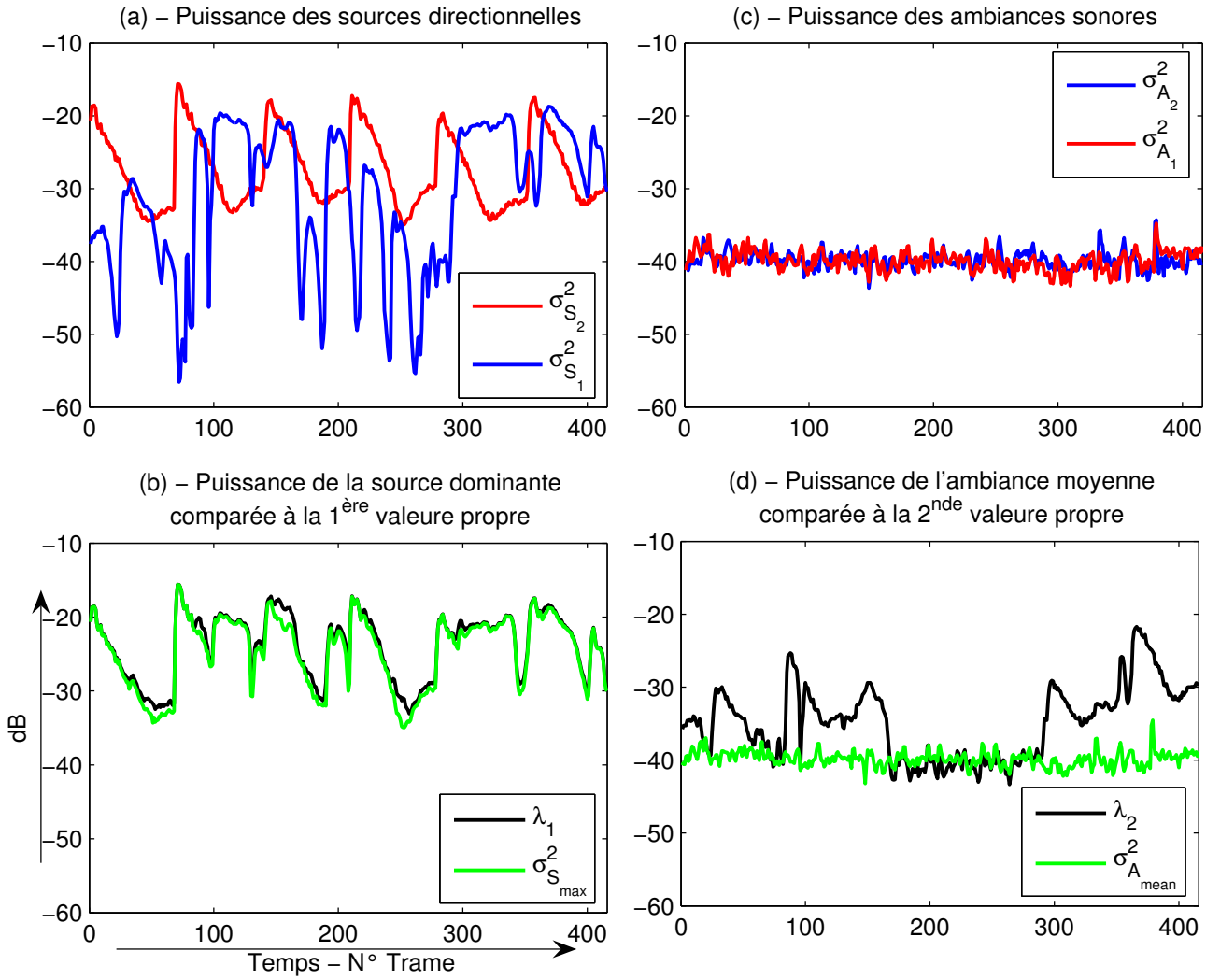

Figure 4.8 **Comparaison des puissances des signaux originaux aux valeurs propres estimées** avec une analyse par fenêtre glissante d'un signal stéréo synthétique avec un RSDA de 15 dB. **(a) -** Puissances des sources directionnelles : glockenspiel (S_2 en rouge) et parole chantée (S_1 en bleu). **(b) -** La plus grande valeur propre correspond à la puissance de la source dominante sommée avec une partie de l'ambiance sonore. **(c) -** Puissances des ambiances sonores A_1 et A_2. **(d) -** La plus petite valeur propre correspond à la puissance moyenne des ambiances plus la puissance des sources secondaires.

Une interprétation plus fine est donnée pour le signal stéréophonique avec un RSDA égal à +15 dB. La **Figure 4.8-(b)** et la **Figure 4.8-(d)** correspondent à une coupe transversale des courbes de la **Figure 4.7**. D'après la **Figure 4.8-(b),** la plus grande valeur propre estimée en temps est équivalente à la puissance de la source dominante additionnée à la puissance de la source secondaire et d'une partie de l'ambiance si elles coïncident spatialement avec la source dominante. La plus faible valeur propre estimée subit également l'influence de la position spatiale des sources directionnelles et par suite de la nature du mélange. En effet, lorsque les sources coïncident en azimut (même couple de gain pour les deux sources), elles fusionnent dans l'espace et le mélange peut être considéré comme sur-déterminé (plus de canaux que de sources). A l'inverse, lorsque les sources ont des azimuts différents, le mélange est déterminé (autant de canaux que de sources). Par conséquent, comme le montre la **Figure 4.8-(d)**, la plus faible valeur propre correspond à la puissance moyenne des

ambiances lorsque les sources directionnelles coïncident dans l'espace (pour les trames d'indices : 180<*l*<280). A l'inverse lorsque les sources directionnelles ont des azimuts différents (*l*<180 et *l*>280), la plus petite valeur propre correspond à la puissance moyenne des ambiances additionnée à la puissance de la source sonore secondaire, par exemple S_2 pour les trames d'indices *l*>370 d'après la **Figure 4.8-(a)** et **Figure 4.8-(d)**.

Cette analyse d'un signal particulier illustre une limite à la décomposition en valeurs propres de la covariance d'un signal stéréophonique. En effet, cette analyse réalisée dans le domaine temporel ne permet pas de dissocier complètement les sources directionnelles des ambiances sonores du mélange. La puissance des sources sonores secondaires, qui rendent le mélange déterminé lorsque l'azimut des sources directionnelles diffère, est sommée avec la puissance moyenne des ambiances dans la plus petite valeur propre.

4.2.2.2 Analyse par sous-bandes de fréquences de dimension deux

Une analyse de la covariance d'un signal stéréophonique dans le domaine fréquentiel par sous-bandes de fréquences est présentée de manière à comparer les distributions des valeurs propres obtenues avec une analyse en temps et une analyse par sous-bandes de fréquences.

Principe

La **Figure 4.9** présente le schéma de principe pour l'extraction des valeurs propres par sous-bandes de fréquences. L'idée principale étant de permettre la comparaison directe des distributions des valeurs propres estimées en temps et en sous-bandes à partir de l'équation (4.20). Pour y parvenir, nous proposons de calculer les valeurs propres à partir des matrices de covariance estimées pour chaque sous-bande. Finalement pour chaque portion de signal glissante, les K_b valeurs propres sont respectivement sommées pour être comparables à celles estimées en temps.

Les signaux sont exprimés dans le domaine fréquentiel par TFCT (*cf.* Annexe A.1.1) avec le paramétrage suivant :

- fenêtrage des signaux par une fenêtre sinus *w*[*l*] de longueur *N* échantillons,
- recouvrement des fenêtres à 50%,
- *N* coefficients spectraux calculés (pas de complément de zéros, *Z*=0).

La TFCT du signal *c*[*n*], $F_c[l,k]$, est exprimée en fonction de l'indice temporel, *l*, des portions glissantes de signal et de l'indice fréquentiel *k* tel que *k* = [1,…,*N*/2+1] en considérant la symétrie hermitienne du spectre. La séparation du spectre en K_b sous-bandes de fréquences (indice des sous-bandes *b* = [1,…K_b]) est réalisée en groupant les coefficients spectraux (*k* = [k_b,…, k_{b+1}-1]) au moyen de l'échelle perceptuelle ERB, définie en Annexe B.1.2, tels que :

$$F_c^b\left[l,k\right] = \left\{F_c\left[l,k_b\right],\ldots,F_c\left[l,k_{b+1}-1\right]\right\} \tag{4.26}$$

Une matrice de covariance réelle est estimée pour chaque sous-bande à partir des spectres et relativement au système d'équations (4.19). Pour cela, la correspondance des puissances des signaux dans les domaines temporel et fréquentiel est admise d'après le théorème de Parseval. En considérant le signal *c*[*n*] fenêtré et centré (*cf.* équation (4.7)); la puissance de $\overline{c}\left[l\right]$ est alors exprimée dans le domaine fréquentiel (pour la portion d'indice *l* et la sous-bande *b*) comme suit :

$$\sigma_C^2\left[l,b\right] = \frac{2}{N^2}\cdot\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{\overline{c}}\left[l,k\right]\right|^2, \tag{4.27}$$

où $F_{\bar{c}}[l,k]$ est la TFCT de $\bar{c}[n]$ c'est à dire un signal de taille N. Les puissances par sous-bandes de fréquences de S_1, S_2, A_1 et A_2 sont donc estimées au regard de l'équation (4.27) en remplaçant le canal C par le signal considéré.

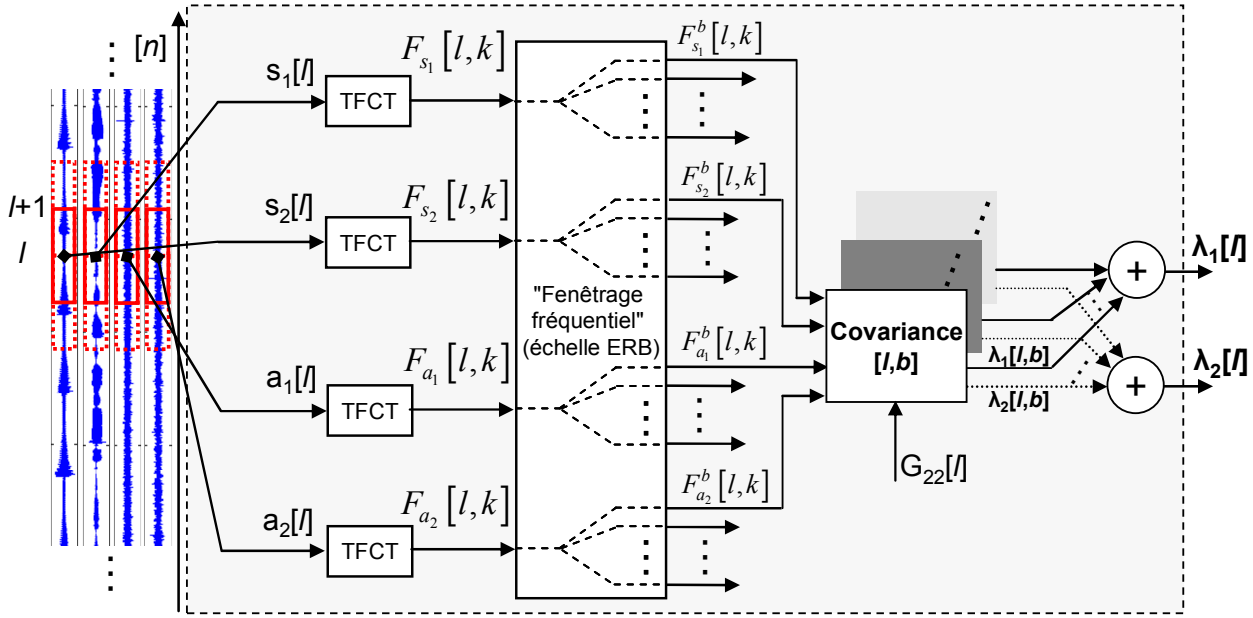

Figure 4.9 Estimation des valeurs propres par sous-bandes de fréquence d'un signal stéréo (*M*=2, *D*=2). La matrice de covariance est estimée à partir des spectres en sous-bandes des signaux analysés pour chaque trame *l* et chaque sous-bande *b*.

L'indice d'inter-corrélation des ambiances est estimé à partir de la partie réelle de l'inter-spectre des d'ambiances centrées tel que :

$$\rho_{A_1A_2}[l,b] = \frac{\Re\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{\bar{a}_1}[l,k]\cdot F_{\bar{a}_2}^{*}[l,k]\right)}{\sqrt{\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{\bar{a}_1}[l,k]\cdot F_{\bar{a}_1}^{*}[l,k]\right)\cdot\left(\sum_{k=k_b}^{k=k_{b+1}-1} F_{\bar{a}_2}[l,k]\cdot F_{\bar{a}_2}^{*}[l,k]\right)}} \quad (4.28)$$

La partie imaginaire de l'inter-spectre est omise pour le calcul de la corrélation des ambiances de manière à calculer une covariance réelle à partir du système d'équations (4.19). Finalement, les valeurs propres en sous-bandes sont estimées à partir de l'équation (4.20). Les valeurs propres estimées directement dans le domaine temporel peuvent alors être comparées à la somme des valeurs propres estimées en sous-bandes définie telle que :

$$\lambda_i[l] = \sum_{b=1}^{K_b} \lambda_i[l,b], \quad i \in [1;2]. \quad (4.29)$$

Expérimentation

De façon à réaliser une comparaison directe avec l'analyse temporelle réalisée au paragraphe 4.2.2.1, le signal de parole, de glockenspiel et l'ambiance stéréo d'aéroport sont analysés par sous-bandes de fréquences.

La **Figure 4.10**[10] présente les valeurs propres estimées en sous-bandes avec K_b=10, 20 et 40. La comparaison des RTF des signaux originaux puis mélangés (*cf.* **Figure 4.6**) avec celles des valeurs propres est menée dans ce paragraphe. Comme le montre la **Figure 4.10**, plus le nombre de sous-bandes est élevé et plus la discrimination est importante. En effet, la séparation des spectres en 10 sous-bandes ne permet pas de dissocier complètement les sources directionnelles des ambiances puisque la plus faible valeur propre contient de l'énergie relative aux sources directionnelles (*cf.* **Figure 4.10-(d)**). Cependant, cette énergie est fortement réduite lorsque K_b augmente comme le montre la **Figure 4.10-(e)** et la **Figure 4.10-(f)** : seules quelques harmoniques du signal de glockenspiel contribuent au niveau d'énergie de la plus faible valeur propre.

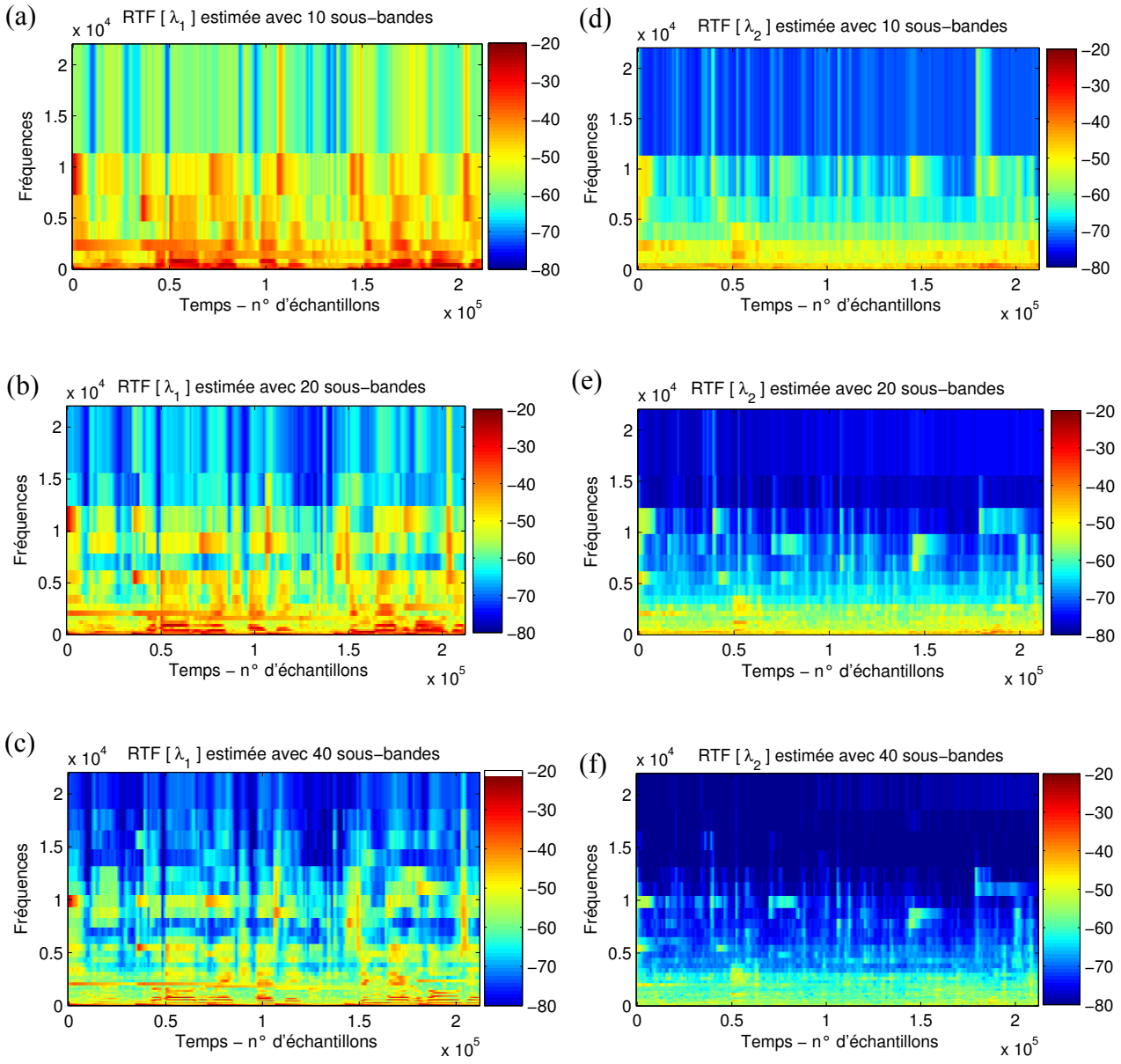

Figure 4.10 RTF (en dB) des valeurs propres [(a)-(b)-(c)- λ_1, (d)-(e)-(f)- λ_2] estimées avec 10, 20 et 40 sous-bandes de fréquences à partir d'un signal stéréo (M=2, D=2) issu du mélange d'un signal de parole, de glockenspiel et d'un signal stéréo d'ambiance d'aéroport.

[10] La représentation des valeurs propres dans le plan temps-fréquence ne correspond pas à celle décrite en Annexe A dans la mesure où aucune TFCT n'est calculée mais simplement un niveau d'énergie suivant une échelle logarithmique en décibels.

La comparaison des valeurs propres estimées en temps (*cf.* **Figure 4.8**), et par sous-bandes de fréquences (K_b = 10, 20 et 40) est présentée sur la **Figure 4.11**. Pour y parvenir, les valeurs propres en sous-bandes (*cf.* **Figure 4.10**) sont sommées comme indiqué par l'équation (4.29).

La plus grande valeur propre estimée en sous-bandes a un niveau d'énergie légèrement supérieur à celui de celle estimée en temps et ceci d'autant plus lorsque les sources directionnelles ont des azimuts différents *i.e.* pour les trames d'indices l<180 et l>280 (*cf.* **Figure 4.11-(a)**). Certaines sources directionnelles considérées comme secondaires par l'analyse temporelle sont considérées comme dominantes pour les sous-bandes où une seule source directionelle apporte de l'énergie. Idéalement, si les sources directionnelles avaient des supports fréquentiels différents, l'analyse en sous-bandes considèrerait ces sources comme dominantes dans leurs sous-bandes respectives. Par conséquent, la plus grande valeur propre estimée en sous-bandes est relative à la somme des puissances de chaque source dominante *i.e.* pour chaque sous-bande, à laquelle s'ajoute une partie de la puissance des signaux d'ambiances coïncidentes.

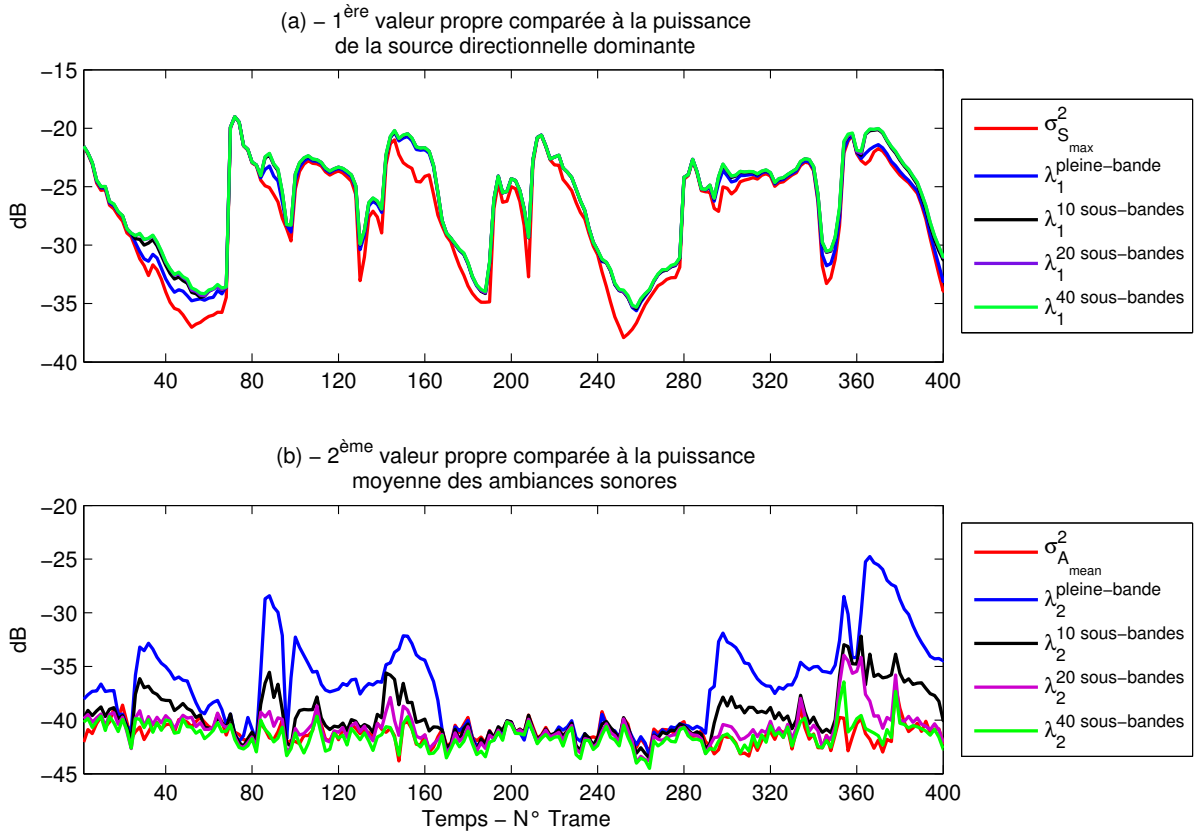

Figure 4.11 Comparaison des valeurs propres [(a) – λ_1, (b) – λ_2] estimées en temps et par sous-bandes de fréquences à partir d'un signal stéréo (M=2, D=2) issu du mélange d'un signal de parole, de glockenspiel et d'un signal stéréo d'ambiance d'aéroport.

D'après la **Figure 4.11-(b)**, plus K_b est grand et plus l'analyse en sous-bandes dissocie les sources directionnelles des ambiances. En effet, la plus petite valeur propre correspond à la puissance moyenne des ambiances plus la puissance des sources secondaires (lorsque les azimuts des sources directionnelles sont différents) qui diminue avec l'analyse en sous-bandes et cela d'autant plus que le nombre de sous-bandes est grand. Finalement, avec une analyse en sous-bandes de fréquences, la plus petite valeur propre est plus proche de la puissance moyenne des ambiances en comparaison à l'analyse temporelle.

4.2.2.3 Conclusion : hiérarchisation des composantes d'un signal stéréophonique

Nous avons présenté les principes de l'analyse en temps et en sous-bandes d'un signal stéréophonique issu du mélange de deux sources directionnelles et d'un signal stéréo d'ambiance. Nous avons expérimenté cette analyse pour établir la distribution des valeurs propres à partir de signaux issus d'enregistrements naturels mélangés par un panoramique d'intensité suivant la loi des sinus.

En comparant les niveaux d'énergie des signaux originaux aux distributions des valeurs propres correspondantes, nous avons, d'une part, mis en évidence l'influence du niveau d'énergie des sources directionnelles introduit par le panoramique d'intensité. En d'autre termes, nous avons observé l'influence de l'azimut des sources sur la hiérarchisation des composantes au sein des valeurs propres. Alors que les puissances des signaux d'ambiance sont distribuées sur chaque valeur propre, la puissance de la source directionnelle dominante est associée à la plus grande valeur propre et il en va de même pour la ou les sources directionnelles secondaires si leur azimut se confond avec l'azimut de la source dominante.

D'autre part, nous avons montré que l'analyse en sous-bandes est plus discriminante qu'une analyse en temps puisqu'elle résulte en la dissociation d'une source dominante du mélange pour chaque sous-bande de fréquences. Autrement dit, plusieurs sources dominantes, au support fréquentiel différent, peuvent être dissociées du mélange à chaque instant. La limite de cette analyse en sous-bandes repose alors sur la nature des signaux mélangés et plus particulièrement sur le recouvrement des supports fréquentiels propres à chaque source directionnelle.

4.3 Analyse en Composante Principale

A partir de la décomposition en valeurs propres de la covariance d'un signal multicanal, nous décrivons maintenant la transformation de Karhunen-Loève (*Karhunen-Loève Transform* - KLT) ou Analyse en Composante Principale (ACP) d'un tel signal. Introduite dans le cas d'une analyse bidimensionnelle par Pearson [PEA00], l'ACP a été étendue au cas multidimensionnel par Hotelling [HOT33]. L'ACP multidimensionnelle permet la caractérisation des données d'un espace à M dimensions vers un sous-espace de dimension P (avec $P<M$) en minimisant les pertes d'informations dues à la projection, c'est-à-dire en maximisant la variance projetée. D'après [HOT33], l'ACP/KLT est une transformation linéaire qui consiste à projeter les données sur la base des vecteurs propres de leur covariance. Les variables corrélées sont remplacées par de nouvelles variables, décorrélées et de variance maximale, établies par combinaisons linéaires des variables initiales.

En considérant le signal multicanal $\mathbf{C_M} = \left(C_1, \cdots, C_m, \cdots, C_M\right)^T$ défini par l'équation (4.4), l'ACP d'un tel signal à M canaux et N_T échantillons discrets (de dimension $M \times N_T$) résulte en un signal multicanal $\mathbf{D_M}$ de même dimension tel que :

$$\mathbf{D_M} = \mathbf{V_M} \times \mathbf{C_M}, \tag{4.30}$$

où $\mathbf{V_M}$ est la matrice orthogonale des vecteurs propres, de dimension $M \times M$, définie par l'équation (4.16).

4.3.1 Propriétés de l'ACP

L'ACP est une transformation linéaire et dépendante de la statistique des signaux puisqu'elle est basée sur une décomposition particulière de la covariance des signaux. De plus, l'ACP

constitue la méthode de décorrélation optimale. En effet, la covariance du signal multicanal transformé $\mathbf{D_M}$ est diagonale :

$$\begin{aligned}
\mathbf{R}_{\mathbf{D_M}} &= \mathrm{E}\left[\mathbf{D_M} \cdot \mathbf{D_M}^T\right] \\
\mathbf{R}_{\mathbf{D_M}} &= \mathrm{E}\left[\left(\mathbf{V_M} \times \mathbf{C_M}\right)\left(\mathbf{V_M} \times \mathbf{C_M}\right)^T\right] \\
\mathbf{R}_{\mathbf{D_M}} &= \mathbf{V_M} \times \mathrm{E}\left[\mathbf{C_M}\mathbf{C_M}^T\right] \times \mathbf{V_M}^T \\
\mathbf{R}_{\mathbf{D_M}} &= \mathbf{V_M} \times \mathbf{R}_{\mathbf{C_M}} \times \mathbf{V_M}^T \\
\mathbf{R}_{\mathbf{D_M}} &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_M \end{pmatrix}
\end{aligned} \tag{4.31}$$

A l'issue de l'ACP, les canaux du signal multicanal $\mathbf{D_M}$ sont donc décorrélés avec une puissance proportionnelle aux valeurs propres telles que : $\lambda_1 > \lambda_2 > \cdots > \lambda_M$. La hiérarchisation des composantes au sein des valeurs propres, décrite au paragraphe 4.2.2, définit, par conséquent, la nature des composantes propres aux canaux transformés du signal $\mathbf{D_M}$.

L'ACP est utilisée pour compacter les données en un minimum de composantes. En projetant les données d'un espace multidimensionel vers un sous-espace (dimension réduite), l'ACP vise à minimiser les pertes d'informations dues à la projection en maximisant la variance projetée. En effet, la combinaison linéaire des données définie par la projection sur le premier vecteur propre correspond à projeter les données sur l'axe au maximum de variance *i.e.* la variance des données est maximale autour de cet axe. De même, la projection des données sur le second vecteur propre définit l'axe au maximum de variance restante. Et ainsi de suite… Finalement, si parmi M valeurs propres, M-P sont nulles ou négligeables, alors l'information contenue dans le signal $\mathbf{C_M}$ à M canaux d'origine peut être représentée par un signal $\mathbf{D_P}$ à P canaux transformés.

La matrice des vecteurs propres utilisée par l'ACP est orthogonale et donc, par nature, inversible par simple transposition. Cette caractéristique de l'ACP permet donc la reconstruction du signal multicanal $\mathbf{C_M}$ par simple opération inverse de l'opération décrite par l'équation (4.30).

4.3.2 Intérêt de l'ACP dans un contexte de codage stéréo et multicanal

L'ACP a été utilisée par un nombre conséquent de chercheurs dans les domaines du codage audio et vidéo. Le codage stéréophonique, décrit au paragraphe 2.2.2, utilise l'ACP pour compacter l'énergie initiale en un signal dominant avec un matriçage adaptatif des canaux. D'après la modélisation des signaux audio que nous avons introduite au paragraphe 4.2.1, nous avons caractérisé le contenu des signaux issus d'une ACP au travers de la distribution des valeurs propres (*cf.* paragraphe 4.2.2). Le matriçage adaptatif basé sur l'ACP permet de regrouper l'information sonore « dominante » : les sources directionnelles et leur ambiance coïncidente. Nous avons également montré l'intérêt d'une analyse en sous-bandes qui permet de mieux discriminer les composantes du signal original pour finalement réduire au maximum le niveau d'énergie des sources directionnelles dans le signal résiduel. Dans un contexte de codage audio, cette hiérarchisation des composantes dans les canaux transformés peut être vue comme un prétraitement qui permet de regrouper l'information prioritaire à coder. La transmission des sources directionnelles dominantes permet la reconstruction d'une

scène sonore où l'information sonore dominante est restituée (*cf.* paragraphe 2.2.2). En outre, si le procédé de codage basé sur l'ACP prévoit la synthèse du signal résiduel (associé aux composantes d'ambiances) alors la reconstruction de la scène originale par ACP inverse sera d'autant plus fidèle. Cette approche, présentée au chapitre 5, permet alors la reconstruction d'une scène sonore avec une qualité variable et fonction de la quantité et de l'ordonnancement des informations extraites et transmises.

Récemment, Yang et *al.* dans [YAN04] ont proposé d'utiliser l'ACP/KLT pour réduire les redondances inter-canal dans un contexte de codage audio multicanal. Les auteurs proposent notamment d'adapter le codeur MPEG-2/4 AAC au codage des signaux issus de l'ACP (voir le paragraphe 2.3.2.2 pour plus de détails). Les auteurs utilisent en particulier le fait que plus le nombre de canaux originaux est grand et plus la hiérarchisation des composantes est importante *i.e.* plus l'énergie est compactée dans les premiers canaux transformés. Basés sur ce principe, nous proposons un codage paramétrique des signaux audio multicanaux. En effet, l'ACP est une transformation linéaire réalisée au moyen d'une matrice orthogonale qui peut être considérée, d'un point de vue mathématique et géométrique, comme une matrice de rotation. L'ACP peut donc être paramétrée par des angles de rotation qui peuvent être, dans certaines conditions, interprétés physiquement comme liés aux positions spatiales ou azimut des sources dominantes. Ainsi, plutôt que de transmettre les éléments de la covariance du signal multicanal à coder (*cf.* paragraphe 2.3.2.2), la transmission d'angles de rotation constitue une approche paramétrique compatible avec un codage audio à bas débit à la manière de la technologie MPEG *surround* (*cf.* paragraphe 2.3.3.2). L'importance perceptive des signaux résiduels issus de l'ACP n'est pas crutiale pour l'intelligibilité (qualité audio basique) mais à l'inverse liée à notre perception spatiale puisqu'ils correspondent principalement aux composantes d'ambiance. Finalement, comme pour le codage stéréophonique basé sur l'ACP, si la méthode de codage prévoit la transmission d'informations utiles à la reconstruction des composantes résiduelles, la reconstruction du signal multicanal sera d'autant plus fidèle et la perception spatiale associée plus proche de l'originale.

4.3.3 L'ACP bidimensionnelle par rotations en sous-bandes

Ce paragraphe s'insère dans un contexte de codage stéréophonique et vise à présenter comment l'ACP bidimensionnelle (M=2) est utilisée dans ce contexte [BRI06a].

L'ACP est utilisée pour transformer les canaux d'un signal stéréophonique $\mathbf{C_2}=(C_1,C_2)^T$ ayant suivi un découpage temporel et fréquentiel propre aux techniques de codage audio perceptuel (*cf.* paragraphe 2.1.2). Le découpage temporel vise à assurer un suivi de l'évolution de la forme d'onde des signaux au cours du temps avec un traitement par fenêtre glissante (ou trame). Le paramétrage de la TFCT ($F_{ci}[l,k]$, $i \in [1,2]$) et la séparation des spectres en sous-bandes au moyen de l'échelle ERB sont identiques à ceux présentés au paragraphe 4.2.2.2. L'ACP est considérée pour chaque portion de signal (indice l) et chaque sous-bande (indice b). La projection du signal stéréo $\mathbf{C_2}$ sur la base des vecteurs propres avec la matrice $\mathbf{V_2}$ définie par l'équation (4.16), est réalisée au moyen d'une matrice de rotation carrée, de dimension M=2 définie par l'équation suivante :

$$\mathbf{V_2} = \mathbf{R_2}\left(\alpha[l,b]\right) = \begin{pmatrix} \cos\left(\alpha[l,b]\right) & \sin\left(\alpha[l,b]\right) \\ -\sin\left(\alpha[l,b]\right) & \cos\left(\alpha[l,b]\right) \end{pmatrix}, \tag{4.32}$$

où $\alpha[l,b]$ est l'angle de rotation estimé à partir de l'expression, tirée de l'équation (4.15), qui caractérise la diagonalisation de la matrice de covariance en sous-bandes telle que :

$$\mathbf{R}_2\left(\alpha[l,b]\right)\cdot\mathbf{\Gamma}^b_{\mathbf{C}_2}\cdot\mathbf{R}_2\left(\alpha[l,b]\right)^T=\begin{pmatrix}\lambda_1[l,b] & 0\\ 0 & \lambda_2[l,b]\end{pmatrix}, \tag{4.33}$$

où : $$\mathbf{\Gamma}^b_{\mathbf{C}_2}=\left(r_{ij}[l,b]\right)_{1\le i,j\le 2}\text{, avec :} \tag{4.34}$$

$$\begin{cases} r_{11}[l,b]=\dfrac{2}{N^2}\cdot\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{\overline{c}_1}[l,k]\right|^2 \\ r_{12}[l,b]=r_{21}[l,b]=\dfrac{2}{N^2}\times\Re\left(\sum_{k=k_b}^{k=k_{b+1}-1}F_{\overline{c}_1}[l,k]\cdot F^*_{\overline{c}_2}[l,k]\right), \\ r_{22}[l,b]=\dfrac{2}{N^2}\cdot\sum_{k=k_b}^{k=k_{b+1}-1}\left|F_{\overline{c}_2}[l,k]\right|^2 \end{cases} \tag{4.35}$$

sans aucune connaissance *a priori* sur les sources directionnelles et les ambiances. Finalement, à partir des équations (4.33) et (4.34), $\alpha[l,b]$ est donné par l'expression suivante :

$$\alpha[l,b]=\arctan\left(\frac{\lambda_1[l,b]-r_{11}[l,b]}{r_{12}[l,b]}\right),\quad \alpha\in\left[-\frac{\pi}{2};\frac{\pi}{2}\right]. \tag{4.36}$$

En tenant compte de l'expression de la plus grande valeur propre λ_1 définie à l'équation (4.20) et de l'équation (4.36), l'estimation de l'angle de rotation devient possible seulement à partir des éléments de la matrice de covariance tel que :

$$\alpha[l,b]=\frac{1}{2}\arctan\left(\frac{2\times r_{12}[l,b]}{r_{11}[l,b]-r_{22}[l,b]}\right),\quad \alpha\in\left[-\frac{\pi}{4};\frac{\pi}{4}\right]. \tag{4.37}$$

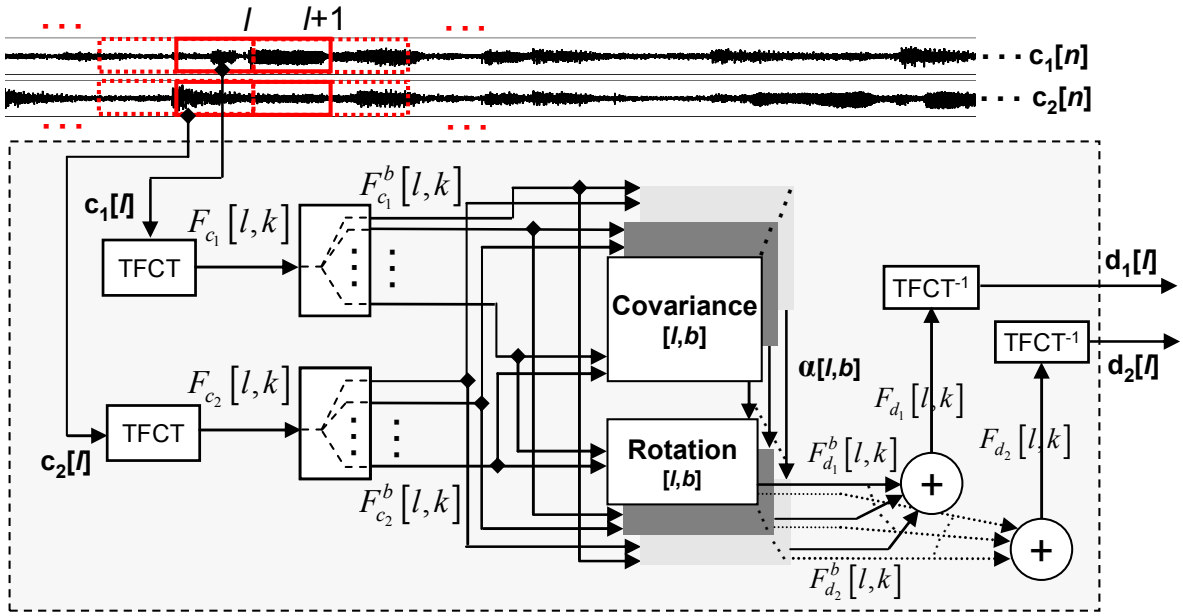

Figure 4.12 ACP bidimensionnelle en sous-bandes d'un signal stéréophonique. La rotation des canaux $c_1[n]$ et $c_2[n]$ est réalisée pour chaque portion glissante *l* et chaque sous-bande *b*. L'angle de rotation est estimé à partir de la covariance déduite des spectres en sous-bandes de chaque canal.

L'ACP en sous-bandes repose donc sur l'estimation de la matrice de covariance à partir des spectres en sous-bandes des signaux centrés. Comme l'indique le système d'équations (4.35) les termes d'auto-covariance (r_{11} et r_{22}) correspondent à la moyenne des densités spectrales de puissance estimées pour chaque sous-bande et le terme d'inter-covariance (r_{12}) est estimé à partir de l'inter-spectre en sous-bandes.

L'angle de rotation extrait pour chaque sous-bande est ensuite utilisé pour projeter les canaux d'origine sur la base des vecteurs propres, projection qui résulte en deux composantes spectrales : la composante $F^{b}_{d1}[l,k]$ relative à $\lambda_1[l,b]$ et la composante $F^{b}_{d2}[l,k]$ relative à $\lambda_2[l,b]$, telles que :

$$\begin{pmatrix} F^b_{d_1}[l,k] \\ F^b_{d_2}[l,k] \end{pmatrix} = \begin{pmatrix} \cos(\alpha[l,b]) & \sin(\alpha[l,b]) \\ -\sin(\alpha[l,b]) & \cos(\alpha[l,b]) \end{pmatrix} \cdot \begin{pmatrix} F^b_{c_1}[l,k] \\ F^b_{c_2}[l,k] \end{pmatrix}. \tag{4.38}$$

Les portions glissantes $d_1[l]$ et $d_2[l]$ sont exprimées dans le domaine temporel en appliquant la TFCT inverse à la combinaison des composantes en sous-bandes, comme présenté sur la **Figure 4.12**. Enfin, les signaux temporels $d_1[n]$ et $d_2[n]$ résultent de la somme au cours du temps via la méthode *overlap-add* (OLA) de toutes les portions glissantes transformées.

4.3.3.1 Pertes d'informations avec la synthèse OLA

Le signe de l'angle de rotation estimé à l'équation (4.37) peut être négatif. Par suite, le signe de la matrice des vecteurs propres ici considérée comme la matrice de rotation $\mathbf{R_2}(\alpha[l,b])$ peut être différent d'une portion glissante l à une autre (l+1). La synthèse des signaux $d_1[n]$ et $d_2[n]$ par la méthode OLA peut alors générer des pertes d'informations.

Par conséquent, il convient d'assurer une continuité du signe de l'angle de rotation $\alpha[l,b]$ sans pour autant modifier son sens physique. En effet, l'angle de rotation utile à l'ACP bidimensionnelle correspond à l'azimut de la source dominante dans la sous-bande considérée (voir le paragraphe 2.2.2 et l'Annexe C.2.1.1).

Conversion de l'angle estimé en azimut

Pour obtenir une méthode d'analyse/synthèse robuste, l'angle de rotation α doit appartenir au premier quadrant du repère des données (*cf.* **Figure 2.4**) soit à l'intervalle $[0;\pi/2]$ et par suite, procurer un signe constant (au cours du temps) aux éléments des matrices de rotation. L'angle de rotation estimé dans cet intervalle peut alors être mis en correspondance, après conversion en degrés, avec l'azimut de la source dominante θ entre les haut-parleurs d'un système stéréophonique (*cf.* **Figure 4.3**), c'est-à-dire dans l'intervalle $[-\theta_0;\theta_0]$, au moyen de l'expression suivante :

$$\theta = \left(\alpha \times \frac{180}{\pi}\right) \times \left(\frac{2 \times \theta_0}{90}\right) - \theta_0 \ \text{degrés}, \tag{4.39}$$

soit :

$$\theta = \theta_0 \left(\frac{4\alpha}{\pi} - 1\right) \ \text{degrés}. \tag{4.40}$$

La **Figure 4.13** illustre la conversion de l'angle estimé α en azimut θ pour trois valeurs d'écartement des haut-parleurs d'un système de reproduction stéréo. Ainsi, un signal stéréo constitué d'une source dominante repérée à gauche (respectivement à droite) de l'image stéréo c'est-à-dire à l'azimut θ=-θ_0 (respectivement θ=θ_0) doit impliquer une estimation de

l'angle de rotation α=0 (respectivement α= π/2) et ainsi générer un signal dominant D_1 issu du canal gauche C_1 (respectivement droit C_2) tel que : $F^b_{d1}[l,k] = F^b_{c1}[l,k]$ d'après l'équation (4.38) (respectivement $F^b_{d1}[l,k] = F^b_{c2}[l,k]$).

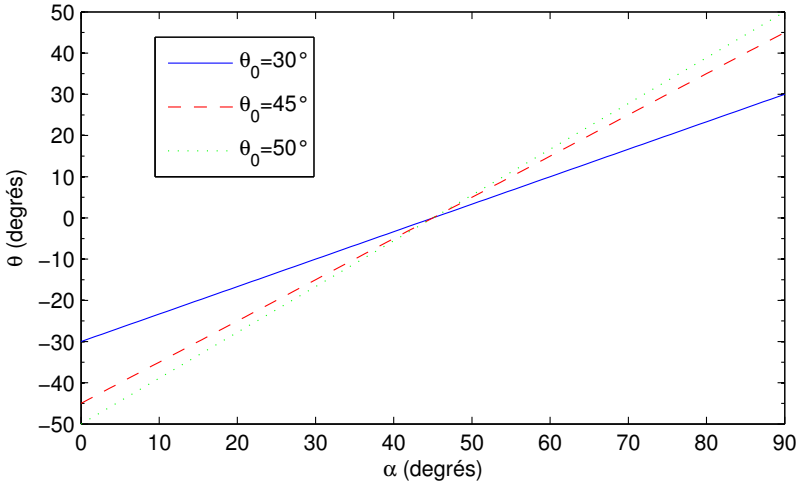

Figure 4.13 Conversion de l'angle α (en degrés), estimé pour la rotation des données stéréo (ACP), en azimut θ en fonction de l'écartement des haut-parleurs θ_0.

Correction de l'estimation sans modifier l'azimut

Plusieurs modifications doivent être apportées pour estimer l'angle de rotation dans le premier quadrant et ainsi éviter la perte d'informations durant la synthèse OLA [BRI06a]. D'après l'expression de l'angle de rotation donnée par l'équation (4.37), lorsque r_{12}=0 c'est-à-dire lorsque les canaux du signal stéréo sont décorrélés aucune direction prédominante ne doit être estimée et un simple matriçage de type somme-différence est préconisé en fixant α=π/4. Cependant, ce cas est valide sous l'hypothèse que les deux canaux soient non nuls ($r_{11} \neq 0$ et $r_{22} \neq 0$) car dans le cas contraire, si un canal est nul (r_{11}=0 ou r_{22}=0), l'angle peut prendre la valeur 0 ou π/2. En outre, le signe de l'inter-covariance r_{12} au numérateur influence le signe de l'angle de rotation estimé. En effet, une inter-covariance négative empêche d'obtenir des valeurs positives de l'angle de rotation même en utilisant un modulo à π/2 près (*cf.* paragraphe 2.2.3.2). Plutôt que de limiter les valeurs minimales de l'inter-corrélation à zéro (comme dans [IRW02], *cf.* Annexe C.2.1.1), utiliser le module de l'inter-corrélation a l'avantage de conserver le sens physique de l'angle de rotation *i.e.* l'azimut des sources dominantes. Une estimation robuste de l'angle de rotation $\alpha \in [0;\pi/2]$ est donnée par l'équation suivante (où les indices l et b des portions de signal et de sous-bande ont été omis pour raison de clarté) :

$$\alpha = \begin{cases} \dfrac{\pi}{4}, & \text{si } r_{12} = 0 \text{ et } \left((r_{11} \neq 0 \text{ et } r_{22} \neq 0) \text{ ou } (r_{11} = r_{22})\right) \\ \dfrac{1}{2}\arctan\left(\dfrac{2\times|r_{12}|}{r_{11}-r_{22}}\right), & \text{si } r_{11} - r_{22} > 0 \\ \dfrac{1}{2}\arctan\left(\dfrac{2\times|r_{12}|}{r_{11}-r_{22}}\right) + \dfrac{\pi}{2}, & \text{sinon} \end{cases} . \quad (4.41)$$

A partir du signal stéréo décrit au paragraphe 4.2.2.1 (panoramique d'intensité appliqué aux sources directionnelles selon la loi des sinus), l'angle de rotation a été estimé au cours du temps (analyse par fenêtre glissante de taille N=1024 échantillons) à partir de l'équation (4.37) en considérant, en outre, un modulo π/2 des valeurs négatives estimées. Deux autres

versions de cet angle estimé au cours du temps sont obtenues en considérant l'inter-corrélation minimale limitée à la valeur zéro ou en utilisant la valeur absolue de l'inter-corrélation comme indiqué par l'équation (4.41). Ces trois angles estimés au cours du temps (pleine bande), et convertis en azimut à partir de l'équation (4.40) avec θ_0=30°, sont présentés à la **Figure 4.14**.

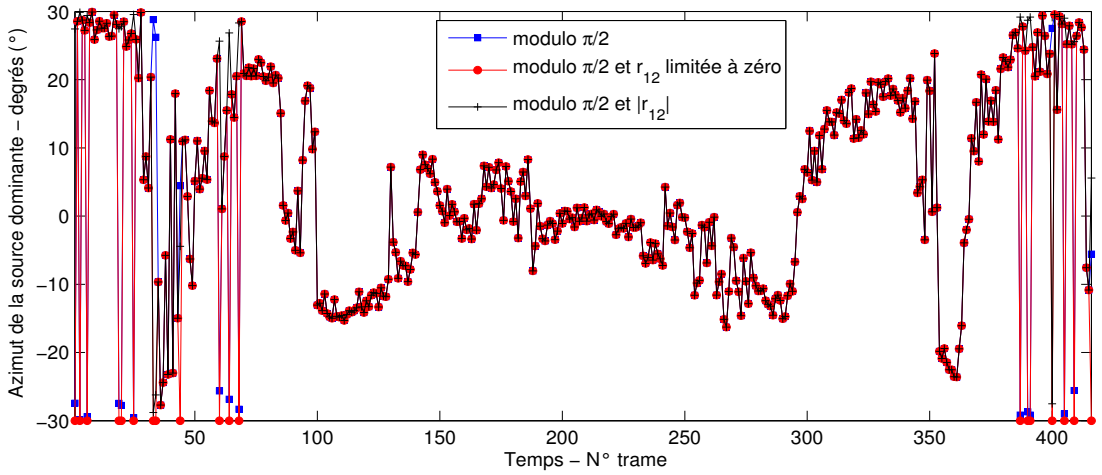

Figure 4.14 Angles de rotation estimés (en temps) et convertis en azimuts (degrés) des sources directionnelles dominantes contenues dans le signal stéréophonique décrit au paragraphe 4.2.2.1.

Au regard des puissances des sources directionnelles présentées à la **Figure 4.8-(a)**, pour les trames d'indices l<35, la source S_2 (à droite de l'image stéréo) a une puissance supérieure à celle de la source S_1 (à gauche de l'image stéréo). Par conséquent la valeur de l'angle de rotation estimé puis converti en azimut doit être proche de 30°. Or, ce n'est pas toujours le cas en considérant une inter-corrélation limitée à zéro puisque certaines valeurs de l'angle de rotation estimé sont nulles (azimut θ=-30°). Un simple modulo π/2 provoque l'estimation d'angles de rotation à valeurs incorrectes qui ne correspondent pas à l'azimut de la source dominante. Ces observations peuvent être également menées pour les trames d'indices l>365 puisque pour ces trames analysées, les sources directionnelles occupent des positions opposées, perçues aux extrémités de l'image stéréo, qui impliquent une très faible corrélation (ponctuellement négative) des canaux analysés.

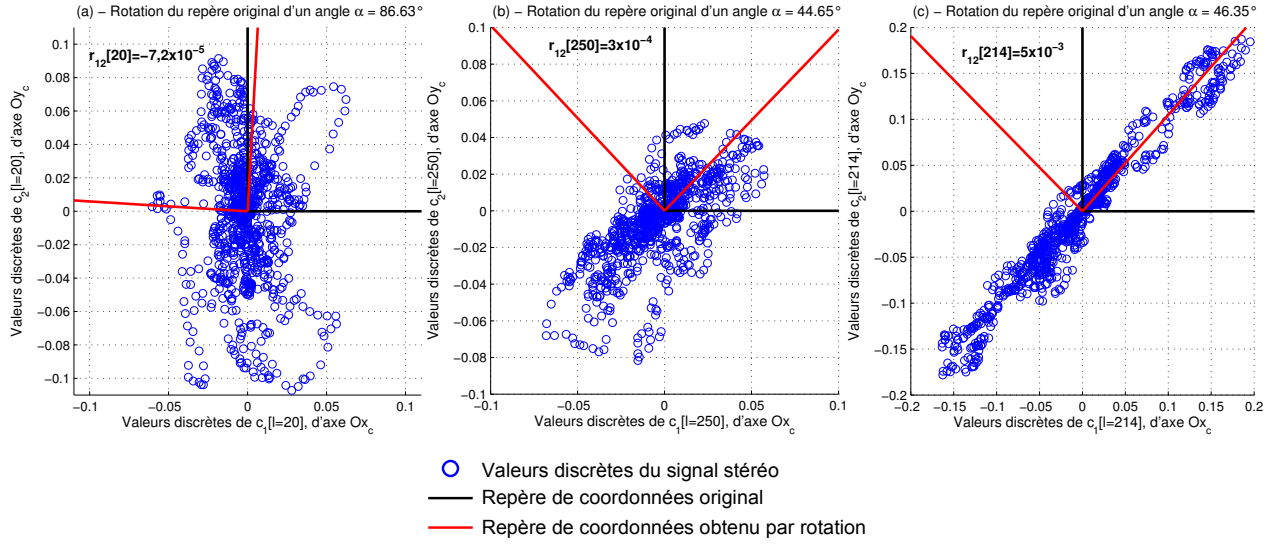

Figure 4.15 Rotation du repère original de coordonnées (x_c,y_c) des canaux C_1 et C_2 du signal stéréo décrit au paragraphe 4.2.2.1 pour les portions d'indice **(a) -** l=20, **(b) -** l=250, **(c) -** l=214.

A titre d'illustration, la **Figure 4.15** présente les valeurs discrètes des portions l=20, 250 et 214 tirées de l'analyse du signal stéréophonique décrit au paragraphe 4.2.2.1. Il apparaît sur la **Figure 4.15-(a)** que le nuage de points, représenté dans le repère original des données (Ox_cy_c), est très diffus dû à une faible inter-corrélation des canaux C_1 et C_2 puisque $r_{12}[20] = -7,2\times10^{-5}$. Dans ce cas, utiliser la valeur absolue de l'inter-corrélation est l'unique moyen d'estimer correctement l'azimut de la source dominante c'est-à-dire le signal de glockenspiel S_2 localisé à droite de l'image stéréophonique. En effet, l'angle de rotation estimé à la trame l=20, à partir de l'équation (4.41) suivi d'une conversion en degrés, vaut 86,63° soit un azimut de la source dominante déduit à partir de l'équation (4.40) à θ=28,87°. Il apparaît à l'inverse sur la **Figure 4.15-(b)** et la **Figure 4.15-(c)** que les nuages de points des données discrètes analysées aux trames l=250 et l=214 sont beaucoup moins dispersés (un axe principal se dessine) et ceci d'autant plus que l'inter-corrélation augmente puisque $r_{12}[250] = 3\times10^{-4}$ et $r_{12}[214] = 5\times10^{-3}$.

En conclusion, seul l'angle de rotation estimé à partir de la valeur absolue de l'inter-corrélation (avec un modulo $\pi/2$) conserve l'azimut de la source directionnelle dominante. L'exemple qui illustre la correction de l'estimation de l'angle de rotation utilise une analyse en temps, mais comme l'indique l'équation (4.41), cette correction est également valable pour un analyse en sous-bandes de fréquences. Enfin, cette estimation de l'angle de rotation est valable sous l'hypothèse que les sources directionnelles soient en phase d'un canal à un autre (*cf.* modèle décrit au paragraphe 4.2.1). En effet, un signal stéréo constitué de sources directionnelles en opposition de phase est caractérisable par un nuage de points orienté dans le quadrant $[-\pi/2;0]$ ou $[\pi/2;\pi]$ (*cf.* **Figure 4.15**) et ne sera convenablement traité que si les signaux sont synchronisés en temps à la manière du procédé décrit au paragraphe 2.2.3.1.

4.3.3.2 Performances de l'ACP bidimensionnelle en sous-bandes

L'ACP est réalisée par rotation paramétrée par un angle estimé directement à partir de la covariance et donc sans connaissance sur les valeurs propres. L'objet de ce paragraphe consiste à évaluer les performances de l'ACP bidimensionnelle directement à partir de l'énergie des signaux transformés. Evaluer la concentration de l'énergie dans les signaux issus de l'ACP constitue une étude complémentaire à celle réalisée au paragraphe 4.2.2 qui discute de la hiérarchisation des composantes d'un signal stéréo au travers de la distribution des valeurs propres.

L'ACP bidimensionnelle d'un signal stéréo (C_1, C_2), réalisée en temps (*cf.* équation (2.2)) ou dans le domaine fréquentiel des sous-bandes (*cf.* équation (4.38)), délivre les signaux (D_1, D_2) dont les niveaux d'énergie diffèrent comme l'illustre la **Figure 4.16** qui présente les RTF (*cf.* Annexe A.1) des signaux issus d'une ACP en temps ou en sous-bandes avec K_b=10 et K_b=40. Ces RTF sont toutes obtenues avec une fenêtre de *Hanning* à N=512 échantillons, 1024 coefficients spectraux sont calculés (Z=512) et un recouvrement de 50% entre les fenêtres glissantes est utilisé. D'après ces RTF et comme l'ont indiqué précédemment les distributions des valeurs propres (*cf.* paragraphe 4.2.2), l'énergie des sources directionnelles dans le signal résiduel D_2 est fortement réduite avec une ACP en sous-bande et cela d'autant plus que K_b est grand. En outre, plus l'énergie des sources directionnelles dans le signal D_2 est faible et plus ce signal résiduel est proche d'une ambiance moyenne des ambiances originales contenues dans le signal stéréo dont les RTF sont présentées à la **Figure 4.6**. Finalement, nous considérons [BRI06a] :

- le signal D_1 comme une « composante principale » constituée des sources directionnelles et de l'ambiance qui coïncide avec ces sources, et
- le signal D_2 comme une « composante ambiance » constituée de l'ambiance résiduelle et des sources sonores secondaires.

Nous proposons d'évaluer les performances de l'ACP ou son aptitude à regrouper les sources directionnelles d'un mélange au sein de la composante principale. Pour cela, nous définissons une mesure de la concentration de l'énergie dans la composante principale avec le rapport

énergétique de la composante principale D_1 à la composante ambiance D_2 ou RCPA_{12} (en dB) donné par :

$$\mathrm{RCPA}_{12}[l] = 10 \times \log_{10}\left(\frac{\sum_{n=1}^{N} d_1^2[l]}{\sum_{n=1}^{N} d_2^2[l]}\right) \text{ dB}. \tag{4.42}$$

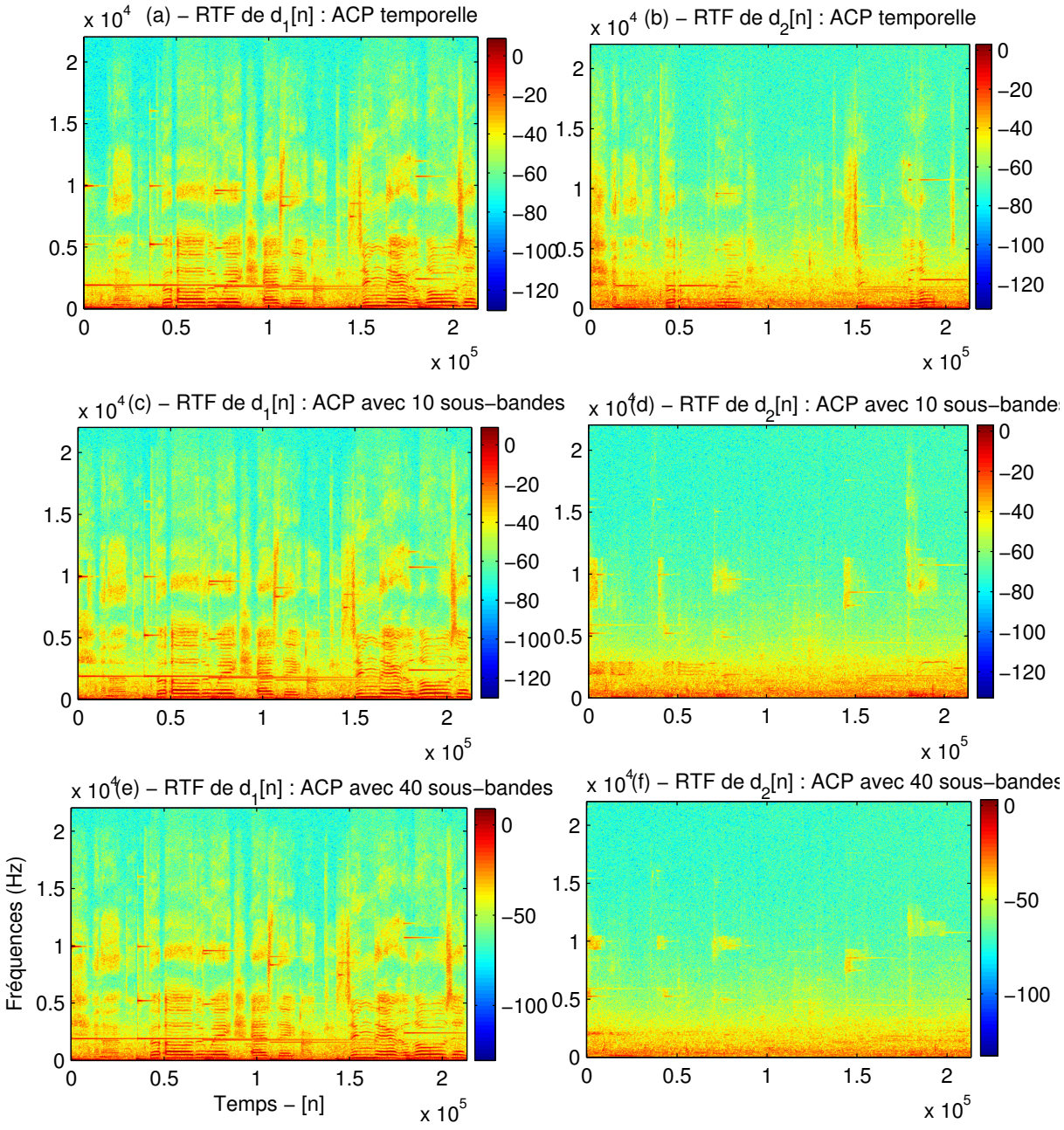

Figure 4.16 RTF (en dB) de [(a)–(c)–(e)- $d_1[n]$], [(b)–(d)–(f)- $d_2[n]$] obtenues avec [(a)-(b)- une ACP en temps], [(c)-(d)- une ACP avec 10 sous-bandes] , [(e)-(f)- une ACP avec 40 sous-bandes] du signal stéréo généré au paragraphe 4.2.2.1 dont les RTF sont présentées à la Figure 4.6.

Par souci d'homogénéité, nous avons choisi d'utiliser la même longueur de fenêtre N pour réaliser l'ACP bidimensionnelle et pour estimer le RCPA au cours du temps. Le RCPA_{12} moyen au cours du temps est alors donné par :

$$\overline{\mathrm{RCPA}_{12}} = \frac{1}{\lceil N_\mathrm{T}/N \rceil} \sum_{l=1}^{\lceil N_\mathrm{T}/N \rceil} \mathrm{RCPA}_{12}[l] \ \mathrm{dB} \tag{4.43}$$

où $\lceil N_\mathrm{T}/N \rceil$ correspond au nombre de portions glissantes résultantes de l'ACP bidimensionnelle puisque $\lceil . \rceil$ désigne la partie entière supérieure (la dernière portion de signal est complétée par des zéros).

La **Figure 4.17** présente une comparaison des RCPA moyens estimés à partir de signaux issus des ACP bidimensionnelles réalisées dans le domaine temporel et des sous-bandes (échelle ERB avec K_b=10, 20 et 40) avec des longueurs de fenêtre d'analyse N=1024, 2048 et 4096 échantillons.

D'après la **Figure 4.17**, la concentration de l'énergie dans la composante principale est plus grande avec une ACP en sous-bandes comparée à une ACP réalisée dans le domaine temporel. En moyenne sur 20 signaux analysés, la différence entre le RCPA des signaux issus d'une ACP en sous-bandes et le RCPA des signaux issus d'une ACP temporelle est de +2 dB. Cette différence s'accroît naturellement avec :

- la diminution de la longueur de la fenêtre d'analyse dans le cas d'une ACP temporelle,
- la diminution de la longueur de la fenêtre d'analyse et l'augmentation du nombre de sous-bandes dans le cas d'une ACP en sous-bandes.

Par exemple, l'ACP en sous-bandes du signal stéréo MPEG « Drama », qui contient plusieurs sources directionnelles nettement dissociables de l'ambiance sonore à l'écoute, résulte en un RCPA maximal supérieur de +5 dB au RCPA minimal obtenu avec une ACP temporelle.

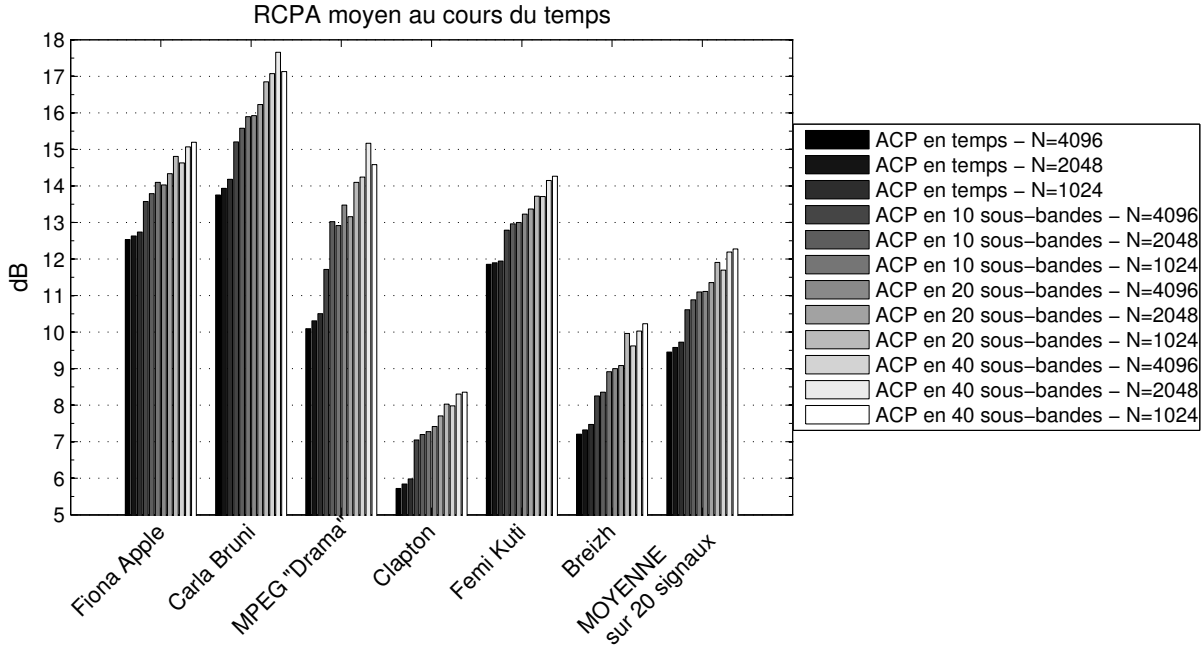

Figure 4.17 RCPA moyens de 6 signaux stéréophoniques obtenus avec une ACP en temps et en sous-bandes (échelle ERB). La moyenne des RCPA est calculée à partir de 20 signaux stéréo.

Le RCPA moyen estimé à partir de la base de signaux stéréo traitée avec une ACP en sous-bandes (de l'ordre de 11 dB) dépend des résolutions temporelle et fréquentielle ainsi

que de la nature des signaux originaux. La **Figure 4.18** présente les coefficients de corrélation (en moyenne au cours du temps), estimés en temps à partir de l'équation (4.11) et en sous-bandes à partir de l'équation (2.7), des canaux de six signaux stéréo dont le RCPA moyen est présenté à la **Figure 4.17**.

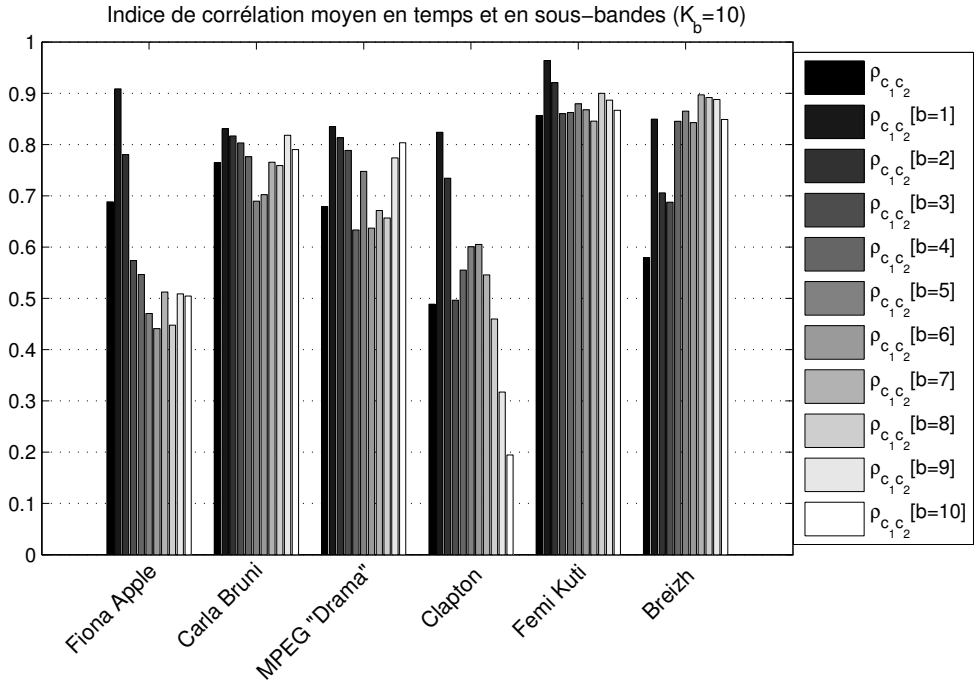

Figure 4.18 Indice de corrélation (moyen au cours du temps) estimé en temps ρ_{c1c2} et en K_b=10 sous-bandes $\rho_{c1c2}[b]$ à partir des canaux de six signaux stéréo.

En règle générale, en comparant la **Figure 4.18** à la **Figure 4.17**, plus l'indice de corrélation moyen estimé en temps ρ_{c1c2} est élevé (hypothèse de corrélation des canaux du modèle présenté au paragraphe 4.2.1) et plus l'ACP est performante en termes de concentration d'énergie. En effet, l'ACP d'un signal stéréo aux canaux non corrélés ne peut concentrer davantage l'énergie des canaux. Dans le cas extrême (r_{12}=0), comme l'indique l'équation (4.41), nous utilisons un simple matriçage de type somme-différence. En pratique, les signaux stéréo dont l'indice de corrélation ρ_{c1c2} est faible comme c'est le cas pour le signal stéréo « Clapton » (*cf.* **Figure 4.18** : $\rho_{c1c2} < 0{,}5$) peuvent bénéficier d'une concentration d'énergie supérieure avec une ACP en sous-bandes puisque la corrélation des bandes $\rho_{c1c2}[b<7]$ dépasse la valeur de ρ_{c1c2}.

Au final, l'ACP bidimensionnelle présentée dans cette section constitue une approche paramétrique pour générer un signal dominant ou composante principale accompagnée d'angles de rotation qui caractérisent les positions spatiales des sources dominantes à chaque instant. En outre, l'ACP en sous-bandes exploite la corrélation des canaux en sous-bandes de fréquences pour compacter davantage l'énergie initiale dans le signal dominant et cela d'autant plus que le nombre de sous-bandes est élevé.

4.3.4 L'ACP tridimensionnelle par rotation d'Euler

Le propos de ce paragraphe se situe dans un contexte de codage audio multicanal avec un nombre de canaux M>2. Une dimension supplémentaire par rapport à l'ACP bidimensionnelle, présentée au paragraphe 4.3.3, est prise en compte pour mettre en œuvre l'ACP tridimensionnelle.

L'ACP d'un signal à trois canaux est donc menée à partir de la diagonalisation de la matrice de covariance $\mathbf{\Gamma}_{\mathbf{C_3}}$ exprimée par :

$$\mathbf{V_3} \times \mathbf{\Gamma}_{\mathbf{C_3}} \times \mathbf{V_3}^T = \mathbf{\Lambda_3} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}. \tag{4.44}$$

L'ACP tridimensionnelle est mise en place de façon à compacter l'énergie des canaux C_1, C_2 et C_3 transformés en un canal dominant D_1, de puissance λ_1 relative aux sources directionnelles dominantes et à l'ambiance coïncidente, et deux canaux résiduels D_2 et D_3, de puissances respectives λ_2 et λ_3 relatives aux sources secondaires et aux signaux d'ambiance, telle que :

$$\begin{aligned} \mathbf{D_3} &= \mathbf{V_3} \cdot \mathbf{C_3} \\ \begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} &= \mathbf{V_3} \cdot \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}. \end{aligned} \tag{4.45}$$

Pour conserver une approche paramétrique pour l'ACP, les sous-espaces propres définis par la matrice des vecteurs propres $\mathbf{V_3}$ peuvent être considérés comme une matrice de rotation orthogonale de dimension trois. Les angles d'Euler α, β, et γ définissent la matrice de rotation tridimensionnelle $\mathbf{R_3}(\alpha,\beta,\gamma)$ comme le produit de trois matrices de rotation bidimensionnelles [DUH01]. Plusieurs conventions peuvent être adoptées selon le choix des axes utilisés (ZYZ, ZXZ, ZYX, etc.) pour réaliser les trois rotations paramétrées par les angles d'Euler. Nous avons adopté la convention ZYX de façon à respecter une hiérarchie énergétique décroissante des signaux issus de l'ACP (D_1 plus énergétique que D_2 qui lui-même sera plus énergétique que D_3). Cette convention définit $\mathbf{V_3}$ comme égale à la matrice $\mathbf{R_3}(\alpha,\beta,\gamma)$ donnée par :

$$\begin{aligned} \mathbf{V_3} &= \mathbf{R_3}(\alpha,\beta,\gamma) = \mathbf{R}_2^x(\gamma) \times \mathbf{R}_2^y(\beta) \times \mathbf{R}_2^z(\alpha) \\ \mathbf{R_3}(\alpha,\beta,\gamma) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & \sin\gamma \\ 0 & -\sin\gamma & \cos\gamma \end{pmatrix} \times \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix} \times \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned} \tag{4.46}$$

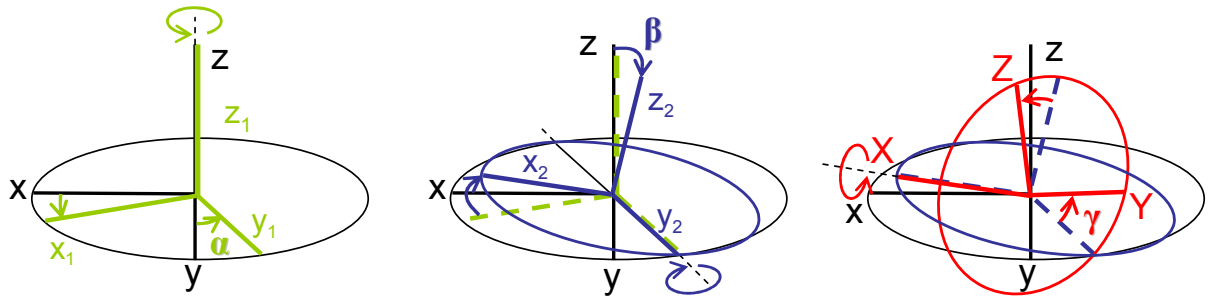

Figure 4.19 Rotations, dans l'espace à 3 dimensions, paramétrées par les angles d'Euler. Le repère original (x,y,z) est transformé en un nouveau repère (X,Y,Z) par rotation d'angle α autour de l'axe z puis par rotation d'angle β autour de l'axe $y_1=y_2$ et enfin par rotation d'angle γ autour de l'axe $x_2=X$.

La matrice de rotation $\mathbf{R_3}(\alpha,\beta,\gamma)$ permet d'obtenir un nouveau repère tridimensionnel au moyen de trois rotations bidimensionnelles réalisées autour des axes du repère original et des repères intermédiaires. La procédure pas à pas est présentée à la **Figure 4.19**. La première rotation d'angle α, $\mathbf{R}^z{}_2(\alpha)$, est réalisée autour de l'axe z du repère original (x,y,z). La seconde rotation d'angle β, $\mathbf{R}^y{}_2(\beta)$, est réalisée autour de l'axe $y_1=y_2$, résultant de la première rotation, dans le repère (x_1,y_1,z_1). Enfin la dernière rotation d'angle γ, $\mathbf{R}^x{}_2(\gamma)$ est réalisée autour de l'axe x_2=X dans le repère (x_2,y_2,z_2) et délivre le nouveau repère (X,Y,Z).

En considérant les données des canaux C_1, C_2 et C_3 comme appartenant au repère original $(x,y,z)=(x_{C1},y_{C2},z_{C3})$, la rotation tridimensionnelle définie par l'équation (4.46) permet de faire coïncider l'axe x_2=X de la **Figure 4.19** avec l'axe principal des données initiales *i.e.* l'axe de plus grande variance. La convention ainsi choisie (rotations selon l'ordonnancement ZYX) concorde avec la définition de l'ACP tridimensionnelle à l'équation (4.45) en considérant la matrice des vecteurs propres $\mathbf{V_3} = \mathbf{R_3}(\alpha,\beta,\gamma)$. Ainsi la composante principale D_1 est relative aux sources directionnelles dominantes puisqu'elle correspond à la projection des canaux d'entrée (C_1,C_2,C_3) sur l'axe des abscisses du repère final (X). Les composantes résiduelles D_2 et D_3, relatives aux sources sonores secondaires et aux ambiances sonores, sont alors obtenues par projection des canaux sur les axes Y et Z du repère final $(X,Y,Z)=(X_{D1},Y_{D2},Z_{D3})$.

L'ACP tridimensionnelle repose alors sur l'estimation des angles d'Euler à partir de l'estimation des valeurs propres et de la matrice de covariance des canaux d'entrée. En effet, l'équation (4.44) devient :

$$\mathbf{R_3}(\alpha,\beta,\gamma)\times\mathbf{\Gamma}_{\mathbf{C_3}} = \mathbf{\Lambda_3}\times\mathbf{R_3}(\alpha,\beta,\gamma), \tag{4.47}$$

qui est une expression équivalente à celle utilisée dans [DUH01] :

$$\mathbf{R_3^T}(\alpha,\beta,\gamma)\times\mathbf{\Lambda_3} = \mathbf{\Gamma}_{\mathbf{C_3}}\times\mathbf{R_3^T}(\alpha,\beta,\gamma), \tag{4.48}$$

où l'auteur considère une convention différente (ordonnancement ZYZ). Le système d'équations (4.48) est alors résolu à partir du système d'équations suivant :

$$\begin{cases} \left[\mathbf{R_3^T}(\alpha,\beta,\gamma)\times\mathbf{\Lambda_3}\right](1,1)-\left[\mathbf{\Gamma}_{\mathbf{C_3}}\times\mathbf{R_3^T}(\alpha,\beta,\gamma)\right](1,1)=0 \\ \left[\mathbf{R_3^T}(\alpha,\beta,\gamma)\times\mathbf{\Lambda_3}\right](2,1)-\left[\mathbf{\Gamma}_{\mathbf{C_3}}\times\mathbf{R_3^T}(\alpha,\beta,\gamma)\right](2,1)=0 \\ \left[\mathbf{R_3^T}(\alpha,\beta,\gamma)\times\mathbf{\Lambda_3}\right](1,2)-\left[\mathbf{\Gamma}_{\mathbf{C_3}}\times\mathbf{R_3^T}(\alpha,\beta,\gamma)\right](1,2)=0 \end{cases} \tag{4.49}$$

qui permettent l'estimation des angles d'Euler (les indices l et b des portions de signal et des sous-bandes ont été omis pour raison de clarté) tels que :

$$\begin{cases} \tan\alpha = \dfrac{r_{12}r_{13}+r_{23}\left(\lambda_1-r_{11}\right)}{r_{12}r_{23}+r_{13}\left(\lambda_1-r_{22}\right)} \\ \tan\beta = \dfrac{-r_{12}\sin\alpha+\left(\lambda_1-r_{11}\right)\cos\alpha}{r_{13}} \\ \tan\gamma = \dfrac{r_{12}\cos\alpha+\left(\lambda_2-r_{11}\right)\sin\alpha}{-\sin\beta\left[r_{13}+\left(-r_{12}\sin\alpha+\left(\lambda_2-r_{11}\right)\cos\alpha\right)\right]} \end{cases} . \tag{4.50}$$

Pour l'analyse bidimensionnelle, nous avons utilisé l'expression analytique des valeurs propres (équation (4.20)) pour simplifier l'équation (4.36) de l'angle de rotation, finalement uniquement dépendant des éléments de la covariance. Dans le cas de l'analyse tridimensionnelle, l'expression analytique des valeurs propres d'une matrice de dimension

trois [DUH01], même symétrique dans le cas de la covariance, ne permet pas de simplifier le système d'équation (4.50). Finalement, notre estimation des angles d'Euler repose à la fois sur l'estimation de la covariance des canaux $\mathbf{\Gamma_{C3}}$ et des valeurs propres de la matrice $\mathbf{\Lambda_3}$ à partir des expressions analytiques données dans [DUH01]. Les angles d'Euler sont estimés à partir de la covariance des signaux exprimés dans le domaine temporel ou en sous-bandes de fréquences comme explicité au paragraphe 4.3.3.

Le système d'équation (4.50) délivre les angles d'Euler dans l'intervalle $[-\pi/2;\pi/2]$. Comme nous l'avons montré au paragraphe 4.3.3.1, la synthèse OLA des canaux décorrélés (D_1,D_2,D_3) peut provoquer, sans correction de l'estimation, des pertes d'informations en suivant le schéma de principe de la **Figure 4.12**. Par conséquent, nous limitons les valeurs des angles d'Euler au premier quadrant $[0;\pi/2]$ en utilisant la valeur absolue des termes d'inter-corrélation $(r_{ij})_{1\leq i,j\leq 3}$ de signaux suivant le modèle décrit au paragraphe 4.2.1. En effet, sous l'hypothèse que les sources directionnelles sont en phase d'un canal à un autre, le couple $(\alpha,\beta)\in[0;\pi/2]$ permet la projection des canaux sur la direction principale des données dans le repère original (x_{C1},y_{C2},z_{C3}) *cf.* **Figure 4.19**. L'utilisation d'un modulo $\pi/2$ n'est pas nécessaire pour corriger l'estimation des angles α et β. Cependant, l'angle γ qui définit la direction secondaire des données dans le plan (Y_{D2},Z_{D3}) nécéssite une correction de son estimation pour éviter la perte d'informations lors de la synthèse OLA (*cf.* paragraphe 4.3.4.1).

La validité du système d'équations (4.50) peut être discutée pour certains cas particuliers, notamment lorsque au moins un des trois signaux analysés est nul c'est-à-dire lorsque plusieurs termes d'inter-corrélation sont nuls. En effet, si le canal C_1 est nul alors $r_{13}=r_{12}=0$ et les angles α et β sont indéterminés d'après le système d'équations (4.50). De la même manière, si $r_{13}=r_{23}=0$ (canal C_3 nul) alors les mêmes indéterminations subsistent. Ces indéterminations peuvent être aisément levées puisque le problème qui était initialement de dimension trois se réduit, dans ces cas particuliers, à un problème de dimension deux. Ces cas particuliers sont alors résolus avec l'utilisation des matrices de rotation suivantes :

$$\mathbf{R}_{\mathbf{3}(C_1=0)}(\alpha)=\begin{pmatrix}0 & \cos(\alpha) & \sin(\alpha)\\ 0 & -\sin(\alpha) & \cos(\alpha)\\ 1 & 0 & 0\end{pmatrix}, \tag{4.51}$$

$$\mathbf{R}_{\mathbf{3}(C_2=0)}(\alpha)=\begin{pmatrix}\cos(\alpha) & 0 & \sin(\alpha)\\ -\sin(\alpha) & 0 & \cos(\alpha)\\ 0 & 1 & 0\end{pmatrix} \tag{4.52}$$

$$\mathbf{R}_{\mathbf{3}(C_3=0)}(\alpha)=\begin{pmatrix}\cos(\alpha) & \sin(\alpha) & 0\\ -\sin(\alpha) & \cos(\alpha) & 0\\ 0 & 0 & 1\end{pmatrix}. \tag{4.53}$$

Il convient alors d'utiliser l'équation (4.41) pour estimer l'unique angle de rotation utile à la résolution de ces problèmes de dimension deux. Remarquons que si le cas où deux canaux parmi trois sont nuls se présente, la résolution du problème est assurée par une des équations (4.51), (4.52) ou (4.53) associée à l'équation (4.41) qui prend en compte ce cas particulier. Enfin, l'angle β est indéterminé si $r_{13}=0$. Dans ce cas, étant donné que les canaux C_1 et C_3 sont décorrélés mais non nuls ($r_{11}\neq 0$ et $r_{33}\neq 0$), il convient de fixer $\beta=\pi/4$ pour assurer la stabilité du système d'équation (4.50).

En conclusion, les matrices de rotation décrites par les équations (4.51), (4.52) et (4.53) seront utilisées pour calculer la matrice des vecteurs propres lorsque au moins deux termes

d'inter-corrélation de la matrice $\mathbf{\Gamma_{C3}}$ sont nuls[11]. Dans tous les autres cas de figure, les équations (4.50) sont utilisées pour en déduire la matrice des vecteurs propres décrite par l'équation (4.46) en tenant compte de l'indétermination sur l'angle β lorsque r_{13}=0 et en utilisant la valeur absolue des termes d'inter-corrélation.

4.3.4.1 Interprétation physique des angles d'Euler

L'ACP bidimensionnelle présentée au paragraphe 4.3.3 est réalisée au moyen d'un angle de rotation pourvu d'un sens physique interprétable dans un contexte d'analyse de signaux stéréophoniques. Cet angle paramètre la rotation du repère original des données en un nouveau repère dont l'axe des abscisses coïncide avec l'axe principal des données (*cf.* **Figure 2.4**). D'après le paragraphe 4.3.3.1, l'angle de rotation estimé pour l'ACP bidimensionnelle α est converti en un azimut θ relatif à la position de la source dominante dans le plan où sont diffusés les signaux.

L'ACP tridimensionnelle, décrite au paragraphe 4.3.4, vise à extraire trois angles de rotation *i.e.* les angles d'Euler, utilisés pour générer un nouveau repère de coordonnées tridimensionnelles dont l'axe des abscisses coïncide avec la direction principale des données (*cf.* **Figure 4.19**). Cet espace d'analyse à trois dimensions ne peut être mis directement en correspondance avec un système de diffusion multicanal dit classique (*cf.* **Figure 4.1**) puisque les haut-parleurs sont disposés dans le plan horizontal. Cependant, on peut imaginer appliquer l'ACP tridimensionnelle à divers types de signaux destinés à la reproduction audio en deux ou trois dimensions. Par exemple, une prise de son ambisonique d'ordre 1 (*cf.* paragraphe 1.2.2.2), délivre notamment trois composantes qui s'expriment suivant les axes d'un repère tridimensionnel [MOR06].

Par conséquent, l'analyse d'un signal multicanal, au moyen de l'ACP tridimensionnelle, peut être interprétée à partir de la connaissance de la technique de prise de son employée et/ou du système de reproduction sonore associé. Le cadre de cette étude se limite à l'interprétation physique des angles d'Euler extraits d'une scène sonore diffusée dans le plan horizontal. L'ACP tridimensionnelle est appliquée aux signaux audio multicanaux de type 5.1 et plus précisément à un ou plusieurs triplets de canaux du signal multicanal de départ.

Conversion des angles d'Euler en azimut

Nous considérons un signal multicanal reproduit sur un système d'écoute 5.0 (*cf.* **Figure 4.2**). Les angles d'Euler sont extraits à partir de l'analyse d'un triplet de canaux (C_1,C_2,C_3) menées au début du paragraphe 4.3.4. Nous cherchons à convertir les angles (α,β,γ) estimés, en un azimut θ repérant la position de la source dominante à chaque instant.

Nous faisons l'hypothèse qu'une source directionnelle perçue à l'azimut θ est reproduite par les haut-parleurs les plus proches de cet azimut quelque soit le type du signal multicanal *i.e.* type I ou II (*cf.* paragraphe 4.1.2.2). Par conséquent, nous considérons des triplets de canaux tels que les haut-parleurs correspondants soient adjacents les uns aux autres. D'après la **Figure 4.2**, nous analysons donc indépendamment les triplets ($c_1[n]$,$c_2[n]$,$c_3[n]$), ($c_4[n]$,$c_1[n]$,$c_2[n]$), ($c_2[n]$,$c_3[n]$,$c_5[n]$), ($c_1[n]$,$c_4[n]$,$c_5[n]$) ou ($c_2[n]$,$c_5[n]$,$c_4[n]$) c'est-à-dire l'image sonore frontale, l'image sonore latérale gauche et droite ainsi que l'image sonore arrière gauche et droite. Finalement, nous traitons trois situations où l'écartement des haut-parleurs diffère ainsi que la symétrie du système de reproduction : l'analyse de la scène sonore frontale, latérale et arrière *i.e.* l'analyse des scènes sonores latérales ou arrières gauche et droite sont équivalente d'un point de vue géométrique. La **Figure 4.20-(a)** présente un système de reproduction générique qui englobe ces trois situations : (θ_0=30°, θ_0'=0°) pour l'analyse d'une scène frontale, (θ_0=55°, θ_0'=25°) pour une scène latérale et (θ_0=110°, θ_0'=30°) pour une scène arrière.

[11] Même si à tout moment l'énergie d'un canal peut être nulle, en pratique, ces cas particuliers sont très rarement rencontrés puisque la présence d'une ambiance sonore pour chaque canal (voir le modèle introduit au paragraphe 4.2.1) implique des inter-corrélations éventuellement faibles mais généralement non nulles.

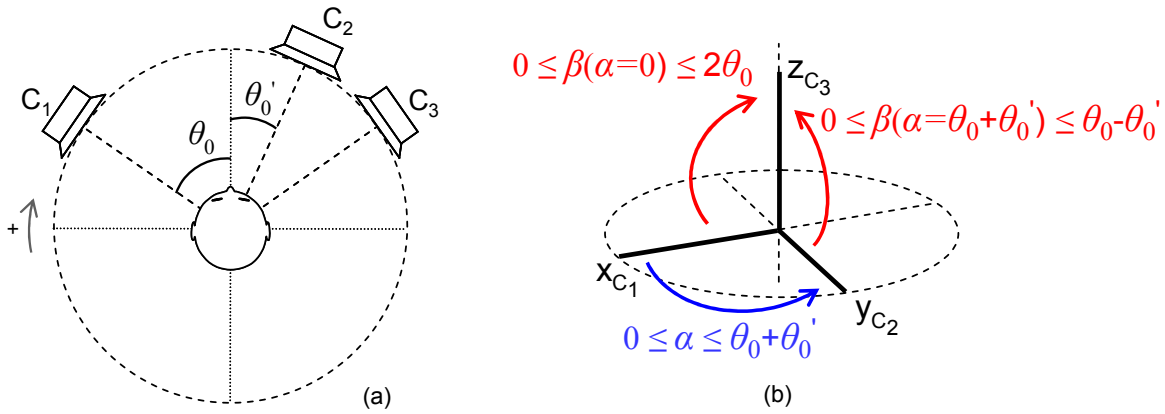

Figure 4.20 (a) - Système de reproduction à trois haut-parleurs dans le plan horizontal. Les deux haut-parleurs aux extrémités sont écartés de $2\theta_0$ et le haut-parleur « central » est repéré par l'angle θ_0'. **(b) – Correspondance entre les angles d'Euler et l'écartement des haut-parleurs.** Les angles α et β sont considérés comme convertis en degrés.

D'autre part, d'après la **Figure 4.19**, la composante principale D_1 est obtenue par projection du triplet de canaux d'entrée sur l'axe des abscisses du repère final (X) dérivé des rotations $\mathbf{R}^z{}_2(\alpha)$ et $\mathbf{R}^y{}_2(\beta)$. Par conséquent, la source directionnelle dominante est repérée dans l'espace tridimensionnel d'origine au moyen de l'angle d'azimut α et de l'angle d'élévation β. D'après la **Figure 4.20-(b)**, sous l'hypothèse que $\alpha,\beta \in [0;\pi/2]$, une estimation de l'azimut de la source dominante dans le plan des haut-parleurs *i.e.* azimut θ dans l'intervalle $[-\theta_0;\theta_0]°$, est donnée par la combinaison des angles α et β, estimés puis convertis en degrés, telle que :

$$\theta = \theta_\alpha + \left(\beta \times \frac{180}{90\times\pi}\right) \times \left(2\theta_0 - \theta_\alpha\right) - \theta_0 \quad \text{degrés} \tag{4.54}$$

avec :

$$\theta_\alpha = \left(\alpha \times \frac{180}{90\times\pi}\right) \times \left(\theta_0 + \theta_0'\right). \tag{4.55}$$

Après simplification, l'azimut de la source dominante θ est donnée à partir des angles d'Euler α, β et des angles repérant les écartements des haut-parleurs tel que :

$$\theta = \frac{2\alpha}{\pi}\left(\theta_0 + \theta_0'\right)\left(1 - \frac{2\beta}{\pi}\right) + \frac{4\beta\theta_0}{\pi} - \theta_0 \quad \text{degrés}. \tag{4.56}$$

La **Figure 4.21-(a)** illustre la conversion des angles α et β en azimut θ pour un écartement (θ_0=30°, θ_0'=0°) des haut-parleurs du système de reproduction présenté à la **Figure 4.20-(a)**. Par exemple, un signal à trois canaux constitué d'une source dominante[12] repérée idéalement à gauche de l'image sonore *i.e.* à l'azimut θ=-θ_0 (le nuage de points des données est orienté sur l'axe x_{C1} du repère original, *cf.* **Figure 4.20-(b)**), implique une estimation des angles d'Euler telle que (α=0, β=0) pour finalement générer un signal dominant tel que $d_1[n]$=$c_1[n]$ d'après l'équation (4.45). L'équation de conversion (4.56) prend en considération le fait qu'une source dominante localisée à l'azimut θ=θ_0 (le nuage de points des données est orienté sur l'axe z_{C3} du repère original, *cf.* **Figure 4.20-(b)**) implique une estimation des angles

[12] Nous considérons le cas le plus général où le signal analysé est constitué de trois canaux non nuls. Par conséquent, la source dominante se réfère à un signal multicanal constitué d'au moins une source directionnelle mélangée à trois canaux d'ambiance.

d'Euler telle que ($\beta=\pi/2, \forall \alpha$) pour finalement générer un signal dominant tel que $d_1[n]=c_3[n]$ d'après l'équation (4.45).

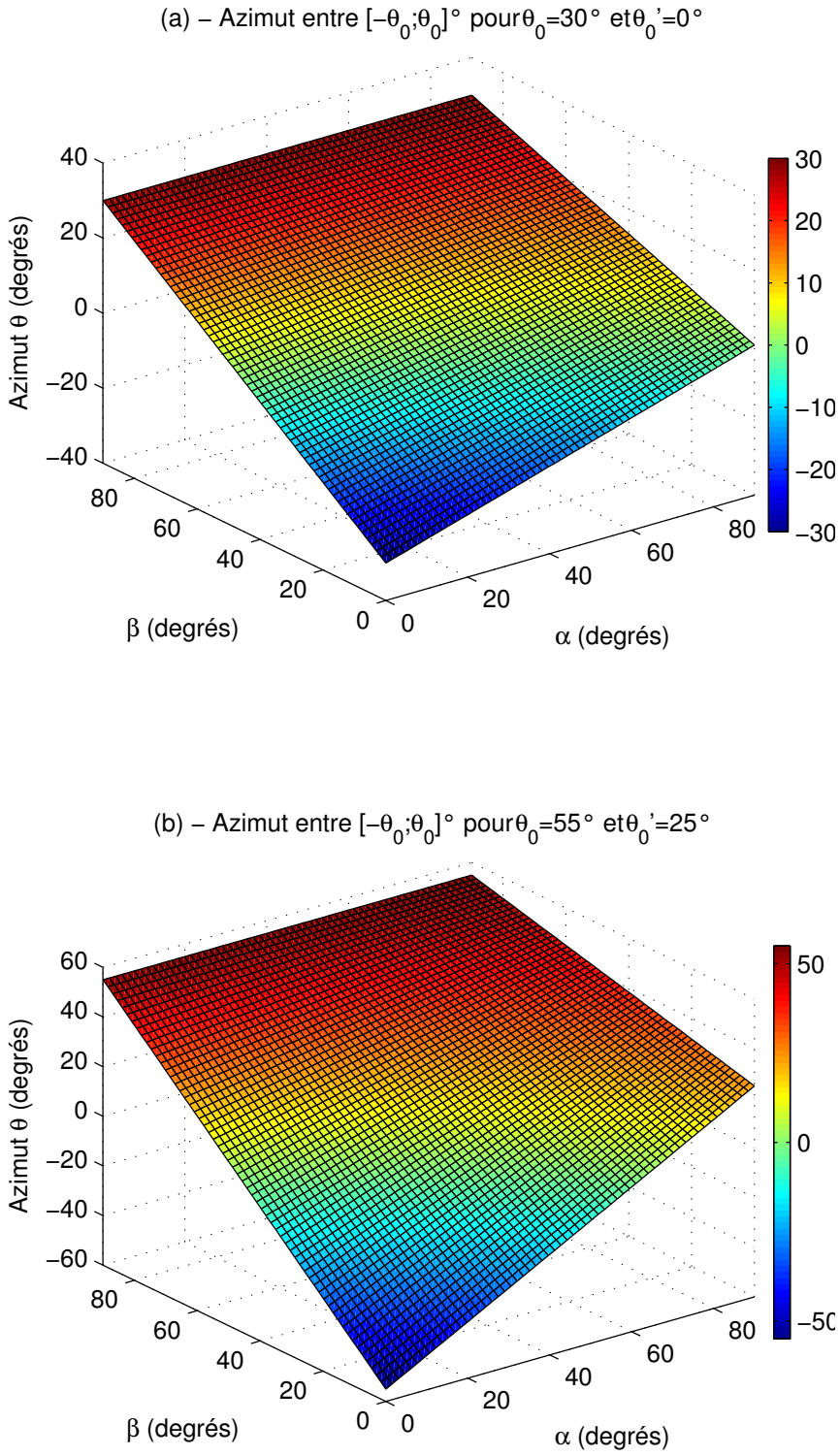

Figure 4.21 Conversion des angles d'Euler α et β (en degrés), estimés pour l'ACP tridimensionnelle, en azimut θ appartenant à l'intervalle [-θ_0,θ_0]. (a) - θ_0=30° et θ_0'=0°. **(b) -** θ_0=55° et θ_0'=25°.

Enfin, pour cette configuration de haut-parleurs symétrique, on remarque sur la **Figure 4.21-(a)** que l'analyse d'un signal constitué d'une source dominante à un azimut $\theta \neq \pm \theta_0$ génère un angle β, non unique, dont la contribution pour la conversion en azimut à partir de

l'équation (4.56) est supérieure à celle de l'angle α également non unique (*cf.* **Figure 4.20-(b)** avec θ_0'=0°). Nous considérons qu'à un azimut perçu correspond plusieurs couples (α,β) puisqu'une source directionnelle peut être perçue au même azimut à partir de signaux différents.

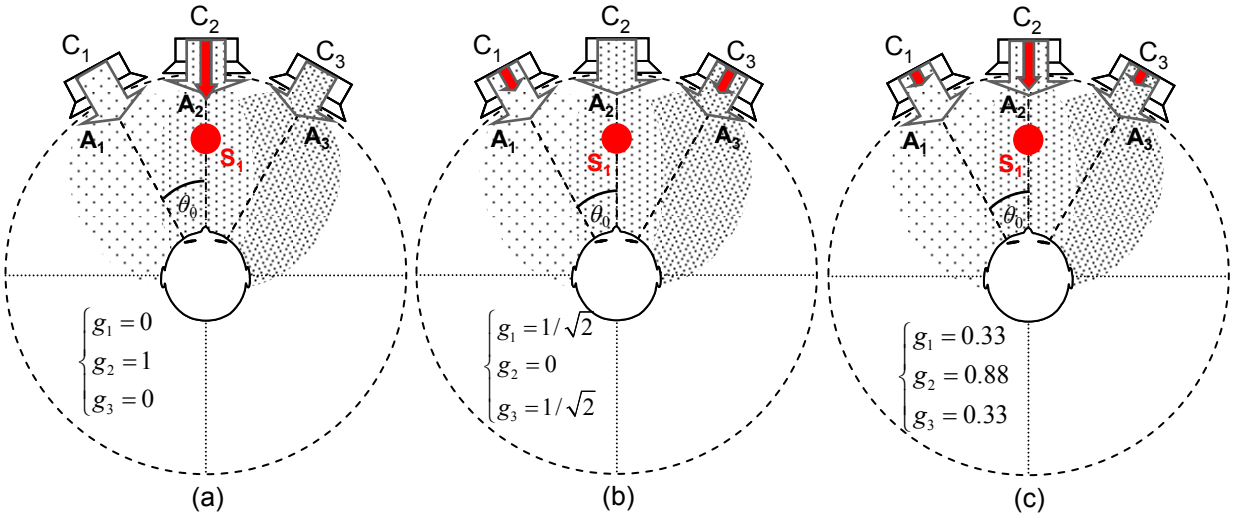

Figure 4.22 Signaux multicanaux issus de différents mélanges de la source directionnelle S_1 aux ambiances A_1, A_2, et A_3 pour une source dominante perçue à l'azimut θ_1=0° (θ_0=30°). (a) - Seul le canal central C_2 délivre la source directionnelle. **(b) –** Les canaux C_1 et C_2 délivrent la source directionnelle. **(c) –** Tous les canaux C_1, C_2 et C_3 délivrent la source S_1 à des niveaux d'énergie différents.

D'après la **Figure 4.22**, une source directionnelle S_1 percue à l'azimut θ_1=0° peut provenir du canal C_2 uniquement, du couple de canaux (C_1,C_3) et éventuellement du triplet de canaux (C_1,C_2,C_3) avec des niveaux d'énergie différents. Les trois signaux multicanaux présentés à la **Figure 4.22** suivent tous le modèle de l'équation (4.5) avec M=3 et D=1. Les gains (g_1,g_2,g_3) respectent le critère de conservation de l'énergie tel que la somme de leur carré vaut un. Le cas de la situation présentée à la **Figure 4.22-(c)** utilise les gains définis par Gerzon dans [GERZ92a].

Nous illustrons la conversion des angles d'Euler en azimut de la source dominante à partir d'un signal multicanal synthétique constitué d'un signal de parole chantée ($s_1[n]$), dont la RTF est présentée à la **Figure 4.6-(a)**, pondéré par les gains (g_1,g_2,g_3) constants au cours du temps. Cette source directionnelle pondérée est finalement respectivement sommée à l'un des trois canaux (A_1,A_2,A_3) d'un signal multicanal d'ambiance provenant d'une prise de son réalisée dans le hall d'un aéroport. Les RTF des signaux $a_1[n]$ et $a_3[n]$ sont présentées à la **Figure 4.6-(c)** et **Figure 4.6-(d)**. Les angles d'Euler sont estimés à partir du système d'équation (4.50) avec un calcul de la covariance et des valeurs propres de signaux exprimés dans le domaine temporel (fenêtre sinus de longueur N=1024 échantillons avec un recouvrement à 50%). L'interprétation des angles d'Euler et de leur conversion en azimut, présentés à la **Figure 4.24**, s'appuient sur la nature des signaux analysés à partir de la **Figure 4.23** qui présente la superposition des puissances de la source S_1 et des canaux d'ambiance (A_1,A_2,A_3). Ces puissances ont été estimé à partir des équations (4.8) et (4.10) à partir des portions glissantes de signaux également utilisées pour le calcul des angles d'Euler. La situation décrite à la **Figure 4.22-(a)** considère la source S_1 diffusée par le canal C_2 et résulte en un couple d'angles (α~90°, β~0°) lorsque la source S_1 a une puissance supérieure aux puissances des ambiances *i.e.* $\sigma_{S_1}^2 > \sigma_{A_i}^2 \ \forall i \in [1;3]$, d'après la **Figure 4.23** et la **Figure 4.24**.

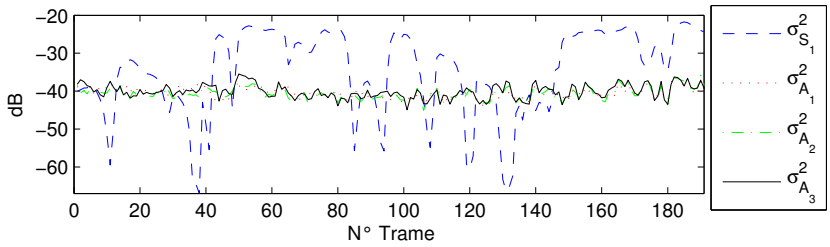

Figure 4.23 Puissance du signal de parole (S_1) superposée à celles des signaux d'ambiance (A_1, A_2 et A_3).

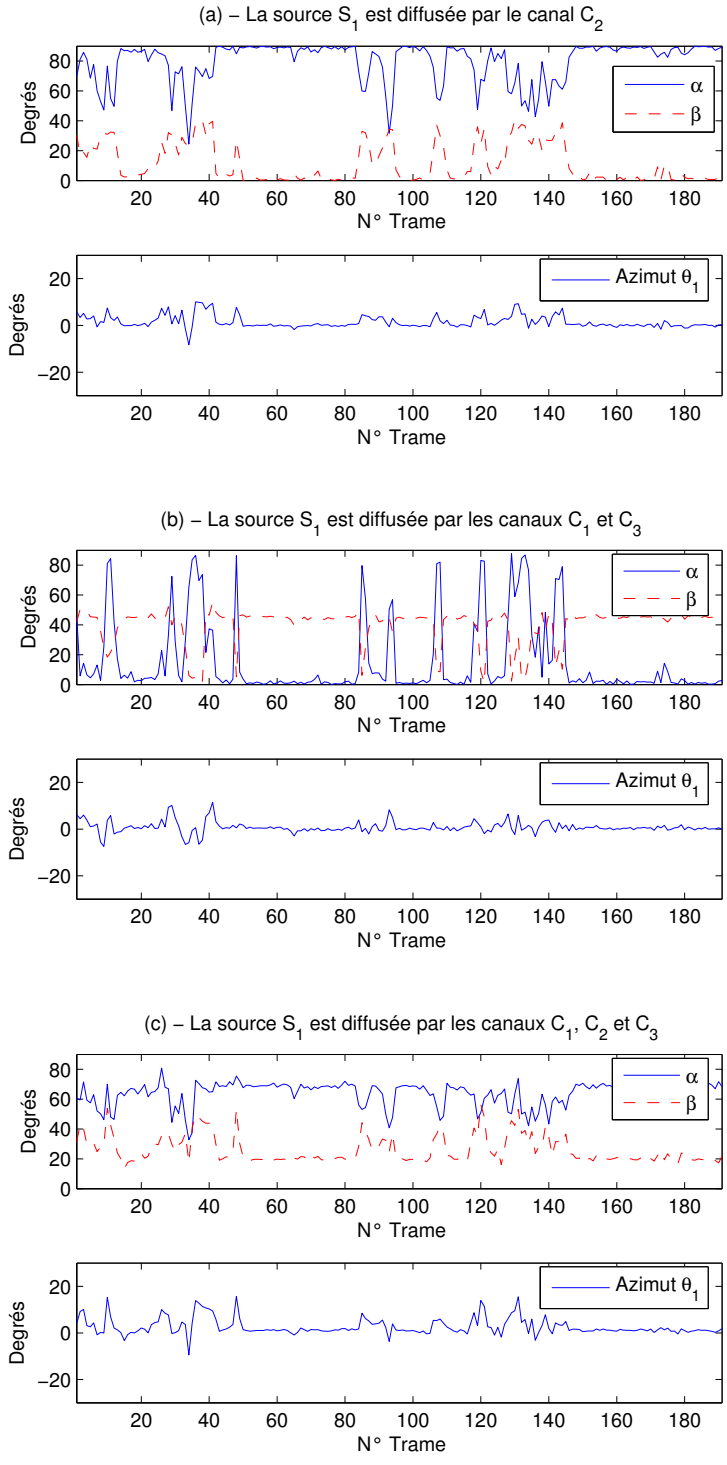

Figure 4.24 Angles d'Euler estimés et convertis en azimuts à partir des situations (a)-(b)-(c) décrites à la Figure 4.22.

La situation décrite à la **Figure 4.22-(b)** considère la source S_1 diffusée par les canaux C_1 et C_3 et résulte en un couple d'angles (α~0°, β~45°) pendant les périodes de forte activité vocale. Enfin, la situation décrite à la **Figure 4.22-(c)** considère la source S_1 diffusée par tous les canaux et résulte en un couple d'angles (α~70°, β~20°) pendant les périodes de forte activité vocale. Les angles α et β ainsi estimés sont ensuite convertis en azimut à partir de l'équation (4.56). D'après la **Figure 4.24**, les trois situations, présentées à la **Figure 4.22**, génèrent trois couples différents d'angles (α,β) pour un même azimut θ_1=0° (pendant les périodes de forte activité vocale) relatif à la source directionnelle dominante (S_1). Notons que l'azimut ainsi calculé s'écarte de la valeur zéro, pour les trois situations, lorsque la puissance de la source S_1 est inférieure aux puissances des ambiances *i.e.* $\sigma_{S_1}^2 < \sigma_{A_i}^2 \ \forall i \in [1;3]$. En effet, les ambiances ne sont pas attachées à une direction particulière d'un point de vue perceptuel. Ce phénomène se traduit par un nuage de point diffus dans l'espace tridimensionnel d'analyse.

Dans le cas où le système de reproduction sonore n'est pas symétrique (*cf.* **Figure 4.20-(a)**), la **Figure 4.21-(b)** illustre la conversion des angles α et β en azimut θ pour un écartement (θ_0=55°, θ_0'=25°). Etant donné que l'écartement des haut-parleurs relatifs aux canaux C_1 et C_2 est supérieur à θ_0, la contribution de l'angle α pour la conversion en azimut augmente (*cf.* **Figure 4.20-(b)**) comparé au cas du système de reproduction symétrique. Dans cette configuration, une source dominante localisée à l'azimut θ=θ_0' à partir d'un signal suivant le modèle de l'équation (4.5) *i.e.* plusieurs triplets de gains appliqués à la source sont admissibles comme pour le cas d'un système symétrique (*cf.* **Figure 4.22**), admet une solution possible telle que (α=π/2, β=0) lorsque la source est délivrée par le canal C_2.

Nous cherchons maintenant à comparer les conversions d'angles, estimés par ACP bi- et tridimensionnelle, en azimut à partir des équations (4.40) et (4.56). Nous proposons l'analyse d'un triplet de signaux, à la fois dans le cas d'un système de reproduction symétrique (θ_0=30°, θ_0'=0°) et non symétrique (θ_0=55°, θ_0'=25°), dont la perception est identique (à la valeur d'écartement θ_0 près) à celle du signal stéréo, présenté au paragraphe 4.2.2.1, dans le cadre de l'ACP bidimensionnelle. Pour y parvenir, deux signaux multicanaux ont été synthétisés à partir du modèle de l'équation (4.5) avec M=3 et D=2. Les RTF des sources directionnelles (S_1 et S_2) sont présentées à la **Figure 4.6-(a)-(b)**. Les RTF des signaux d'ambiance $a_1[n]$ et $a_3[n]$ sont présentées à la **Figure 4.6-(c)-(d)**. A la manière du cas bidimensionnel, un panoramique suivant la loi des sinus a été choisi pour définir la matrice de gains $\mathbf{G_{32}}$ dans les deux configurations présentées à la **Figure 4.25**. Les gains de panoramique $g_{md}[n]$ sont définis pour chaque source d'indice d en fonction de son azimut θ_d et ceci pour chaque couple de canaux adjacents *i.e.* (g_{1d},g_{2d}) et (g_{2d},g_{3d}), à partir de l'équation (4.22) où θ_0 correspond à l'écartement des haut-parleurs relatifs au couple de canaux considérés. Les trajectoires opposées des sources directionnelles simplifient la matrice de gains $\mathbf{G_{32}}$ seulement dans le cas d'un système de reproduction symétrique (*cf.* **Figure 4.25-(a)**), telles que :

$$\begin{cases} g_{11}[n] = g_{32}[n] = g_1[n] \\ g_{21}[n] = g_{22}[n] = g_2[n] \\ g_{31}[n] = g_{12}[n] = g_3[n] \end{cases}. \tag{4.57}$$

Les gains de panoramique dans le cas où le système de reproduction n'est pas symétrique sont présentés à la **Figure 4.25-(b)** respectivement pour la source S_1 et pour la source S_2. Comme pour le signal stéréo présenté au paragraphe 4.2.2.1, les signaux d'ambiance, qui ont des puissances originales équivalentes $\sigma_{A_1}^2 \approx \sigma_{A_2}^2 \approx \sigma_{A_3}^2$ (*cf.* **Figure 4.23**), ont été pondérées par un coefficient défini par l'équation (4.21) tel que RSDA=15 dB. Finalement, les sources pondérées par les gains de la matrice $\mathbf{G_{32}}$ sont sommées aux signaux d'ambiance d'après l'équation (4.5). La **Figure 4.25-(a)-(b)** illustre les scènes sonores ainsi synthétisées.

Les angles d'Euler sont estimés à partir du système d'équation (4.50) avec un calcul de la covariance et des valeurs propres de signaux exprimés dans le domaine temporel (fenêtre

sinus N=1024 échantillons avec un recouvrement à 50%). L'évolution temporelle des angles d'Euler estimés à partir du signal multicanal décrit à la **Figure 4.25-(a)** (respectivement à la **Figure 4.25-(b)**) est présentée à la **Figure 4.26-(a)** (respectivement à la **Figure 4.27-(a)**). La comparaison des azimuts établis à partir d'une ACP bidimensionnelle (*cf.* **Figure 4.14**) et tridimensionnelle à partir de l'équation (4.56) et du signal multicanal décrit à la **Figure 4.25-(a)** (respectivement à la **Figure 4.25-(b)**) est présentée à la **Figure 4.26-(b)** (respectivement à la **Figure 4.27-(b)**).

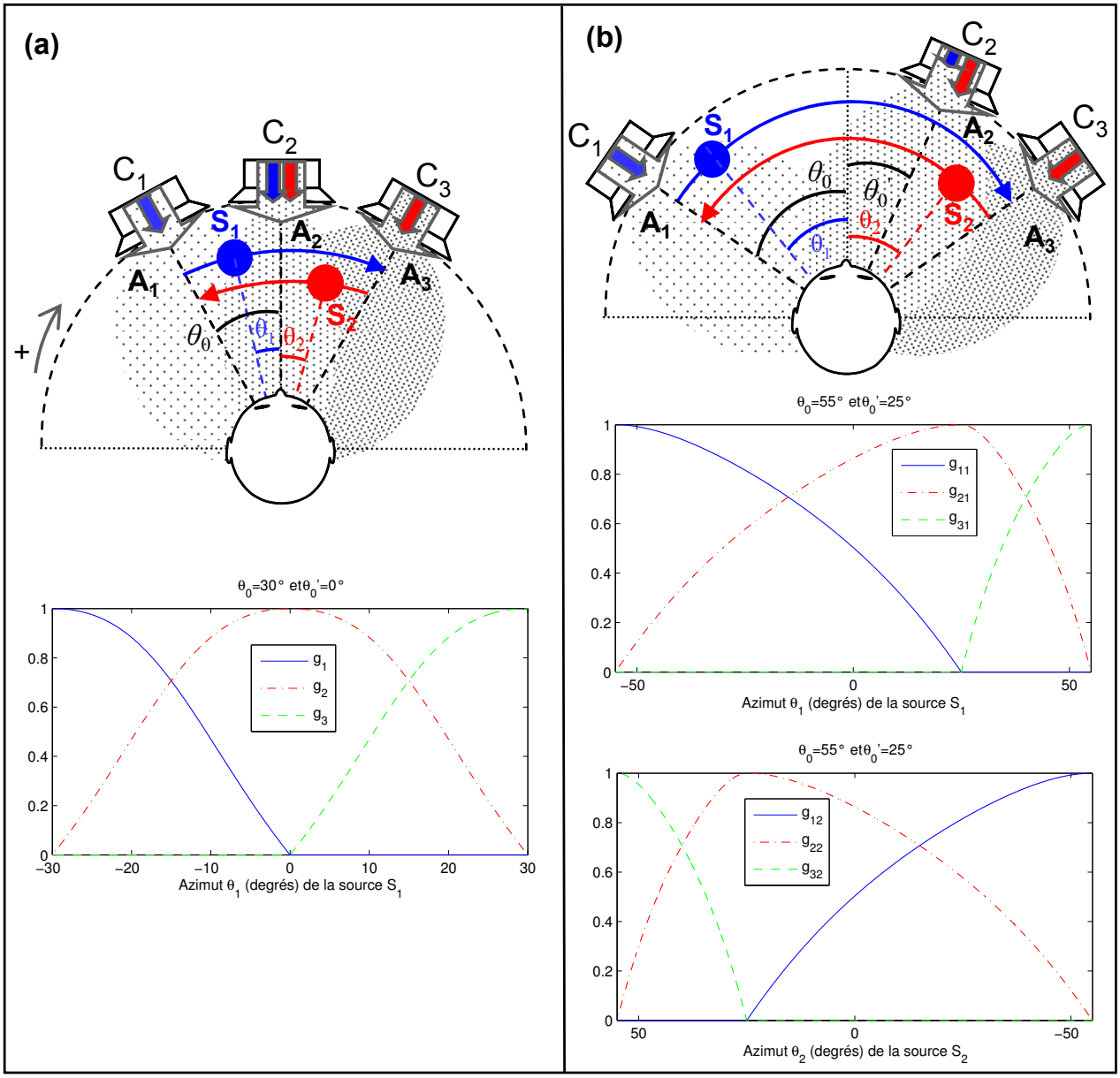

Figure 4.25 Signaux multicanaux (M=3) issus du mélange de D=2 sources directionnelles (signal de parole chantée S_1 en bleu et de glockenspiel S_2 en rouge) et d'un signal d'ambiance multicanal (A_1, A_2, A_3 en textures grises). Les gains de panoramique sont définis par la loi des sinus avec **[(a) –** un système de reproduction symétrique (θ_0=30°, θ_0'=0°), **(b) –** un système de reproduction non symétrique (θ_0=55°, θ_0'=25°)**]**.

D'après la **Figure 4.26**, la combinaison des angles d'Euler α et β définie par l'équation (4.56) permet de retrouver globalement le même azimut de la source dominante qui avait été obtenu par l'ACP bidimensionnelle.

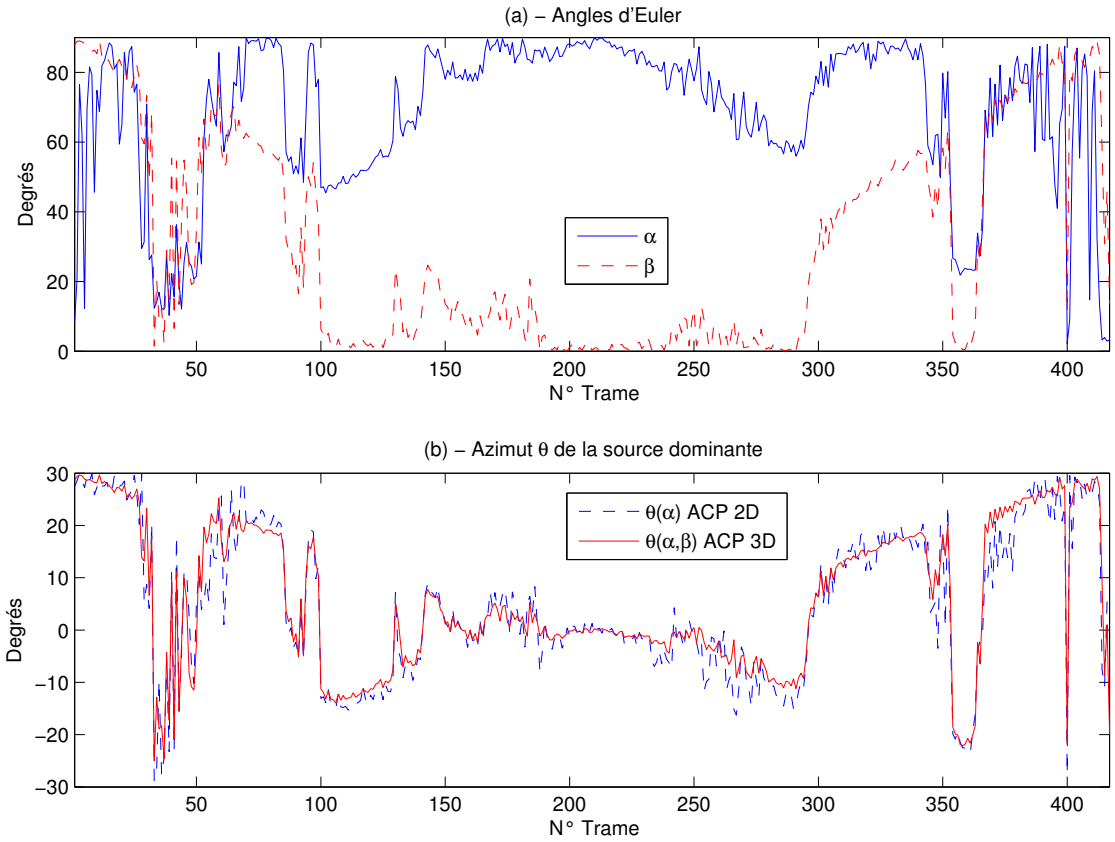

Figure 4.26 (a) - Angles d'Euler (α et β en degrés) estimés à partir du signal multicanal présenté à la Figure 4.25-(a). **(b) –** Azimuts $\theta(\alpha)$ et $\theta(\alpha,\beta)$ calculés à partir d'une ACP bi- et tridimensionnelle.

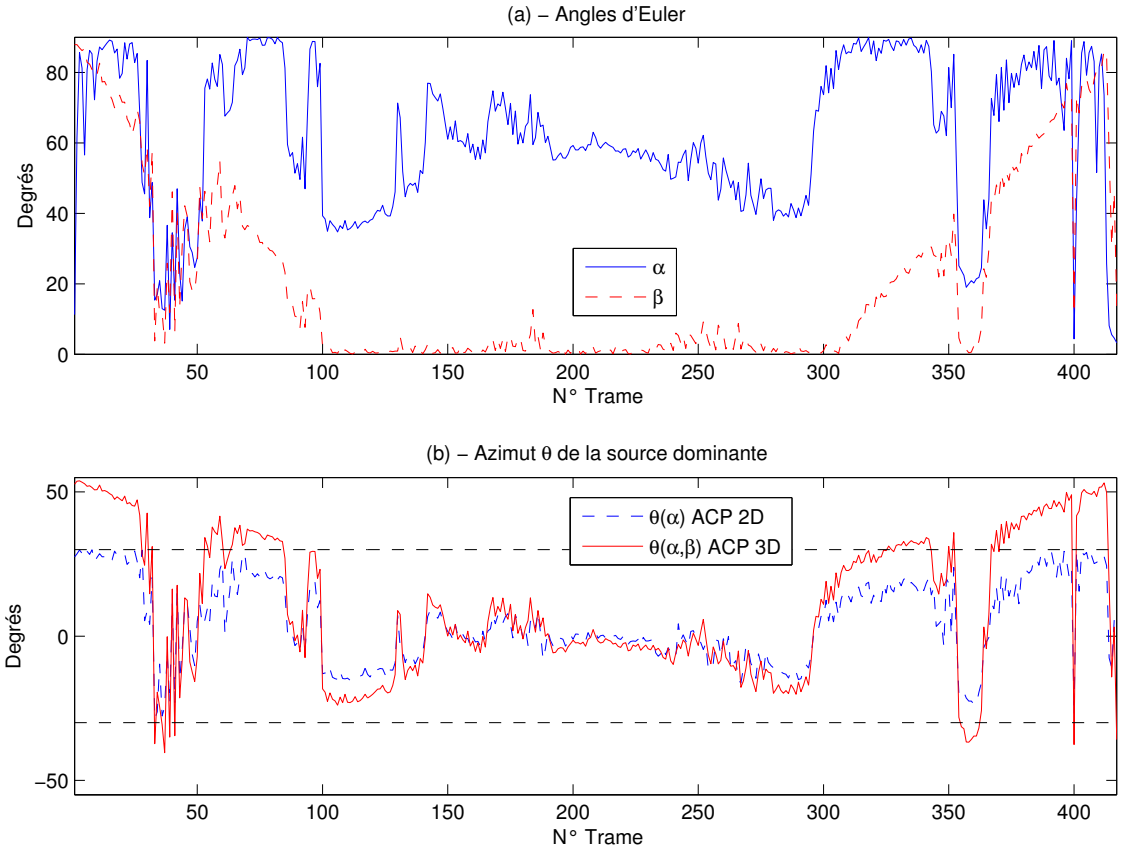

Figure 4.27 (a) - Angles d'Euler (α et β en degrés) estimés à partir du signal multicanal présenté à la Figure 4.25-(b). **(b) –** Azimuts $\theta(\alpha)$ et $\theta(\alpha,\beta)$ calculés à partir d'une ACP bi- et tridimensionnelle.

A partir des signaux présentés à la **Figure 4.25-(a)** et à la **Figure 4.25-(b)**, les angles d'Euler estimés et présentés sur la **Figure 4.26-(a)** et la **Figure 4.27-(a)** sont différents étant donné la nature différente des canaux analysés. Cependant, d'après **Figure 4.27-(b)**, la combinaison des angles d'Euler définie par l'équation (4.56) permet de retrouver l'azimut de la source dominante qui évolue dans le système de reproduction non-symétrique *i.e.* entre $-\theta_0$ et θ_0=55°.

En conclusion, les angles d'Euler α et β permettent d'estimer la position dans l'espace de la source dominante d'un signal multicanal (à trois canaux) à partir de la connaissance du système de reproduction sonore. Nous avons établi une opération de conversion dans le cas d'un système de reproduction dans le plan horizontal. Comme nous l'avons remarqué au début de cette analyse, cette opération est directement transposable à d'autres systèmes de reproduction et différents formats de signaux audio.

Deux rotations pour extraire l'information principale

D'après notre définition de la matrice de projection tridimensionnelle à partir des angles d'Euler (convention ZYX, *cf.* équation (4.46)), nous avons proposé une interprétation physique des angles α et β. Basé sur cette interprétation, nous pouvons définir le contenu du signal D_1 issu du matriçage adaptatif défini par l'équation (4.45). En effet, ce matriçage fait intervenir les trois canaux d'origine et les angles α et β pour définir le signal D_1 c'est-à-dire la projection des canaux (C_1,C_2,C_3) sur l'axe X du repère obtenu à partir des rotations $\mathbf{R}^z{}_2(\alpha)$ et $\mathbf{R}^y{}_2(\beta)$. Par conséquent, ces deux rotations permettent d'extraire l'information principale (sources dominantes et ambiances coïncidentes) du signal multicanal d'origine dans le signal D_1.

De manière à déterminer les composantes des signaux D_2 et D_3, l'étude suivante vise à évaluer l'influence de la troisième rotation d'angle γ, $\mathbf{R}^x{}_2(\gamma)$, qui permet de réaliser l'ACP tridimensionnelle complète. D'après la **Figure 4.28-(a)**, la rotation $\mathbf{R}^x{}_2(\gamma)$ définie l'orientation des axes Y_{D2} et Z_{D3} du repère final à partir du repère (x_2,y_2,z_2) obtenu par les rotations $\mathbf{R}^z{}_2(\alpha)$ et $\mathbf{R}^y{}_2(\beta)$ du repère original $(x,y,z) = (x_{C1},y_{C2},z_{C3})$. Alors que l'orientation des données dans le repère original est définie par le couple d'angle $(\alpha,\beta)\in[0;\pi/2]$, l'orientation ou direction des données dans le plan (Y_{D2},Z_{D3}) est complètement définie par le demi-plan repéré par l'angle $\gamma\in[-\pi/2;\pi/2]$. De façon à assurer la conservation de l'énergie lors de la synthèse OLA, nous limitons les valeurs de l'angle γ à l'intervalle $[0;\pi/2]$.

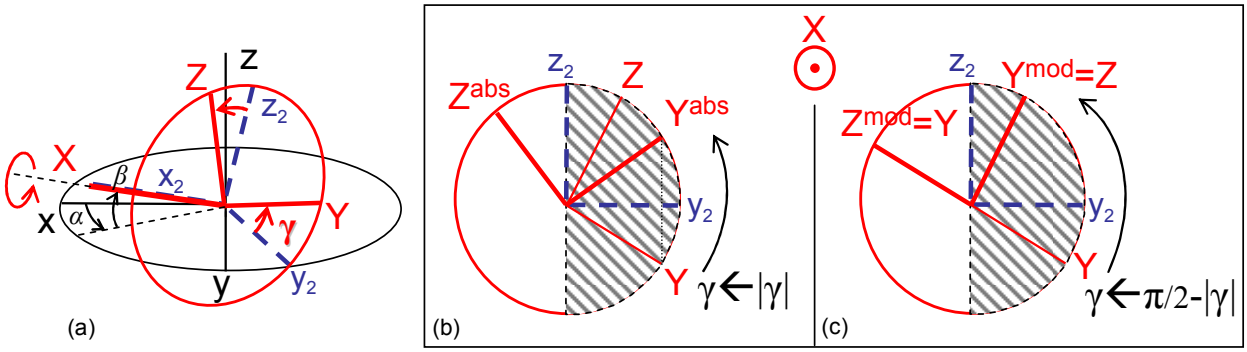

Figure 4.28 (a) – Troisième rotation d'Euler d'angle γ autour de l'axe x_2=X. **Correction de l'estimation de l'angle γ tel que $\gamma\in[0;\pi/2]$ en utilisant [(b) – une valeur absolue, (c) – un modulo $\pi/2$].**

D'après la **Figure 4.28-(c)**, lorsque $\gamma\in[-\pi/2;0]$, l'utilisation d'un modulo $\pi/2$ génère une inversion des axes originellement obtenu par la rotation $\mathbf{R}^x{}_2(\gamma)$ tel que : $(Y^{mod}{}_{D2},Z^{mod}{}_{D3}) = (Z_{D3},Y_{D2})$. Par conséquent, il conviendrait d'inverser l'ordre des canaux D_2 et D_3 dans l'équation (4.45) pour chaque portion de signal analysée où l'angle γ est négatif de façon à respecter la hiérarchisation énergétique des signaux. Cependant, cette manipulation n'est

toujours pas compatible avec la synthèse finale (OLA) des signaux. Nous proposons donc une correction de l'angle γ en utilisant une valeur absolue illustrée à la **Figure 4.28-(b)**. Bien que cette correction soit sous-optimale en termes de hiérarchisation des données : les sources secondaire sont réparties dans les deux canaux D_2 et D_3 obtenus par projection sur les axes $({Y^{abs}}_{D2},{Z^{abs}}_{D3}) \neq (Y_{D2},Z_{D3})$ lorsque γ est négatif et d'autant plus lorsque γ tend vers $\pi/4$; cette correction ne nécessite pas de manipulation des données et respecte globalement la hiérarchie indiquée par l'équation (4.45).

Etant donné l'influence de la décomposition en sous-bandes de fréquences de la covariance (*cf.* paragraphes 4.2.2.3 et 4.3.3.2), nous proposons d'évaluer à la fois l'influence de la troisième rotation et d'un découpage des signaux en sous-bandes de fréquences. Autrement dit, nous considérons la transformation décrite par l'équation (4.45) avec :

$$\mathbf{V_3} = \begin{cases} \mathbf{R_3}(\alpha,\beta,\gamma), \text{ avec } (\alpha,\beta,\gamma)[l] \text{ ou } (\alpha,\beta,\gamma)[l,b] \\ \text{ou} \\ \mathbf{R_3}(\alpha,\beta,0), \text{ avec } (\alpha,\beta)[l] \text{ ou } (\alpha,\beta)[l,b] \end{cases} \tag{4.58}$$

Les angles d'Euler sont estimés à partir du système d'équation (4.50) avec un calcul de la covariance et des valeurs propres de signaux exprimés dans le domaine temporel (fenêtre sinus N=4096 échantillons avec un recouvrement à 50%) ou des sous-bandes (K_b=20). Le principe des calculs est rigoureusement le même que celui utilisé au paragraphe 4.3.3.

L'évolution temporelle des angles d'Euler estimés en temps à partir du signal multicanal décrit à la **Figure 4.25-(a)** est présentée à la **Figure 4.29**.

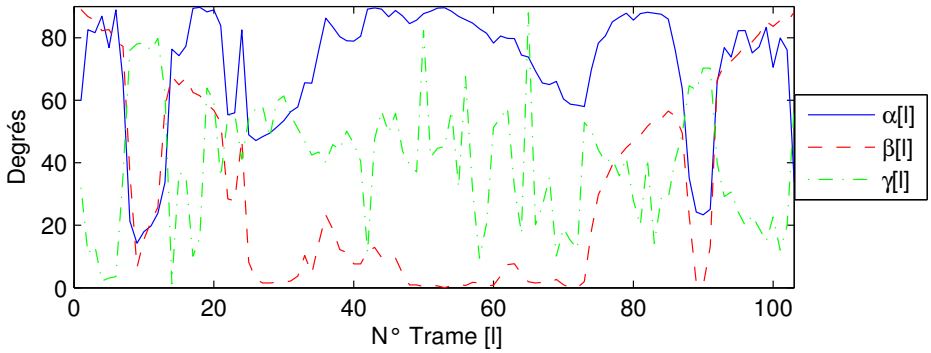

Figure 4.29 Angles d'Euler (α, β et γ en degrés) estimés en temps (portions glissante de N=4096 échantillons) à partir du signal multicanal présenté à la Figure 4.25-(a).

Alors que les angles α et β évoluent dans le temps au regard des gains appliqués aux sources directionnelles (*cf.* paragraphe précédent pour la combinaison de ces angles en azimut de la source dominante), l'angle γ (en degrés) présente une forte variance au cours du temps. Cela s'explique par le fait que cet angle repère la direction secondaire du nuage de point dans le plan orthogonal à l'axe principal des données. Etant donné que le nuage de point (dans ce plan) présente une moins grande variance (plus compact) comparé à celui contenant l'ensemble des données (sources directionnelles dominantes notamment), l'angle γ repère la direction des données énergétiquement secondaires c'est-à-dire les ambiances décorrélées d'un canal à un autre et les sources secondaires. Par conséquent, cet angle de rotation peut être mis en correspondance avec l'azimut de la source secondaire seulement lorsque les sources directionnelles (multiples) occupent des azimuts différents dans le plan horizontal. En effet, d'après la **Figure 4.29**, lorsque les sources directionnelles occupent leur position initiale géométriquement opposées (*cf.* **Figure 4.25-(a)**) *i.e.* l<20 et l>80, l'angle γ voit sa variance réduire et ses valeurs pourraient être mises en correspondance avec l'azimut de la source secondaire. Cependant, la variance des ambiances peut être supérieure à celles des sources secondaires notamment lorsque les sources directionnelles occupent la même

position dans l'espace de reproduction et d'analyse. D'après la **Figure 4.29**, la variance de l'angle γ est plus importante pour les portions glissantes d'indices $30<l<70$, c'est-à-dire lorsque les sources directionnelles occupent des positions spatiales identiques ou proches. En outre, la puissance de la source secondaire peut être, à tout instant, équivalente ou inférieure à celle de l'ambiance moyenne et par conséquent perturber l'estimation de l'azimut de la source secondaire. Finalement, le troisième angle de rotation utilisé pour réaliser la projection des données sur la base des vecteurs propres n'est pas complètement assimilable à la position d'une source sonore ponctuelle et localisable.

L'évolution temporelle des angles d'Euler estimés en sous-bandes de fréquences (K_b=10) à partir du signal multicanal décrit à la **Figure 4.25-(a)** est présentée à la **Figure 4.30**. La même interprétation peut-être menée à partir des composantes fréquentielles des signaux originaux dont les RTF sont présentées à la Figure 4.31**-(a)**. Les angles d'Euler estimés en sous-bandes sont donc dépendant du contenu fréquentiel des signaux et par conséquent de la finesse de l'analyse en sous-bandes (échelle fréquentielle et nombre de sous-bandes). Encore une fois, les angles α et β évoluent dans le temps au regard des gains appliqués aux sources directionnelles alors que l'angle γ présente globalement une forte variance excepté pour les dernières sous-bandes (b=9,10) de fréquences $7500<f<22050$ Hz puisqu'à ces fréquences l'énergie des signaux d'ambiance est très faible ou nulle (*cf.* les RTF des signaux d'ambiance gauche et droit présentés à la **Figure 4.6-(c)-(d)**).

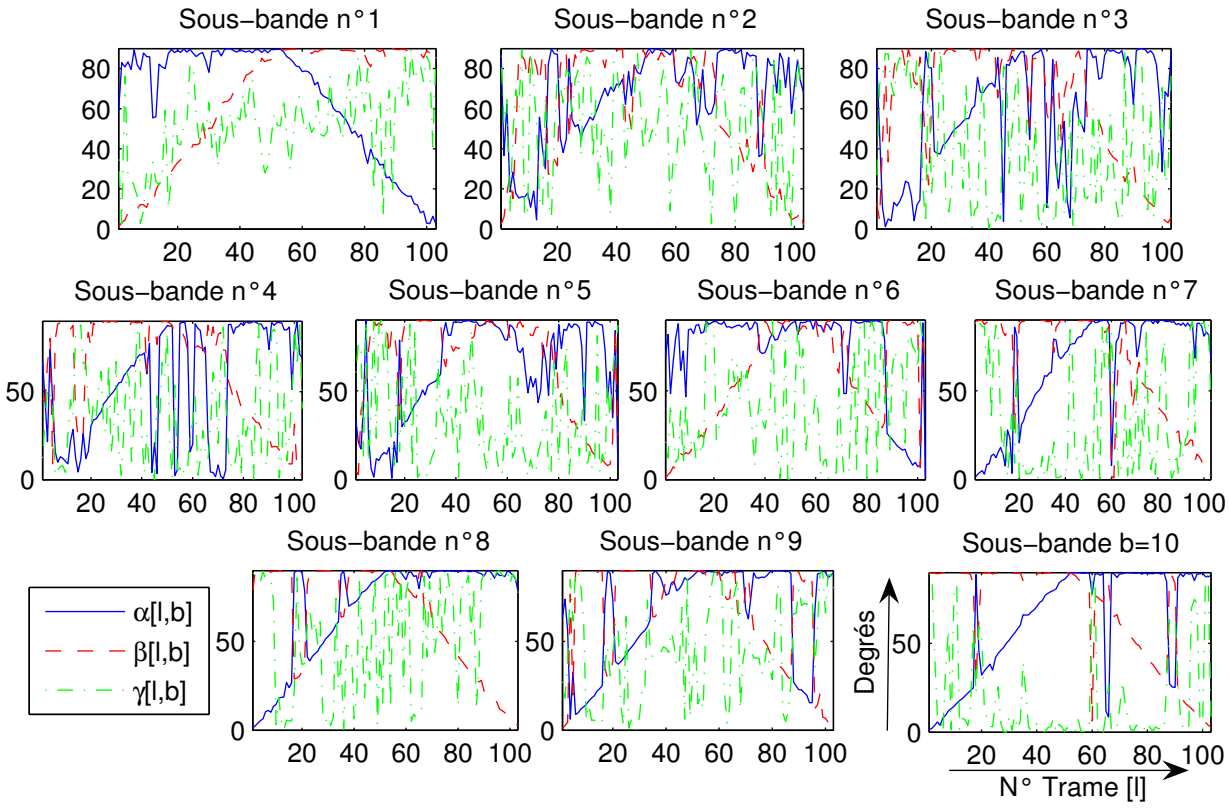

Figure 4.30 Angles d'Euler (α, β et γ en degrés) estimés en sous-bandes de fréquences (portions glissante de N=4096 échantillons et K_b=10 sous-bandes) à partir du signal multicanal présenté à la Figure 4.25-(a).

A partir de l'estimation des angles d'Euler, la matrice de transformation est établie avec seulement deux ou trois angles comme indiqué par le système d'équations (4.58). Les RTF des canaux originaux et des canaux obtenus avec deux ou trois rotations dans le domaine temporel ou des sous-bandes sont présentés à la **Figure 4.31**. Ces RTF sont toutes obtenues avec une fenêtre de *Hanning* à N=512 échantillons, 1024 coefficients spectraux sont calculés (Z=512) et un recouvrement de 50% entre les fenêtres glissantes est utilisé. D'après la **Figure**

4.31-(b), la RTF du signal d_1[n] obtenu avec trois rotations en temps est dépourvue de certaines composantes fréquentielles des sources directionnelles réparties dans les signaux d_2[n] et d_3[n]. Les RTF des signaux d_1[n] obtenus avec deux ou trois rotations en dix sous-bandes sont identiques puisqu'une rotation en azimut suivie d'une rotation en élévation sont suffisantes pour faire coïncider l'axe des abscisses du repère de projection avec la direction principale des données (*cf.* **Figure 4.28-(a)**). Par contre, d'après la **Figure 4.31-(c)** et la **Figure 4.31-(d)**, les RTF des signaux d_2[n] et d_3[n] obtenus avec deux ou trois rotations en sous-bandes sont différentes. En effet, l'ACP réalisée au moyen de trois rotations en sous-bandes permet de compacter au maximum l'énergie résiduelle au sein du signal d_2[n] qui contient les sources sonores secondaires et une partie de l'ambiance. Le signal d_3[n] ne contient alors que l'information provenant des signaux d'ambiance. Si l'ACP n'est réalisée qu'au moyen des deux premières rotations en sous-bandes, il apparaît que les signaux d_2[n] et d_3[n] contiennent chacun un mélange de sources secondaires et d'ambiances.

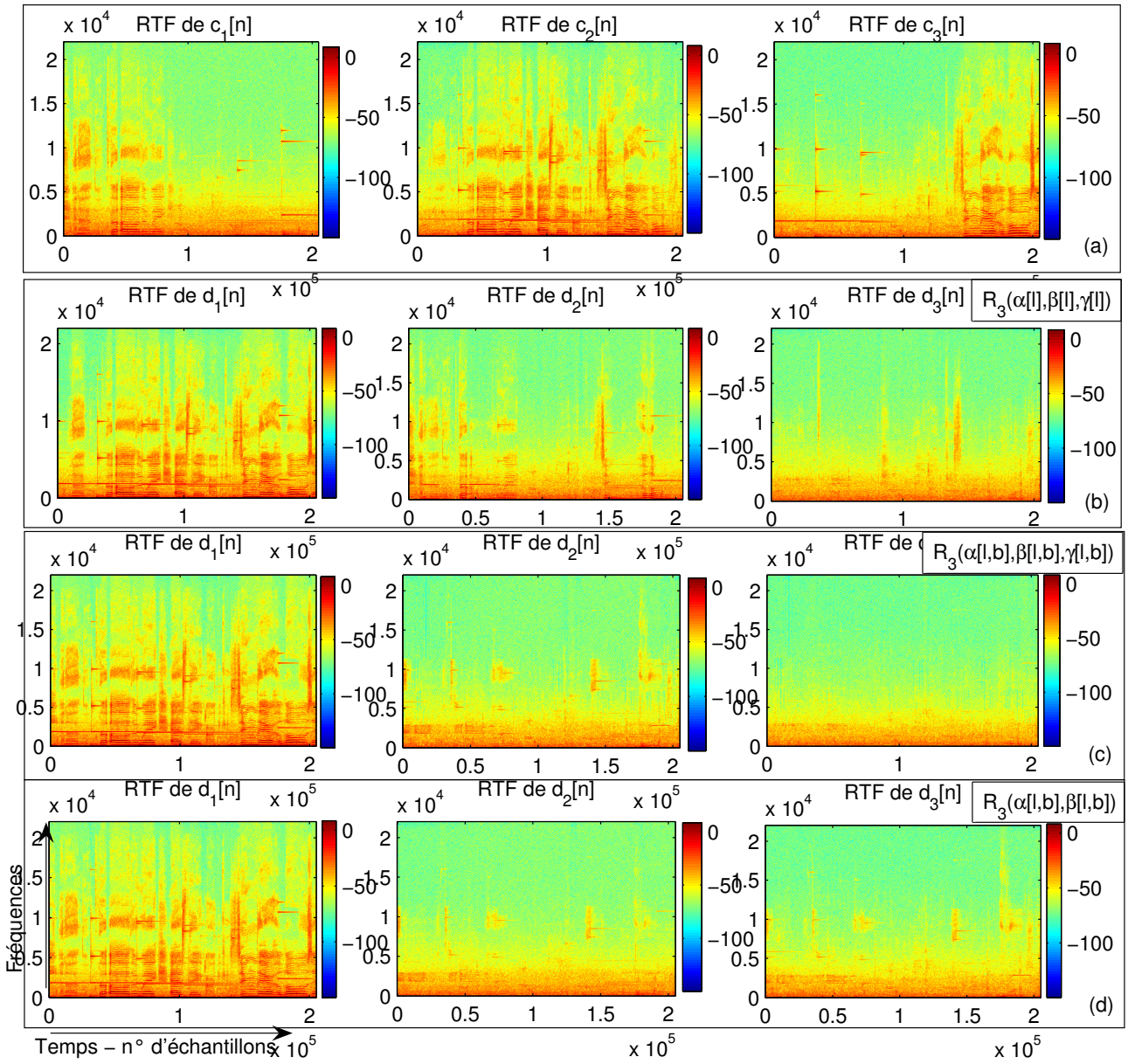

Figure 4.31 **(a) -** RTF (en dB) des canaux (C_1,C_2,C_3) du signal présenté à la Figure 4.25-(a). RTF des canaux (D_1,D_2,D_3) issus de l'ACP tridimensionnelle : **[(b) -** en temps au moyen de trois rotations d'angle $\alpha[l]$, $\beta[l]$, et $\gamma[l]$, **(c) -** en 10 sous-bandes au moyen de trois rotations d'angle $\alpha[l,b]$, $\beta[l,b]$ et $\gamma[l,b]$, **(d) -** en 10 sous-bandes au moyen de deux rotations d'angle $\alpha[l,b]$ et $\beta[l,b]$**]**.

D'un point de vue énergétique, l'ACP tridimensionnelle, réalisée au moyen de trois rotations, hiérarchise les signaux obtenus au regard de la distribution des valeurs propres. La **Figure 4.32** présente la comparaison des pourcentages d'énergies des canaux analysés (C_1, C_2, et C_3) et transformés (D_1, D_2 et D_3), au moyen de deux ou de trois rotations d'Euler, par rapport à l'énergie totale des canaux analysés. Les pourcentages d'énergie des signaux présentés sur la **Figure 4.32** montrent que l'ACP tridimensionnelle réalisée avec seulement deux rotations d'Euler ne hiérarchise pas complètement les signaux d'un point de vue énergétique. Les composantes D_2 et D_3 ont alors un pourcentage d'énergie quasi-équivalent.

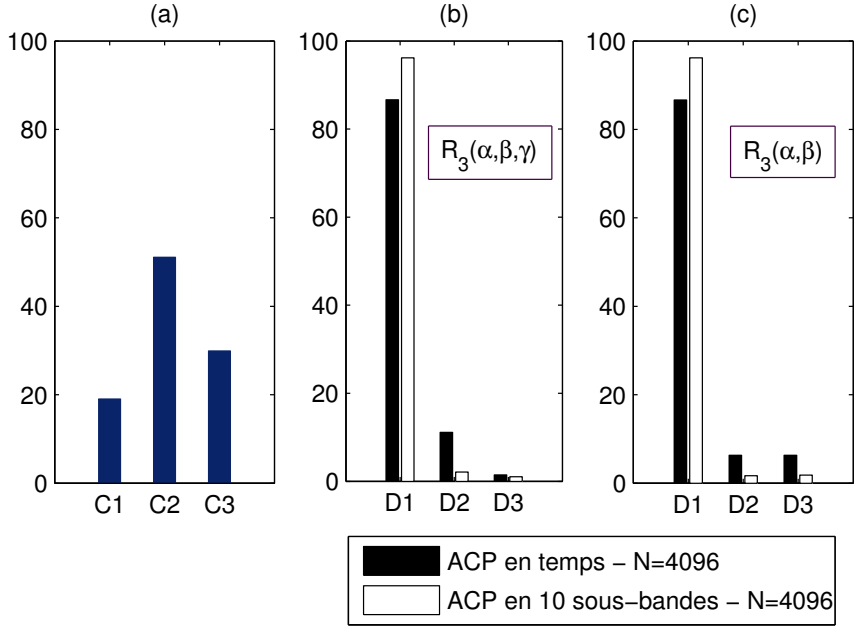

Figure 4.32 Pourcentage d'énergie, par rapport à l'énergie initiale, des signaux : [(a) - C_1, C_2 et C_3, **(b)** - D_1, D_2 et D_3 obtenus avec trois rotations en temps et en dix sous-bandes, **(c)** - D_1, D_2 et D_3 obtenus avec deux rotations en temps et en 10 sous-bandes**].**

En conclusion, le contenu des composantes orthogonales à D_1 (D_2 et D_3) est naturellement influencé par la troisième rotation d'Euler d'angle γ. Cette troisième rotation d'angle γ permet de concentrer au maximum l'énergie des signaux et plus précisément celle des sources directionnelles secondaires au sein de la seconde composante issue de l'ACP tridimensionnelle (D_2). En outre, le matriçage adaptatif avec deux ou trois rotations en sous-bandes pour réaliser l'ACP permet de concentrer au maximum l'énergie initiale dans la composante D_1 et, par suite, de réduire sensiblement, d'un point de vue énergétique et perceptif, la présence des sources directionnelles dans les composantes D_2 et D_3. Finalement, la différence entre un matriçage basé sur deux ou trois matrices de rotation s'amenuise avec un traitement en sous-bandes de fréquences (*cf.* **Figure 4.32**).

4.3.4.2 Performances de l'ACP tridimensionnelle en sous-bandes

De manière à comparer les performances de l'ACP tridimensionnelle réalisée dans le domaine temporel ou en sous-bandes de fréquences, une mesure de la concentration de l'énergie est obtenue en calculant les rapports énergétiques de la composante principale D_1 aux composantes ambiances D_2 et D_3. L'expression des $RCPA_{1i}$ en décibels (dB) est donnée par :

$$\mathrm{RCPA}_{1i}[l] = 10 \times \log_{10}\left(\frac{\sum_{n=1}^{N} d_1^2[l]}{\sum_{n=1}^{N} d_i^2[l]}\right) \text{ dB}, \; i \in [2;3]. \tag{4.59}$$

Comme pour le cas de l'analyse bidimensionnelle, nous avons choisi d'utiliser la même longueur de fenêtre N pour réaliser l'ACP tridimensionnelle et pour estimer les RCPA_{1i} au cours du temps. Le RCPA_{1i} moyen au cours du temps est alors donné par :

$$\overline{\mathrm{RCPA}_{1i}} = \frac{1}{\lceil N_\mathrm{T}/N \rceil} \sum_{l=1}^{\lceil N_\mathrm{T}/N \rceil} \mathrm{RCPA}_{1i}[l] \text{ dB}, \; i \in [2;3]. \tag{4.60}$$

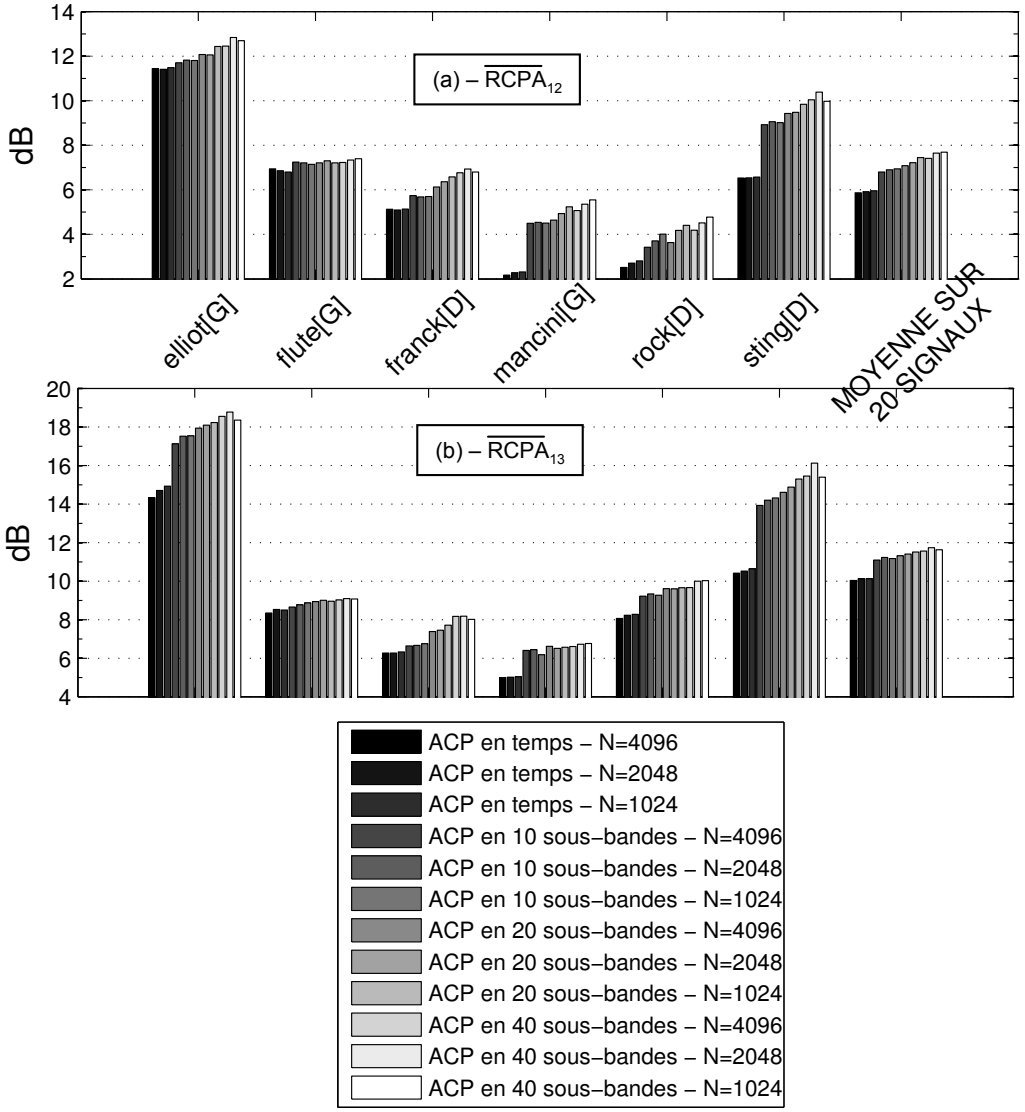

Figure 4.33 [(a) - RCPA_{12} et **(b) -** RCPA_{13}**]** moyens estimés à partir des signaux issus de l'ACP tridimensionnelle (en temps et en sous-bandes ave une échelle ERB) de 20 triplets de signaux au format 5.1 (images sonores gauche [G] et droite [D]). Les RCPA_{1i} issus de l'analyse de 6 triplets de signaux représentatifs sont présentés ainsi que la moyenne des RCPA_{1i} estimés à partir des 20 triplets.

La **Figure 4.33** présente une comparaison des RCPA_{1i} moyens, estimés à partir des signaux issus d'ACP tridimensionnelles (complètes avec trois rotations d'Euler) réalisées dans le

domaine temporel et des sous-bandes (K_b=10, 20 et 40 sous-bandes) avec des longueurs de fenêtre d'analyse telles que N=1024, 2048 et 4096 échantillons. Les vingt signaux analysés correspondent à des triplets de signaux provenant tous de scènes sonores au format 5.1 (*cf.* paragraphe 1.2.2) de types I et II (*cf.* paragraphe 4.1.2.2). La symétrie gauche droite du système de restitution sonore a été prise en compte puisque les triplets analysés sont constitués soit des canaux gauche, centre et arrière gauche (scène sonore latérale gauche) soit des canaux droit, centre et arrière droit (scène sonore latérale droite) soit C_1=C, C_2=L/R et C_3=Ls/Rs, *cf.* **Figure 4.34-(a)** *i.e.* séparation des canaux selon le plan médian.

De façon générale, d'après la **Figure 4.33**, la concentration de l'énergie dans la composante principale D_1 est plus grande avec une ACP en sous-bandes comparée à une ACP réalisée dans le domaine temporel. En moyenne sur 20 triplets de signaux analysés, la différence entre les $RCPA_{1i}$ des signaux issus d'une ACP en sous-bandes et les $RCPA_{1i}$ des signaux issus d'une ACP temporelle est de l'ordre de 2 dB. Cette différence s'accroît naturellement avec :

- la diminution de la longueur des fenêtres d'analyse dans le cas d'une ACP temporelle,
- la diminution de la longueur des fenêtres d'analyse et l'augmentation du nombre de sous-bandes dans le cas d'une ACP en sous-bandes.

De plus, le $RCPA_{12}$ est naturellement inférieur au $RCPA_{13}$ pour chaque triplet de canaux traité, la différence en moyenne est de l'ordre de 4 dB : le $RCPA_{12}$ est en moyenne de l'ordre de 7 dB et le $RCPA_{13}$ de l'ordre de 11 dB. Cette mesure traduit donc bien la hiérarchie énergétique des signaux issus de l'ACP tridimensionnelle. En outre, l'ordre de grandeur du $RCPA_{13}$ est équivalent à celui du $RCPA_{12}$ obtenu sur une autre base de signaux dans le cas bidimensionnel (*cf.* paragraphe 4.3.3.2). On a donc bien une correspondance énergétique des composantes principales D_1 et résiduelles D_2 ou D_3 issues d'une ACP bi- ou respectivement tridimensionnelle.

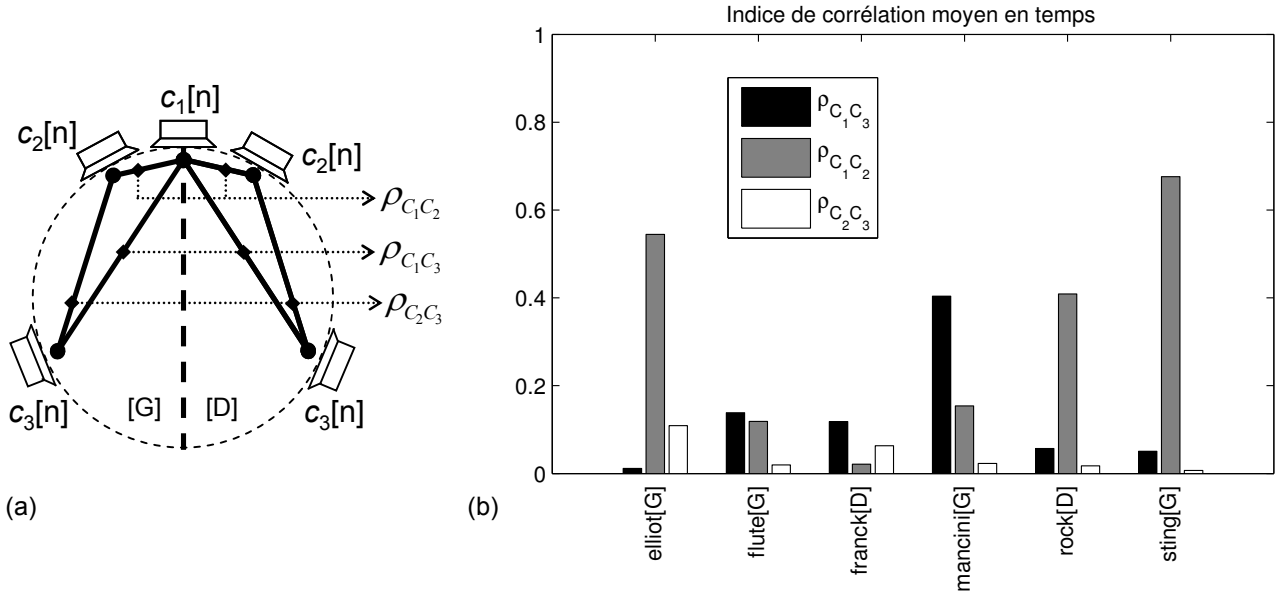

Figure 4.34 (a) – Séparation des canaux d'un signal 5.1 en deux triplets pour l'analyse. (b) - Indice de corrélation (moyen au cours du temps) estimé en temps {ρ_{C1C2}, ρ_{C1C3}, ρ_{C2C3}} à partir des canaux de six triplets de signaux au format 5.1.

De façon plus précise, on remarque sur la **Figure 4.33**, que les $RCPA_{1i}$ des signaux issus de l'ACP tridimensionnelle augmentent avec un traitement en sous-bandes notamment pour l'échantillon « sting[D] » de type II dont le contenu spectral est très riche et varié d'une sous-bande à une autre avec la présence de plusieurs instruments (guitare basse, percussions, guitare, trompette, etc.). A l'inverse le signal multicanal « flute[G] » de type I reproduit une

flûte dans un environnement réverbéré par conséquent les canaux contiennent une seule source directionnelle mélangée à une ambiance sonore relative à l'effet de salle (réverbération). Par conséquent, le traitement en sous-bandes ne permet pas d'améliorer la concentration de l'énergie pour ce signal au contenu spectral relativement pauvre.

Comme nous l'avons remarqué dans le cas de l'analyse bidimensionnelle (au paragraphe 4.3.3.2), les $RCPA_{1i}$ moyens peuvent être mis en correspondance avec les indices de corrélation des canaux originaux. La **Figure 4.34-(b)** présente les coéfficients de corrélation (en moyenne au cours du temps), estimés en temps à partir de l'équation (4.11), des canaux de six triplets dont les $RCPA_{1i}$ moyens sont présentés à la **Figure 4.33**. En règle générale, en comparant la **Figure 4.34-(b)** à la **Figure 4.33**, plus les indices de corrélation moyens sont grands (hypothèse de corrélation des canaux du modèle présenté au paragraphe 4.2.1) et plus l'ACP est performante en termes de concentration d'énergie.

A partir du même corpus de signaux multicanaux, la **Figure 4.35** présente la comparaison des pourcentages d'énergies moyens des canaux analysés (C_1, C_2, et C_3) et transformés (D_1, D_2 et D_3), au moyen de deux ou de trois rotations d'Euler, par rapport à l'énergie totale des canaux analysés.

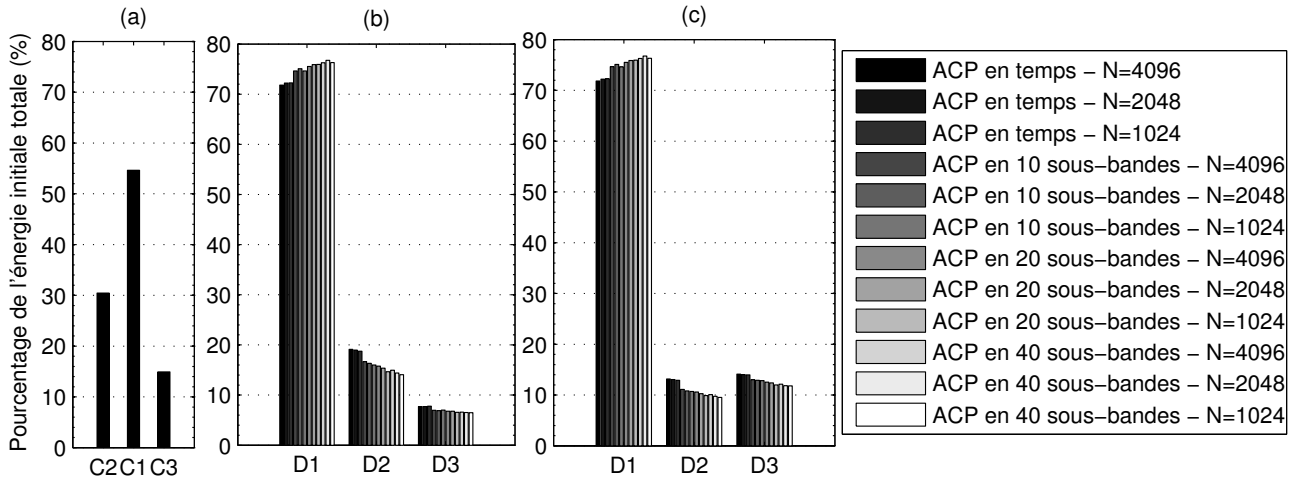

Figure 4.35 (a) - Pourcentages d'énergie moyens (sur le corpus) des signaux analysés C_2 (canal gauche ou droit), C_1 (canal central) et C_3 (canal arrière gauche ou arrière droit) par rapport à l'énergie totale initiale. Pourcentage de l'énergie initiale des signaux traités D_1, D_2 et D_3 obtenus par ACP avec **: [(b)** - trois rotations d'Euler, **(c)** - deux rotations d'Euler**]**.

La **Figure 4.35-(a)** présente les pourcentages énergétiques moyens, par rapport à l'énergie totale initiale, des canaux analysés (d'après la **Figure 4.34-(a)** C_1=C, C_2=L/R, C_3=Ls/Rs), tirés de vingt signaux audio au format 5.1 (corpus identique à celui utilisé pour l'estimation des $RCPA_{1i}$ présentés à la **Figure 4.33**). Ainsi, le canal central (C_1) représente à lui seul 54% de l'énergie diffusée à gauche ou à droite de l'auditeur lors d'une écoute sur un système de diffusion multicanal 5.0 (en faisant abstraction du canal basse fréquence). L'énergie du canal central a d'ailleurs été mesurée comme supérieure à l'énergie des canaux gauche ou droit frontaux (31%). De plus, il apparaît logiquement que l'énergie des canaux arrières ne représente qu'environ 15% de l'énergie diffusée à gauche ou à droite de l'auditeur puisque la plupart du temps, les haut-parleurs arrières délivrent l'ambiance de la scène sonore (naturellement moins énergétique que le son direct). La proportion de signaux de type I, au sein du corpus, est donc supérieure à celle des signaux de type II : le corpus choisi est donc représentatif de la nature des signaux audio multicanaux (format 5.1) au jour d'aujourd'hui.

Comme pour l'analyse des $RCPA_{1i}$, la **Figure 4.35** montre que les pourcentages énergétiques (de l'énergie initiale) des signaux issus de l'ACP tridimensionnelle sont fonction de la résolution temps-fréquence employée. Une ACP temporelle à faible résolution (fenêtre d'analyse à 4096 échantillons) génère une composante principale D_1 (identique avec deux ou trois rotations) qui représente, en moyenne sur les vingt composantes D_1 générée (pour

chaque triplet), 72% de l'énergie totale des signaux de départ. Avec une ACP réalisée en 10 sous-bandes de fréquences mais avec la même résolution temporelle, la composante principale D_1 représente alors environ 75% de l'énergie totale. De plus, en améliorant la résolution temps-fréquence (réduction de la taille de fenêtre et augmentation du nombre de sous-bandes) ce chiffre peut encore croître jusqu'à environ 78%. A l'inverse, les composantes ambiances ou résiduelles D_2 et D_3 ont un pourcentage énergétique qui décroit avec la précision de la résolution temps-fréquence. En effet, plus l'ACP est discriminante en fréquence et plus la concentration de l'énergie dans la composante principale est grande et par suite, plus l'énergie dans les composantes ambiances est faible. Enfin, l'ACP tridimensionnelle réalisée au moyen de trois rotations permet de hiérarchiser au mieux les signaux résultants de l'ACP alors qu'une ACP réalisée uniquement au moyen des deux premières rotations résulte en des signaux D_2 et D_3 qui ont un pourcentage de l'énergie initiale quasi-équivalent. L'information qui pourrait être mieux répartie avec trois rotations se trouve alors mélangées dans ces deux composantes.

Au final, l'ACP tridimensionnelle, présentée dans cette section, constitue une approche paramétrique pour générer un signal dominant ou composante principale accompagnée d'angles de rotation qui caractérisent les positions spatiales des sources dominantes à chaque instant. Comme dans le cas de l'ACP bidimensionnelle, l'ACP en sous-bandes exploite la corrélation des canaux en sous-bandes de fréquences pour compacter davantage l'énergie initiale dans le signal dominant et cela d'autant plus que le nombre de sous-bandes est élevé. En outre, l'utilisation de la troisième rotation est discutable suivant le contexte d'utilisation de cette méthode. En effet, en utilisant un traitement en sous-bandes de fréquences, la différence entre les composantes générées par une ACP avec deux ou trois rotation est fortement réduite (d'un pointe de vue énergétique et perceptif).

4.4 Conclusion

Dans ce chapitre, nous avons, en premier lieu, rappelé les méthodes de génération des signaux audio multicanaux de façon à les catégoriser par leur contenu, souvent lié à leur contexte d'utilisation.

A partir de la description des composantes des signaux multicanaux, nous avons dérivé, en second lieu, un modèle de mélange instantané de ces composantes : les sources directionnelles et les ambiances. Une analyse en temps et en sous-bandes de fréquences de la distribution des valeur propres de signaux stéréo suivant ce modèle a ensuite été menée : elle nous a notamment permis de définir la répartition des composantes originales au sein des valeurs propres. Nous avons relevé l'influence de la position des sources (azimut dans le plan horizontal) et du pouvoir discriminant de l'analyse en sous-bandes pour la hiérarchisation des composantes originales.

Etant donné que la distribution des valeurs propres caractérise la puissance des signaux projetés sur la base des vecteurs propres correspondants, nous avons ensuite considéré l'ACP de signaux à deux et trois composantes suivant notre modèle. Nous avons tout d'abord justifié le choix d'une approche paramétrique pour réaliser l'ACP dans un contexte de codage audio multicanal. Nous avons ensuite proposé une méthode pour extraire un ou plusieurs angles de rotation utiles à l'ACP, méthode compatible avec un schéma de codage classique utilisant une analyse/synthèse par recouvrement-addition (méthode OLA). Une interprétation physique des angles de rotation a été présentée d'après la connaissance du système de reproduction sonore : le ou les angles de rotation peuvent être convertis en un azimut lié à la position de la source directionnelle dominante contenue dans le signal multicanal analysé. Enfin, nous avons évalué les performances de l'ACP bi- et tri-dimensionnelle en termes de concentration de l'énergie initiale au sein des composantes transformées. Il est ressort notamment que la nature (nombre de sources par exemple), la corrélation des canaux ainsi que le type de traitement (en temps ou en sous-bandes de fréquences) influencent considérablement la hiérarchisation des composantes au sein des canaux transformés.

Bien que nous nous soyons positionnés dans un contexte de codage audio, l'ACP bidimensionnelle est notamment utilisée dans un contexte de conversion entre formats dit *upmix* décrit en Annexe C.2. D'après notre analyse, la limite donnée à la hiérarchisation des composantes permet de définir précisément le contenu des signaux qui seront reproduit par le système multi-haut-parleurs, notamment pour les canaux arrières. En outre, l'ACP tridimensionnelle, présentée dans ce chapitre, pourrait par exemple être utilisée pour extraire un ou deux canaux d'ambiance à partir d'une scène sonore frontale (canaux avant gauche, central et droit).

4.5 Perspectives : mélange convolutif et ACI

Le modèle de signaux, décrit au paragraphe 4.2, est dit instantané dans la mesure où les sources sont simplement pondérées par des fonctions liées à la position des sources dans l'espace. Autrement dit, les sources sont considérées comme captées par des microphones au même instant mais avec des niveaux d'énergie différents. De plus, les sources sonores secondaires, d'un point de vue perceptuel et énergétique, et l'effet de salle qui s'applique aux sources directionnelles sont considérés comme appartenant aux signaux d'ambiances. L'expression du modèle de mélange instantané, décrit par l'équation (4.5), diffère des modèles plus évolués dits « convolutifs ». Par exemple, le modèle utilisé par Asano et *al.* dans [ASA00] considère des sources directionnelles convoluées avec des réponses impulsionnelles de salle. Ces modèles sont naturellement homogènes entre eux puisque l'effet de salle sur les sources directionnelles a été considéré, par le modèle de l'équation (4.5), comme distribué sur au moins une partie des canaux au travers des ambiances sonores.

4.5.1 Mélange convolutif de sources directionnelles et d'ambiances

Les problématiques de séparation aveugle de sources (SAS ou BSS pour *Blind Source Separation*) considèrent les signaux audio multicanaux avec des modèles de mélange plus ou moins évolués en fonction des conditions environnementales ou expérimentales. L'étude récapitulative des méthodes de SAS menée par O'Grady et *al.* dans [OGR05] fait référence aux modèles de mélange instantané, convolutif anéchoïque et convolutif échoïque.

4.5.1.1 Mélange convolutif anéchoïque de sources et d'ambiances

Le modèle de mélange convolutif anéchoïque constitue une extension du modèle de mélange instantané dans la mesure où les délais des sources captées par les microphones sont pris en compte. Autrement dit, un modèle de mélange instantané qui prend en compte les retards inter-canaux des sources peut être qualifié de convolutif anéchoïque [OGR05]. Par conséquent, l'équation (4.5) du modèle de mélange instantané devient dans le cas d'un modèle de mélange convolutif anéchoïque :

$$c_m[n] = \sum_{d=1}^{D} \left[g_{md}[n] \cdot s_d[n - \delta_{md}] \right] + a_m[n], \tag{4.61}$$

où δ_{md} correspond au temps de propagation de la source d jusqu'au microphone m.

4.5.1.2 Mélange convolutif échoïque de sources et d'ambiances

Une extension du mélange convolutif anéchoïque consiste à prendre en compte, en outre, les réflexions du son direct c'est-à-dire à considérer la source sonore comme provenant de plusieurs directions [OGR05]. Autrement dit, chaque microphone qui capte le son direct

d'une source capte également ses réflexions sur l'espace environnant. Cette hypothèse supplémentaire qualifie alors le mélange de convolutif échoïque. Dans ce cas, l'équation (4.5), devient :

$$c_m[n] = \sum_{r=1}^{R} \sum_{d=1}^{D} \left[g_{md}^{r}[n] \cdot s_d\left[n - \delta_{md}^{r} \right] \right] + a_m[n], \tag{4.62}$$

où, r correspond à l'indice des réflexions qui peuvent atteindre un nombre maximal de R réflexions.

Un tel modèle de mélange appliqué aux signaux audio multicanaux, décrits au paragraphe 4.1, considère les signaux d'ambiances comme dépourvu des réflexions (réverbération) des sources directionnelles. Les ambiances sonores peuvent alors être considérées comme relatives aux sources sonores secondaires de l'environnement sonore (musique de fond, applaudissements par exemples) et non plus à l'effet de salle.

4.5.2 Analyse en Composante Indépendante

4.5.2.1 L'ACI pour la séparation de sources

Les problématiques de SAS, étudiées depuis une vingtaine d'années, couvrent l'analyse de signaux acoustiques, médicaux (électroencéphalogrammes par exemple), de données financières et bien d'autres encore. La première approche fut proposée par Hérault et Jutten, dans [HER86], avec l'intention de séparer des mélanges linéaires instantanés sous-déterminés de sources indépendantes et non-gaussiennes. La solution consiste à utiliser un réseau de neurones artificiel pour séparer les sources inconnues avec l'hypothèse forte que ces sources sont indépendantes. Linsker (1989) mis en place des règles d'apprentissages basées sur la théorie de l'information qui maximisent l'information mutuelle moyenne entre les observations et les sources estimées avec un réseau de neurones artificiel.

En se basant sur le fait que l'information mutuelle est la mesure la plus naturelle de l'indépendance, Comon, dans [COM94], démontra que maximiser la non-gaussianité des sources est équivalent à minimiser l'information mutuelle entre elles. Ainsi, le concept de la séparation de sources en maximisant l'indépendance *i.e.* Analyse en Composante Indépendante (ACI), fut mis en place. Bell et Sejnowski, dans [BEL95], ont développé un algorithme de SAS, appelé *Infomax*, qui est similaire aux propositions de Linsker en utilisant toutefois une règle basée sur un gradient stochastique. Le concept de la non-gaussianité des sources a été utilisé par Hyvärinen et Oja avec l'algorithme *fastICA* [HYV97]. Une alternative à la séparation de sources basée sur l'information mutuelle mais qui repose sur l'estimation du maximum de vraisemblance a été proposée par Gaeta et Lacoume (1990) puis élaborée par la suite par Pham et *al.* (1992) bien que Cardoso démontra plus tard, dans [CAR97], que l'algorithme *Infomax* et l'estimation du maximum de vraisemblance constituent des approches équivalentes. Une solution à la séparation de mélanges sous-déterminés fut proposée par Belouchrani et Cardoso dans [BEL94] avec une approche basée sur la probabilité du maximum à posteriori. Enfin, la séparation de mélanges convolutifs anéchoïques a été abordée par l'algorithme DUET (*Degenerate Unmixing Estimation Technique*), de Jourjine et *al.* dans [JOU00], avec l'hypothèse que les supports fréquentiels des sources originales ne se recouvrent pas.

4.5.2.2 L'ACI dans un contexte de codage audio

Le mélange décrit par l'équation (4.5) a été utilisé au paragraphe 4.3 pour regrouper les sources directionnelles $\mathbf{S_D} = \left(S_1, \cdots, S_d, \cdots, S_D\right)^T$ et une partie de l'ambiance sonore $\mathbf{A_M} = \left(A_1, \cdots, A_m, \cdots, A_M\right)^T$ en une seule composante au moyen de l'ACP réalisée par rotation en sous-bandes de fréquences. Loin d'une séparation des sources directionnelles entre elles, cette approche constitue plutôt une solution possible aux problématiques de SAS de type *cocktail party* qui cherchent à réduire les effets gênants (pour l'intelligibilité par exemple) d'un environnement bruité ou à extraire un signal de parole parmi une multitude de signaux de parole parasites (d'où le nom).

L'intérêt de l'ACP dans un contexte de codage audio multicanal a été justifié au paragraphe 4.3.2. Dans ce contexte de codage, plutôt que séparer individuellement des sources qui devront être toutes transmises, il apparaît plus intéressant de regrouper ces sources dans un même canal qui doit être prioritairement transmis.

La question qui vient alors naturellement à l'esprit est la suivante: les modèles convolutifs d'ACI et les techniques avancées utilisées en SAS peuvent-ils permettre une meilleure séparation des sources directionnelles et des ambiances sonores que ne le permet l'ACP?

Jusqu'alors, aucune des solutions proposées par les méthodes de SAS basées sur l'ACI, énumérées au paragraphe 4.5.2.1, ne prennent en compte la non-stationnarité des signaux au cours du temps. Par non-stationnarité, nous entendons ici l'évolution de la position spatiale des sources au cours du temps. En effet, les positions des sources sont habituellement considérées comme fixes par rapport aux microphones. Ainsi les algorithmes d'ACI existants s'appliquent à l'intégralité d'un signal dans le domaine temporel ou fréquentiel selon le modèle et la méthode de séparation employée. Par suite, la résolution de ce problème doit logiquement passer par un schéma d'analyse qui segmente le signal selon des portions temporelles. Par exemple, un signal de parole est admis comme stationnaire selon des trames de longueur moyenne égale à 20 ms, qui correspond d'ailleurs à la taille de trame classiquement utilisée en codage de parole ou plus généralement en codage audio. Ainsi, un algorithme d'ACI devra être appliqué pour chacune de ces trames.

Les premières expérimentations menées à l'aide des algorithmes *fastICA* et *Jade* (de Jean-François Cardoso) ont montré des résultats sensiblement différents compte tenu de la nature des signaux traités. En effet, à partir d'un signal stéréo dont les canaux délivrent des sources (signaux de parole) aux positions spatiales fixes, la séparation s'applique à un cas favorable et fonctionne bien. Une autre expérience a été menée à partir d'un signal stéréo également synthétique et constitué du mélange de signaux de parole dont les gains de pondération respectifs évoluent au cours du temps *i.e.* les positions spatiales des sources perçues évoluent au cours du temps. Dans ce cas, la séparation ne fonctionne pas même en appliquant ces algorithmes d'ACI aux portions du signal découpé au cours du temps. Pour assurer une bonne séparation il faudrait d'une part considérer des portions de signal d'une longueur supérieure (de l'ordre de la seconde au minimum) et d'autre part s'assurer que les positions des sources n'évoluent pas trop vite au cours du temps. Dans ce cas, l'évolution au cours du temps de la matrice de séparation pourrait être contrôlée pour éviter les permutations (l'ordonnancement des sources extraites n'est pas constant) inhérentes aux méthodes de séparation de sources [OGR05]. Des solutions à ce problème de permutation ont été notamment proposées par un certain nombre de méthodes œuvrant dans le domaine fréquentiel. Il conviendrait donc d'adapter ces méthodes à la résolution du problème de permutation des sources extraites au cours du temps.

En conclusion, l'état actuel des performances obtenues par les méthodes de SAS appliquées aux signaux audio multicanaux font que cette voie constitue une perspective de recherche dans notre contexte d'étude. En effet, l'objectif de regouper l'information dominante (d'un point de vue énergétique et perceptif), atteint par l'ACP en sous-bandes, diffère des objectifs de SAS. Pour parvenir à cet objectif, les hypothèses et modèles employés par la SAS doivent être étendus en considérant non seulement un mélange

échoïque de sources mais un tel mélange sommé à une ambiance sonore (*cf.* équation (4.62)). Même si certaines méthodes de SAS considèrent des modèles avec un bruit additif au mélange convolutif de sources, les objectifs à atteindre sont différents puisque l'ambiance est une information utile dans notre contexte *i.e.* par excès : une multitude de sources secondaires dans le cas des applaudissements par exemple, à la différence du bruit. Enfin, le modèle convolutif anéchoïque constitue une extension directe du mélange instantané utilisé par notre méthode paramétrique pour réaliser l'ACP. Ainsi ce modèle de mélange pourrait être pris en considération pour permettre une représentation efficace des signaux audio multicanaux susceptibles de contenir des sources directionnelles aux positions spatiales estimées par différences d'intensité et/ou retards inter-canaux.

5. Codage paramétrique basé sur l'Analyse en Composante Principale

Les propriétés de l'Analyse en Composante Principale (ACP), décrites au paragraphe 4.3.1, font que cette transformation, qualifiée de transformation optimale dans le cas de signaux stationnaires, est naturellement utilisée dans une multitude de contextes. Les applications touchent aussi bien au domaine de la compression audio et vidéo qu'au domaine de la reconnaissance de formes (objet 3D) ou encore de l'analyse de données multidimensionnelles telles que les données financières.

En matière de codage audio, l'ACP est utilisée pour le codage stéréophonique dit d'intensité (*cf.* paragraphe 2.2.2) et le codage audio multicanal (*cf.* paragrpahe 2.3.2.2) pour sa capacité à concentrer l'énergie des signaux en un nombre de canaux réduit. Ainsi, l'allocation binaire pour le codage des signaux issus de l'ACP est établi comparativement à l'énergie moyenne de chaque signal [YAN04].

De manière à proposer une solution à la transmission des signaux audio multicanaux (stéréo, 5.1, etc.) pour les applications à débit contraint (communications mobiles, par internet, etc.), les paragraphes suivants présentent une nouvelle méthode de codage audio multicanal à bas débit. L'intérêt de l'utilisation de l'ACP dans ce contexte ayant déjà été présenté au paragraphe 4.3.2, ce chapitre s'attache à décrire la mise au point de la méthode de compression qui consiste à la fois à utiliser l'ACP, pour compacter au maximum l'énergie initiale, et à réaliser un codage paramétrique des signaux issus de l'ACP des canaux originaux.

Nous présentons au paragraphe 5.1 le principe et l'implémentation de notre méthode de codage audio paramétrique des signaux stéréophoniques. L'identification, l'extraction et le codage des paramètres utiles à la méthode de compression sont présentés. Nous présentons également les résultats de l'évaluation subjective de cette implémentation de la méthode de codage stéréo. Enfin, nous proposons une discussion qui traite des avantages et des inconvénients de cette méthode comparée aux méthodes de codage audio paramétrique existantes. Nous présentons ensuite, au paragraphe 5.2, une extension de cette méthode de codage pour la compression des signaux audio multicanaux notamment au format 5.1. Le principe de la méthode basée sur l'association de plusieurs modules d'ACP bi- et/ou tridimensionnelle est donnée tout en conservant la compatibilité avec les systèmes de reproduction mono et stéréo.

5.1 Codage paramétrique des signaux stéréophoniques

5.1.1 Principe de la méthode

Le principe de la méthode de codage paramétrique présentée à la **Figure 5.1** repose sur la représentation compacte des données obtenues par une ACP bidimensionnelle réalisée en sous-bandes de fréquences (*cf.* paragraphe 4.3.3). Basée sur cette représentation, la méthode consiste, d'une part, à compresser la composante principale D_1 par un codeur audio monophonique (perceptuel par transformée par exemple) et, d'autre part, à coder les angles de rotation relatifs à l'ACP en sous-bandes ainsi que les paramètres utiles à la synthèse de la composante résiduelle ou d'ambiance D_2 (au décodage présenté à la **Figure 5.2**) [BRI06b].

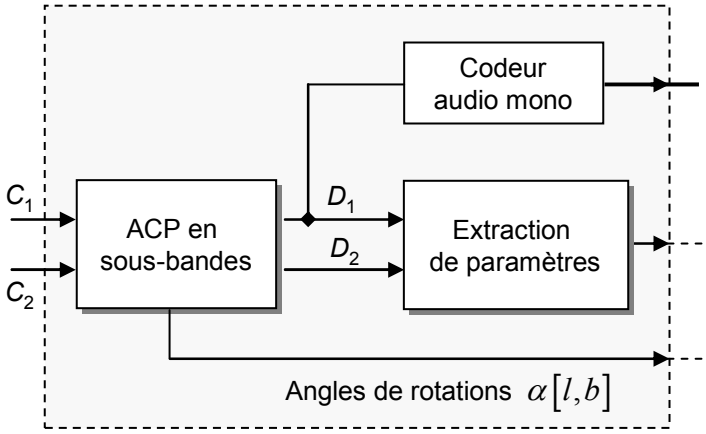

Figure 5.1 Schéma de principe de l'encodeur stéréo paramétrique basé sur l'ACP réalisée en sous-bandes de fréquences.

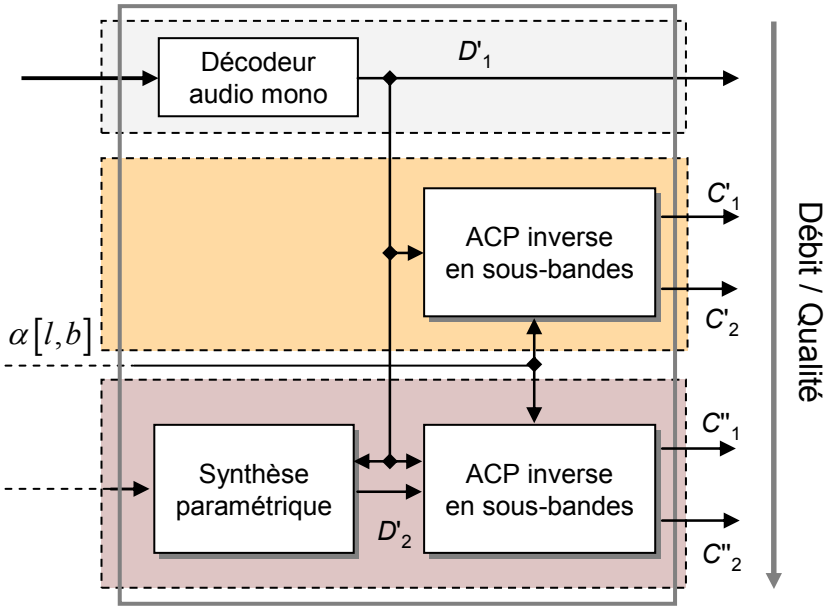

Figure 5.2 Schéma de principe du décodeur stéréo paramétrique basé sur l'ACP réalisée en sous-bande de fréquences.

La méthode de codage audio paramétrique, proposée dans [BRI06b], assure la compatibilité avec les systèmes de reproduction monophonique en restituant la composante principale décodée (D'_1) lors de l'opération de décodage présentée à la **Figure 5.2**. Avec un débit légèrement supérieur, relatif à la transmission des angles de rotation en sous-bandes, le décodeur peut générer, par ACP inverse en sous-bandes, un signal stéréophonique (C'_1,C'_2) de qualité intermédiaire : les sources directionnelles seront correctement repositionnées mais l'ambiance ne sera pas totalement restituée. Finalement, le décodeur assurera une meilleure reconstruction subjective de la scène avec le signal stéréo (C''_1,C''_2) obtenu à partir de la composante principale décodée, des angles de rotation et des paramètres utiles à la synthèse de l'ambiance sonore (D'_2). Par conséquent, une part importante de la résolution du problème ainsi considéré repose donc sur l'extraction de paramètres subjectivement pertinents pour la synthèse de la composante ambiance.

5.1.1.1 Paramètres subjectivement pertinents pour la synthèse de l'ambiance

La méthode de codage audio paramétrique ainsi définie traite les signaux dans le domaine des sous-bandes, autrement dit dans le domaine fréquentiel. Le décodeur doit régénérer l'enveloppe spectrale de la composante ambiance au moyen de la synthèse paramétrique. La question qui se pose alors naturellement est la suivante : la phase de la composante ambiance D_2, constituée principalement des ambiances originales décorrélées (*cf.* paragraphe 4.2.2.3), a-t-elle une importance d'un point de vue subjectif?

Pour répondre à cette question, un test subjectif d'évaluation par catégories de comparaison (*Comparison Category Rating* CCR), décrit dans la recommandation P.800 de l'UIT-T [UIT800], a été mené. L'objet du test est de comparer subjectivement des signaux stéréophoniques non traités (non dégradés) à leur version traitée telle que les canaux originaux subissent une ACP et une ACP inverse (en sous-bandes) à partir de la composante principale D_1 et de la composante ambiance D_2 filtrée par un filtre passe-tout à phase aléatoire (*cf.* **Figure 5.3**). Rappelons que l'ACP est une transformation orthogonale linéaire et donc inversible par simple transposition (*cf.* équation (4.38) pour réaliser l'ACP bidimensionnelle inverse). Par conséquent, le traitement décrit à la **Figure 5.3** serait objectivement et subjectivement transparent sans l'opération de filtrage. En d'autres termes, cette expérimentation permet d'évaluer l'influence du filtrage passe-tout décorrélateur sur la phase de la composante ambiance issue de l'ACP.

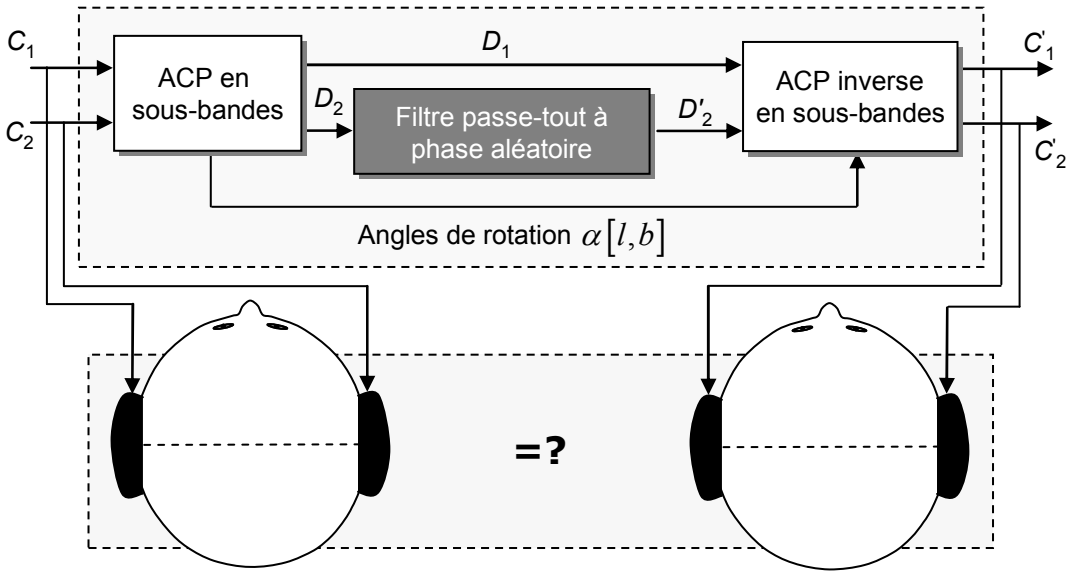

Figure 5.3 Schéma de principe du traitement appliqué aux signaux stéréo utilisés pour réaliser un test subjectif de comparaison de type CCR.

L'utilisation de filtres de décorrélation à réponse impulsionnelle infinie (introduction d'un décalage temporel constant) est décrite en Annexe C.2.3 dans un contexte d'*upmix* d'un signal stéréo en un signal 5.0. D'une manière générale, un filtre passe-tout $H(\omega)$ à phase aléatoire $\varphi(\omega)$ est défini, pour chaque valeur de la pulsation $\omega=2\pi f$, par :

$$H(\omega) = A(\omega) \cdot e^{j\varphi(\omega)} \quad \text{avec :} \begin{cases} A(\omega) = 1 \\ \varphi(\omega) \in [-\pi; \pi] \end{cases}. \tag{5.1}$$

Un tel filtre est qualifié de filtre de décorrélation par Kendall dans [KEN95b] et est habituellement utilisé pour créer une sensation d'élargissement de la scène sonore perçue.

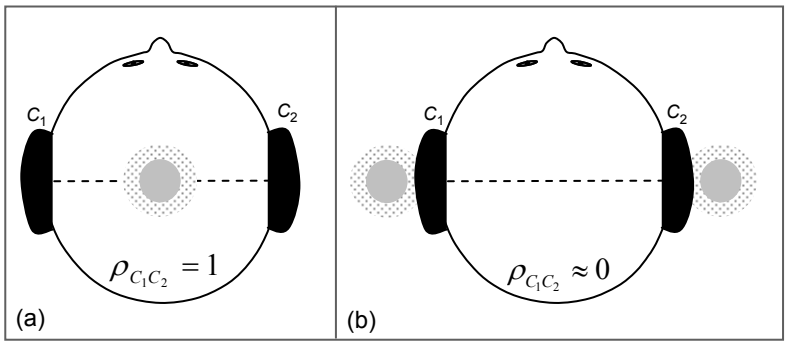

Figure 5.4 Internalisation et externalisation de l'image sonore perçue (écoute au casque) à partir de signaux respectivement (a) – corrélés, et (b) – décorrélés, d'après [KEN95b].

Comme l'illustre la **Figure 5.4**, l'écoute au casque d'un signal stéréo dont les canaux sont décorrélés ($\rho_{C_1C_2} \approx 0$) implique la sensation d'externalisation (les sons proviennent de l'extérieur de la tête) alors que l'écoute d'un signal stéréo aux canaux corrélés ($\rho_{C_1C_2} \approx 1$) implique une perception des sons comme provenant de l'intérieur de la tête. La définition de l'indice de corrélation des canaux d'un signal stéréo $\rho_{C_1C_2}$, ou ICC, est donnée au paragraphe 2.2.3.1. Cet indice peut être comparé à l'IACC qui caractérise la corrélation des signaux atteignant les oreilles de l'auditeur *i.e.* les signaux binauraux (*cf.* paragraphe 1.1.2.2).

Dans le cadre de notre expérimentation, les composantes issues de l'ACP en sous-bandes et plus particulièrement la composante ambiance D_2 a été filtrée par un filtre passe-tout à phase aléatoire défini dans le domaine fréquentiel par l'équation (5.1). La réponse impulsionnelle du filtre $h[n]$ a été établie, comme indiquée par Kendall dans [KEN95b], par transformée de Fourier inverse de la réponse initialement définie dans le domaine fréquentiel $H(\omega)$. Finalement, les composantes ambiances D_2, issues des ACP bidimensionnelles des signaux stéréo considérés pour notre expérimentation, ont été convoluées par des réponses impulsionnelles différentes ($h_i[n]$). Nous avons négligé l'influence sur notre perception du choix de la réponse impulsionnelle du filtre en faisant l'hypothèse que la perception d'un même signal audio $d_2[n]$ convolué par deux réponses impulsionnelles ($h_1[n]$ et $h_2[n]$) aux phases aléatoires différentes ($\varphi_1(\omega)$ et $\varphi_2(\omega)$) n'est pas significativement différente.

Les quinze sujets qui ont participé au test, dont le principe est présenté à la **Figure 5.3**, ont été tenus de comparer subjectivement à la fois la qualité audio et l'image stéréophonique (spatialisation, effet de salle) des signaux stéréo (C_1,C_2) et (C'_1,C'_2). Le test de comparaison CCR a été réalisé à partir de quatre *stimuli* ou échantillons de musique stéréo. Les signaux choisis sont critiques pour cette expérimentation dans la mesure où ils sont tous caractérisables par un effet stéréo (spatialisation) marqué. En outre, certains échantillons présentent une ambiance sonore perceptuellement significative en termes de largeur d'image stéréo avec une forte réverbération par exemple. Un des *stimuli* provient du *downmix* (les coefficients définis par l'UIT ont été utilisés, *cf.* Annexe C.1) d'un signal 5.1 de type II (*cf.*

paragraphe 4.1.2.2) et procure, par conséquent, une large image stéréo (à l'écoute au casque) avec de nombreuses sources directionnelles positionnées dans cet espace. Pour chaque extrait sonore, les sujets ont écouté les signaux (C_1, C_2) et (C'_1, C'_2) dans un ordre aléatoire et ont donné leur préférence entre les *stimuli* au moyen de l'échelle de notations définie par la recommandation P.800 de l'UIT-T [UIT800]. Les résultats du test de comparaison CCR sont présentés à la **Figure 5.5**.

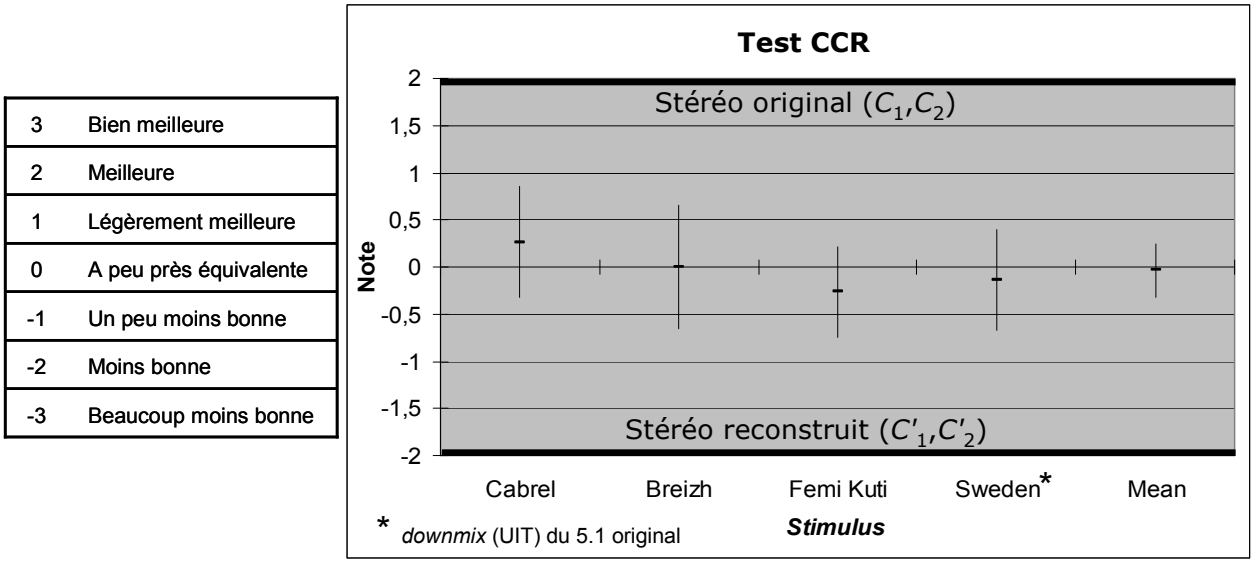

Figure 5.5 Moyenne des notes données par les quinze sujets (par stimulus et en moyenne sur tous les *stimuli* à droite) et intervalles de confiance à 95%.

D'après la **Figure 5.5**, les notes moyennes sur tous les sujets attribuées à chaque échantillon stéréo sont très proches de zéro. En moyenne sur tous les items et sur tous les sujets, la valeur moyenne est de -0,033 ce qui signifie qu'en moyenne les signaux stéréo obtenus à partir de la composante ambiance à phase aléatoire ont été très légèrement préférés. De plus, tous les intervalles de confiance passent par la valeur zéro ce qui traduit bien l'incapacité des sujets à établir une nette préférence subjective entre les signaux jugés. Globalement, les signaux stéréo originaux et reconstruits ne sont pas statistiquement différents d'un point de vue subjectif. Une analyse plus précise peut être menée en considérant l'impression spatiale délivrée par les échantillons stéréo testés.

D'après la **Figure 5.5**, la version originale du *stimulus* intitulée « Cabrel » a été légèrement préférée, en moyenne sur tous les sujets, comparé à la version traitée. Ce résultat se justifie parfaitement étant donné que l'impression spatiale délivrée par ce signal stéréo est relativement pauvre : les sources directionnelles ont des azimuts différents mais les canaux ne présentent pas ou très peu d'effet de salle (pas de réverbération). Par conséquent, la composante ambiance D_2, issue de l'ACP en sous-bandes, est relative aux sources directionnelles secondaires qui n'occupent pas la même position spatiale que les sources directionnelles dominantes extraites dans D_1. Finalement, le filtre passe-tout décorrélateur appliqué à la composante D_2 résulte en une légère coloration des sources directionnelles (modification du timbre de certains instruments par effet de filtrage passe-tout décrit en Annexe C.2.3) et donc en une dégradation sensiblement perceptible.

A l'inverse, la moyenne des scores attribués aux *stimuli* intitulés « Femi Kuti » et « Sweden » est légèrement inférieure à zéro *i.e.* le signal stéréo traité a été légèrement préféré. Ce résultat s'explique par le fait que ces signaux stéréo sont pourvus d'un effet de salle qui procure une large image stéréo agrémentée de nombreuses sources directionnelles spatialisées à des positions azimutales différentes (azimuts opposés aux positions extrêmes notamment). Par conséquent, la composante ambiance D_2, issue de l'ACP en sous-bandes, est relative à la fois aux sources directionnelles secondaires et aux composantes réverbérées des

sources directionnelles (effet de salle sur les sources directionnelles) ainsi qu'aux sources secondaires qui participent à l'effet de salle (bruit environnant non localisable par exemple). Dans ce cas, le filtre de décorrélation appliqué à la composante D_2 résulte, à l'issue de l'ACP inverse, en un signal stéréo dont la coloration des sources directionnelles (modification de timbre introduite par le filtrage) est rendue moins audible par la présence de l'effet de salle : l'effet de salle recouvre l'énergie décroissante des sources dans le temps (*cf.* paragraphe 1.1.2). Finalement, l'effet du filtre de décorrélation sur ce type de signaux stéréo réside dans la perception de l'effet de salle qui peut générer une largeur d'image stéréo légèrement supérieure à celle de la scène originale sans pour autant introduire de dégradations du timbre des sources directionnelles.

En conclusion, les résultats de ce test subjectif donnent une tendance sur l'importance perceptive de la phase de la composante D_2 issue de l'ACP bidimensionnelle réalisée en sous-bandes de fréquences. La phase de cette composante n'est que faiblement perceptible, du moins, la perturbation de cette information n'entraîne pas une dégradation gênante et cela d'autant plus que le signal stéréo considéré présente une ambiance originale pourvue d'un effet de salle. En effet, l'ambiance du signal stéréo original est par définition décorrélée d'un canal à un autre (*cf.* paragraphe 4.2.1). Autrement dit, notre perception de l'ambiance diffuse correspond à un processus de localisation auditive ambiguë qui ne s'attache à aucune direction particulière. Par conséquent, l'information de phase d'un tel contenu n'est pas un paramètre essentiel d'un point de vue subjectif.

Finalement, l'extraction des paramètres subjectivement pertinents pour la synthèse paramétrique de la composante ambiance (*cf.* **Figure 5.2**) peut se limiter à des paramètres relatifs à l'enveloppe spectrale de l'ambiance. La solution proposée à l'encodage consiste à extraire le niveau d'énergie de la composante ambiance $E_{D_2}[l,b]$ en sous-bandes ou le rapport d'énergie entre la composante principale et la composante ambiance, c'est-à-dire le $RCPA_{12}[l,b]$ défini au paragraphe 4.3.3.2. Ainsi pour chaque trame d'indice l et chaque sous-bande d'indice b, l'encodeur doit extraire un paramètre énergétique qui permettra au décodeur de régénérer l'enveloppe spectrale de la composante ambiance à partir de la composante principale décodée. Les composantes D'_1 et D'_2 générées au décodage (*cf.* **Figure 5.2**) par synthèse paramétrique dans le domaine des sous-bandes seront corrélées alors qu'à l'origine ces composantes sont décorrélées, puisque issues de l'ACP en sous-bandes (voir les propriétés de l'ACP présentées au paragraphe 4.3.1).

5.1.1.2 Synthèse paramétrique et filtrage décorrélateur

Outre la synthèse paramétrique, le procédé de décodage doit assurer une très faible corrélation des composantes qui seront utilisées pour réaliser l'ACP inverse en sous-bandes (*cf.* **Figure 5.2**). Pour cela, un filtre passe-tout décorrélateur, comme celui décrit par l'équation (5.1), doit être appliqué à la composante ambiance D'_2 issue de la synthèse paramétrique, comme indiqué par la **Figure 5.6**.

Le schéma de décodage présenté à la **Figure 5.6** propose de synthétiser une ambiance artificielle D''_2 qui permet de réaliser au mieux l'ACP inverse en sous-bandes de fréquences. Cette synthèse est rendue possible à partir de la composante principale décodée D'_1 et des paramètres énergétiques transmis ($RCPA_{12}[l,b]$ ou $E_{D_2}[l,b]$). Une fois l'enveloppe spectrale de l'ambiance appliquée à D'_2, la méthode proposée consiste à décorréler ce signal de la composante principale D'_1. Comme nous l'avons évoqué au paragraphe 2.2.3.2 et en Annexe C.2.3, plusieurs méthodes de décorrélation sont disponibles : filtrage passe-tout, filtre de réverbération tardive.

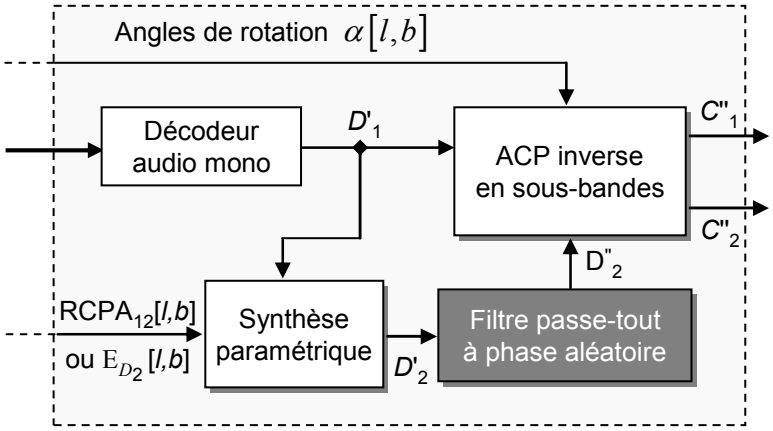

Figure 5.6 Décodeur stéréo paramétrique basé sur l'ACP en sous-bandes. Les paramètres énergétiques (rapport d'énergie RCPA[*l*,*b*] ou énergie $E_{D_2}[l,b]$ de la composante ambiance en sous-bandes) sont appliqués à la composante principale décodée D'_1 pour chaque sous-bande de fréquences. La composante ambiance D'_2, mise en forme spectralement, est filtrée par un filtre passe-tout décorrélateur de façon à réduire l'inter-corrélation des composantes D''_2 de D'_1.

Le choix du type de filtre décorrélateur influence naturellement la forme d'onde du signal d'ambiance synthétisé et par suite la qualité audio du signal stéréo issu de l'ACP inverse. Les codeurs audio sont évalués sur la qualité audio des signaux décodés et plus particulièrement sur la fidélité du rendu sonore comparé au rendu du signal original. Si le signal original à coder présente des sources directionnelles fortement réverbérées (colorées par l'effet de salle), le signal D_2 extrait à l'encodeur sera alors constitué des composantes réverbérées (par définition décorrélées d'après le modèle introduit au paragraphe 4.2.1) des canaux du signal original. En conséquence, plutôt que de réaliser un filtrage passe-tout décorrélateur de la composante principale (faiblement réverbérée puisqu'essentiellement constituée des sources directionnelles), le décodeur devrait utiliser un filtrage de type réverbération artificielle. Cependant, cet emploi n'est pas toujours justifié notamment lorsque le signal à coder est dépourvu d'effet de salle (réverbération). Par conséquent, si l'analyse faite par l'encodeur est capable de déceler la présence de réverbération dans le signal original alors un paramètre de type indice de réverbération pourrait être transmis au décodeur pour lui permettre de choisir le type de filtre le plus adapté à la nature du signal. Etant donné la difficulté de cette tâche (*cf.* paragraphe 5.3.2.1), nous avons opté pour l'utilisation d'un filtre de décorrélation de type passe-tout à phase aléatoire comme celui décrit dans [KEN95b] et utilisé au paragraphe 5.1.1.1. De plus, nous n'avons pas retenu la méthode de décorrélation par filtrage passe-tout à phase non-linéaire, décrite au paragraphe 2.2.3.2, du fait de la coloration (synthétique) ajoutée pour les hautes fréquences.

5.1.2 Implémentation de la méthode de codage stéréo paramétrique

Nous présentons dans cette section notre implémentation de la méthode de codage stéréo paramétrique, basée sur l'ACP en sous-bandes, présentée au paragraphe 5.1.1. Tout d'abord, nous rappelons brièvement les principales opérations affectées à l'encodeur et au décodeur. Ensuite nous nous focalisons, d'une part, sur la mise en place de la quantification des paramètres spatiaux et énergétiques (la transmission de la structure fine de l'ambiance n'est pas prise en compte). D'autre part, nous présentons la mutualisation des opérations dans le domaine fréquentiel au décodeur pour réaliser la synthèse énergétique à partir des spectres décodés et décorrélés par filtrage passe-tout à phase aléatoire.

5.1.2.1 Opérations affectées à l'encodeur et au décodeur

L'encodeur stéréo, présenté à la **Figure 5.7**, procède par TFCT (*cf.* Annexe A.1.1 avec l'utilisation de l'algorithme rapide *Fast Fourier Transform* FFT) des canaux originaux ($c_1[n]$, $c_2[n]$) avant de regrouper les coefficients spectraux en sous-bandes de fréquences selon l'échelle perceptuelle ERB (voir le détail de la procédure au paragraphe 4.2.2.2). Nous utilisons typiquement une fenêtre sinus de taille N=4096 échantillons et K_b=20 sous-bandes de fréquences. La covariance des spectres en sous-bandes est ensuite calculée pour pouvoir estimer l'angle de rotation utile au matriçage adaptatif *i.e.* rotations en sous-bandes (*cf.* paragraphe 4.3.3). De manière à utiliser les mêmes valeurs des angles de rotation à l'encodeur et au décodeur, les angles sont quantifiés (Q), selon la procédure décrite au paragraphe 5.1.2.2, avant d'être utilisés pour le matriçage. L'encodeur extrait et quantifie ensuite les paramètres énergétiques utiles au décodeur pour la synthèse de l'ambiance. Le choix du paramètre $\mathrm{RCPA}_{12}[l,b]$ défini par :

$$\mathrm{RCPA}_{12}[l,b] = 10 \times \log_{10}\left(\frac{\sum_{k=k_b}^{k=k_{b+1}-1} F_{d_1}[l,k]}{\sum_{k=k_b}^{k=k_{b+1}-1} F_{d_2}[l,k]}\right) \ \mathrm{dB}, \qquad (5.2)$$

est discuté au paragraphe 5.1.2.2. Enfin, la TFCT inverse (TFCT^{-1}) de la somme des composantes en sous-bandes du signal dominant obtenu par matriçage délivre la composante principale $d_1[l]$. La méthode OLA est ensuite utilisée pour sommer les portions glissantes du signal $d_1[n]$ encodé par un codeur audio monophonique.

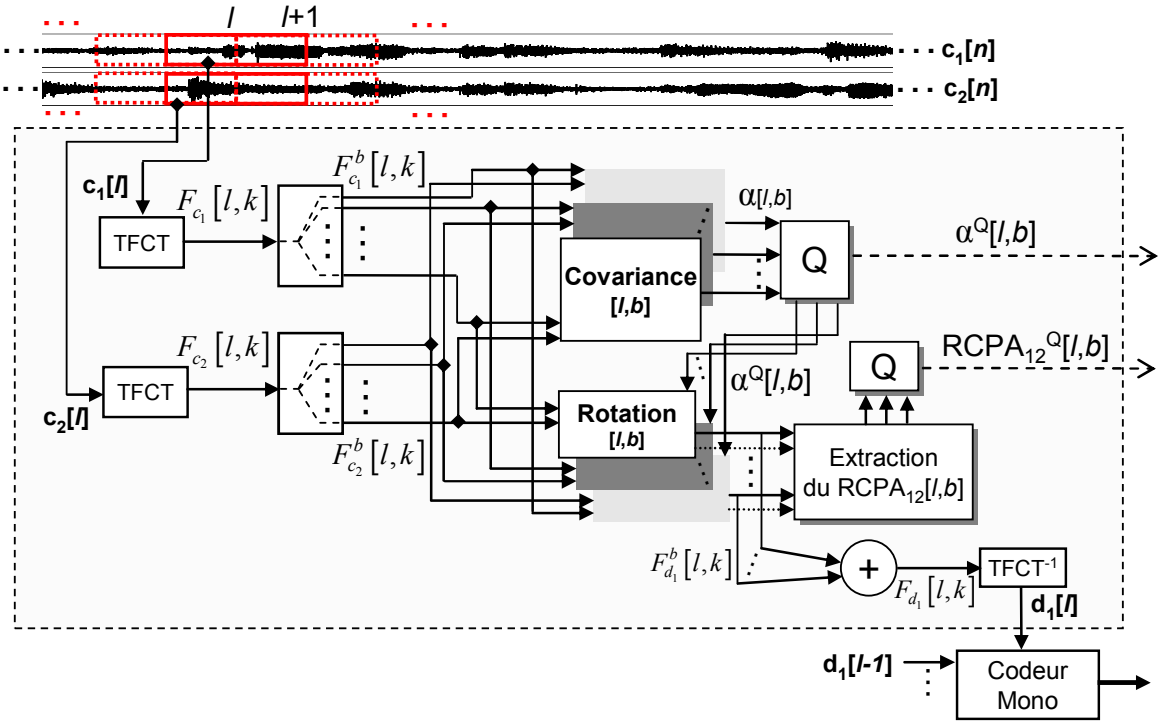

Figure 5.7 Codeur stéréo paramétrique basé sur un matriçage adaptatif en sous-bandes des canaux. Les paramètres énergétiques ($\mathrm{RCPA}_{12}[l,b]$) et spatiaux ($\alpha[l,b]$) sont quantifiés et transmis avec la composante principale codée par un codeur audio monophonique.

La **Figure 5.8** présente le décodeur stéréo paramétrique associé au décodeur audio monophonique qui délivre la composante principale $d'_1[n]$. L'opération de décodage utilise le même découpage temps-fréquence que celui utilisé à l'encodage. L'opération de quantification inverse (Q^{-1}) délivre les paramètres énergétiques utilisés pour régénérer

l'enveloppe spectrale du signal d'ambiance extrait à l'encodage. L'étape de décorrélation, détaillée au paragraphe 5.1.2.3, se déroule dans le domaine fréquentiel et génère un spectre en sous-bandes décorrélé du spectre en sous-bandes de la composante principale. Ces signaux en sous-bandes sont ensuite matriçés au moyen de la matrice de rotation inverse (par transposition) définie par les angles de rotation déquantifiés. Finalement, les portions glissantes du signal stéréo sont obtenues par TFCT inverse de la somme des composantes en sous-bandes respectives à chaque canal. Le signal stéréo ($c'_1[n]$, $c'_2[n]$) est reconstruit dans le domaine temporel par la méthode OLA de toutes les portions glissantes.

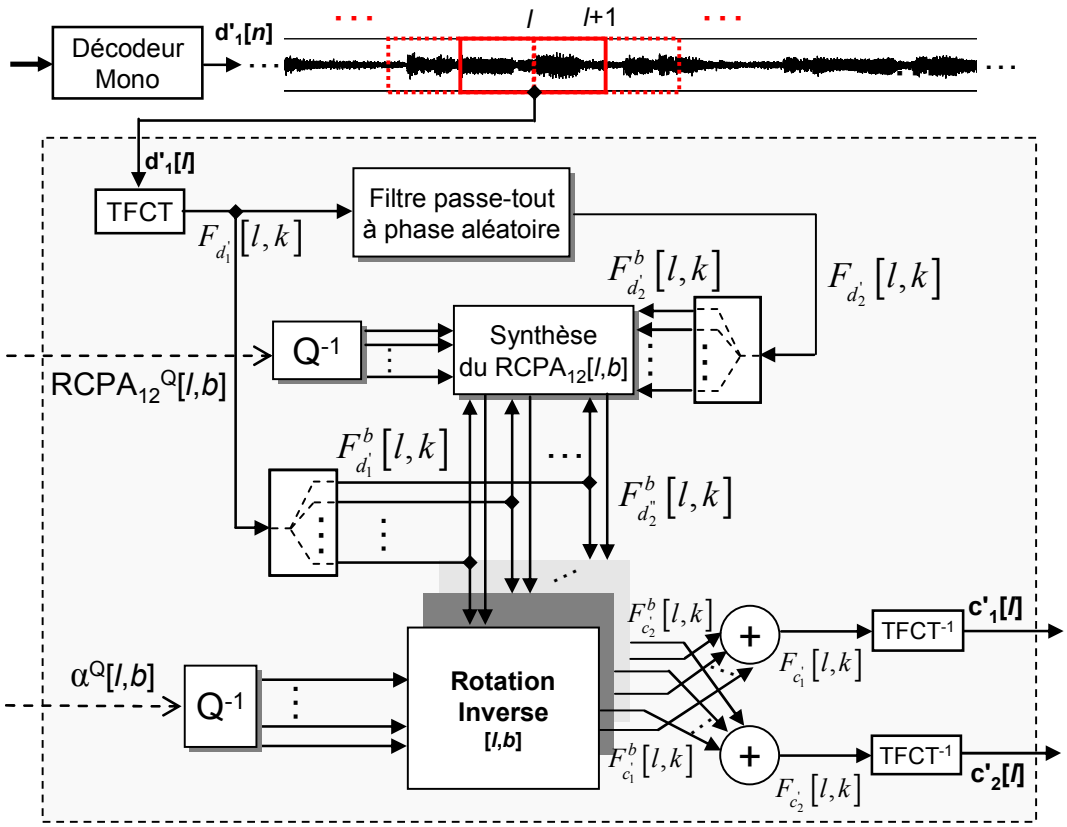

Figure 5.8 Décodeur stéréo paramétrique basé sur les paramètres spatiaux ($\alpha[l,b]$) **utiles au matriçage adaptatif en sous-bandes et les paramètres énergétiques** ($RCPA_{12}[l,b]$) **utiles à la synthèse de la composante d'ambiance.**

5.1.2.2 Quantification des paramètres spatiaux et énergétiques

Notre intention est de réaliser une quantification uniforme [GER92] des paramètres en sous-bandes qui apparaissent comme des vecteurs corrélés, de dimension égale au nombre de sous-bandes K_b. En effet, d'une sous-bande de fréquence à une autre, les paramètres extraits à partir des signaux (paramètres énergétiques) ou de leur covariance (paramètres spatiaux) peuvent être considérés comme corrélés. En conséquence, la redondance d'informations intrinsèque aux paramètres peut être réduite avec un codage différentiel en sous-bandes d'après [GER92]. Après avoir présenté les résultats de cette méthode de codage pour les deux types de paramètres, nous présentons les critères subjectifs utilisés pour la quantification de façon à ce que cette opération n'introduise pas de dégradation perceptive. La mise en application de ces principes consiste à extraire les paramètres d'un large corpus d'apprentissage puis à les quantifier au regard de leur distribution et des précisions fixées par les critères subjectifs énumérés. L'étape de quantification est finalement suivie d'un codage de Huffman [GER92] de façon à estimer un débit moyen du flux de paramètres. Ce débit moyen est alors comparé au débit estimé à partir d'une base différente de la base d'apprentissage, la base de signaux MPEG (stéréo).

Codage différentiel en sous-bandes des paramètres spatiaux

De façon à proposer une méthode de codage à débit variable et scalable dans la mesure où seule une partie du train binaire permettrait de réaliser un décodage « basique », nous proposons la transmission d'un angle de rotation moyen défini par :

$$\bar{\alpha}[l] = \frac{1}{K_b} \sum_{b=1}^{K_b} \alpha[l,b]. \tag{5.3}$$

$\bar{\alpha}[l]$ représente la position spatiale moyenne (azimut moyen) des sources directionnelles contenues dans la portion de signal stéréo analysée (indice de trame l). La position moyenne des sources directionnelles, estimée sur la base d'apprentissage (décrite en Annexe D.1) et présentée sur la **Figure 5.9** en degrés, est de l'ordre de 0° après conversion à partir de l'équation (4.40). Par conséquent, ce paramètre, transmis au décodeur à très bas débit, assure un repositionnement basique des sources directionnelles entre les haut-parleurs.

Avec la transmission supplémentaire des angles de rotation en sous-bandes, à un débit globalement plus élevé, les positions des sources directionnelles en sous-bandes peuvent alors être reproduites par ACP inverse. Plus précisément, les angles de rotation à moyenne retranchée, définis tels que :

$$\alpha^{mr}[l,b] = \alpha[l,b] - \bar{\alpha}[l], \tag{5.4}$$

peuvent être considérés comme les valeurs complémentaires à l'angle de rotation moyen.

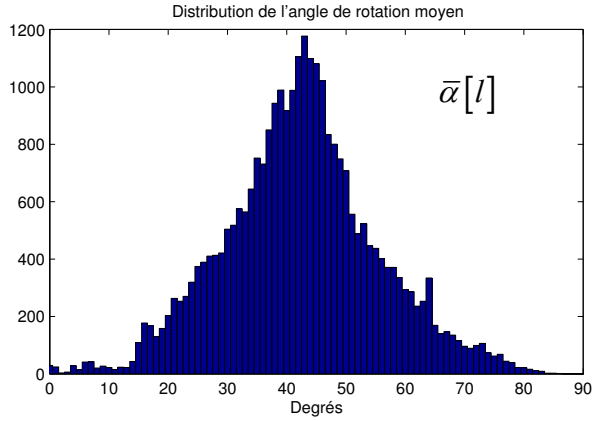

Figure 5.9 Distribution de l'angle de rotation moyen estimé à partir d'un corpus de signaux stéréo (20 signaux) d'une durée totale de 30 minutes. La longueur de la fenêtre d'analyse est de N=2048 échantillons et le nombre de sous-bandes K_b=20.

Cependant, la méthode de codage utilisée prend en compte la redondance d'informations entre les angles estimés en sous-bandes. Pour cela, un codage différentiel en sous-bandes (inter-bande) des paramètres spatiaux est mis en place. Les angles de rotation différentiels sont définis tels que :

$$\alpha^{d}[l,b] = \alpha[l,b] - \alpha[l,b-1], \text{ si } b > 1. \tag{5.5}$$

L'analyse des distributions des angles de rotation en sous-bandes, relatifs aux positions des sources directionnelles dominantes (*cf.* paragraphe 4.3.3.1), calculées sur la base d'apprentissage est présentée en Annexe D.2. D'après la **Figure 5.10**, il apparaît nettement plus avantageux de quantifier les angles de rotation différentiels plutôt que les angles de rotation à moyenne retranchée.

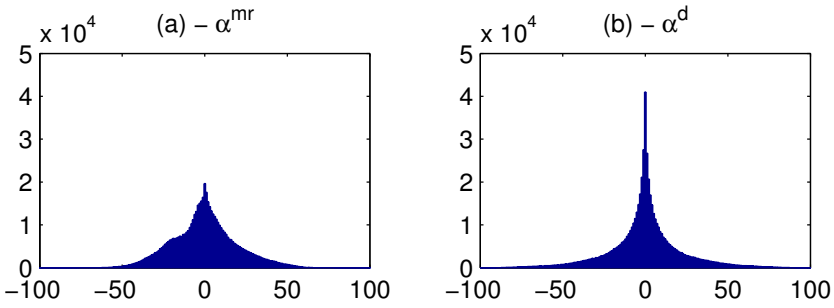

Figure 5.10 Distributions des angles de rotation [(a) - à moyenne retranchée α^{mr}, **(b) -** différentiels α^{d}**]** pour toutes portions glissantes et toutes les sous-bandes analysées à partir d'un corpus de 20 signaux stéréo.

Finalement, les angles de rotation originaux en sous-bandes $\alpha[l,b]$ peuvent être obtenus par combinaison de l'angle de rotation moyen $\overline{\alpha}\,[l]$ associé à l'angle de rotation à moyenne retranchée pour la première sous-bande $\alpha^{mr}[l,1]$ et des angles de rotation différentiels $\alpha^{d}[l,b]$ pour les autres sous-bandes comme indiqué par le schéma de principe du processus de quantification des paramètres spatiaux (*cf.* **Figure 5.11**). La transmission des paramètres spatiaux est possible avec un débit scalable et représentatif de la qualité de l'image spatiale des signaux reconstruits.

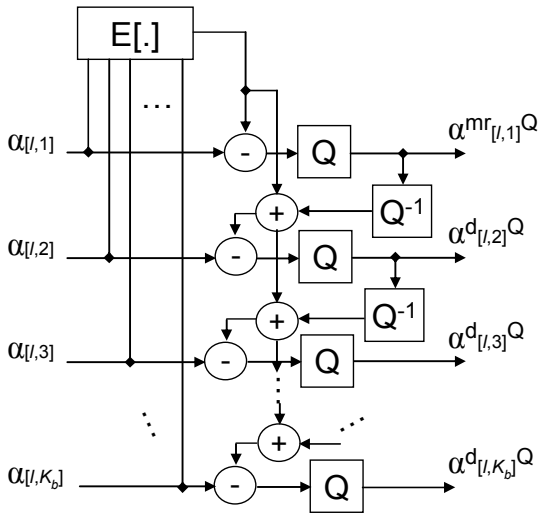

Figure 5.11 Schéma de principe du codage inter-différentiel des angles de rotation en sous-bandes.

Codage différentiel en sous-bandes des paramètres énergétiques

Cette étude vise à établir un choix pour le paramètre énergétique à quantifier et par suite à transmettre avec les paramètres spatiaux et le flux audio mono encodé. Puisque l'ambiance sonore extraite par ACP a une énergie moyenne faible, il est possible de transmettre soit directement cette énergie de la composante ambiance ($E_{D_2}[l,b]$) en sous-bandes ou le rapport d'énergie entre la composante principale et la composante ambiance, c'est-à-dire le $RCPA_{12}[l,b]$.

A la différence du codage des paramètres spatiaux qui se base sur la transmission d'un angle de rotation moyen, les paramètres énergétiques doivent être extraits et quantifiés pour chaque sous-bande de fréquences pour assurer la reconstruction énergétique des canaux stéréo par ACP inverse. Autrement dit, une énergie moyenne serait insuffisante avec la présence de multiples sources directionnelles (donc secondaires) dans le signal original. Par contre, identiquement au cas du codage des paramètres spatiaux, la méthode de codage des

paramètres énergétiques prend en compte la corrélation des paramètres énergétiques en sous-bandes. Les paramètres énergétiques différentiels sont alors définis tels que :

$$\begin{cases} \mathrm{E}^{d}_{D_2}[l,b] = \mathrm{E}_{D_2}[l,b] - \mathrm{E}_{D_2}[l,b-1] \\ \mathrm{RCPA}^{d}_{12}[l,b] = \mathrm{RCPA}_{12}[l,b] - \mathrm{RCPA}_{12}[l,b-1] \end{cases}, \ \forall\ b > 1. \qquad (5.6)$$

L'analyse des distributions des paramètres énergétiques différentiels en sous-bandes calculés sur la base d'apprentissage (décrite en Annexe D.1) est présentée en Annexe D.3. En outre, d'après la **Figure 5.12**, les distributions des paramètres différentiels RCPA^{d}_{12} et $\mathrm{E}^{d}_{D_2}$ (pour toutes les sous-bandes) sont relativement proches excepté leur valeur moyenne (0 dB pour le rapport d'énergie et -2 dB pour l'énergie de la composante D_2). Cependant l'écart-type de la distribution des rapports d'énergie différentiels est inférieure (de 1 dB) à celui de la distribution des énergies différentielles en sous-bandes de la composante D_2.

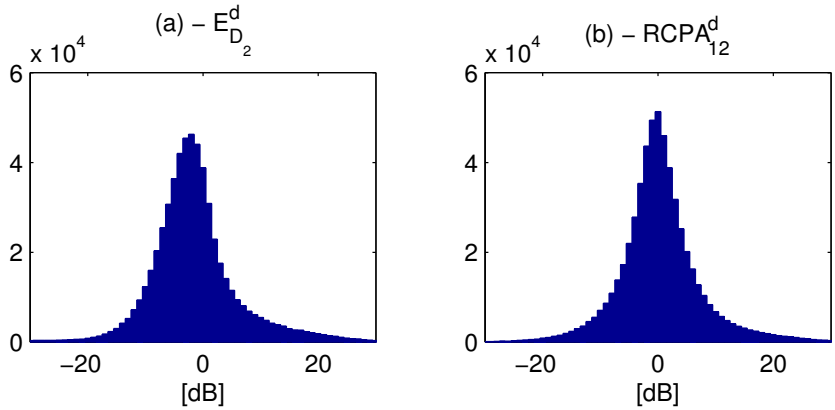

Figure 5.12 Distributions [(a) - des énergies en sous-bandes de la composante ambiance, **(b)** - des RCPA_{12}**]** pour toutes les sous-bandes et toutes les portions de signal analysées de la base d'apprentissage (30 minutes de signaux stéréo).

A partir sur cette indication, nous avons choisi le paramètre énergétique RCPA_{12} pour la synthèse paramétrique de la composante ambiance.

Quantification des paramètres selon des critères subjectifs

Les paramètres spatiaux (angles de rotation) et énergétiques (RCPA_{12}) en sous-bandes sont quantifiés au regard de critères perceptuels. De cette manière, le processus de quantification vise à introduire une erreur de quantification inaudible.

Le critère perceptuel pris en compte pour attribuer le pas de quantification des paramètres spatiaux *i.e.* les angles de rotation, est établi au regard de l'angle audible minimal ou *minimum audible angle* (MAA) décrit par Moore dans [MOO03]. La performance de localisation des sources sonores dans l'espace est une fonction complexe qui dépend à la fois de la nature du stimulus mais également de sa position dans l'espace vis-à-vis de l'auditeur. Par exemple, une sinusoïde ou fréquence pure à 1 kHz spatialisée à une position centrale *i.e.* en face de l'auditeur (azimut à 0 degré), peut être localisée par l'auditeur avec une précision de +/-1 degré. Toujours d'après [MOO03], la précision de la localisation décroît lorsque la source s'éloigne de la position centrale (azimut à 0°) et également lorsque la fréquence de la source sonore augmente. En effet, une fréquence pure à 2 kHz spatialisée à un azimut de 30° par rapport à l'auditeur peut être localisée par l'auditeur avec une précision de +/-5°.

Par conséquent, en considérant des sources sonores réelles avec des positions spatiales appartenant à l'intervalle [-30;30]° *i.e.* système de haut-parleurs stéréo conventionnel, la précision allouée à l'angle de rotation moyen, converti en degrés, $\overline{\alpha}\,[l] \in [0;90]°$ a été

choisie, dans le pire des cas, égale à $\Delta_{\bar{\alpha}} = \pm 3°$. D'après l'Annexe D.1, les distributions de l'angle de rotation à moyenne retranchée α^{mr} pour la première sous-bande et des angles de rotation différentiels α^{d} pour toutes les autres sous-bandes possèdent des domaines de variations différents avec une dynamique décroissante suivant les fréquences. De plus, ces angles de rotation en sous-bandes ne nécessitent pas la même précision que l'angle de rotation moyen qui est primordial pour assurer une bonne reconstruction spatiale de la scène sonore. Ces angles de rotation en sous-bandes sont utilisés de manière à affiner la position spatiale des sources directionnelles plus ou moins bien positionnées par l'angle de rotation moyen. Le processus de quantification perceptuel vise à attribuer différents niveaux de précisons pour ces angles de rotation en sous-bandes. La précision de la localisation auditive en basse fréquence étant meilleure, l'angle de rotation à moyenne retranchée $\alpha^{mr}[l,1]$ se voit attribuer une précision égale à $\Delta_{\alpha^{mr}} = \pm 3°$. La précision attribuée aux angles de rotation différentiels pour les autre sous-bandes est établie à la fois au regard des résultats académiques présentés dans [MOO03] et des distributions présentées en Annexe D.1. Ainsi, cette précision décroît avec la fréquence telle que :

$$\Delta_{\alpha^{d}} = \pm \begin{cases} 3°, & f \le 1 \text{ kHz} \\ 7{,}5°, & f \le 5 \text{ kHz} \\ 10°, & 5 \le f \le f_s \text{ kHz} \end{cases} \tag{5.7}$$

Les paramètres énergétiques définis par le *RCPA* exprimés en décibels (dB) ne procurent aucune différence audible entre les signaux stéréo originaux et les signaux stéréo reconstruits avec une précision de l'ordre de $\Delta_{\mathrm{RCPA}_{12}^{d}} = \pm 3 \text{ dB}$. Précisément, la précision allouée au RCPA_{12} de la première sous-bande est supérieure à la précision allouée aux $\mathrm{RCPA}^{\mathrm{d}}{}_{12}$ différentiels définis pour les autres sous-bandes.

Estimation du débit moyen du flux des paramètres

Basé sur les dynamiques et les précisions allouées à chaque paramètre, le nombre de bits (*nbits*) nécessaires à la quantification des paramètres spatiaux et énergétiques se déduit naturellement au moyen de l'équation suivante :

$$\frac{\max_P - \min_P}{2^{nbits_P - 1}} = \Delta_P, \tag{5.8}$$

où p correspond au paramètre à quantifier, Δ_p au pas de quantification/précision et le numérateur à la dynamique du paramètre p. Ainsi $nbits_p$ correspond au nombre de bits nécessaire à la quantification du paramètre p estimé pour chaque trame et chaque sous-bande.
Finalement, le débit moyen du flux de paramètres utilisés par cette méthode de codage stéréo est donné par :

$$d_{parametres} = \left(d_{\alpha} + d_{RCPA}\right) \times \frac{2 \times f_S}{N}, \text{ tel que :} \tag{5.9}$$

$$d_{RCPA} = \left(nbits_{RCPA} + nbits_{RCPA_d} \times \left(K_b - 1\right)\right), \text{ et} \tag{5.10}$$

$$
\begin{aligned}
& d_{\alpha} = d_{\bar{\alpha}} + d_{\alpha^{FB}} + d_{\alpha^{FM}} + d_{\alpha^{FH}}, \\
& \begin{cases}
d_{\bar{\alpha}} = nbits_{\bar{\alpha}}, \\
d_{\alpha^{FB}} = nbits_{\alpha_{mr}} + nbits_{\alpha_d^{FB}} \times K_b^{FB}, \\
d_{\alpha^{FM}} = nbits_{\alpha_d^{FM}} \times K_b^{FM}, \\
d_{\alpha^{FH}} = nbits_{\alpha_d^{FH}} \times K_b^{FH}.
\end{cases}
\end{aligned}
\tag{5.11}
$$

Ainsi, le débit du flux de paramètres est fonction de la taille de la fenêtre d'analyse (N), de la fréquence d'échantillonnage (f_S), du nombre de sous-bandes (K_b) et des fréquences limites qui séparent les fréquences basses (FB) des fréquences médiums (FM) et des fréquences hautes (FH).

Paramètre	*Dynamique / Précision*			nbits		
$\bar{\alpha}[l]$	90° / 3°			5		
$\alpha^{mr}[l,1]$	140° / 3°			6		
$\alpha^{d}[l,b]$	140° / 3° $f < f^{FB}$=1 kHz	120° / 7.5° $f^{FB} < f < f^{FM}$=5 kHz	80° / 10° $f^{FM} < f < f_s/2$	6	4	3
$\mathrm{RCPA}_{12}[l,1]$	60 dB / 2 dB			5		
$\mathrm{RCPA}_{12}^{d}[l,b]$	60 dB / 3.75 dB			4		

Tableau 5.1 Récapitulatif des dynamiques des paramètres spatiaux et énergétiques tirés de leur distribution. Les précisions allouées à chaque paramètre pour la quantification sont fixées par des critères perceptuels. Le nombre de bits nécessaire pour la quantification de chaque paramètre est présenté à droite.

D'après le **Tableau 5.1** qui présente les fréquences limites séparatrices des domaines fréquentiels utilisés pour la quantification des angles de rotation différentiels, les nombres de sous-bandes propres à chacun de ces domaines fréquentiels sont donnés par :

$$
\begin{cases}
K_b^{FB} = \arg\limits_{b}\left(f^{FB} \times \left(\frac{N}{2} + 1 \right) \times \frac{2}{f_S} \right) \\
K_b^{FM} = \arg\limits_{b}\left(f^{FM} \times \left(\frac{N}{2} + 1 \right) \times \frac{2}{f_S} \right) - K_b^{FB}, \\
K_b^{FH} = K_b - \left(K_b^{FB} + K_b^{FM} \right)
\end{cases}
\tag{5.12}
$$

où b représente le numéro de la sous-bande qui contient la fréquence calculée.

Finalement, à partir du **Tableau 5.1** et des équations (5.8), (5.9), (5.10), (5.11) et (5.12), le débit maximal estimé est de 4265 bps soit 4,2 kbps pour un échantillonnage des signaux à f_S=48000 Hz, K_b=20 sous-bandes de fréquences, une longueur de fenêtre N=4096 échantillons et un taux de recouvrement de 50%.

L'étape de quantification des paramètres spatiaux décrite à la **Figure 5.11** (transposable au cas des paramètres énergétiques) est suivie d'un codage Huffman. Ce codage, décrit dans [GER92], vise à réduire la longueur moyenne des mots de code (octets sur huit bits) des paramètres quantifiés. Par suite, le débit moyen du flux des paramètres peut être estimé. A partir d'un large corpus de divers signaux stéréo (base d'apprentissage décrite en Annexe D.1), ce débit moyen a été estimé à 3 kbps pour un codage paramétrique des signaux (avec une fréquence d'échantillonnage f_S=44100 ou 48000 Hz) au moyen d'une fenêtre d'analyse de 4096 échantillons et de K_b=20 sous-bandes de fréquences.

De manière à vérifier cette estimation du débit moyen, des paramètres ont été extraits et quantifiés à partir de la base de signaux stéréo MPEG. Les débits moyens estimés pour chaque signal et en moyenne sur tous les signaux sont présentés à la **Figure 5.13**. Le débit moyen estimé à partir de la base MPEG est comparable à celui estimé à partir de la base d'apprentissage. En conclusion, le procédé de quantification mise en place est peu sensible à la base d'apprentissage puisque le débit moyen estimé à partir de la base de signaux MPEG (2,8 kbps) est équivalent au débit moyen (3 kbps) estimé avec la base d'apprentissage. Cette légère différence s'explique notamment par le fait que les signaux de la base MPEG présentent parfois moins d'effets stéréo (spatialisation) que ceux de notre base d'apprentissage. Par exemple l'échantillon « Guitar+Castanets » ne présente pas de spatialisation (les sources sonores sont perçues comme fixes au centre de l'image stéréo) mais en revanche d'importantes attaques si critiques pour les procédés de codage.

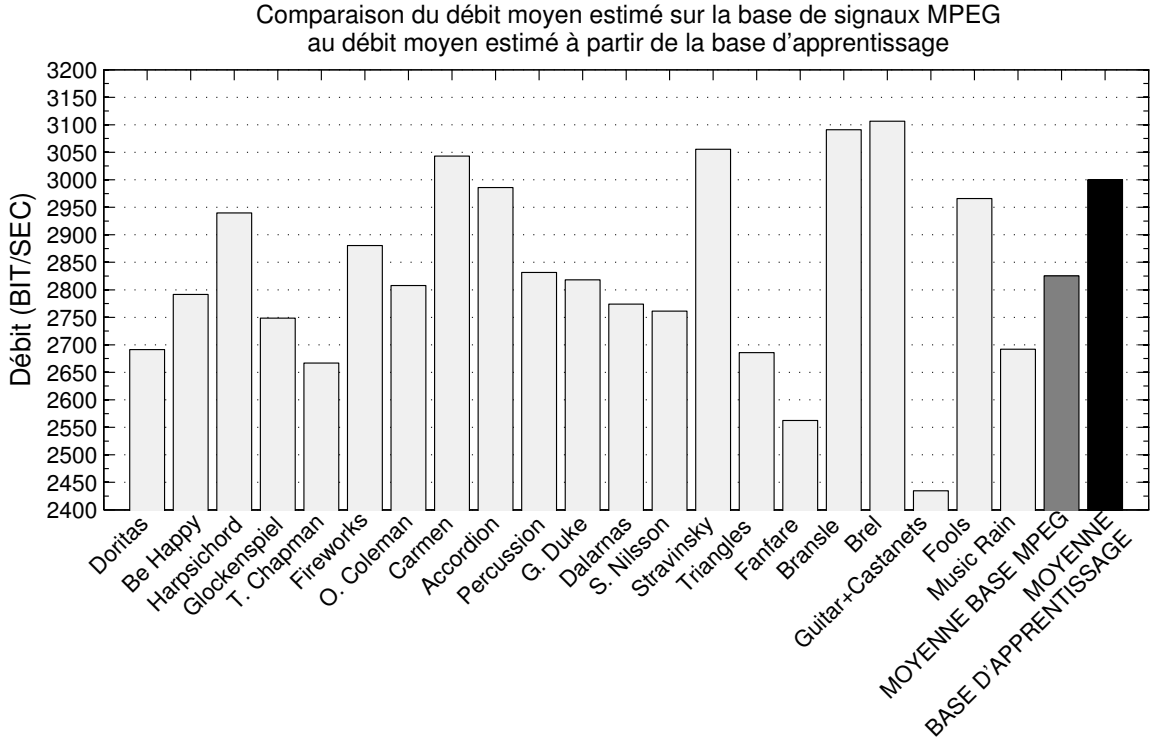

Figure 5.13 Débit des paramètres estimés pour 21 signaux tirés de la base de signaux stéréo MPEG. Le débit moyen estimé à partir de la base MPEG est légèrement inférieur (2,82 kbps) à celui estimé à partir de la base d'apprentissage (3 kbps).

Par conséquent, la quantification mise en place peut encore être améliorée notamment pour les angles de rotation qui peuvent avoir des valeurs constantes au cours du temps. En outre, une quantification polaire peut être imaginée en combinant la quantification des angles et des rapports d'énergie (voir [VAF05] pour le principe de cette quantification).

5.1.2.3 Synthèse de l'ambiance dans le domaine spectral

L'opération de décodage, dont le schéma de principe est présenté à la **Figure 5.8**, vise à mutualiser les traitements dans le domaine fréquentiel. Nous avons d'abord présenté au paragraphe 5.1.1.1, l'utilisation du filtrage passe-tout (décorrélation à phase aléatoire) pour évaluer l'influence de l'information de phase de la composante d'ambiance D_2 en réalisant la convolution du signal dans le domaine temporel. Etant donné que le filtre passe-tout à phase aléatoire est défini dans le domaine fréquentiel par l'équation (5.1), nous pouvons réaliser l'opération de filtrage directement dans le domaine fréquentiel par simple multiplication du spectre (établi par TFCT) de la composante principale D'_1 décodée soit $F_{d'_1}[l,k]$ tel que :

$$F_{d'_2}[l,k] = F_{d'_1}[l,k] \cdot \mathrm{e}^{j\varphi[k]}, \text{ avec } \varphi[k] \in [-\pi;\pi] \text{ et } k=[1,N/2+1]. \tag{5.13}$$

Précisons que nous utilisons la même réponse en phase au cours du temps c'est-à-dire pour chaque portion glissante, de façon à être homogène avec la synthèse finale par la méthode OLA. Autrement dit, l'addition de portions glissantes qui se recouvrent de moitié mais potentiellement en opposition de phase (avec des séquences aléatoires différentes) provoquerait des pertes d'informations.

La seconde étape de l'opération de décodage repose sur la synthèse énergétique à partir du spectre décorrélé en sous-bandes (échelle ERB, K_b=20) $F^b_{d'2}[l,k]$ et des paramètres énergétiques déquantifiés (quantification uniforme et codage différentiel inverses) $\mathrm{RCPA}_{12}^{\mathrm{Q}}[l,b]$ telle que :

$$F^b_{d''_2}[l,k] = \frac{1}{10^{\mathrm{RCPA}^{\mathrm{Q}}_{12}[l,b]/20}} . F^b_{d'_2}[l,k], \quad \forall k \in [k_b; k_{b+1}-1]. \tag{5.14}$$

Enfin, le décodeur procède par rotation inverse en sous-bandes au moyen des angles de rotation déquantifiés $\alpha^{\mathrm{Q}}[l,b]$ et des spectres en sous-bandes $F^b_{d'1}[l,k]$ et $F^b_{d''2}[l,k]$. La rotation inverse en sous-bandes est tirée de l'équation (4.38) avec la matrice de rotation transposée.

5.1.3 Evaluation de la méthode de codage stéréo paramétrique

Les méthodes d'évaluation objectives proposées par le standard de l'UIT-R (recommendation BS.1387) : PEAQ - *Perceptual Evaluation of Audio Quality*, peuvent être utilisées pour évaluer les performances des codecs audio à haute-qualité mais elles ne sont pas encore suffisamment matures et fiables pour évaluer les codecs audio à bas débit qui délivrent des qualités intermédiaires. De même, la méthode d'évaluation subjective standardisée par l'UIT-R en 1997 figurant dans la recommandation BS.1116 [UIT1116] permet l'évaluation de faibles dégradations dans les systèmes audio et n'est donc pas fiable pour évaluer des dégradations plus conséquentes pour les systèmes audio à qualité intermédiaire.

La méthode standard utilisée pour réaliser des tests d'évaluation de la qualité audio et spatiale délivrée par les codecs audio « bas-débit » est la méthode MUSHRA (*MUlti Stimulus test with Hidden Reference and Anchors*), décrite dans [STO00]. La méthode MUSHRA issue du standard de l'UIT-R en 2003 [UIT1534] présente l'avantage, contrairement aux méthodes figurant dans la recommandation UIT-R BS.1116, de présenter tous les *stimuli* en même temps, ce qui permet au sujets du test de faire instantanément toutes les comparaisons auditives possibles (basculement entre les *stimuli* instantané). La cohérence des résultats s'en trouve accrue, ce qui conduit à des intervalles de confiance plus petits. La durée du test, lorsqu'on utilise la méthode MUSHRA, peut être sensiblement réduite par rapport à la méthode exposée dans la Recommandation UIT-R BS.1116.

De façon à évaluer notre implémentation de la méthode de codage stéréo paramétrique basée sur l'ACP (*cf.* paragraphe 5.1.2), un test subjectif suivant la méthodologie MUSHRA a

été mis en œuvre. Il convient de rappeler que l'intérêt des méthodes de codage audio paramétrique comparé aux méthodes de codage audio dites purement perceptuelles (codeur MPEG-1 Couche III par exemple) a déjà été démontré par les travaux de Faller et Breebaart notamment (*cf.* paragraphe 2.2.3.3).

5.1.3.1 Objectif du test subjectif

L'objectif de ce test subjectif est de comparer la méthode de codage stéréo paramétrique basée sur l'ACP à la méthode de codage stéréo paramétrique (PS décrit au paragraphe 2.2.3) utilisée par le codec standardisé MPEG-4 HE-AAC-PS ou HE-AACv2 d'après l'organisation *3rd Generation Partnership Project* 3GPP [TGP05] (*cf.* paragraphe 2.3.2.2).

Pour y parvenir, un corpus de signaux stéréo encodés par le codec HE-AACv2 à 24 kbps est comparé au même corpus de signaux encodés par la méthode paramétrique basée sur l'ACP associée à un codage audio de la composante principale par le profil monophonique du codec MPEG-4 HE-AAC (codec mono AAC associé à l'outil SBR décrit au paragraphe 2.3.2.2). Le flux de paramètres (α^{Q} et $RCPA^{Q}_{12}$) est transmis à un débit de 3 kbps, d'après le paragraphe 5.1.2.2, et la composante principale est encodée à un débit de 22 kbps. Un débit inférieur (typiquement 21 kbps pour obtenir exactement le même débit global) du codeur mono HE-AAC ne permet pas d'obtenir des signaux décodés avec la même bande passante (16 kHz) que celle des signaux issus d'un encodage/décodage par l'HE-AACv2 à 24 kbps. Par suite, le débit moyen global estimé pour la méthode de codage stéréo paramétrique basée sur l'ACP (PCA-HE-AAC selon une terminologie anglaise) est de 25 kbps avec une longueur de fenêtre d'analyse fixe et égale à N=4096 échantillons et K_b=20 sous-bandes de fréquences suivant l'échelle ERB.

Le codec HE-AACv2 utilise, lui, 20 sous-bandes de fréquences (échelle ERB) et une taille de fenêtre variable. L'encodeur ne transmet pas d'ICPD ni d'OPD *i.e.* seuls les ICLD et ICC sont transmises. Le débit des informations spatiales du codec HE-AACv2 varie entre 1 et 8 kbps, nous l'avons estimé à 2 kbps en moyenne.

5.1.3.2 Déroulement du test subjectif

Treize sujets ont participé au test subjectif d'évaluation de la méthode de codage audio paramétrique basée sur l'ACP. Tous les sujets avaient déjà de l'expérience en matière d'évaluation de codecs audio et ont été particulièrement sensibilisés pour tenir compte à la fois de la qualité audio spatiale (image stéréo, effet de salle, etc.) et des dégradations audio introduites principalement par le codage audio monophonique. Le test d'écoute a été réalisé dans une salle insonorisée au moyen d'un ordinateur (déporté de la salle) muni d'un convertisseur numérique/analogique *Denon* et d'un casque d'écoute professionnel *Stax*. Il est important de rappeler que les conditions d'écoute au casque sont beaucoup plus précises et favorables à un jugement plus sévère (écoute critique) que celui obtenu avec une écoute sur haut-parleurs. En effet, la séparation des canaux avec une écoute au casque est accrue comparée à une écoute sur haut-parleurs puisque cette dernière implique un mélange des signaux atteignant les oreilles de l'auditeur (effet de *cross-talk* ou trajets croisés décrits au paragraphe 1.2.1.1). Les sujets ont dû noter la qualité perçue de huit signaux traités (par les deux codecs) sur une échelle à 100 points avec une référence cachée et deux ancres correspondant à la référence filtrée à 3,5 kHz et 7 kHz. Cinq signaux stéréo ont été extraits de la base de signaux stéréo MPEG (classiquement utilisée en normalisation) et deux autres signaux stéréo ont été choisis pour l'originalité de leur image spatiale.

5.1.3.3 Analyse des résultats

Une comparaison des notes MUSHRA moyennées sur tous les sujets en fonction de l'échantillon audio évalué et des deux méthodes de codage est présentée sur la **Figure 5.14**.

Les signaux stéréo encodés par le codec HE-AACv2 ont été globalement légèrement mieux notés que les signaux stéréo encodés par le codec dénommé PCA-HE-AAC.

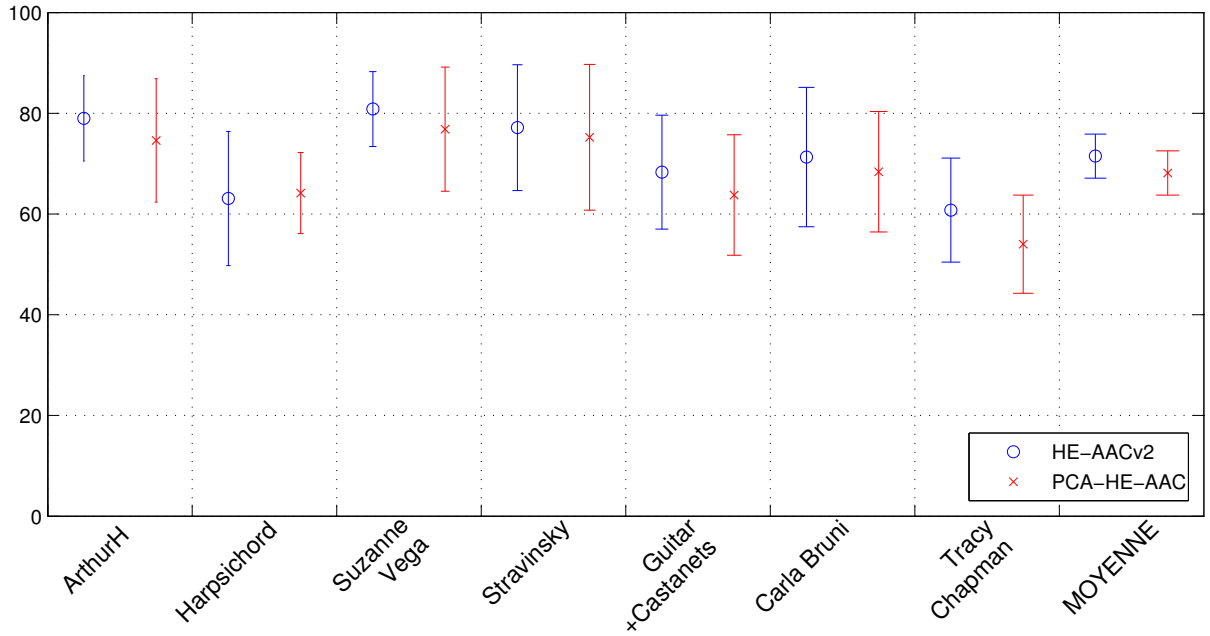

Figure 5.14 Intervalles de confiances à 95% et notes MUSHRA moyennées sur tous les sujets en fonction de l'échantillon audio et des méthodes de codage stéréo paramétrique évaluées. La moyenne des notes sur tous les sujets et tous les items est présentée à droite.

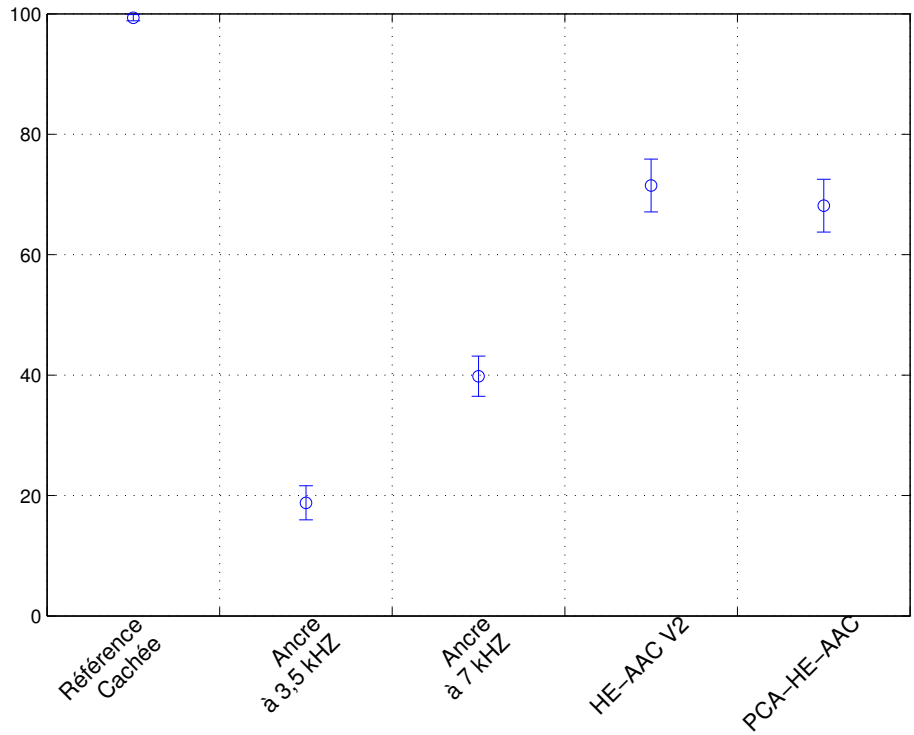

Figure 5.15 Intervalles de confiances à 95% et moyennes des notes MUSHRA (moyennées sur tous les sujets et tous les items audio évalués) en fonction du type de signal évalué.

En moyenne sur tous les signaux évalués et sur tous les sujets, la différence entre les notes moyenne est seulement de 3,3 points et les intervalles de confiance à 95% se recouvrent comme indiqué par la **Figure 5.14**. Cette faible différence exprimée par les sujets peut être

expliquée par le fait que la méthode de codage stéréo paramétrique basée sur l'ACP met à jour les paramètres seulement tous les 42,5 ms (pour une fenêtre d'analyse à 4096 échantillons et une fréquence d'échantillonnage à 48000 Hz). Ainsi, l'encodage par le codec PCA-HE-AAC de signaux constitués d'attaques, comme l'item « Guitar+Castanets », résulte en un signal décodé dont les attaques sont dégradées (phénomène de pré-écho décrit en Annexe B.1.3.2). Par conséquent, la comparaison subjective avec les signaux encodés avec le codec HE-AACv2 qui utilise un taux de rafraîchissement des paramètres inférieur (23 ms, d'après [BRE05a], voire moins lors d'un basculement de fenêtre) a tendance à mettre en avant le codec HE-AACv2.

Bien que la résolution temps-fréquence doive être prise en compte et adaptée aux contenus des signaux, il apparaît, suite aux notes analysées et aux commentaires des sujets experts, que l'impression spatiale donnée par les signaux décodés du codec PCA-HE-AAC est très proche de l'originale et considérée comme plus stable que celle donnée par le codec HE-AACv2. Ce qui explique qu'en moyenne, il n'y ait pas de différence significative entre les deux méthodes de codage stéréo paramétriques comme indiqué par la **Figure 5.15**.

5.2 Extension au codage des signaux au format 5.0

Les méthodes de codage audio multicanal paramétriques (BCC, MPEG *surround*), décrites au paragraphe 2.3.3, proposent la transmission d'informations spatiales auxiliaires pour accompagner le flux audio (mono ou stéréo) directement utilisable par les systèmes audio classiques. Ainsi, suivant les moyens de reproduction audio dont dispose le récepteur (PC, décodeur HD-TV, etc.), le décodeur audio multicanal peut fournir à la fois un flux audio mono/stéréo ou un flux audio enrichi, par les paramètres spatiaux, notamment au format 5.1.

L'extension de la méthode de codage stéréo paramétrique basée sur l'ACP pour le codage des signaux au format 5.1 consiste à utiliser les « modules » d'ACP bidimensionnelle et tridimensionnelle pour générer un flux audio, compatible avec les systèmes de diffusion audio classiques, accompagné des informations nécessaires à l'expansion de ce flux en une scène sonore multicanale. Nous considérons dans cette partie du document des signaux au format 5.0 *i.e.* le canal basses fréquences n'est pas directement pris en compte. Par conséquent, la dimension du signal à coder permet de résoudre le problème posé à partir de plusieurs combinaisons possibles de modules d'ACP. En effet, les canaux d'entrée au procédé de codage peuvent être traités à partir de plusieurs paires ou triplets de canaux suivant un ordonancement à définir. Nous présentons dans cette section l'utilisation de modules d'ACP trimidensionnelles selon la symétrie (séparation) gauche-droite du système de reproduction 5.1 pour obtenir une image stéréo cohérente avec la scène de départ.

5.2.1 Modules d'ACP et débits des paramètres associés

Un module d'ACP 2D (bidimensionnelle) ou 3D (tridimensionnelle) est défini (*cf.* **Figure 5.16**) comme un matriçage dynamique (suivant l'évolution de la covariance des signaux) qui, à partir de deux ou trois signaux d'entrée, délivre une composante principale et un flux de paramètres utiles à la synthèse des composantes ambiances (paramètres énergétiques) et par suite à l'ACP 2D ou 3D inverse (angles de rotation en sous-bandes).

Le principe du traitement pour réaliser l'ACP 2D ou l'ACP 2D inverse suit rigoureusement celui exposé au paragraphe 5.1.2. Pour réaliser l'ACP 3D, nous faisons le choix d'utiliser seulement deux angles d'Euler, définis au paragraphe 4.3.4, étant donné que le matriçage adaptatif avec deux ou trois rotations en sous-bandes permet de concentrer au maximum l'énergie initiale dans la composante principale D_1. En outre, la différence, d'un point de vue énergétique et perceptif, entre un matriçage basé sur deux ou trois matrices de rotation s'amenuise avec un traitement en sous-bandes de fréquences (*cf.* 4.3.4.2). Par conséquent, l'ACP 3D inverse sera également réalisée en sous-bandes de fréquences à partir des angles

d'Euler (α,β) et des signaux résiduels ou d'ambiances D'_2 et D'_3 générés à partir de la composante principale décodée D'_1 et des paramètres énergétiques $RCPA_{12}$ et $RCPA_{13}$ définis au paragraphe 4.3.4.2.

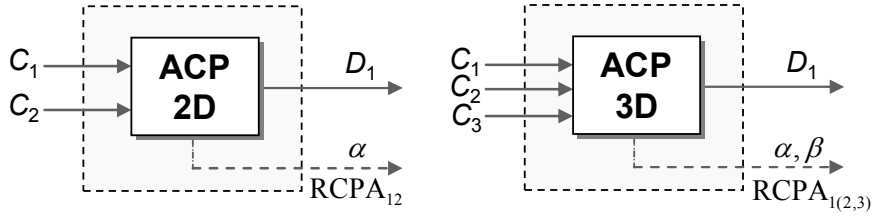

Figure 5.16 Modules d'ACP 2D et 3D. A partir de deux ou trois canaux, ces modules génèrent une composante principale et un flux de paramètres (angles de rotation et rapport énergétique de la composante principale aux ambiances).

Finalement, les informations en sortie des modules d'ACP 2D et 3D (*cf.* **Figure 5.16**) constituent une représentation concise et efficace des canaux d'entrée au moyen :

- d'une composante principale (D_1) établie par matriçage adaptatif (*cf.* équation (4.30)),
- d'angles de rotation (α dans le cas bidimensionnel ou (α,β) dans le cas tridimensionnel) en sous-bandes et,
- des rapports d'énergie ($RCPA_{12}$ ou ($RCPA_{12}$ et $RCPA_{13}$)) entre les spectres en sous-bandes de la composante principale et de la ou des composante(s) d'ambiance(s).

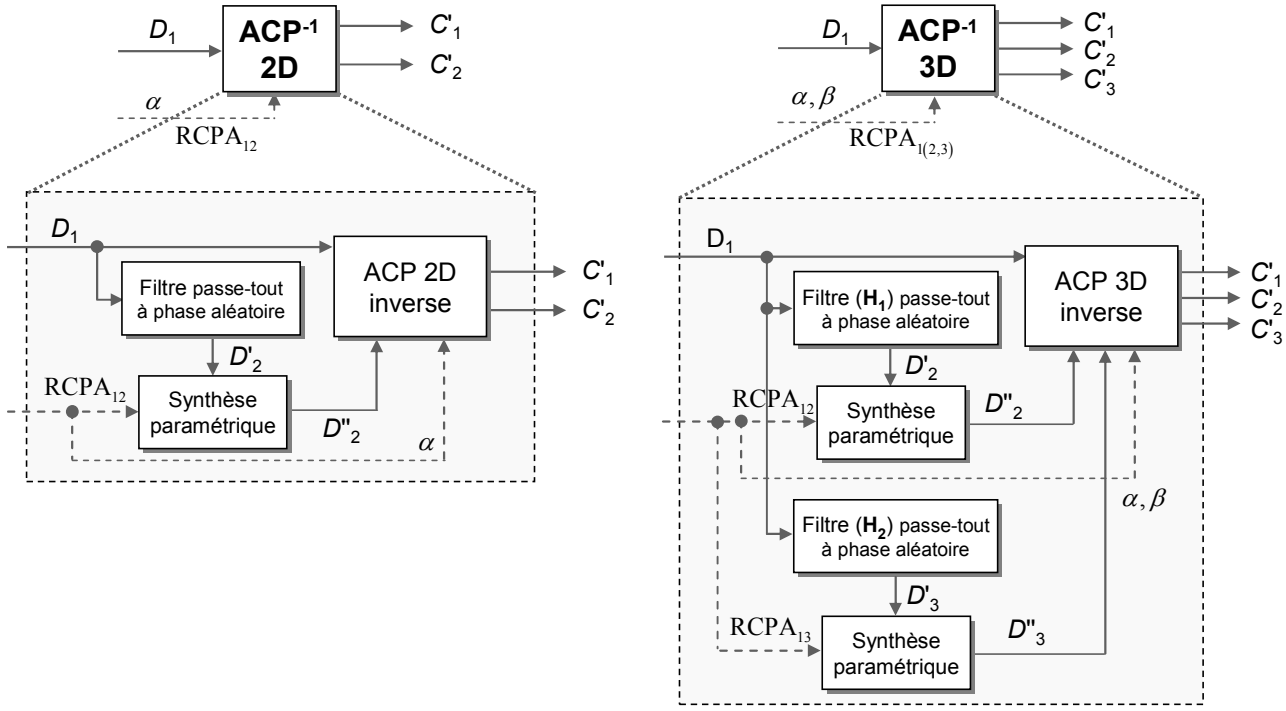

Figure 5.17 Modules d'ACP 2D et 3D inverses.

Pour simplifier les notations, nous considérons que les modules d'ACP inverse (ACP^{-1} 2D et 3D) correspondent à la fois à l'étape de filtrage décorrélateur suivie de la synthèse paramétrique en sous-bandes (*cf.* paragraphe 5.1.2.3) et à la fois à l'opération d'ACP inverse (2D ou 3D) proprement dite comme indiqué par la **Figure 5.17**. Dans le cas d'une ACP 3D

inverse, deux filtres passe-tout à phase aléatoire (réponses différentes H_1 et H_2) sont utilisés pour obtenir trois composantes (D_1, D''_2 et D''_3) décorrélées les unes des autres.

D'après le paragraphe 5.1.2.2, le flux de paramètres quantifiés issus d'un module d'ACP 2D en sous-bandes est transmis à un débit moyen estimé à 3 kbps (*cf.* paragraphe 5.1.2.2) avec une résolution temps-fréquence de l'ACP fixée par N=4096 échantillons et K_b=20 sous-bandes de fréquences. Ainsi, on déduit que le débit de paramètres issus d'une ACP 3D sera au maximum le double de celui obtenu avec une ACP 2D, soit 6 kbps pour la même résolution temps-fréquence d'analyse.

5.2.2 Réduction du nombre de dimensions par séparation selon le plan médian

La méthode de codage audio paramétrique (basée sur l'ACP) des signaux au format 5.0 vise dans un premier temps à scinder la scène sonore originale en paires et/ou triplets de signaux. La séparation des canaux d'un signal multicanal au format 5.0 selon le plan médian (plan qui sépare la gauche de la droite pour l'auditeur, *cf.* **Figure 1.1**) constitue une approche similaire à celle proposée dans la recommandation UIT-R BS.775-1 [UIT775]. En effet, cette recommandation préconise des équations d'encodage *i.e.* matriçage ou *downmix*, d'un signal au format 5.1 pour assurer la compatibilité des signaux multicanaux avec les systèmes de diffusion stéréo et mono (*cf.* Annexe C.1). A la différence des équations de matriçage établies par l'UIT-R [UIT775] qui sont constantes au cours du temps, la méthode basée sur l'ACP se base sur la statistique des signaux, au travers de l'estimation de la covariance des canaux, pour définir un matriçage adaptatif qui compacte au maximum l'énergie des signaux dans la composante principale résultante (au sens de l'EQM d'après le paragraphe 2.3.2.2).

5.2.2.1 Encodeur multicanal paramétrique basé sur l'ACP avec une image stéréo cohérente

L'intérêt principal d'utiliser une séparation des canaux selon le plan médian est de permettre la génération d'un signal stéréo, constitué des canaux dominants issus des deux modules d'ACP 3D, dont l'image sonore est cohérente avec le signal multicanal d'origine.

L'ACP 3D des canaux relatifs à l'image sonore provenant de la gauche de l'auditeur *i.e.* le triplet de canaux (L,C,Ls), délivre une composante principale $D_1^{\,L}$ (*cf.* **Figure 5.18**) qui correspond à l'information principale qui provient de la gauche de l'auditeur. De la même manière, la composante principale $D_1^{\,R}$ extraite par ACP 3D du triplet (R,C,Rs) correspond à l'information principale provenant de la droite de l'auditeur. Ainsi le signal stéréo ($D_1^{\,L}$, $D_1^{\,R}$) constitue un matriçage adaptatif des canaux du signal au format 5.0 (le canal basses fréquences LFE n'étant pas considéré ici) en respectant l'image spatiale originale gauche-droite séparée par le plan médian. Naturellement, la profondeur de l'image spatiale originale avant-arrière ne peut être conservée avec cette séparation des canaux ainsi matricés. Finalement, le signal original au format 5.0 peut être représenté par un signal stéréo et un flux de paramètres extraits par chaque module d'ACP 3D *i.e.* les angles de rotation (α,β) et les $RCPA_{12}$ et $RCPA_{13}$ en sous-bandes. Le signal au format 5.0 peut donc être encodé à un débit équivalent au débit du codeur audio stéréophonique additionné au débit du flux de paramètres codés et quantifiés, soit le $Débit_1$ présenté sur la **Figure 5.18**.

Une orientation plus bas-débit du codeur multicanal paramétrique consiste à ajouter un étage de compression qui correspond au coin inférieur droit en grisé de la **Figure 5.18**. L'ACP 2D du signal stéréo ($D_1^{\,L}$, $D_1^{\,R}$) extrait par double ACP 3D délivre alors une composante principale $D_1^{\,LR}$ représentative de l'information dominante de la scène sonore originale. Ainsi, le signal original au format 5.0 peut être représenté par un signal mono et un flux de paramètres extraits par chaque module d'ACP 3D et un flux de paramètres extraits par le module d'ACP 2D *i.e.* les angles de rotation (α^{LR}) en sous-bandes et le $RCPA^{LR}_{12}$. Le signal au format 5.0 peut donc être encodé à un débit équivalent au débit du codeur audio

monophonique additionné au débit du flux de paramètres codés et quantifiés, soit le $Débit_2$ présenté sur la **Figure 5.18**.

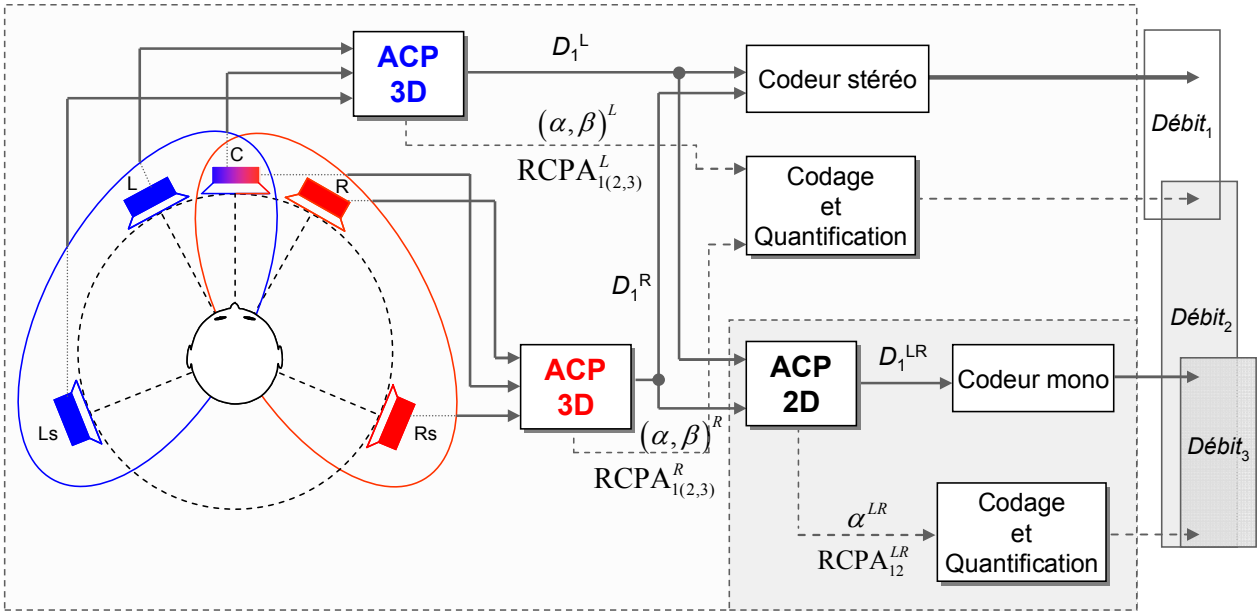

Figure 5.18 Codage audio multicanal paramétrique basé sur l'ACP 3D appliquée aux triplets de canaux gauche et droit du signal au format 5.0.

Enfin, une dernière alternative bas débit qui assure uniquement une compatibilité mono et stéréo consiste à transmettre le flux mono encodé par le codeur audio mono et le débit du flux de paramètres issus du module d'ACP 2D, soit le $Débit_3$ présenté sur la **Figure 5.18**.

5.2.2.2 Décodeurs multicanaux paramétriques basés sur l'ACP

En considérant le flux audio encodé par un codeur audio mono ou stéréo associé à un flux de paramètres spatiaux comme indiqué par le schéma de principe de la **Figure 5.18**, les décodeurs correspondants sont présentés à la **Figure 5.19**.

Dans le cas où le décodeur audio est monophonique (orientation plus bas débit : $Débit_2$ et $Débit_3$), deux possibilités sont envisageables au regard de la quantité de paramètres transmis. Les paramètres décodés et déquantifiés peuvent être relatifs à la reconstruction :

- d'un signal stéréo par ACP 2D inverse ($Débit_3$),
- d'un signal 5.0 à partir du signal stéréo reconstruit et de deux modules d'ACP 3D inverse ($Débit_2$).

Dans le cas où le décodeur audio est stéréophonique ($Débit_1$), les paramètres décodés et déquantifiés permettent la reconstruction d'un signal 5.0 à partir du signal stéréo décodé et de deux modules d'ACP 3D inverse.

Ainsi, un tel système de transmission laisse envisager la possibilité de reconstruire une scène sonore avec une qualité audio et spatiale fonctions du débit (variable) des informations auxiliaires (paramètres énergétiques et spatiaux).

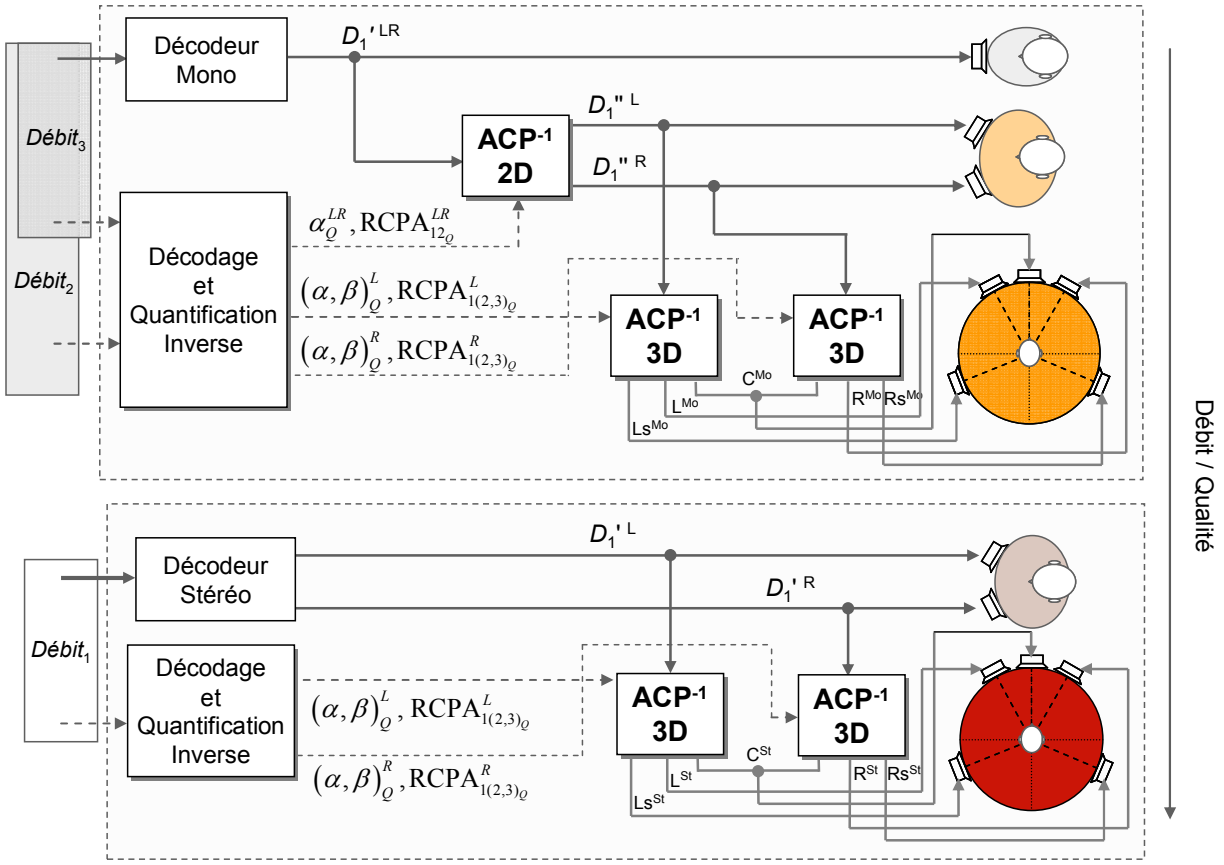

Figure 5.19 Décodeur audio multicanal basé sur un décodeur audio mono ou stéréo et sur le flux de paramètres utiles à la synthèse de signaux résiduels puis aux matriçages adaptatifs inverses.

5.3 Conclusion et perspectives

Nous avons introduit dans ce chapitre une nouvelle méthode de codage audio multicanal compatible avec une transmission à bas débit puisqu'elle repose sur une décomposition paramétrique de la covariance des canaux. En conclusion à ce chapitre, nous proposons d'une part de discuter l'intérêt de cette méthode par rapport aux méthodes existantes et d'autre part de donner des éléments pour améliorer ses performances actuelles.

5.3.1 Discussion : comparaison des approches

Les méthodes de codage audio paramétrique (BCC et PS), présentées au paragraphe 2.2.3, exploitent les propriétés de notre perception spatiale du son avec l'extraction de paramètres liés aux mécanismes de la localisation auditive. Les différences de temps et d'intensité entre les canaux d'origine définissent un modèle de mélange convolutif (*cf.* paragraphe 4.5.1) pour représenter les signaux stéréo et multicanaux. Notre modèle, introduit au paragraphe 4.2.1, devrait donc intégrer ces différences inter-canal de temps pour complètement correspondre au modèle utilisé par les méthodes existantes. Cependant, la technologie référence en la matière *i.e.* MPEG *surround* (*cf.* paragraphe 2.3.3), qui se base sur les procédés BCC et PS, ne juge pas indispensable de prendre en considération ces différences de temps entre les canaux dans le sens où ces paramètres n'améliorent pas les performances actuelles de cette technologie [ISON8324] pour la compression des signaux au format 5.1. Néanmoins cette remarque n'est pas valable dans le cas de signaux stéréo pour lesquels les différences de temps entre les canaux sont plus fréquemment détectées.

Une autre différence entre les schémas de codage des procédés BCC/PS et notre méthode repose sur le processus de matriçage qui délivre le signal utile à la transmission du contenu audio basique. Les procédés BCC et PS génèrent un signal somme (mono ou stéréo) qui représente le signal multicanal original en respectant le critère de conservation de l'énergie. Comme nous l'avons indiqué au paragraphe 2.2.3.1, les différences inter-canal de temps et d'intensité jouent un rôle important pour éviter les pertes d'énergie souvent provoquées par la simple somme des canaux. Notre méthode de codage repose sur un matriçage adaptatif qui maximise l'énergie du point de vue de l'EQM (*cf.* paragraphe 2.3.2.2) et génère un signal représentatif de l'information corrélée entre les canaux de départ *i.e.* dominante d'après le paragraphe 4.2.2.3. Par conséquent, en prenant en considération un modèle de mélange convolutif, une solution hybride serait d'utiliser un matriçage adaptatif pour compacter au maximum l'énergie des canaux tout en tenant compte des différences de temps d'arrivée des sources directionnelles entre les différents canaux.

Du point de vue de la synthèse spatiale proprement dite, la méthode de codage paramétrique basée sur l'ACP ne nécessite pas la transmission d'indices de cohérence *i.e.* corrélation, à la manière des procédés BCC et PS. En réalité, le matriçage adaptatif délivre des signaux très faiblement corrélés (*cf.* propriétés de l'ACP présentées au paragraphe 4.3.1). Par conséquent, l'opération de décodage repose sur l'utilisation d'un filtre de décorrélation pour réduire la corrélation entre la composante principale décodée et l'ambiance finalement synthétisée à partir des paramètres énergétiques. Ainsi, l'ACP inverse est apte à générer un signal stéréo (ou à trois canaux dans le cas de l'ACP tridimensionnelle) dont la corrélation des canaux appromixe celle des canaux d'origine.

Le procédé PS utilise, au décodage, une décomposition en valeurs propres de la matrice de covariance qui est définie à partir des paramètres énergétiques et de corrélation (*cf.* paragraphe 2.2.3.2). Cependant, à la différence du procédé PS, notre méthode utilise un décodeur qui effectue les opérations inverses de celles réalisées par l'encodeur. En outre, nous utilisons un jeu de paramètres (quantifiés) qui sont directement liés à la position perçue des sources directionnelles (*cf.* paragraphes 4.3.3.1 et 4.3.4.1) et au niveau d'énergie des composantes résiduelles ou d'ambiances. Les angles de rotation permettent finalement (au décodeur) le positionnement des sources directionnelles c'est-à-dire la reconstruction de signaux avec des différences d'intensité qui approximent celles des canaux originaux.

Enfin, un autre intérêt de cette méthode de codage paramétrique basée sur l'ACP est de pouvoir dissocier les sources directionnelles dominantes des sources secondaires et de l'information d'ambiance du signal original. Cette hiérarchisation des composantes des canaux (*cf.* paragraphe 4.2.2.3) peut être utilisée par un système de compression où une partie ou la totalité des canaux résiduels ou d'ambiances sont codés et transmis. Cette possibilité, décrite au paragraphe 5.3.2.2, laisse alors la possibilité d'améliorer la qualité de la reconstruction jusqu'à la transparence (sans la prise en compte des dégradations propres au codage audio classique) moyennant un débit variable.

5.3.2 Perspectives

5.3.2.1 Basculement entre différents types de filtres décorrélateurs

La méthode de codage paramétrique basée sur l'ACP sera en mesure de restituer au mieux l'impression spatiale de la scène originale sans pour autant colorer les sources sonores (effet de filtrage en peigne décrit en Annexe C.2.3) si la technique de décorrélation employée est adaptée au contenu du signal à coder. En effet, si l'analyse faite par l'encodeur est capable de déceler la présence de réverbération dans le signal original alors cette information (paramètre de type indice de réverbération) transmise au décodeur permettrait au décodeur de choisir le type de filtre le plus adapté à la nature du signal (*cf.* **Figure 5.20**).

Une multitude de techniques de filtrage de type réverbération existe avec différentes solutions et niveaux de complexité. Les travaux de Moorer, dans [MOO79], ou de Jot, dans [JOT92], présentent notamment des solutions de filtrage de type réverbération basé sur des filtres à réponse impulsionnelle finie (RIF) ou infinie (RII) établie par la combinaison de filtres en peigne et de filtres passe-tout (*cf.* Annexe C.2.3).

Cependant, le processus d'identification de la réverbération dans un signal mono ou multicanal n'est pas trivial. Les expériences menées par Baskind et *al.* dans [BAS03] permettent d'estimer le temps de réverbération (défini au paragraphe 1.1.2) à partir des périodes d'inactivité vocale d'un signal de parole réverbéré c'est-à-dire lorsque l'effet de salle prédomine sur le son direct. Cependant, l'utilisation de cette méthode sur un signal stéréo de musique constitué de multiples sources directionnelles reste délicate puisque dans un contexte de codage (d'une portion glissante de signal à une autre), il serait nécessaire de pouvoir détecter la présence de réverbération à tout instant.

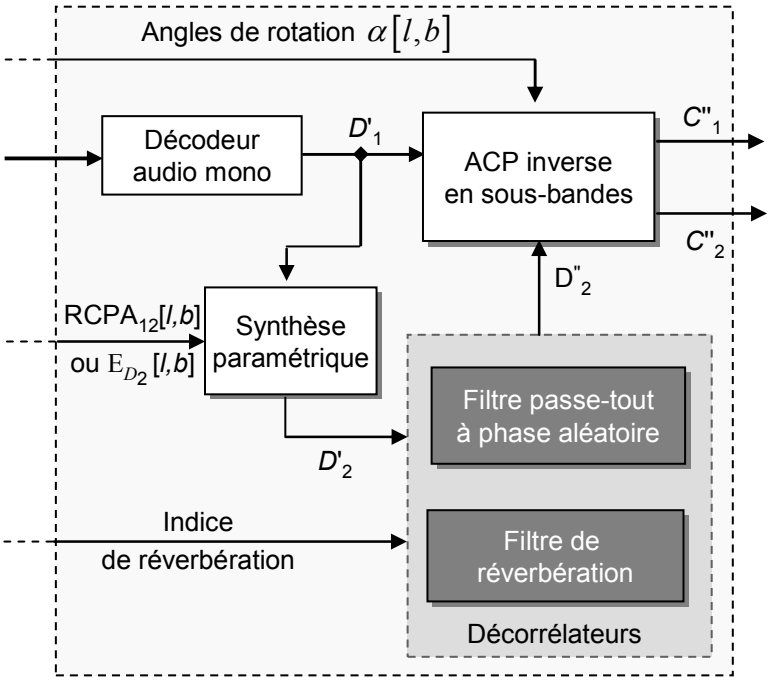

Figure 5.20 Décodeur stéréo paramétrique basé sur l'ACP en sous-bandes. Outre les paramètres énergétiques et les angles de rotation propres à l'ACP, un indice de réverbération reçu par le décodeur permettrait de choisir le type de filtre décorrélateur à utiliser pour synthétiser la composante ambiance D''_2 la plus adaptée au contenu du signal stéréo original.

5.3.2.2 Débit scalable pour la stéréo haute qualité

L'extension du procédé de codage bas débit basé sur l'ACP en un procédé de codage avec un débit scalable (graduel) permettrait d'obtenir plusieurs niveaux de qualité audio et spatiale. En d'autres termes, la transmission d'informations supplémentaires avec un débit global plus élevé permettrait d'obtenir une meilleure reconstruction de la scène sonore originale.

Le schéma de principe du décodeur stéréo paramétrique présenté à la **Figure 5.2** permet déjà d'obtenir divers niveaux de qualité audio/spatiale intermédiaire. En effet, au regard de la quantité de paramètres transmis *i.e.* les angles de rotation et éventuellement les paramètres énergétiques, le décodeur est en mesure de reconstituer un signal stéréo perceptuellement plus ou moins proche du signal stéréo original. La **Figure 5.21** présente plusieurs solutions de décodage stéréo paramétrique à partir du flux audio de la composante principale décodée D'_1 et du flux de paramètres à débit scalable.

La composante principale décodée constitue un signal monophonique représentatif de l'information relative aux sources directionnelles dominantes et à une partie de l'ambiance originale qui coïncide spatialement avec ces sources directionnelles.

Au moyen des angles de rotation en sous-bandes $\alpha[l,b]$ et de la composante principale décodée D'_1, un signal stéréophonique (C'_1, C'_2), dont la position spatiale des sources directionnelles est restituée, peut être reconstruit par ACP inverse en sous-bandes. Cependant, l'ambiance sonore ne peut qu'être partiellement reconstruite.

Les paramètres énergétiques ($RCPA_{12}[l,b]$ ou $E_{D_2}[l,b]$) appliqués à la composante principale décodée D'_1 permettent la synthèse d'une ambiance sonore D''_2 décorrélée de la composante principale par filtrage passe-tout décorrélateur. Ensuite, au moyen des angles de rotation, l'ACP inverse en sous-bandes des composantes principale et ambiance permet la reconstruction d'un signal stéréo (C''_1, C''_2) de qualité intermédiaire : l'évaluation des qualités audio et spatiale délivrées par cette méthode de codage est présentée au paragraphe 5.1.3.

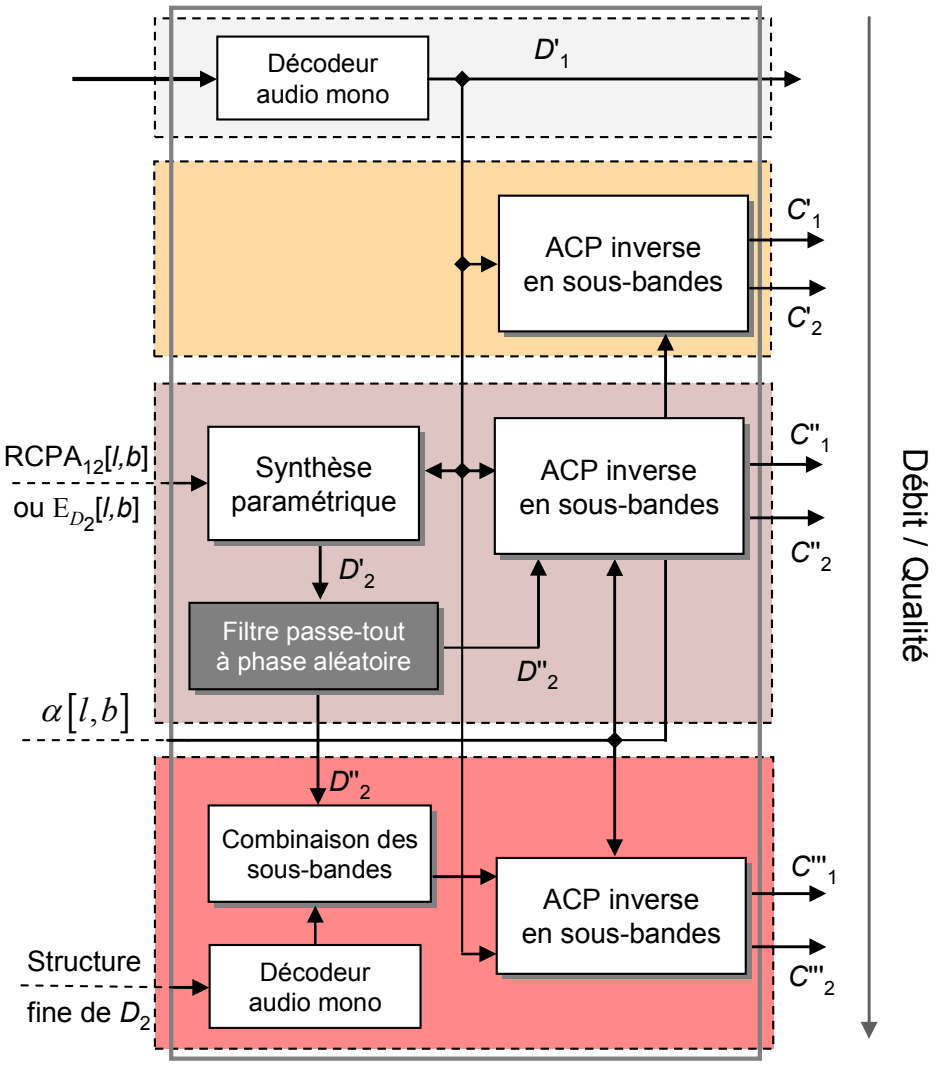

Figure 5.21 Schéma du principe du décodeur stéréo paramétrique basé sur l'ACP en sous-bandes avec un débit de paramètres scalable.

Une autre variante du procédé de décodage, présentée à la **Figure 5.21**, consiste à utiliser la « structure fine » de l'ambiance sonore D_2 codée et transmise par l'encodeur à un débit globalement plus élevé. La structure fine de la composante ambiance correspond à un codage monophonique classique (perceptuel par transformée par exemple) d'au moins une partie du spectre (en sous-bandes) du signal d'ambiance. Par conséquent, dans le cas d'une réception

partielle de la structure fine (quelques sous-bandes de fréquences particulières), l'ambiance synthétique (D''_2), délivrée par les modules de synthèse paramétrique/filtrage passe-tout décorrélateur, peut être utilisée pour compléter les sous-bandes absentes de la structure fine (*cf.* **Figure 5.21**).

La transmission de la structure fine de l'ambiance D_2 offre la possibilité de s'approcher d'une reconstruction asymptotiquement parfaite du signal stéréo original *i.e.* seul le codage audio des composantes principales et de la structure fine ainsi que la quantification des paramètres (*cf.* paragraphe 5.1.2.2) peuvent dégrader les signaux.

Un ordre de transmission de la structure fine de la composante ambiance peut être mis en place de façon à hiérarchiser les informations à transmettre. En effet, certaines bandes de fréquences de la structure fine peuvent être transmises en priorité. Nous faisons référence aux méthodes de transmission décrites dans [KOV04] qui préconisent un ordonnancement des bandes de fréquences d'une enveloppe spectrale quantifiée. Cet ordonnancement peut être prédéfini (ordre croissant ou autre) ou être fonction de l'énergie des sous-bandes (à l'aide du paramètre $RCPA_{12}$ par exemple) ou de l'importance perceptuelle des sous-bandes ou encore être fonction de la corrélation des signaux originaux. Cet ordonnancement peut également être une combinaison de certains de ces critères. L'intérêt d'un ordonnancement des sous-bandes selon un critère énergétique est de pouvoir ainsi réduire l'erreur quadratique moyenne entre les signaux originaux et reconstruits. En utilisant un critère perceptuel, à l'aide d'un modèle psychoacoustique (*cf.* Annexe B) par exemple, l'intérêt est de pouvoir minimiser la dégradation perceptuelle de la reconstruction *i.e.* l'ambiance (impression spatiale) reconstruite sera d'autant plus proche de l'originale, d'un point de vue subjectif, que les sous-bandes de la structure fine perceptuellement importantes seront transmises en priorité.

Finalement, les performances d'un tel système de codage peuvent être directement reliées aux débits des paramètres (spatiaux et énergétiques) et de la structure fine utiles à la reconstruction du signal stéréo. En outre, la résolution temps-fréquence utilisée pour la découpe des signaux (en temps et en fréquence) influence également la quantité d'informations transmises et la qualité de la reconstruction.

5.3.2.3 Extension de la méthode paramétrique au codage d'autres signaux multicanaux

Une perspective intéressante serait d'étudier la généralisation de notre approche paramétrique pour réaliser une ACP de dimension M et l'utiliser dans un contexte de codage multicanal. Etant donné que le nombre de rotations nécessaires à l'ACP de dimension M augmente de manière non-linéaire (6 rotations pour M=4 et 10 rotations pour M=5) [GOL96], il conviendrait de vérifier l'intérêt d'une méthode de codage paramétrique basée sur l'ACP de dimension M sans rejeter la contrainte de débit. En outre, l'étude consisterait à dégager une interprétation physique pour déterminer les angles pertinents en se basant sur des critères objectifs (concentration d'énergie par exemple) et subjectifs (importance subjective des canaux résiduels issus de l'ACP?).

Cependant, notre approche permet déjà l'extension de la méthode au codage d'autres signaux multicanaux en combinant l'ACP bi- et tri-dimensionnelle. Nous avons présenté le principe de l'extension de notre méthode de codage stéréo paramétrique pour la compression des signaux au format 5.0 (*cf.* paragraphe 5.2). Il en va de même pour coder tous types de signaux multicanaux comme par exemple les signaux aux formats 7.1, 10.1, 22.2, etc. En se basant sur la géométrie du système de reproduction sonore (positions des haut-parleurs par rapport à l'auditeur), la méthode consiste à analyser et traiter les canaux par paires ou triplets selon un axe de symétrie comme le plan médian pour conserver la compatibilité stéréo par exemple. De façon générale, la méthode peut être utilisée pour rendre compatible les formats multicanaux par exemple pour la conversion d'un signal 7.1 en un signal 5.1/stéréo/mono et inversement. Enfin, il serait intéressant d'étudier les performances de la méthode pour le codage de signaux audio spatialisés issus d'une prise de son ambisonique (*cf.* paragraphe 1.2.2.2) dont la représentation de base s'appuie sur une décomposition du champ acoustique selon les trois axes de l'espace (compatibilité avec l'ACP tridimensionnelle).

Conclusion

Contributions de la thèse

Fondés sur les orientations de la thèse définies en introduction à ce document, ces travaux de thèse ont, d'une part, cherché à améliorer les procédés de codage audio paramétrique existants et, d'autre part, permis l'élaboration d'un modèle à partir duquel nous avons dérivé une nouvelle méthode de codage paramétrique.

Pour améliorer les méthodes de codage audio paramétrique, nous avons orienté nos recherches vers une extraction de paramètres de spatialisation adaptée au contenu fréquentiel des signaux (*cf.* chapitre 3). Autrement dit, plutôt qu'extraire les paramètres de spatialisation avec une résolution fréquentielle fixe, nous avons cherché à définir un paritionnement temps-fréquence adapté aux composantes des signaux. Pour cela, nous avons abordé deux approches qui reposent sur un processus automatique de segmentation dans le plan temps-fréquence. La première approche, basée sur ce processus, a consisté à déceler les motifs spectraux caractéristiques de l'information spatiale contenue dans les canaux d'un signal stéréophonique (différences inter-canal de temps ou d'intensité). Cependant la comparaison des RTF segmentées, même issues de canaux très corrélés, ne nous a pas permis de caractériser les similarités/différences spectrales de signaux stéréo synthétiques ou issus d'une prise de son naturelle. La seconde approche que nous avons abordée a consisté à utiliser le processus de segmentation pour localiser les zones temps-fréquence où l'erreur de reconstruction générée par un codage/décodage paramétrique (de type BCC) n'est pas négligeable (audible) à l'écoute des signaux reconstruits. Cependant, d'après nos expérimentations, l'erreur de reconstruction établie par différence entre les signaux originaux et reconstruits est répartie énergétiquement sur l'ensemble du plan temps-fréquence. Par conséquent, la RTF d'une erreur de reconstruction ainsi établie ne reflète pas d'une manière fiable les dégradations qui peuvent être perçues à l'oreille. Finalement, l'identification de zones temps-fréquence critiques au moyen du processus de segmentation ne nous a pas permis d'améliorer les procédés de codage paramétrique existants qui s'appuient sur la perception auditive et non sur une représentation visuelle des signaux.

Etant donné que le modèle utilisé par le processus de segmentation temps-fréquence est très général, nous avons ensuite orienté nos travaux vers l'élaboration d'un modèle de mélange adapté aux signaux audio multicanaux (*cf.* chapitre 4). Pour cela, nous avons défini les composantes de tels signaux en se basant sur les principes de prise de son naturelle et de mixage artificiel. Nous avons finalement défini un modèle de mélange instantané des canaux, constitués de sources directionnelles et d'ambiances sonores. L'analyse en temps et en sous-bandes de fréquences de la distribution des valeurs propres de signaux stéréo suivant ce modèle nous a permis de décrire la hiérarchisation des composantes originales. Nous avons

notamment relevé l'influence de la position des sources (azimut dans le plan horizontal) et du pouvoir plus discriminant de l'analyse en sous-bandes comparée à l'analyse en temps. Puisque la distribution des valeurs propres caractérise la puissance des signaux projetés sur la base des vecteurs propres correspondants, nous avons ensuite considéré l'ACP/KLT de signaux à deux et trois composantes suivant notre modèle. Nous avons évalué les performances de l'ACP bi- et tri-dimensionnelle en termes de concentration de l'énergie initiale au sein des composantes transformées. Il en ressort notamment que la nature (nombre de sources par exemple) et la corrélation des canaux ainsi que le type de traitement (en temps ou en sous-bandes de fréquences) influencent considérablement la hiérarchisation des composantes au sein des canaux transformés. Pour répondre à l'objectif principal de compression multicanale, nous avons fait le choix d'utiliser une approche paramétrique pour réaliser l'ACP bi- et tridimensionnelle. Nous avons donc proposé une méthode pour extraire un ou plusieurs angles de rotation utiles à l'ACP et compatibles avec un schéma de codage classique utilisant une analyse/synthèse par recouvrement-addition (méthode OLA). Une interprétation physique des angles de rotation a été présentée d'après la connaissance du système de reproduction sonore : le ou les angles de rotation peuvent être convertis en un azimut lié à la position de la source directionnelle dominante contenue dans le signal multicanal analysé.

En se basant sur l'hypothèse que cette hiérarchisation des composantes puisse nous permettre d'omettre la transmsission de canaux résiduels ou d'ambiance, nous avons utilisé cette décomposition paramétrique de la covariance au sein d'une nouvelle méthode de codage paramétrique (*cf.* chapitre 5). Un premier test subjectif nous a permis de mesurer la faible influence de la phase de la composante ambiance issue d'une ACP bidimensionnelle. Par conséquent, nous avons mis en œuvre une méthode de codage stéréo paramétrique qui repose à la fois sur la concentration de l'information dominante dans la composante principale et sur l'extraction de paramètres (énergétiques) utiles à la synthèse de l'ambiance. L'extraction des paramètres énergétiques et spatiaux (caractéristiques de l'ACP) a été suivie d'une étape de codage et de quantification suivant des critères perceptuels éventuellement adaptables au système de reproduction sonore. Finalement, nous avons donné les détails de notre implémentation de cette méthode de codage paramétrique ainsi que les résultats de son évaluation subjective. Grâce à l'approche paramétrique utilisée pour réaliser l'ACP tridimensionnelle, nous avons également décrit le principe de l'extension de la méthode de codage stéréo pour le codage de signaux au format 5.0. Même si nous n'avons pas encore évalué les performances de son implémentation, le principe de notre méthode de codage multicanal offre la compatibilité avec les systèmes de reproduction mono et stéréo grâce à un encodage hiérarchique (en arbre) de la scène sonore pour différents débits qui pourraient être mis en correspondance avec la qualité audio et spatiale de la reconstruction finale.

Perspectives de recherche

Concernant l'utilisation du processus de segmentation temps-fréquence, la marge de manœuvre reste encore importante. En effet, d'après nos expérimentations, nous suggérons notamment d'approfondir le principe du processus pour permettre l'analyse directe d'un signal multicanal. Autrement dit, la comparaison des motifs spectraux pourrait être réalisée au cœur d'un processus de segmentation multicanal et non à l'issue de segmentations monocanales réalisées indépendamment les unes des autres.

Nous avons établi et utilisé un modèle de mélange instantané pour représenter les signaux audio multicanaux. Bien évidemment, ce modèle de mélange ne procure pas une représentation exhaustive des signaux multicanaux notamment issus d'une prise de son non-coïncidente (différences de temps introduites par les microphones). Par conséquent, il conviendrait de remplacer cette modélisation par un mélange convolutif plus représentatif des signaux audio issus de prises de son naturelles. Néanmoins, certains types de signaux nécessitent une attention particulière pour pouvoir être modélisés à des fins de compression. C'est le cas des signaux qui présentent une faible corrélation inter-canal *i.e.*

applaudissements par exemple, puisque même modélisés par un mélange convolutif, ces signaux nécéssitent une reconstruction très précise de leurs formes d'ondes temporelle et spectrale. En effet, le standard MPEG *surround* est actuellement poussé à ses limites pour permettre une reconstruction subjective convenable de ce type de signaux. Autrement dit, la transmission des paramètres de spatialisation est insuffisante et doit être accompagnée de canaux résiduels pour assurer une bonne qualité de reconstruction.

Pour répondre à la problématique générale du codage audio multicanal, nous avons choisi d'utiliser une méthode paramétrique pour concentrer l'énergie des canaux d'origine en un minimum de canaux. L'état actuel de nos travaux permet le matriçage adaptatif (par ACP) de deux ou trois composantes simultanément. Par conséquent, une perspective inéressante serait de pouvoir matricer directement un nombre quelconque de canaux en généralisant notre approche. Sous l'hypothèse que cette généralisation (espace de dimension M) soit plausible (stable avec la synthèse OLA) et physiquement interprétable, il conviendrait ensuite d'évaluer l'importance subjective des canaux résiduels ou d'ambiances. Combien de canaux perceptuellement importants doivent être transmis par le procédé de codage tout en maintenant la contrainte de compatibilité avec les systèmes audio classiques? Autrement dit, combien de canaux perceptuellement non significatifs, c'est-à-dire modélisables au moyen de paramètres énergétiques, seraient issus d'un matriçage adaptatif à l'ordre M? Nos expérimentations, basées sur notre modèle de canaux corrélés, ont toutes fait l'hypothèse que ces canaux résiduels soient subjectivement non significatifs. C'est d'ailleurs pour cette raison que notre implémentation actuelle de la méthode de codage des signaux au format 5.0 est moins performante pour compresser des signaux faiblement corrélés. Dans ce cas, l'approche par filtrage décorrélateur est insuffisante et l'importance subjective des canaux résiduels ou d'ambiances doit être prise en considération. C'est pourquoi, pour accroître la qualité finale de la reconstruction, il serait nécessaire de lever cette hypothèse de forte corrélation des canaux originaux pour associer aux paramètres auxiliaires la structure fine d'un ou plusieurs canaux résiduels. Finalement, les perspectives de recherche directement liées à notre approche consistent d'une part à généraliser la décomposition paramétrique de la covariance à l'odre M puis à évaluer l'importance subjective des canaux résiduels au regard de la corrélation des canaux originaux. Cette généralisation de notre approche avec la transmission d'informations auxiliaires devra également tenir compte de la contrainte forte du débit.

ANNEXES

A. Représentation et Segmentation dans le plan Temps-Fréquence

A.1 Résolution de la RTF

A.1.1 Transformée de Fourier à Court Terme

La Transformée de Fourier à Court Terme (TFCT) du signal est une collection de Transformée de Fourier (TF) de portions du signal analysé. La dimension temporelle est introduite dans l'analyse fréquentielle en multipliant le signal par des versions décalées d'une fenêtre glissante.

Un signal $x[n]$ à N_T échantillons discrets (n=1,…,N_T) est tronqué en L portions de signal par une fenêtre glissante à N échantillons discrets $w[n]$ d'énergie unité et centrée en zéro. La TFCT du signal $x[n]$, $F_x[l,k]$, aux instants l=1,…,L (indice des fenêtres glissantes) et aux fréquences d'indice k est donnée par (voir l'illustration à la **Figure A.1**) :

$$F_x[l,k] = \sum_{n=l-\frac{N-1}{2}}^{l+\frac{N-1}{2}} x[n] \times w[n-l] \times e^{-2i\pi k \frac{n}{N+Z}} . \qquad \text{(A.1)}$$

Les performances d'estimation fréquentielle (pouvoir séparateur par exemple) découlent du choix de la longueur et du type de fenêtre d'analyse spectrale $w[n]$ (rectangulaire, Hanning, Blackman, Gaussienne, Hamming, etc.), *cf.* Annexe A.1.2. De plus le recouvrement des fenêtres et le nombre de zéros Z en complément du signal (portion de signal) pour calculer la TF ont également une influence très importante.

La TFCT du signal x est une représentation à valeurs complexes qui est inversible (TFCT inverse). On appelle spectrogramme son module carré :

$$S_{F_x}[l,k] = \left| F_x[l,k] \right|^2 . \qquad \text{(A.2)}$$

Le spectrogramme donne alors une estimation de la répartition énergétique du signal x le long des axes temporel et fréquentiel (*cf.* **Figure A.2**).

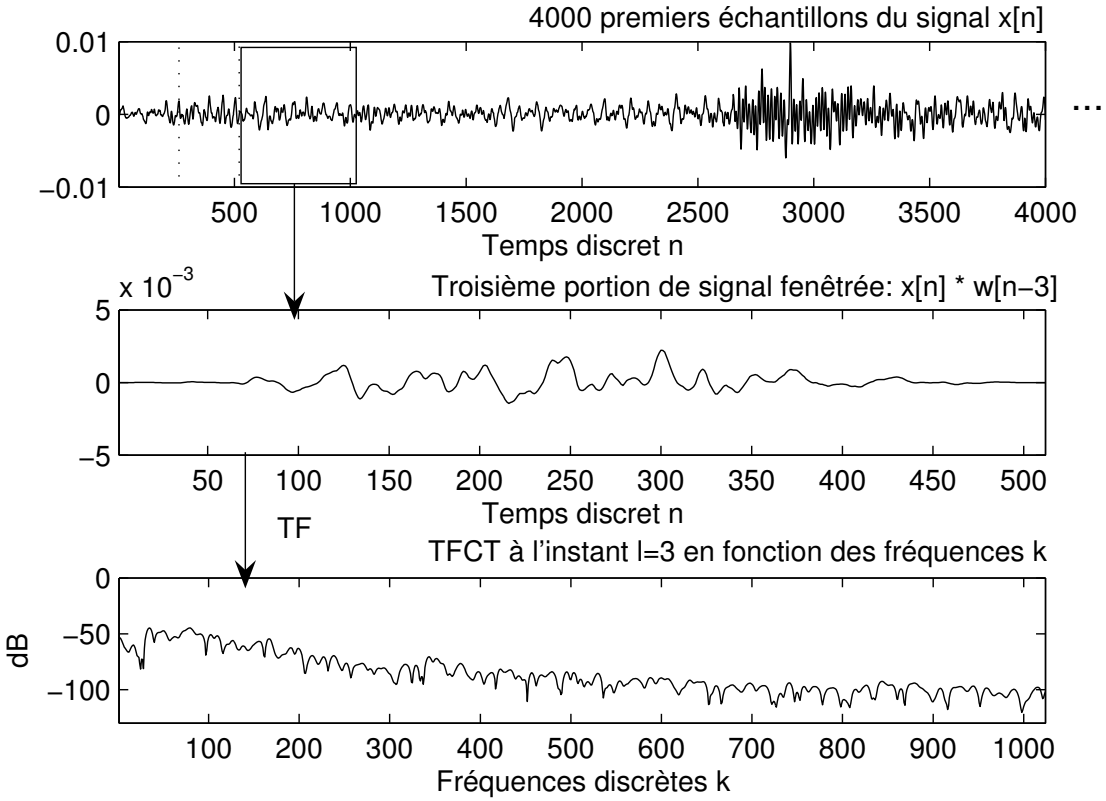

Figure A.1 **TFCT d'un signal de piano *x*[*n*]** à (N_T=4000 échantillons) fenêtré par une fenêtre de Hanning à N=512 points avec un recouvrement de 50% entre les fenêtres glissantes. La TFCT à l'instant l=3 et aux fréquences k correspond à la TF de la troisième portion de signal obtenue par fenêtrage, TF[x[n].w[n-3]], et complétée par Z=512 zéros.

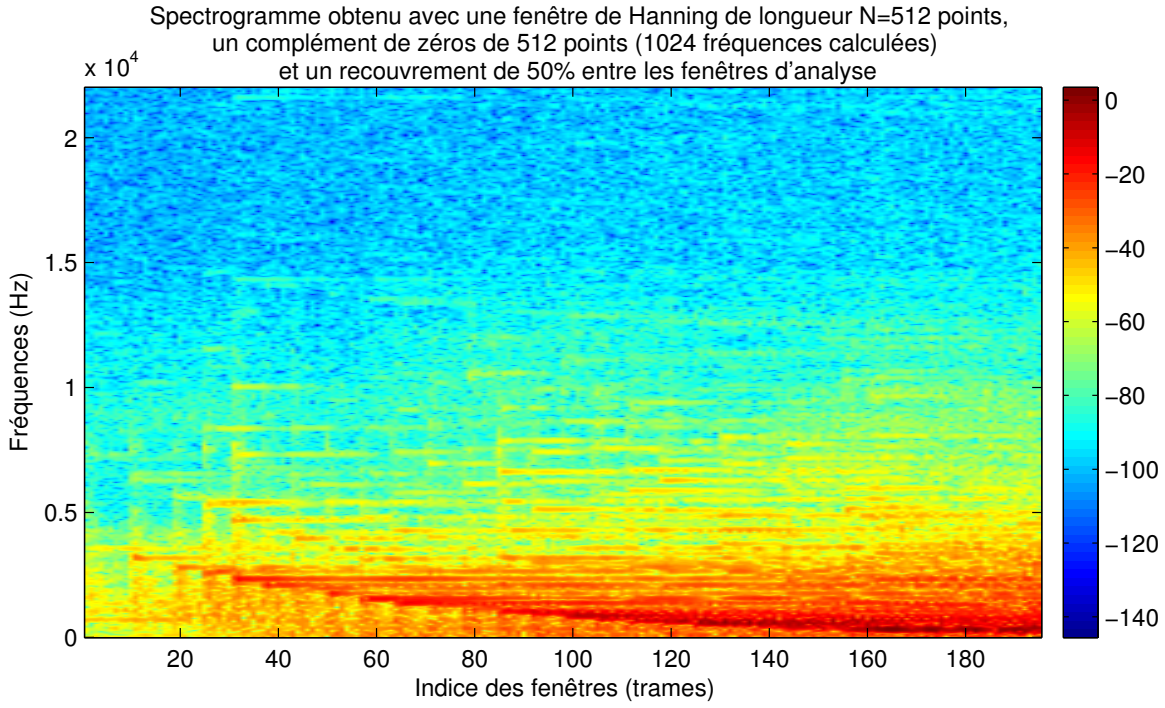

Figure A.2 **Représentation Temps-Fréquence (RTF) ou spectrogramme avec une échelle logarithmique en décibels (dB).** Cette RTF présente l'évolution temporelle des fréquences d'un signal de piano (descente de notes d'une durée de 1,1 s) échantillonné à 44100 Hz. La longueur de la fenêtre d'analyse correspond à une durée de 11,6 ms.

A.1.2 Choix de la fenêtre d'analyse spectrale

Le choix de la fenêtre d'analyse spectrale pour la TFCT (*cf.* équation (A.3)) repose sur un compromis entre la résolution temporelle et la résolution fréquentielle. Le type de fenêtre et sa longueur influence sensiblement la finesse de l'analyse spectrale. En effet, plus la taille de la fenêtre est petite et plus on aura tendance à approximer le spectre du signal analysé. Cependant, la taille de la fenêtre (N) doit être suffisamment petite pour préserver la structure temporelle du signal analysé (transitoires comme les attaques d'un signal percussif par exemple).

Nous traitons ici le cas de différentes fenêtres (*cf.* **Figure A.3**) utilisées en traitement de signal: Rectangulaire, Hanning, Hamming, Blackman, Gaussienne et Kaiser (avec β = 4).

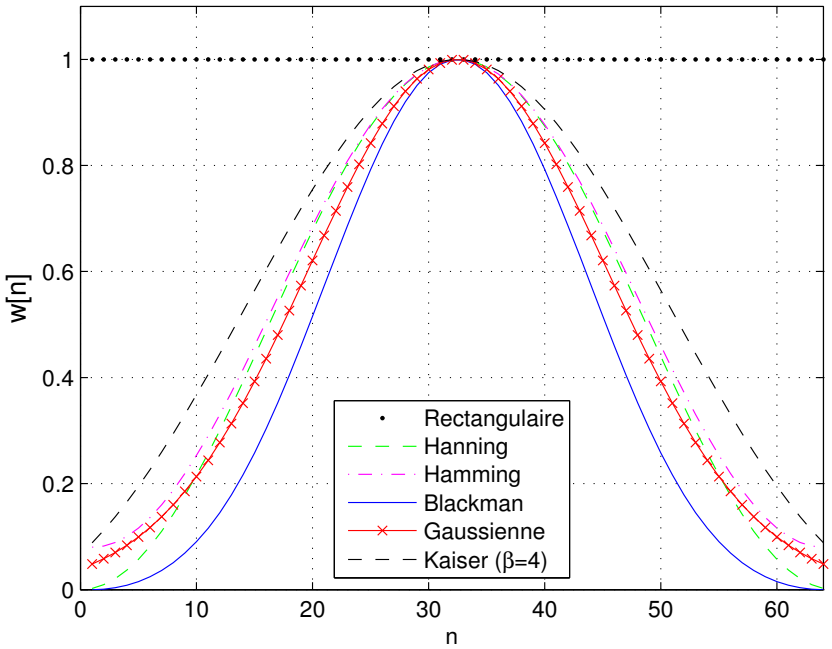

Figure A.3 Représentations temporelles de six fenêtres d'analyse spectrale *w*[*n*] à *N*=64 échantillons : Rectangulaire, Hanning, Hamming, Blackman, Kaiser (β=4) et Gaussienne.

A partir d'une fenêtre $w[n]$ définie dans le domaine temporel, nous définissions sa Transformée de Fourier Discrète (TFD) $F_w[k]$ comme :

$$F_w[k] = TFD\left(w[n]\right) = \frac{1}{N}\sum_{n=1}^{N} w[n] \times e^{-j2\pi k \frac{n}{N+Z}} . \qquad \text{(A.3)}$$

La **Figure A.4** présente les TFD des fenêtres, présentées à la **Figure A.3**, obtenues avec un complément de zéros Z=1984, soit un nombre total de coefficients spectraux calculés égal à N+Z=2048 (prise en compte de la symétrie hermitienne sur la **Figure A.4** soit N/2+1 fréquences positives).

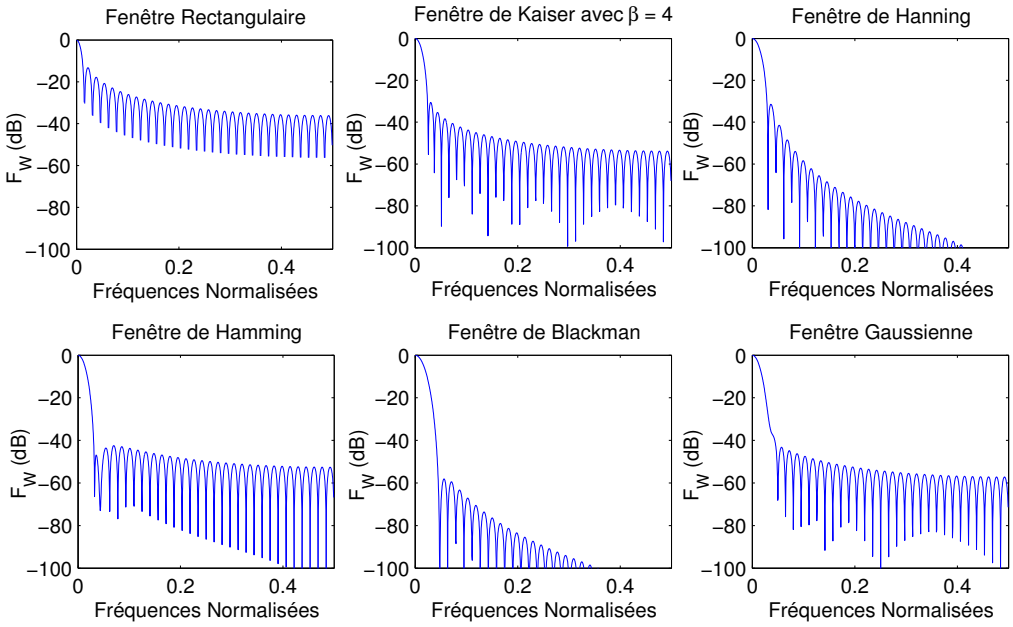

Figure A.4 **Représentations fréquentielles** (TFD($w[n]$) logarithmique en dB) **de six fenêtres d'analyse spectrale** (Rectangulaire, Hanning, Hamming, Blackman, Kaiser, Gaussienne).

D'après les réponses fréquentielles tracées à la **Figure A.4**, on voit que la largeur du lobe principal, l'amplitude du plus grand lobe secondaire ainsi que la pente de ces lobes secondaires diffèrent d'une fenêtre à une autre. Par exemple, la fenêtre rectangulaire possède le lobe principal le plus étroit (qui procure une analyse fine en fréquence) mais le lobe secondaire le moins atténué (-13 dB). On s'aperçoit alors qu'en utilisant une fenêtre avec une faible rupture de pente (en temporel, *i.e.* Hamming, Blackman ou Kaiser), on peut réduire de manière importante l'amplitude des lobes secondaires parasites.

La TFCT d'un signal et par suite le spectrogramme (*cf.* équation (A.1)) est fonction de la fenêtre d'analyse. La multiplication du signal par la fenêtre dans le domaine temporel est équivalente, dans le domaine fréquentiel, à la convolution de la TFD du signal par la TFD de la fenêtre. Pour illustrer l'influence des lobes secondaires, nous considérons un signal purement sinusoïdal x de fréquence f_0 multiple du pas d'échantillonnage et d'amplitude A. L'opération de fenêtrage est alors présentée à la **Figure A.5**. En fenêtrant ce signal, on voit apparaître deux nouvelles fréquences dues aux lobes secondaires caractéristiques de la fenêtre spectrale (cf. **Figure A.4**).

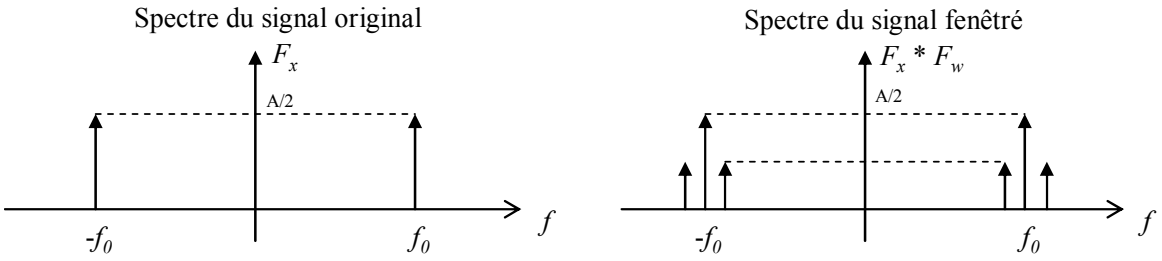

Figure A.5 **Effet du fenêtrage sur le calcul de la TFD d'un signal sinusoïdal** de fréquence f_0 et d'amplitude A.

Pour mesurer la résolution fréquentielle obtenue on utilise la bande équivalente B_{eq} qui est fonction de la largeur du lobe principal de la fenêtre d'analyse (représentation fréquentielle). Cette mesure dépend donc directement du choix de la fenêtre d'analyse (largeur du lobe

différente pour chaque fenêtre) et de la longueur (temporelle) de cette fenêtre puisque B_{eq} est proportionnelle à $1/N$ (voir [DUR99] pour plus de détails).

La résolution fréquentielle des fenêtres d'analyse spectrale considérées est finalement évaluée à partir de la largeur de bande équivalente (B_{eq}) à -3 dB et de l'atténuation des lobes secondaires (*ALS*) avec une longueur de fenêtre (N) variable et une fréquence d'échantillonnage f_S = 44,1 kHz.

Caractéristiques spectrales	Rectangulaire	Hanning	Hamming	Blackman	Kaiser	Gaussienne
B_{eq} (Hz), N=64	603	947	861	1120	775	861
B_{eq} (Hz), N=256	151	237	215	280	194	215
B_{eq} (Hz), N=512	75,36	118	108	140	96,9	108
B_{eq} (Hz), N=1024	37,7	59,2	53,8	69,9	48,4	53,8
ALS (dB),$\forall N$	-13,3	-31,5	-43	-58	-30	-43,3

Tableau A.1 B_{eq} et ALS de six fenêtres d'analyse spectrale (Rectangulaire, Hanning, Hamming, Blackman, Kaiser, Gaussienne).

D'après le **Tableau A.1**, la fenêtre de Hanning, Hamming et la fenêtre Gaussienne proposent le meilleur compromis entre faible largeur du lobe principal et faible rapport d'amplitude entre le lobe principal et secondaire. Par conséquent, ces trois fenêtres peuvent être utilisées pour estimer la densité spectrale du signal par exemple.

En pratique, les procédés de codage audio (musique et parole) utilisent habituellement une fenêtre d'analyse d'une durée égale à 20 ms (soit 882 échantillons à une fréquence d'échantillonnage f_S = 44100 Hz et 960 échantillons à f_S = 48000 Hz) sous l'hypothèse qu'une telle portion de signal audio est stationnaire. La majorité des codeurs audio utilisent des trames temporelles de 1024 échantillons, soit la puissance de deux la plus proche d'une durée de 20 ms. Les codeurs audio évolués utilisent également un outil de basculement de fenêtre (*dynamic window switching* présenté dans [SPO92] et décrit en Annexe B.1.3.2) qui permet d'adapter la longueur de la fenêtre d'analyse à la nature du signal. Le principe consiste à détecter les attaques présentes dans le signal à coder pour éviter le phénomène de pré-écho engendré par le codage d'une portion de signal (contenant une impulsion) de trop grande taille. Ainsi, la taille de la fenêtre d'analyse peut être réduite aux instants critiques où une résolution temporelle très fine est nécessaire pour suivre les variations rapides du signal.

La périodicité des signaux de parole (sons voisés particulièrement) ou de musique rythmique peut apparaître sur la RTF de ces signaux. Ce phénomène apparaît lorsque la taille de la fenêtre d'analyse est inférieure à la période du signal comme le montre la **Figure A.6**. Cette période apparaît dans le domaine fréquentiel comme la fréquence fondamentale du signal de parole de l'ordre de 150 Hz chez l'homme et comprise entre 150 et 450 Hz chez la femme, d'après [BOI87]. L'exemple présenté à la **Figure A.6** illustre le cas d'un signal de voix de femme dont la périodicité, évaluée à T_0=180 échantillons, correspond à une fréquence fondamentale égale à F_0=266 Hz. C'est pour cette raison qu'avec une fenêtre d'analyse de N=128 échantillons, soit $N<T_0$, l'analyse par fenêtres glissantes a tendance à révéler ce phénomène de périodicité temporelle sur la RTF. Ce phénomène serait d'ailleurs encore accentué avec une fenêtre de cette taille ou supérieure (N=256 échantillons) et une voix d'homme dont la fondamentale est plus basse (périodicité de l'ordre de T_0=300 échantillons à f_S=44100 Hz). Cependant, nous ne cherchons pas à représenter (ou extraire) cette information de périodicité qui pourrait perturber l'algorithme de segmentation et engendrer une sur-segmentation de ces motifs spectraux. Par conséquent, une fenêtre d'analyse d'une taille minimale de N=512 échantillons permettra d'éviter ce phénomène si l'on considère le cas

extrême d'une voix d'homme grave (F_0=100 Hz) puisque N=512>T_0=480 échantillons pour f_S=48000 Hz.

Finalement, en tenant compte de la taille des trames habituellement utilisées pour le codage audio, de façon à rendre compatible le processus de segmentation à notre procédé de codage, et du phénomène périodique qui apparaît sur certaines RTF, nous devons considérer des fenêtres d'analyse de taille N=512 ou 1024 échantillons. D'après le **Tableau A.1,** la conséquence directe de l'augmentation de la taille de fenêtre est la diminution de la bande équivalente qui se traduit par une meilleur séparation des fréquences mais aussi un moins suivi de l'évolution temporelle du signal. Une taille de fenêtre de N=512 points assure donc un bon compromis entre richesse spectrale et évolution temporelle du signal.

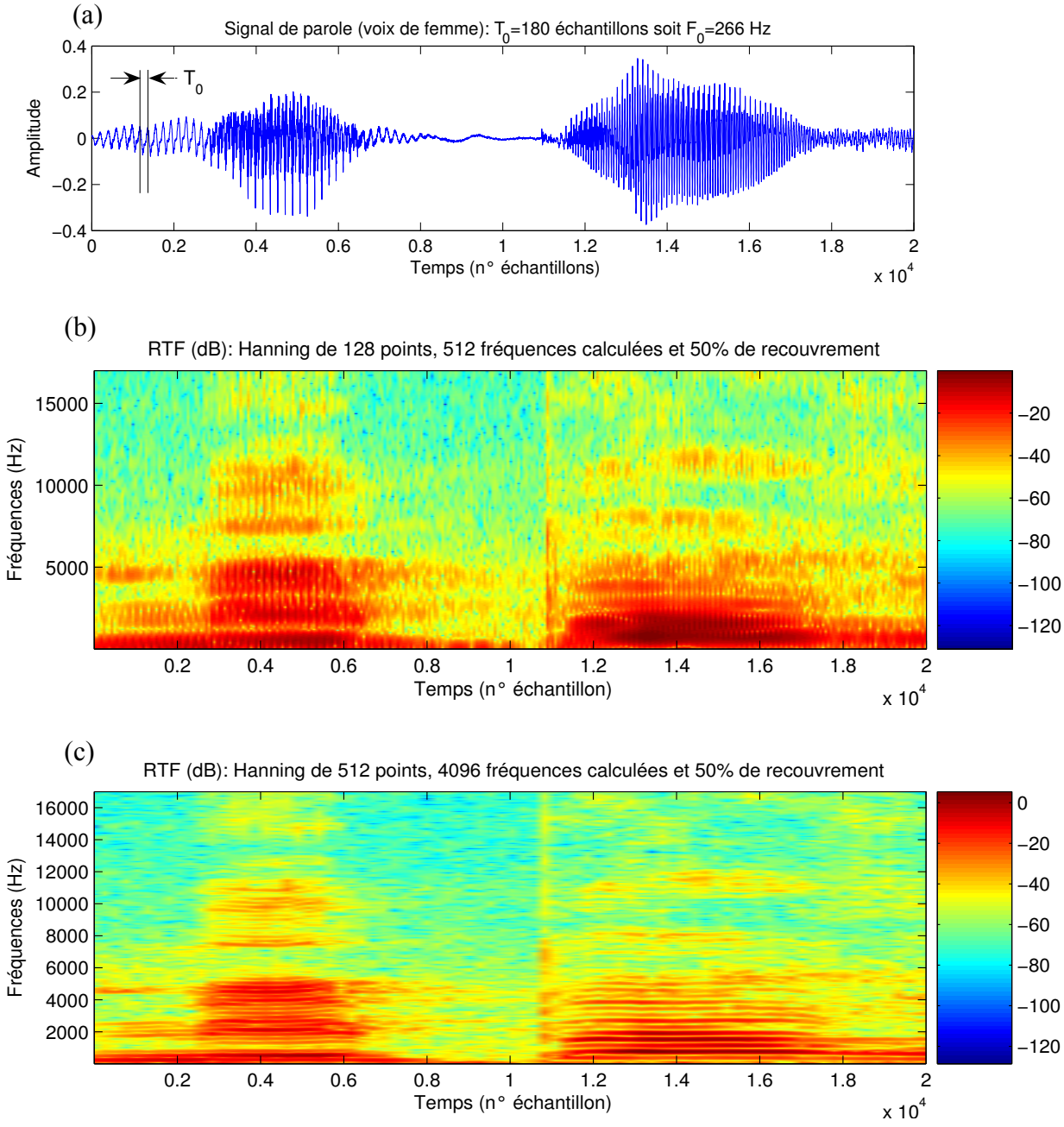

Figure A.6 **(a) -** Amplitude du signal de parole (voix de femme) de fréquence fondamentale F_0=266 Hz soit une période T_0=180 échantillons. **(b) -** Agrandissement selon l'axe des fréquences [1-17kHz] de la RTF (en dB) du signal de parole obtenue avec une fenêtre de Hanning de N=128 échantillons, soit N<T_0. **(c) -** Agrandissement selon l'axe des fréquences [1-17kHz] de la RTF (en dB) du signal de parole obtenue avec une fenêtre de Hanning de N=512 échantillons, soit N>T_0.

A.1.3 Résolution de la RTF adaptée au processus de segmentation

Notre objectif est de calculer le module carré de la TFCT au moyen d'une fenêtre de Hanning de longueur N=128, 256, 512 ou 1024 points en s'assurant d'obtenir une résolution en fréquence suffisante pour que le processus de segmentation puisse extraire les informations pertinentes de la RTF.

Le signal analysé par le processus de segmentation est un signal de parole (voix de femme : « … mes parents sont nos … ») échantillonné à 48000 Hz et d'une durée de 1,1 seconde. L'objet du paragraphe vise à comparer les résultats de segmentation obtenus avec des RTF à différentes résolutions temps-fréquences (taille de la fenêtre variable).

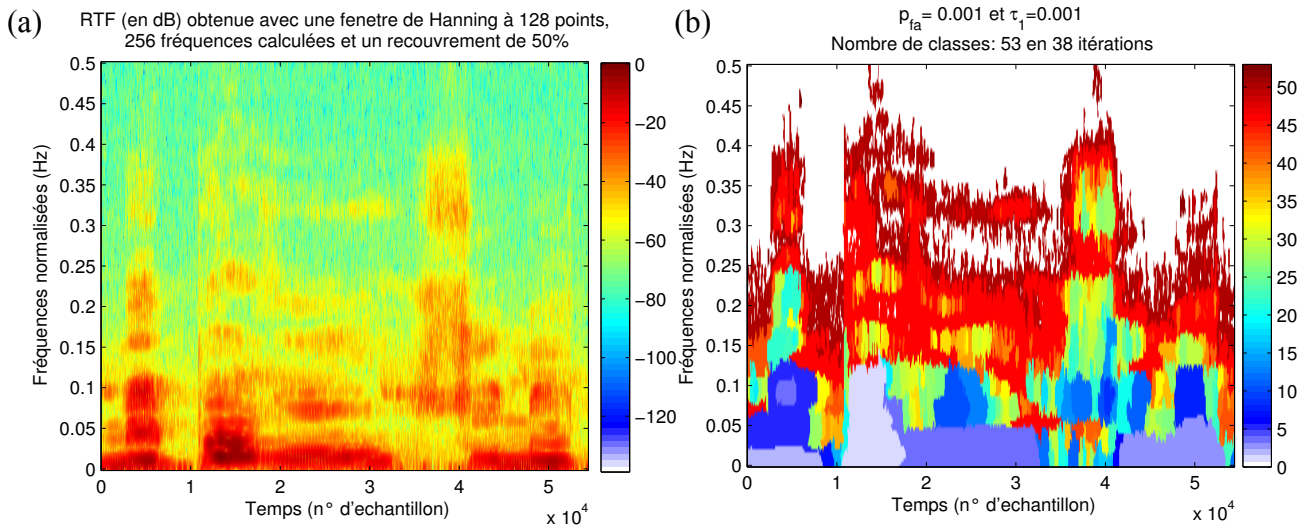

Figure A.7 **(a) -** RTF du signal de parole. **(b) -** La RTF segmentée comporte 53 classes.

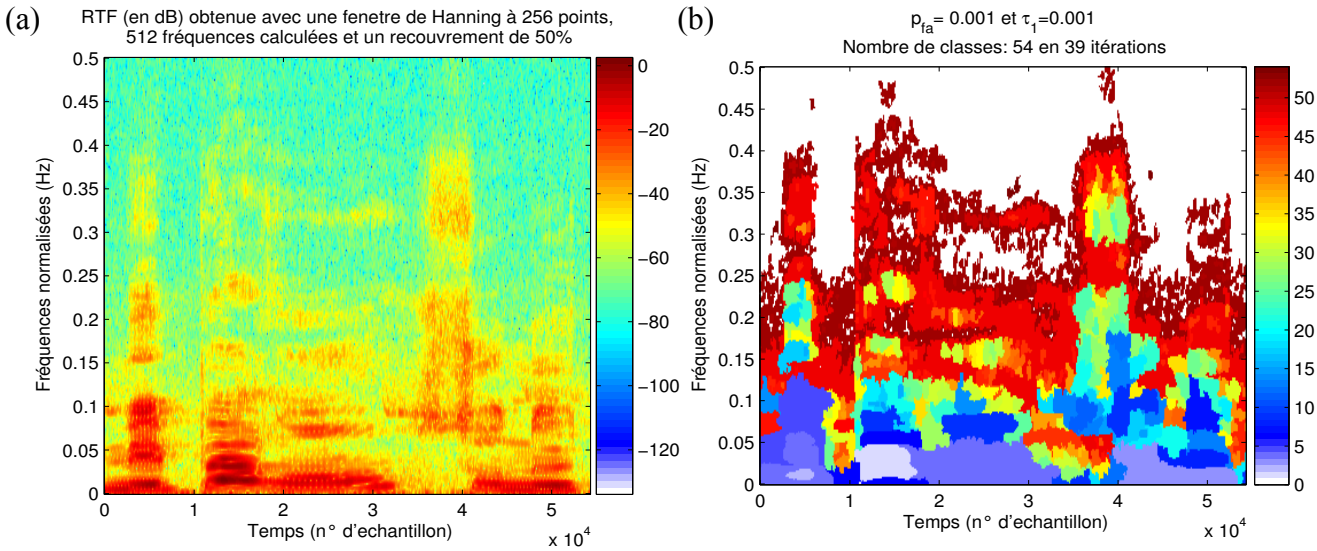

Figure A.8 RTF du signal de parole. La RTF segmentée comporte 54 classes.

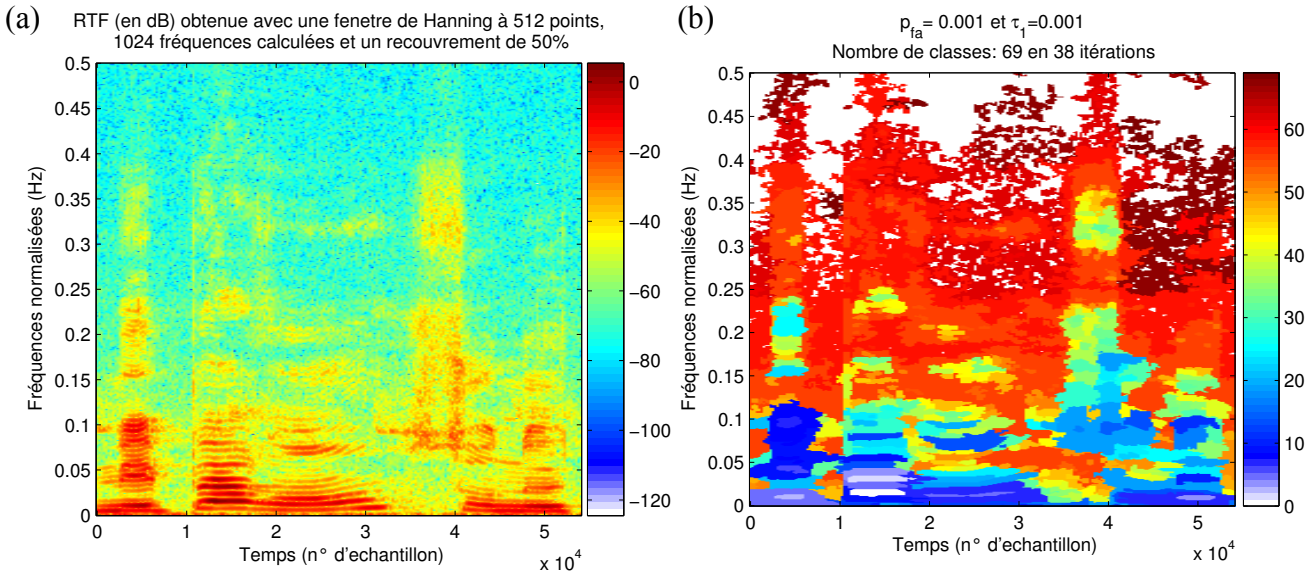

Figure A.9 **(a) -** RTF du signal de parole. **(b) -** La RTF segmentée comporte 69 classes.

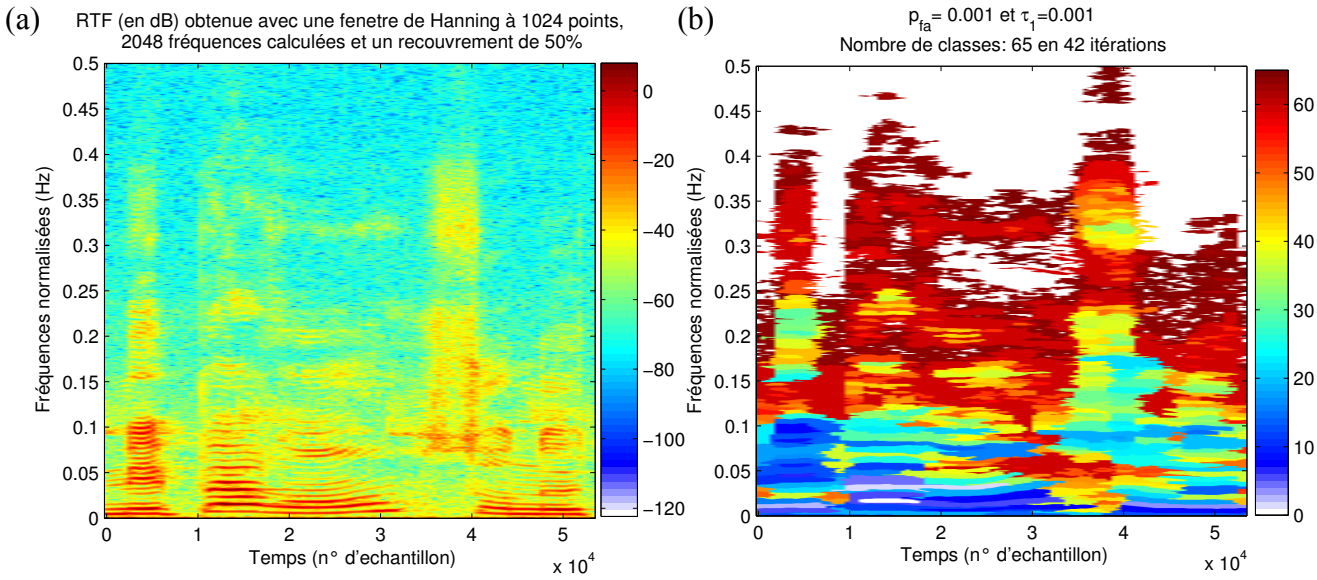

Figure A.10 **(a) -** RTF du signal de parole. **(b) -** La RTF segmentée comporte 65 classes.

On s'aperçoit que la résolution temps-fréquence influence considérablement la segmentation obtenue. En effet, plus la taille de la fenêtre est petite (*cf.* **Figure A.7** et **Figure A.8**) et plus la segmentation offre un bon suivi de l'évolution temporelle du signal mais une mauvaise séparation des fréquences (dans le cas de ce signal de parole, on peut parler des formants). A l'inverse lorsque la taille de fenêtre augmente (*cf.* **Figure A.9** et **Figure A.10**), le suivi fréquentiel est accru et les harmoniques du signal de parole sont mieux dissociées. De plus, le nombre de classes extraites augmente avec la taille de la fenêtre alors que le nombre d'itérations du processus reste presque constant. On s'aperçoit également que les dernières classes extraites (derniers numéros de classe), des points de l'EC les plus proches du bruit, sont les plus étendues (*cf.* **Figure A.9-(b)**). On préférera plutôt extraire plus de classes, même si elles sont proches du bruit, plutôt que de ne pas prendre en considération certains motifs spectraux. D'après ces RTF segmentées, la fenêtre de taille N=512 ou 1024 points réalise le compromis nécessaire entre une évolution temporelle fidèle et une séparation en fréquence suffisante sachant qu'il est possible d'évincer les dernières classes extraites trop proche du bruit.

Le nombre de classes extraites reste cependant trop élevé (69 classes pour 1,1 seconde de signal *cf.* **Figure A.9-(b)**) pour pouvoir réaliser une interprétation précise des classes. Rappelons également qu'à une classe correspond en règle général à de nombreux motifs qui rendent encore plus délicate la comparaison entre les classes extraites.

A.2 Exemples de segmentation de signaux audio stéréophoniques

A.2.1 Segmentation d'un signal de parole stéréo

Le signal stéréo analysé par le processus de segmentation a été synthétisé à partir d'un signal de parole (voix d'homme d'une durée d'une seconde : « … le front froid … ») dupliqué sur chaque canal puis pondéré de manière à introduire une ICLD, entre les canaux, qui varie au cours du temps.

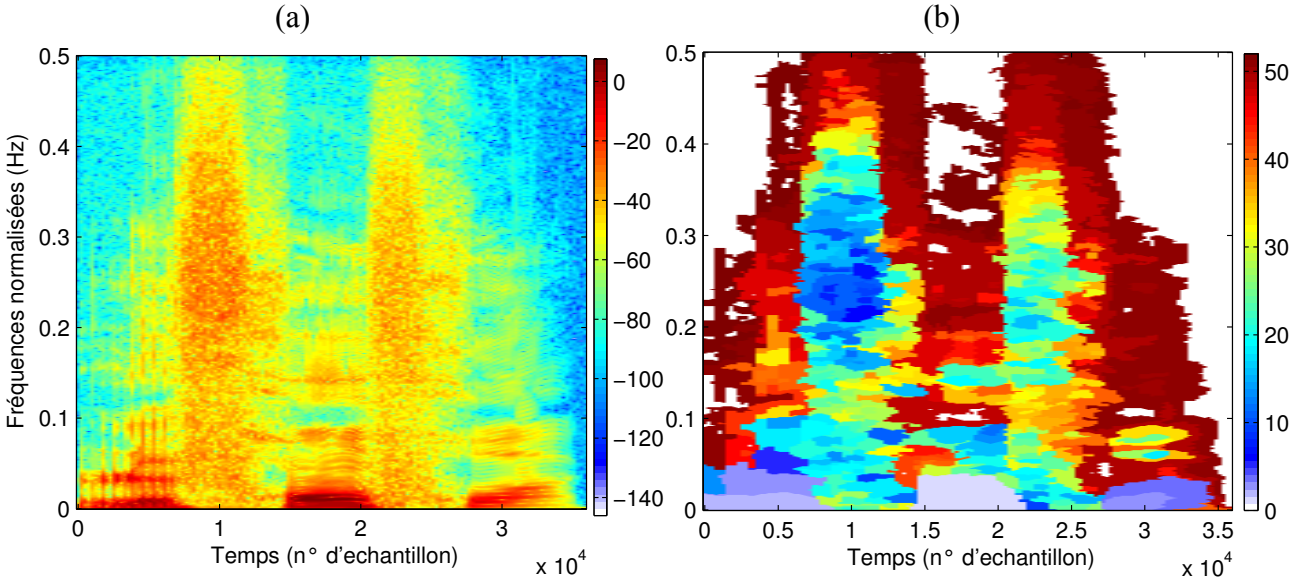

Figure A.11 (a) - RTF du canal gauche obtenue avec une fenêtre de Hanning de 512 points, un recouvrement de 50% et 2048 coefficients spectraux calculés. **(b) -** La RTF segmentée contient 52 classes qui ont été extraites en 48 itérations.

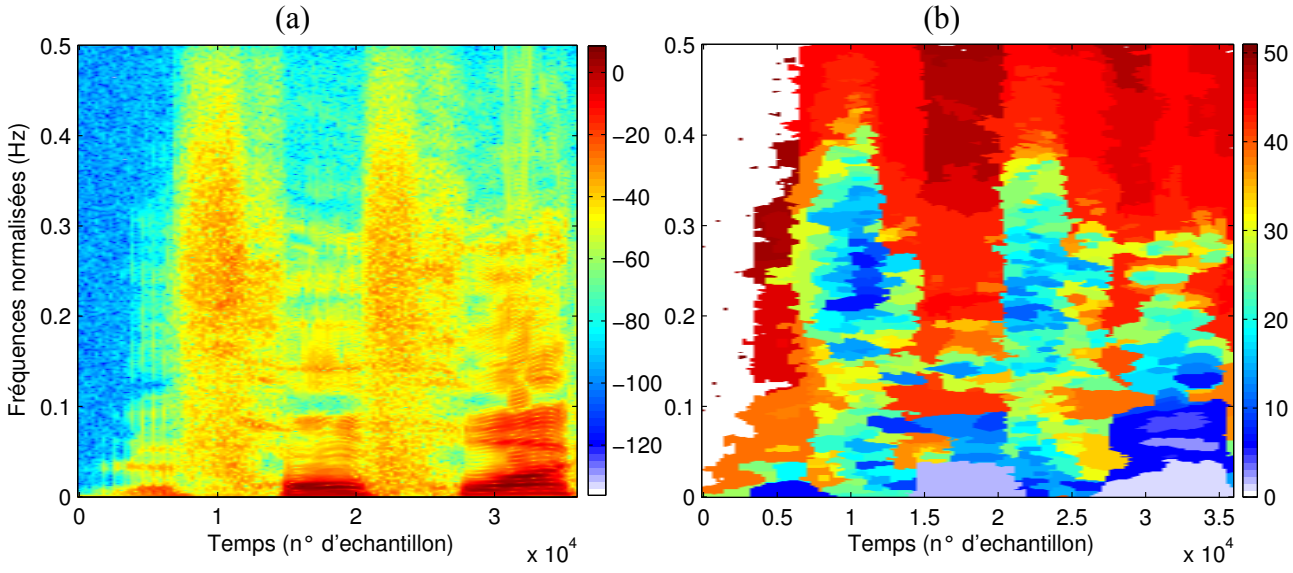

Figure A.12 (a) - RTF du canal droit obtenue avec une fenêtre de Hanning de 512 points, un recouvrement de 50% et 2048 coefficients spectraux calculés. **(b) -** La RTF segmentée contient 51 classes qui ont été extraites en 45 itérations.

Les gains appliqués au signal de parole suivent la loi des sinus (*cf.* paragraphe 1.2.1.2) et procurent au son perçu une position allant du haut-parleur gauche au haut-parleur droit. Ce déplacement correspond au phénomène de latéralisation décrit au paragraphe 1.1.1.

Etant donné que les canaux du signal stéréo sont identiques à un ICLD près, les RTF de ces canaux originaux, présentées aux **Figure A.11-(a)** et **Figure A.12-(a)**, sont identiques lorsque l'ICLD entre les canaux est nulle c'est-à-dire entre les échantillons 15000 et 20000. Les RTF segmentées indépendamment par le processus automatique sont présentées aux **Figure A.11-(b)** et **Figure A.12-(b)**. On constate que la classification des portions de signal complètement identiques (ICLD nulle) sont différentes notamment pour les classes les plus proches de la classe du bruit. On remarque également que les fricatives (« f » du mot front et du mot froid) ont générés un très grand nombre de classe et de motifs spectraux pour une composante fréquentielle large bande *i.e.* une unique structure qui s'étale sur toutes les fréquences.

A.2.2 Segmentation d'un signal de musique stéréo

Le signal stéréo analysé par le processus de segmentation est extrait d'un morceau de musique africaine qui contient plusieurs instruments dont deux guitares (une à gauche et l'autre à droite de l'image stéréo), une guitare basse (au centre), une batterie et des percussions réparties sur les deux canaux. Ce signal exploite pleinement les possibilités offertes par la stéréophonie puisque les canaux sont composés de composantes corrélées (c'est le cas de la guitare basse située au centre de l'image stéréo) et décorrélées d'un canal à un autre (certains instruments ne sont présents que sur un canal).

Les RTF des canaux analysés sont présentées aux **Figure A.13-(a)** et **Figure A.14-(a)**, on remarque principalement la présence de la guitare isolée sur le canal droit et celle de l'autre guitare, dont les notes sont discontinues au cours du temps, répartie sur les deux canaux. Les RTF segmentées sont présentées aux **Figure A.13-(b)** et **Figure A.14-(b)**, on note qu'une classe dite de signal a été généré par le processus à partir d'un bruit hautes fréquences et cela essentiellement sur le canal droit (*cf.* **Figure A.14-(b)**). Ce « bruit » haute fréquence est inaudible à l'écoute du signal stéréo original de qualité CD (échantillonnage à 44100 Hz et résolution des échantillons sur 16 bits) mais a été malencontreusement détecté par le procédé de segmentation comme appartenant aux composantes du signal déterministe. Réduire la valeur de la probabilité de fausse alarme permettrait d'évincer ces classes de bruit mais signifie également la perte de l'approche automatique. Le nombre de classes extraites est encore une fois très important pour une portion de signal d'une durée d'une seconde.

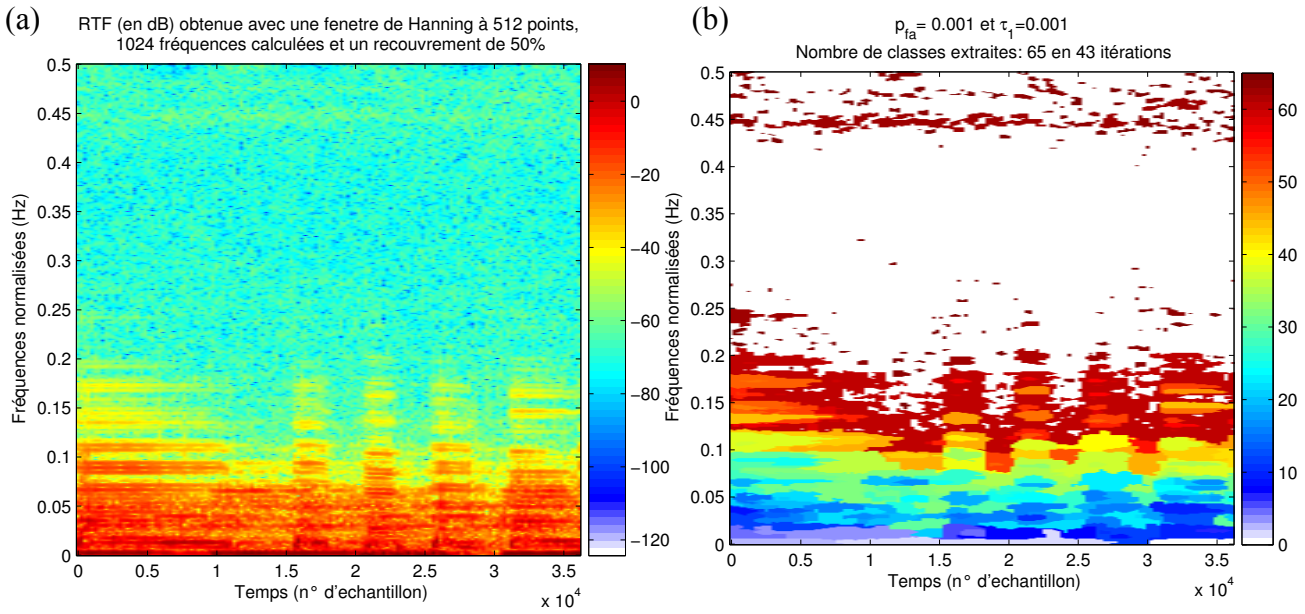

Figure A.13 (a) - RTF du canal gauche. **(b) -** La RTF segmentée contient 65 classes.

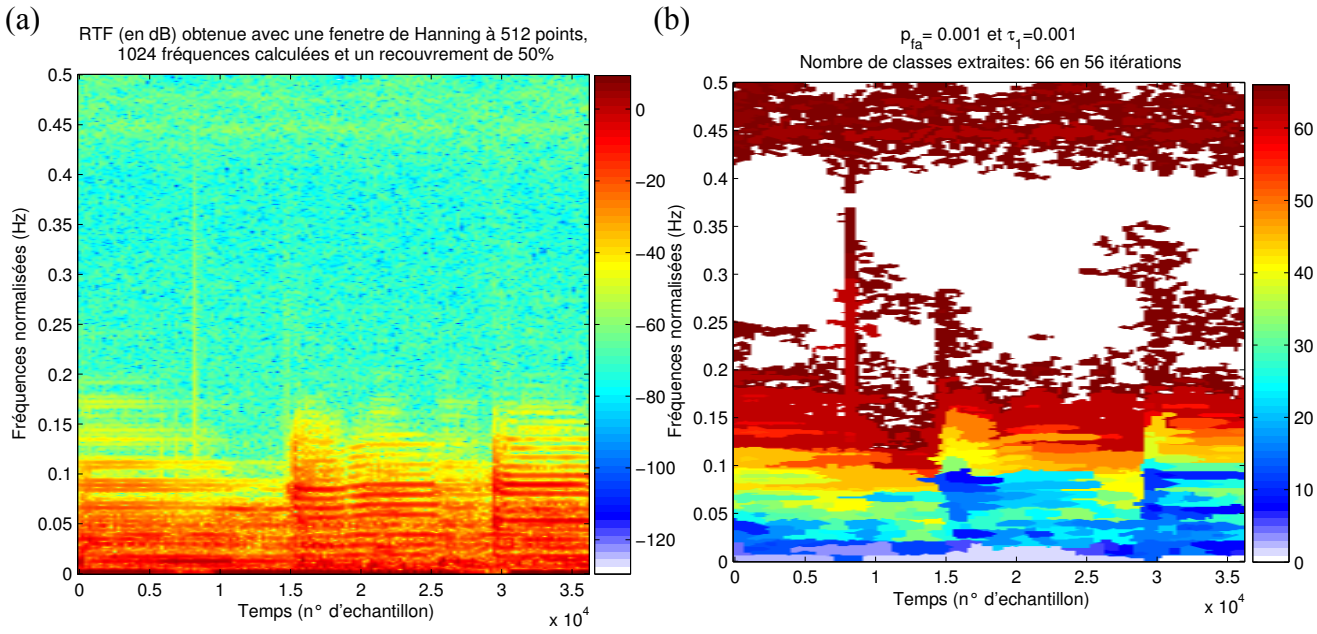

Figure A.14 (a) - RTF du canal droit. **(b) -** La RTF segmentée contient 66 classes.

B. Modèle psychoacoustique pour le codage audio perceptuel

Les codeurs audio perceptuels tirent profit de notre perception auditive pour limiter les informations à coder et ensuite à transmettre. L'idée étant que le bruit de quantification peut être placé dans des zones du spectre du signal où la dégradation sera la moins perceptible. Pour cela, les codeurs perceptuels utilisent les propriétés de notre perception auditive à partir de notions fondamentales empruntées à la psychoacoustique.

B.1 Caractéristiques de l'oreille humaine

B.1.1 Seuil absolu d'audition

La perception des sons par notre système auditif n'est possible qu'à l'intérieur de certaines limites qui dépendent de la fréquence et de l'intensité du signal. La limite inférieure ou seuil d'audition absolu est fixée par la sensibilité de l'oreille. Au-dessous de cette limite les sons ne peuvent pas être détectés. L'échelle logarithmique qui caractérise le niveau d'audition d'une pression acoustique p, de la limite inférieure d'audition au seuil de douleur, est appelée niveau de pression ou *Sound Pressure Level* - *SPL*, exprimé en décibels par:

$$SPL = 10 \times \log_{10}\left(\frac{p}{p_0}\right)^2 \quad \text{dB} \tag{B.1}$$

avec p_0 = 20 μPa, la pression acoustique au seuil d'audition et à la fréquence f = 2 kHz [BOS02]. L'oreille n'est pas un récepteur parfait puisqu'elle n'est pas sensible uniformément à toutes les fréquences. On considère que l'ensemble des fréquences audibles s'étend de 20 à 20000 Hz. Les fréquences inférieures (respectivement supérieures) à 20 Hz sont qualifiées d'infrasons (respectivement ultrasons). Le *seuil absolu d'audition* est alors défini comme l'intensité à laquelle un son pur (sinusoïdale) de fréquence f est juste audible. Le seuil d'audition absolu $T(f)$ (*cf.* **Figure B.**) est exprimé en termes de dB *SPL* et approximé par la fonction non-linéaire de l'équation (B.2) représentative d'un jeune auditeur avec une bonne audition [PAI00]. Ce seuil varie peu entre les individus (+/-3 dB d'après [MOO03]) parmi ceux ayant une audition normale et à tranche d'âge donnée.

$$T(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \quad \text{dB } SPL \tag{B.2}$$

Cette description du seuil d'audition absolu montre qu'un signal doit être suffisamment puissant pour passer à travers le tympan et exciter la cochlée qui est située dans l'oreille interne (*cf.* **Figure B.2**).

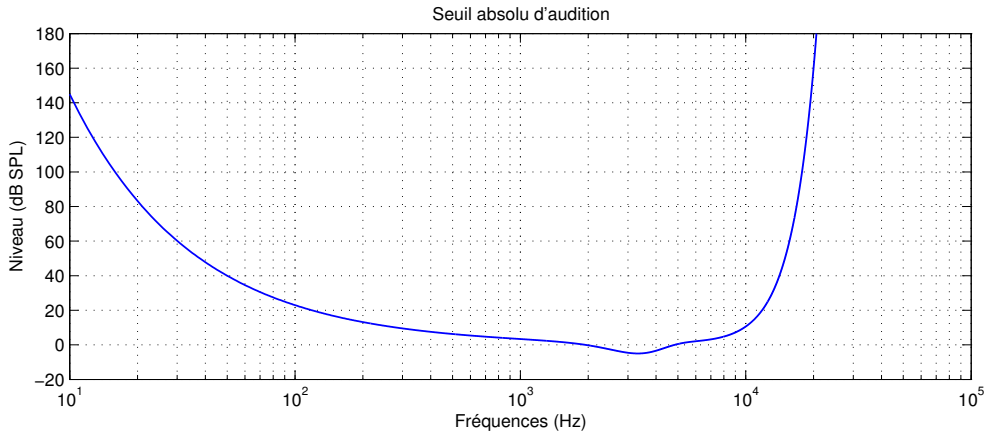

Figure B.1 **Approximation du seuil d'audition absolu.** Il représente le niveau (dB *SPL*) nécessaire pour qu'à chaque fréquence, un auditeur détecte un son pur dans un environnement non bruité.

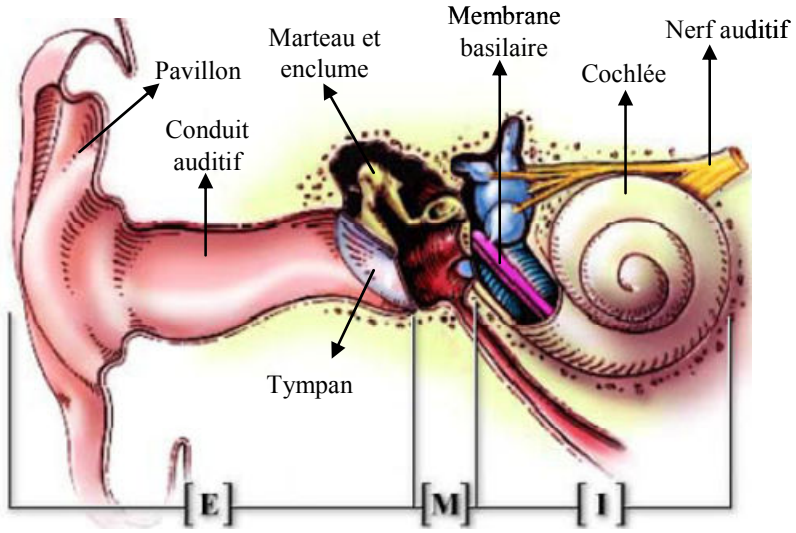

Figure B.2 **L'oreille humaine** se décompose en 3 parties (de gauche à droite): l'oreille Externe, l'oreille Moyenne et l'oreille Interne - tiré de [INSERM].

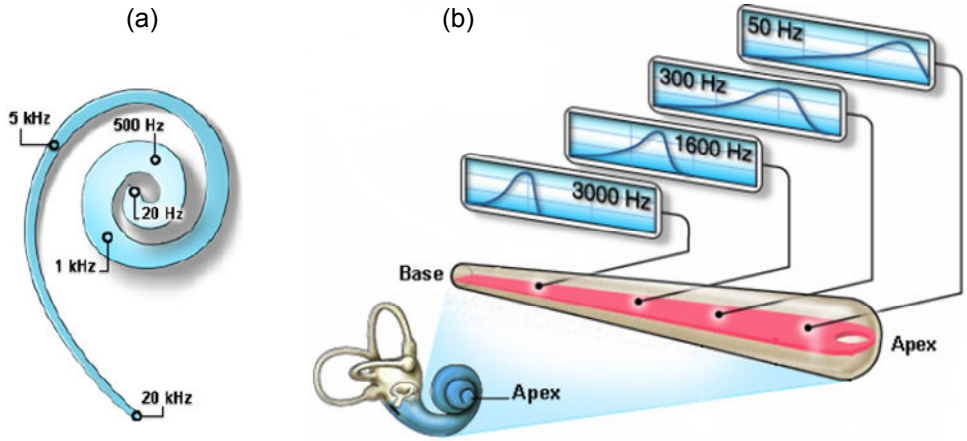

Figure B.3 **La cochlée -** tiré de [INSERM]. **(a) -** La cochlée fonctionne comme un analyseur de spectre en séparant les sons en fréquence grâce à sa membrane basilaire. **(b) -** La Base de la membrane basilaire détecte les sons aigus alors que son extrémité (Apex) détecte les sons graves.

Les études de Canévet [CAN00] présentent l'oreille externe (pavillon et conduit auditif) comme un amplificateur sélectif (le pavillon joue le rôle d'un filtre) et l'oreille moyenne (tympan, enclume, marteau, etc.) comme un adaptateur d'impédance.

L'oreille interne et tout particulièrement la cochlée (*cf.* **Figure B.2** et **Figure B.3**) constitue le récepteur auditif qui renseigne sur l'information fréquentielle du signal. En effet, le long de la cochlée se trouve la membrane basilaire (*cf.* **Figure B.3**) qui porte les cellules sensorielles réagissant à certaines fréquences. Il existe une sorte de bijection fréquence-abscisse (position des cellules sensorielles) qui équivaut à une véritable affectation spatiale des fréquences dans la cochlée [CAN00]. C'est la déformation de la membrane basilaire, en fonction de la fréquence du signal, qui permet de considérer notre oreille comme un puissant « analyseur de spectre » (*cf.* **Figure B.3**).

La localisation spatiale de la zone de vibration de la membrane basilaire dépend des fréquences présentes dans le son (*cf.* **Figure B.3-(a)**). Il est à noter que la relation entre la fréquence du son et la distance de l'apex à la zone de vibration le long de la membrane basilaire (*cf.* **Figure B.3-(b)**) n'est pas triviale: celle-ci est quasi-linéaire pour les fréquences f < 500 Hz et logarithmique pour les fréquences $f \geq 500$ Hz. Il devient alors commode d'exprimer les fréquences selon une échelle, dite *échelle des Barks*, dérivée de manière non-linéaire par rapport à l'échelle naturelle en Hz. Ainsi, la relation entre les fréquences en Bark et la localisation spatiale le long de la membrane basilaire devient linéaire.

B.1.2 Les bandes critiques

De nombreuses études en psychoacoustique ont montré que la cochlée se comporte, d'un point de vue « traitement du signal », comme un banc de filtres fréquentiels qui se recouvrent. L'échelle des Barks correspond en fait à la largeur des bandes passantes des « filtres cochléens » qui est non uniforme puisqu'elle augmente avec la fréquence. On appelle ces bandes passantes les *bandes critiques* qui caractérisent la résolution fréquentielle de l'oreille. Dans l'intervalle de fréquence 50 Hz à 16 kHz, on trouve 24 bandes critiques [CAN00] qui sont listées dans le **Tableau B.2**.

Plusieurs expériences menées notamment par Scharf [SCH70], ont permis d'établir les bandes critiques. Une expérience a consisté à augmenter l'intensité perçue d'un bruit à bande étroite avec une isosonie (en phones) constante *i.e.* intensité constante (*cf.* [CAN00] pour plus de détails sur cette grandeur). L'expérience prouve qu'en augmentant la largeur de la bande du bruit, l'intensité perçue augmente lorsque plusieurs bandes critiques sont « excitées » (*cf.* **Figure B.4**). Il est alors possible d'estimer précisément la largeur des bandes critiques en modifiant la fréquence du bruit et en relevant à chaque fois la largeur de la bande du bruit qui modifie l'intensité perçue.

C'est d'après ce type d'expérience qu'a pu être établi la largeur des bandes critiques. Pour un auditeur moyen, la largeur des bandes critiques [CAN00] en fonction de la fréquence centrale f_c de la bande de fréquence tel que :

$$BC\left(f_c\right) = 25 + 75\left[1 + 1.4\left(\frac{f_c}{1000}\right)^2\right]^{0.69} \text{ Hz}. \tag{B.3}$$

Plus communément, on définit la largeur d'une bande critique comme un Bark. L'équation (B.4) permet alors de passer de l'échelle des fréquences en Hertz à l'échelle des fréquences en Barks :

$$z\left(f\right) = 13\arctan\left(0.00076 f\right) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{ Bark}. \tag{B.4}$$

Numéro	Fréquence centrale (Hz)	Bande Critique (Hz)	Fréquence de coupure basse (Hz)	Fréquence de coupure haute (Hz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500

Tableau B.2 Bandes critiques - tiré de [CAN00]. Chaque ligne donne, pour une bande critique, la fréquence centrale, la largeur de bande et les fréquences de coupure inférieure et supérieure.

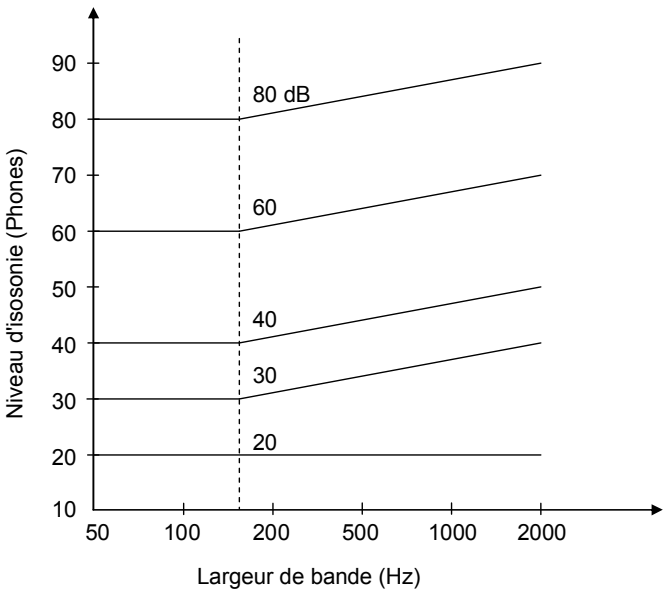

Figure B.4 Expérience de Scharf - tiré de [SCH70]. Niveaux d'isosonie en phones d'une bande de bruit ayant une fréquence centrale de 1000 Hz à différents niveaux de pression acoustique en fonction de sa largeur de bande. Au dessus de 20 dB la sonie commence à augmenter lorsque la largeur de bande du bruit dépasse 160 Hz (largeur de la 9ème bande critique *cf.* **Tableau B.2**.

Une approximation des bandes critiques est donnée par Moore et Glasberg [MOO83] en considérant qu'une bande critique peut être représentée par un filtre équivalent de forme rectangulaire, dont la largeur en Hz (*Equivalent Rectangular Bandwidth* - ERB) est exprimée par :

$$ERB(f_c) = 24.7\left(4.37\left(\frac{f_c}{1000}\right)+1\right) \text{ Hz} \tag{B.5}$$

La **Figure B.5** présente la différence qui réside entre la largeur des bandes critiques définie par l'équation (B.3) de la largeur des bandes de fréquences définie par l'échelle ERB de l'équation (B.5).

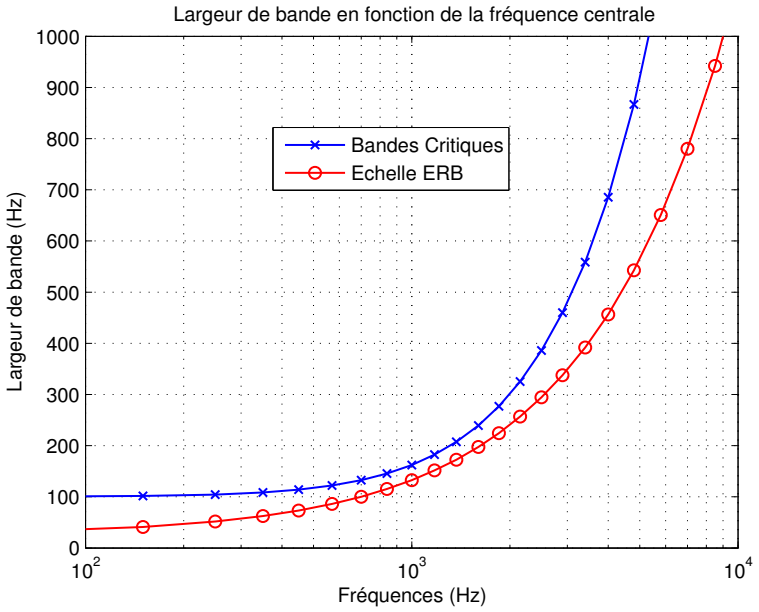

Figure B.5 **Largeur des bandes critiques** (en bleu) comparées à la largeur des bandes définies par l'échelle ERB (en rouge) calculées pour chaque fréquence centrale données par le **Tableau B.2**.

La largeur des bandes suivant l'échelle ERB ne correspond pas tout à fait à la largeur des bandes critiques (*cf.* **Figure B.5**). En particuliers, l'échelle ERB diminue la largeur des bandes en-dessous de 500 Hz. Cette meilleure sélectivité fréquentielle en basses fréquences a un intérêt notable en codage audio pour la conception des bancs de filtres et également pour la stratégie d'allocation des bits de manière perceptuelle.

B.1.3 Information rendue inaudible par masquage

Outre la limite fixée par le seuil absolu d'audition, les cellules sensorielles de la membrane basilaire présentent leurs propres limites pour capter plusieurs sons simultanément. Le *masquage simultané ou fréquentiel* intervient lorsque deux sons de fréquence proche sont en concurrence. L'excitation des cellules sensorielles à une certaine fréquence provoque alors l'inaudibilité d'une fréquence proche d'amplitude plus faible.

De plus, lorsque deux *stimuli* atteignent le système auditif à deux instants différents mais proches, les cellules sensorielles seront « opérationnelles » seulement après un certains temps. Ce phénomène se réfère au *masquage temporel* et plus précisément au post-masquage.

Plus généralement, on peut définir le *masquage* comme la baisse d'audibilité d'un son causée par la présence d'un autre son et ainsi parlé de masquage partiel [CAN00] si le son masqué reste audible tout en ayant diminué d'intensité apparente.

B.1.3.1 Masquage simultané dit fréquentiel

Le masquage fréquentiel intervient lorsque un son d'amplitude élevée excite un groupe de cellules sensorielles de la membrane basilaire qui ne pourront pas être excités par un son de fréquence proche et d'amplitude plus faible. En conséquence, le son d'amplitude plus faible sera masqué par le son d'amplitude plus élevé (*cf.* **Figure B.6**).

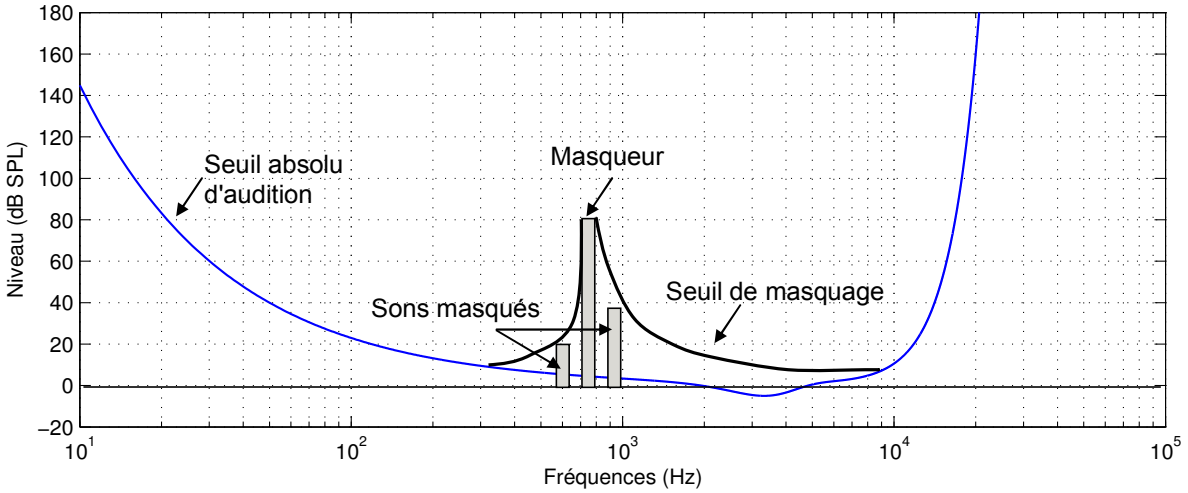

Figure B.6 Exemple de masquage fréquentiel « total », les sons masqués sont inaudibles.

On peut illustrer ce phénomène de masquage en fréquence par l'exemple suivant: une discussion (parole de fréquence moyenne) n'est pas perturbée par le chant des oiseaux (fréquences élevées). On est par contre obligé de hausser la voix si l'on discute dans un lieu où tout le monde parle en même temps (fréquences semblables perçues simultanément). Ce cloisonnement en fréquence a bien sûr ses limites: si un avion décolle, le niveau sonore est tel que tout le spectre s'en trouve masqué.

Masquage tonal et non-tonal

Le masquage fréquentiel est défini comme le phénomène qui rend un son inaudible par la présence d'un autre son de fréquence proche. Les signaux audio présentent au système auditif plusieurs *stimuli* (de fréquences plus ou moins proches) au même instant ce qui provoque l'apparition de masquage simultanés complexes et de différents types. Deux types de masquages simultanés interviennent dans le domaine fréquentiel [PAI00]: le *masquage tonal et non-tonal* ou « *tone-masking-noise* » et « *noise-masking-tone* ». Le cas du « *noise-masking-noise* » étant un scénario plus difficile à caractériser et peu répandu, nous ne nous pencherons donc pas sur ce cas.

Le masquage tonal/non-tonal peut être décris comme suit à partir d'un exemple simple tiré de [PAI00]. Le masquage tonal intervient lorsqu'un son pur ou fréquence pure masque un bruit à bande étroite présent dans la même zone de fréquence (*cf.* **Figure B.7-(a)**). A l'inverse le masquage non-tonal (*cf.* **Figure B.7-(b)**) intervient lorsqu'un bruit à bande étroite masque une fréquence pure située dans la bande de fréquence du bruit. Le Rapport Signal à Masque (RSM) traduit la différence minimale d'intensité (en dB *SPL*), entre le masqueur et le masqué, résultante du phénomène de masquage tonal/non-tonal. En comparant la **Figure B.7 (a)** et **(b)**, il est important de constater l'asymétrie du masquage s'il est tonal ou non-tonal. En effet, le masquage non-tonal produit un RSM plus faible (4 dB) que le masquage tonal (24 dB). Autrement dit, le pouvoir de masquage est grand (respectivement faible) lorsqu'il est non-tonal (respectivement tonal).

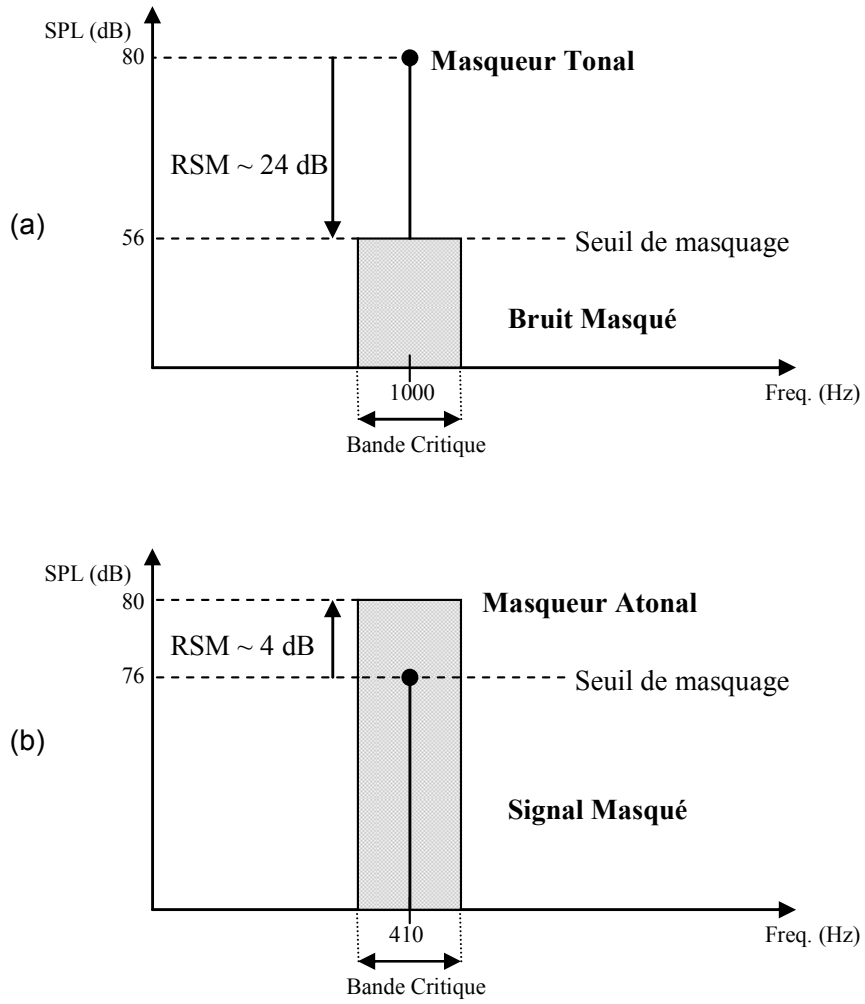

Figure B.7 **Exemple de masquage tonal (a) et non-tonal (b) tiré de** [PAI00]. **(a)** - Au seuil de détection, une fréquence pure à 1 kHz d'intensité 80 dB *SPL* masque (limite de masquage) un bruit de bande équivalente à la bande critique de fréquence centrale 1 kHz et d'intensité 56 dB *SPL*. Le RSM minimal vaut 24 dB et augmente lorsque la fréquence pure est déplacée autour de 1 kHz. **(b)** - Au seuil de détection, une fréquence pure à 410 Hz d'intensité 76 dB *SPL* est masquée (limite de masquage) par un bruit de bande équivalente à la bande critique de fréquence centrale 410 Hz et d'intensité 80 dB *SPL*. Le RSM minimal vaut 4 dB et augmente lorsque la fréquence pure est déplacée autour de 410 Hz.

La diffusion du masquage

Le masquage fréquentiel caractérisé par le masquage tonal et non-tonal n'est pas un phénomène à bande limitée contrairement au cas de l'exemple précédent. En effet, l'influence d'un masqueur, qu'il soit tonal ou non-tonal, ne se limite pas à la bande critique à laquelle il appartient mais également à un certain nombre de bandes critiques concernées par la diffusion du masquage.

Un masqueur appartenant à une bande critique a une influence sur le seuil de détection de ses bandes critiques adjacentes. Un masqueur a une influence importante sur le seuil de détection de la bande critique à laquelle il appartient et cette influence se propage, en s'atténuant, sur les bandes critiques adjacentes. Cette influence du masqueur sur plusieurs bandes critiques constitue son degré d'excitation sur ces bandes critiques communément représenté par la forme ou courbe d'excitation (*cf.* **Figure B.8**).

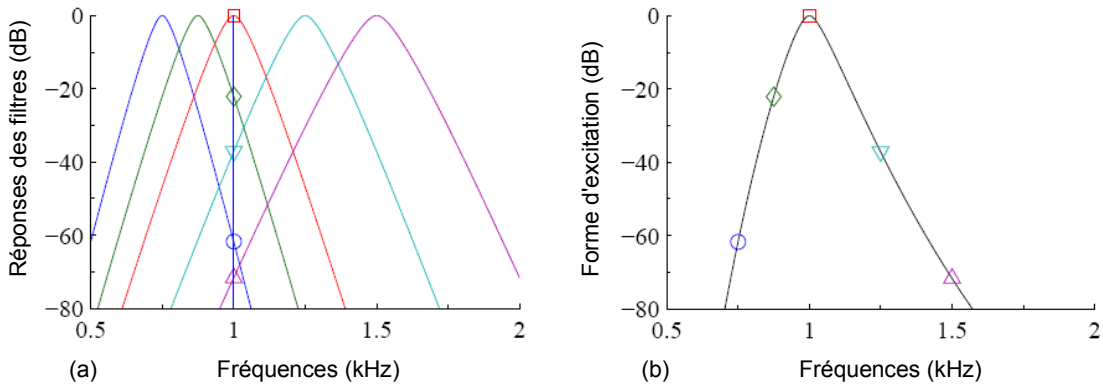

Figure B.8 **Calcul d'une courbe d'excitation pour une fréquence pure à 1 kHz** - tirée de [MOO83]. La fréquence pure est filtrée par le banc de filtres décrit en (a). La courbe d'excitation en (b) correspond au signal filtré représenté en fonction des fréquences centrales du banc de filtres. La fréquence pure a une influence sur 5 bandes de fréquences.

Cette courbe d'excitation ou de diffusion du masquage a une allure qui est également fonction de l'intensité du masqueur (*cf.* **Figure B.8**). On s'aperçoit sur la **Figure B.9** que plus l'intensité du masqueur augmente et plus sa courbe de diffusion de masquage s'étale et ceci surtout vers les hautes fréquences.

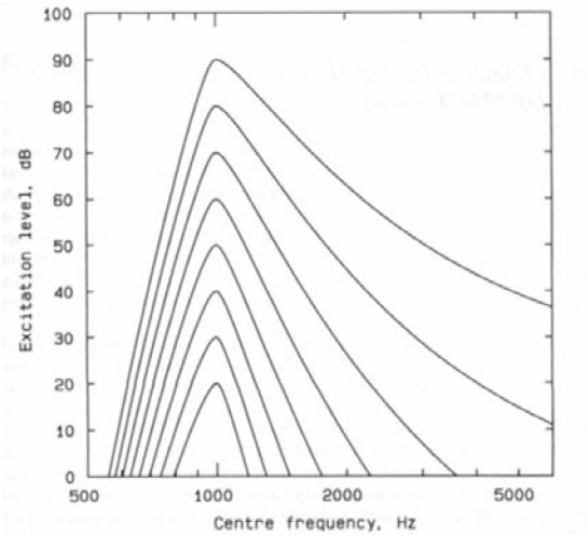

Figure B.9 **Courbes d'excitation** - tiré de [MOO83], à partir d'une fréquence pure à 1 kHz d'intensité variable allant de 20 à 90 dB *SPL* par pas de 10 dB.

Ce phénomène de diffusion du masquage a été modélisé, pour le codage audio perceptuel, par une approximation de forme triangulaire. L'équation (B.6), tirée de [PAI00], propose une expression analytique de la diffusion du masquage (*DM*) établie par Schroeder tel que :

$$DM(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2} \ \text{dB}, \qquad \text{(B.6)}$$

où x à pour unité le Bark. La fonction adoptée par l'ISO/IEC *MPEG Psychoacoustic Model* 1 est définie par l'équation (B.7) d'après [BOS02].

$$DM(dz, Lm) = \begin{cases} -17dz + 0.15Lm(dz - 1)\Theta(dz - 1) & , dz \geq 0 \\ -(6 + 0.4Lm)|dz| - (11 + 0.4Lm)(|dz| - 1)\Theta(|dz| - 1) & , dz < 0 \end{cases} \qquad \text{(B.7)}$$

dz correspond à la différence en Bark de la fréquence du masqueur et du masqué tel que $dz = z\left(f_{masqué}\right) - z\left(f_{masqueur}\right)$, Lm est l'intensité (dB SPL) du masqueur et $\Theta\left(dz\right)$ est défini par :

$$\Theta\left(dz\right) = \begin{cases} 0 & , dz < 0 \\ 1 & , dz \geq 0 \end{cases} \tag{B.8}$$

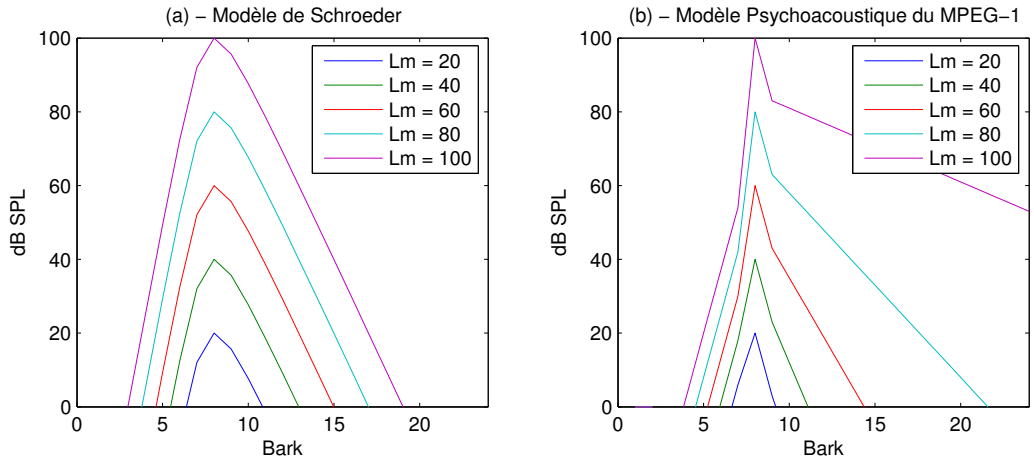

Figure B.10 Courbes de diffusion de masquage (a)- établies par la modélisation de Schroeder. **(b)**- utilisées par le modèle psychoacoustique du standard MPEG-1.

Les courbes de diffusion de masquage (*cf.* **Figure B.10-(b)**) sont presque symétriques lorsque le niveau du masqueur est bas (*Lm* faible). Lorsque le niveau du masqueur augmente, le pouvoir de masquage augmente et ceci de façon plus importante pour les hautes fréquences. Ce phénomène tend à se rapprocher de la réalité c'est-à-dire des courbes de diffusion de masquage mesurées, présentées à la **Figure B.10**, qui traduisent la baisse de sélectivité fréquentielle de l'oreille lorsque l'intensité du masqueur augmente.

B.1.3.2 Masquage temporel

Un phénomène de masquage intervient dans le domaine temporel lorsqu'un signal d'amplitude élevé excite les cellules sensorielles de la membrane basilaire. En effet, il faudra un certain temps avant lequel une nouvelle excitation de la membrane basilaire sera impossible. Ce phénomène est appelé post-masquage. Le phénomène inverse existe également, le pré-masquage qui entre en jeu lors de la détection de sons percussifs notamment.

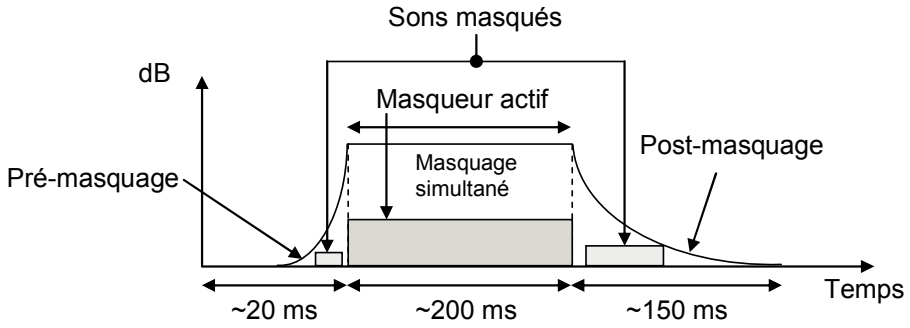

Figure B.11 Exemple de masquage temporel. Un pré-masquage de 20 ms et un post-masquage de 150 ms interviennent pour un son masquant d'une durée de 200 ms.

Comme indiqué par la **Figure B.11**, le post-masquage est actif pour un intervalle de temps significatif qui dépend du niveau du son masquant, de sa durée et de sa fréquence; alors que le pré-masquage est actif seulement quelques millisecondes avant le son masquant (forte décroissance de la courbe de pré-masquage).

Bien que le post-masquage soit un phénomène plus significatif, le pré-masquage constitue un problème couramment traité pour la conception des codeurs audio perceptuels. En effet, ce phénomène de pré-masquage est relié aux effets de pré-écho qui apparaissent suite au traitement/codage de portions de signal fenêtrées. Le pré-écho est une distorsion du signal qui apparaît lorsque la taille de la fenêtre d'analyse est trop grande pour permettre un suivi temporel des composantes transitoire ou « attaques » du signal (la portion de signal analysée est non-stationnaire).

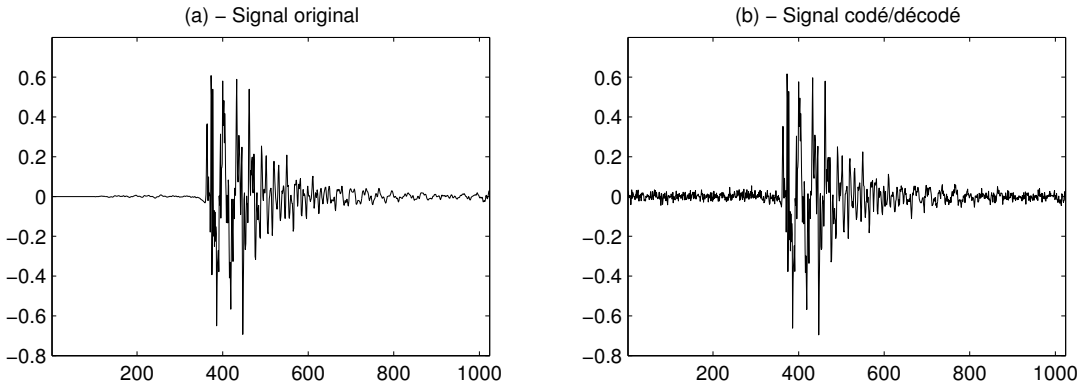

Figure B.12 Phénomène de pré-écho. (a) - signal percussif (castagnette) original. **(b)** - signal codé/décodé : le bruit de quantification s'étale sur toute la portion de signal analysée (N=1024 échantillons).

Dans ce cas où la stationnarité n'est pas vérifiée (*cf.* **Figure B.12**), le bruit de quantification engendré par le procédé de codage est surestimé pour la zone temporelle qui précède l'attaque du signal. Même si un pré-masquage peut limiter la perception de ce bruit de quantification, ce phénomène ne peut à lui seul éviter la perception de cette dégradation du signal.

Des mécanismes d'adaptation de la taille de la fenêtre d'analyse (outil *window/block switching* dans [ISO11172]) ont été introduits dans le standard MPEG-1 couche III notamment, de façon à réduire ce phénomène de pré-écho. Pour cela, lorsqu'une composante transitoire est détectée, le procédé de codage utilise une fenêtre de transition (*cf.* **Figure B.13**) avant de réduire considérablement la taille de la fenêtre (cf. [BOS02] pour l'expression numérique des différentes fenêtres) pour assurer un meilleur suivi temporel du signal à coder.

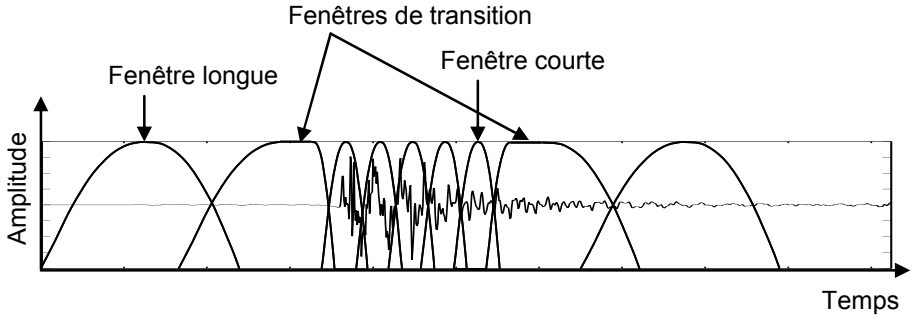

Figure B.13 Exemple d'adaptation de la taille de la fenêtre d'analyse (fenêtre sinus) pour le codage d'un signal percussif (castagnette).

On perçoit cependant ici les limites de la modélisation du masquage temporel puisqu'aucune solution technique n'existe pour assurer le codage d'une transitoire mélangée avec un signal stationnaire.

Cependant, l'erreur de quantification peut être contrôlée de manière continue, par opposition au mécanisme adaptatif de réduction de la taille de fenêtre, avec la combinaison d'un banc de filtre et d'un filtre prédictif adaptatif en référence à l'outil de mise en forme de l'enveloppe temporelle du bruit de quantification (*temporal noise shaping*) utilisé par le standard MPEG-2/4 Advanced Audio Coding décrit dans [ISO13818] et au paragraphe 2.3.2.2. Le procédé de codage utilise un banc de filtre dont la résolution fréquentielle s'adapte dynamiquement au contenu du signal. De plus, un codage prédictif est appliqué au spectre du signal de manière à obtenir une erreur de quantification dont l'enveloppe est adaptée à l'enveloppe temporelle du signal (*cf.* [BOS97] pour plus de détails).

B.2 Implémentation du modèle psychoacoustique du MPEG-1

L'utilisation du modèle psychoacoustique au sein d'un codeur audio perceptuel et le schéma de principe du modèle psychoacoustique 1 du MPEG-1 sont représentés à la **Figure B.14**. L'entrée du modèle psychoacoustique est la représentation temporelle du signal suivant un certain intervalle (longueur des fenêtres d'analyse). La sortie du modèle contient le seuil de masquage et les RSM par sous-bandes de fréquences relatifs à la portion de signal analysée. Basé sur ces informations, l'allocation des bits définit l'opération de quantification finalement suivie par le codage (sans perte) des coefficients spectraux quantifiés (*cf.* **Figure B.14**). L'étape initiale du modèle psychoacoustique consiste à analyser le signal d'entrée $x[n]$. Cette analyse est réalisée au moyen d'une fenêtre de Hanning de longueur $N = 512$ points pour la couche I et 1024 points pour les couches II et III. Le recouvrement entre les fenêtres adjacentes est fixé à $N/16$ points.

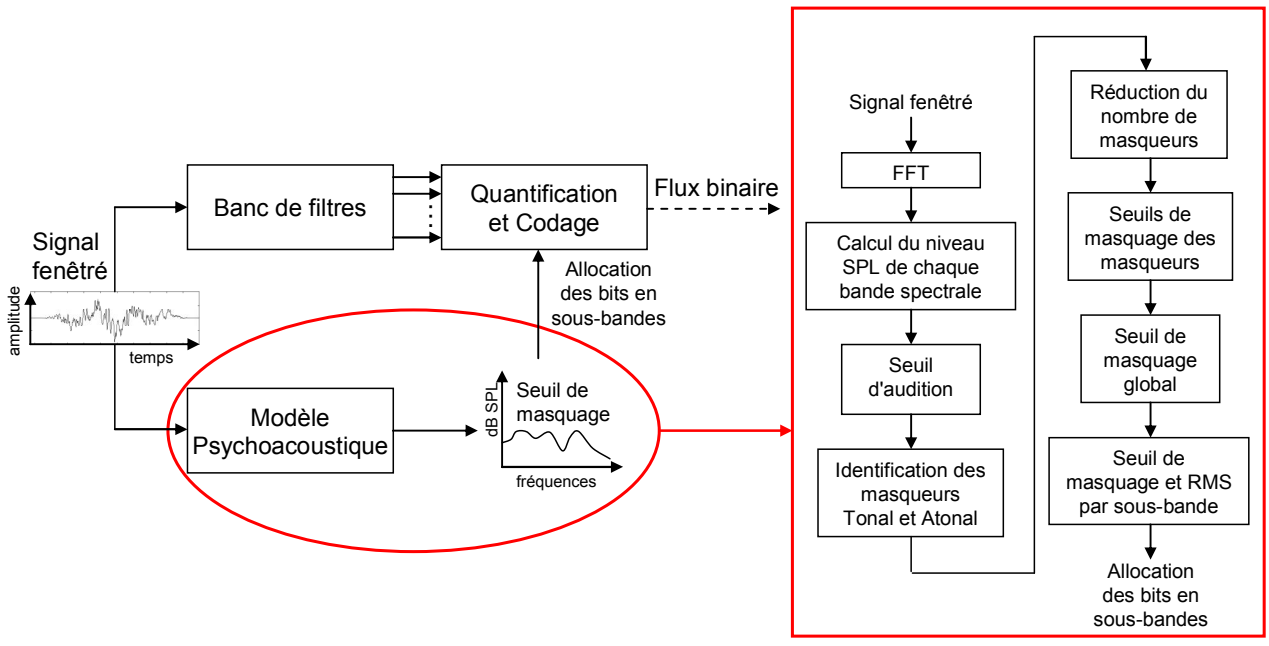

Figure B.14 Principe du modèle psychoacoustique 1 du standard MPEG-1.

Le signal temporel fenêtré par $w[n]$ est ensuite transformé dans le domaine fréquentiel (indice des fréquences k) par Transformation de Fourier Discrète (TFD) tel que :

$$X[k] = \sum_{n=0}^{N-1} w[n] \times x[n] \times e^{-j2\pi \frac{nk}{N+Z}}, \; k \in \left[0, \cdots, \frac{N}{2} - 1\right], \tag{B.9}$$

où Z correspond au taux de bourrage de points. A partir de la TFD du signal fenêtré, $X[k]$, le niveau du signal est ajusté au niveau absolu de la pression du son (SPL) de manière à aligner le seuil d'audition (*cf.* équation (B.2)) et les courbes de masquages (*cf.* paragraphe B.1.3.1) avec la densité spectrale de puissance (DSP) du signal analysé. Une attention particulière doit être donnée pour la normalisation utilisée lors du passage du domaine temporel au domaine fréquentiel. En effet, le choix de la fenêtre d'analyse w affecte le gain de la TFD et par suite de la DSP. Le modèle psychoacoustique 1 du MPEG-1, décrit dans [BOS02]-[ISO11172], propose une mise au niveau SPL de la DSP du signal, $L[k]$, tel que :

$$L[k] = 96 + 10 \times \log_{10}\left(\frac{8}{3} \times \frac{4}{N^2} \times \left|X[k]\right|^2\right) \text{ dB}, \; k = \left[0, \cdots, \frac{N}{2} - 1\right] \tag{B.10}$$

où $|X[k]|^2$ représente l'estimation de la DSP du signal fenêtré.

B.2.1 Distinction des masqueurs tonaux et non-tonaux

Avant de pouvoir distinguer les masqueurs tonaux et non-tonaux à partir de la DSP du signal, la première étape consiste à extraire les maxima de $L[k]$ de façon à caractériser les coefficients spectraux les plus énergétiques susceptibles de masquer les coefficients spectraux aux fréquences voisines. Les maxima de la DSP sont extraits au moyen de l'expression suivante :

$$L_{\max} = \left\{L[k] \middle| L[k] > L[k \pm 1]\right\}, \; k \in \left[1, \cdots, \frac{N}{2} - 2\right] \tag{B.11}$$

Les masqueurs tonaux M_T sont alors extraits en considérant les coefficients spectraux voisins, suivant une certaine distance fréquentielle Δ_k, des maxima de la DSP tel que :

$$M_T = \left\{L_{\max}[k] \middle| L_{\max}[k] > L[k \pm \Delta_k] + 7 \text{ dB}\right\}, \tag{B.12}$$

$$\Delta_k \in \begin{cases} 2 \times \dfrac{N+Z}{N} & \text{pour} \quad 170 \times \dfrac{N+Z}{f_S} < k < 5500 \times \dfrac{N+Z}{f_S} \\ [2,\ 3] \times \dfrac{N+Z}{N} & \text{pour} \quad 500 \times \dfrac{N+Z}{f_S} \le k < 11000 \times \dfrac{N+Z}{f_S} \\ [2,\ 6] \times \dfrac{N+Z}{N} & \text{pour} \quad 11000 \times \dfrac{N+Z}{f_S} \le k \le 20000 \times \dfrac{N+Z}{f_S} \end{cases}, \tag{B.13}$$

où f_S représente la fréquence d'échantillonnage du signal d'entrée. La partie entière des termes rationnels a été prise en compte mais n'est pas représentée dans l'équation (B.13) par soucis de clarté. Plus la fréquence k est grande et plus le nombre voisins considérés par Δ_k est grand. Si un maximum local L_{MAX} vérifie la condition de l'équation (B.12), il sera considéré comme un masqueur tonal. Le voisinage fréquentiel considéré correspond à celui présenté dans [PAI00] avec la prise en compte du nombre de zéros en complément du signal (Z) pour le calcul de la TFD qui peut alors résulter en une analyse spectrale plus discriminante.

La puissance d'un masqueur tonal [PAI00] correspond à la somme des puissances du masqueur et de ses deux plus proches voisins tel que :

$$P_{MT}[k] = 10 \times \log_{10} \sum_{j=-1}^{1} 10^{\frac{L[k+j]}{10}} \text{ dB}, \text{ avec : } L[k] \in M_T \,. \tag{B.14}$$

Un masqueur non-tonal est ensuite extrait pour chaque bande critique suivant l'échelle Bark définie par l'équation (B.4). La puissance P_{MN} de chaque masqueur non-tonal correspond, d'après [PAI00], à la somme des puissances des coefficients spectraux de la sous-bande excepté ceux appartenant à l'ensemble des masqueurs tonaux et de leurs voisinages :

$$P_{MN}\left[\overline{k}\right] = 10 \times \log_{10} \sum_{j} 10^{\frac{L[j]}{10}} \text{ dB}, \quad \forall L\left[j\right] \notin \left\{P_{MT}\left[k, k \pm 1, k \pm \Delta_k\right]\right\} \tag{B.15}$$

où $\overline{k}$ correspond à la moyenne géométrique des indices fréquentiels pour chaque sous-bande de fréquence [PAI00]. L'équation (B.15) revient à associer la puissance spectrale résiduelle pour chaque sous-bande, qui n'est pas associée à un masqueur tonal, à un masqueur non-tonal. D'après l'exemple de la **Figure B.15**, nous considérons ici 25 bandes critiques et non 24 comme indiqué par le **Tableau B.2**.

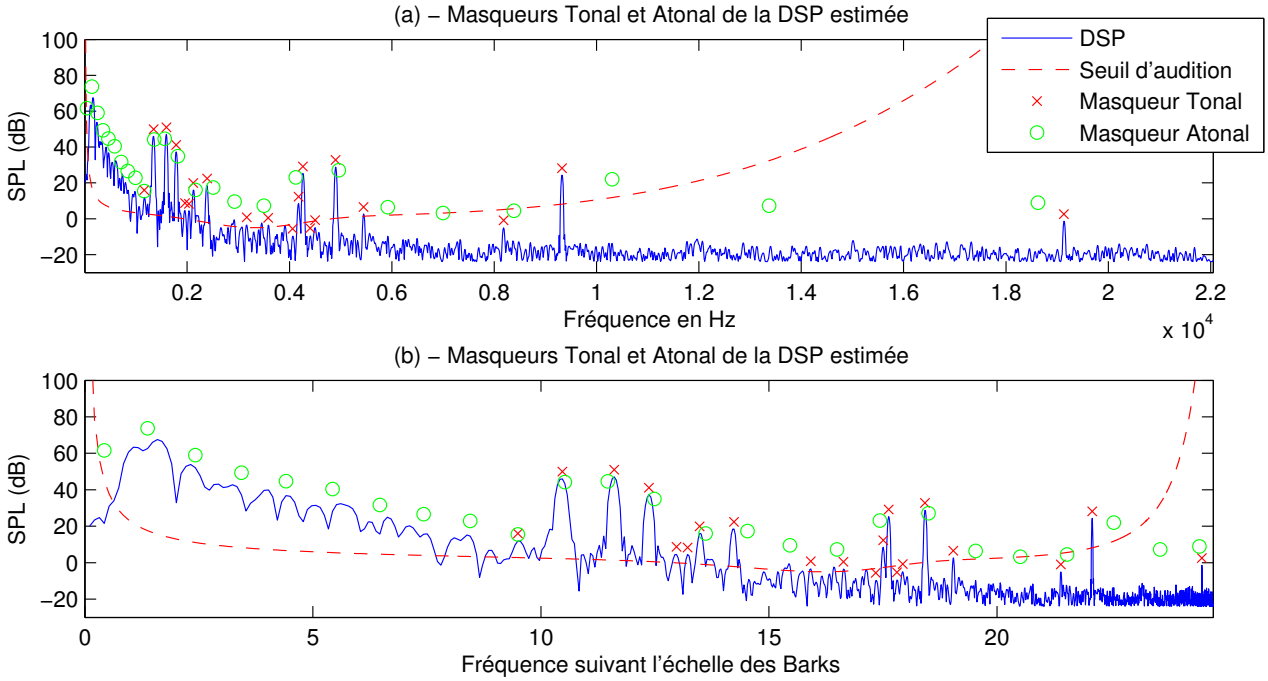

Figure B.15 Superposition du seuil absolu d'audition et des masqueurs tonaux et non-tonaux extraits à partir de la DSP d'un signal mélange de glockenspiel et de grosse caisse. **(a) -** représentation fréquentielle selon l'échelle linéaire des Hertz. **(b) -** représentation fréquentielle selon l'échelle non-linéaire des Barks.

B.2.2 Réduction du nombre de masqueurs

Deux critères sont utilisés pour réduire le nombre des masqueurs. Le premier critère consiste à éliminer les masqueurs qui sont situés en-dessous du seuil d'audition absolu :

$$P_{M(T,N)}\left[k\right] \geq T\left[k\right] \tag{B.16}$$

avec $T[k]$ le seuil absolu d'audition défini par l'équation (B.2).

Le second critère relève d'une comparaison des puissances des masqueurs au sein d'une même bande critique [PAI00]. Pour chaque masqueur (tonal et non-tonal) localisé à la fréquence z en Bark, une recherche de masqueurs voisins est réalisée dans une zone de fréquence allant de la position (z–0,5) à (z+0,5) Bark *i.e.* soit la largeur d'une bande critique.

Si plusieurs masqueurs sont présents dans cette zone alors le masqueur qui à la puissance la plus grande sera le seul masqueur conservé.

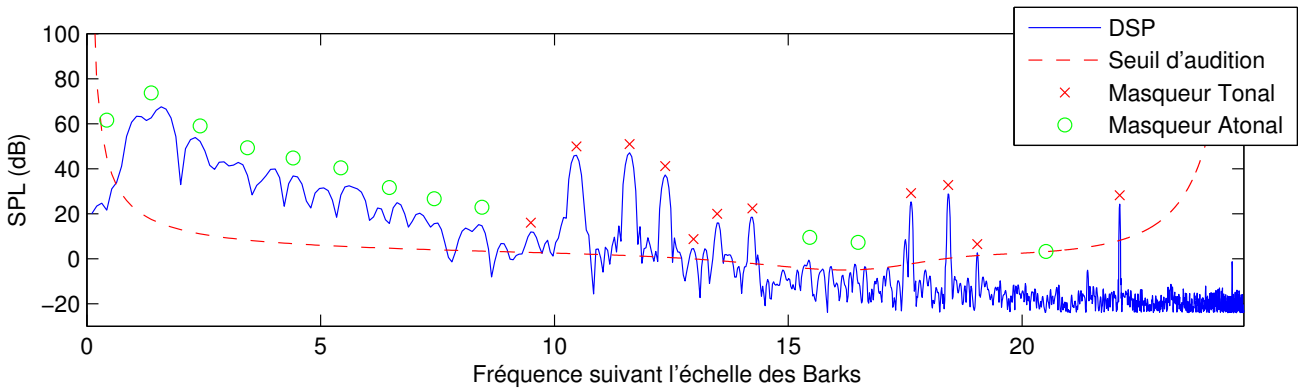

Figure B.16 Superposition du seuil absolu d'audition et des masqueurs tonaux et non-tonaux décimés après avoir été extraits à partir de la DSP d'un signal mélange de glockenspiel et de grosse caisse (représentation fréquentielle selon l'échelle des Barks).

B.2.3 Seuils de masquage individuels

Les courbes de diffusion du masquage fréquentiel *DM* sont données, d'après [PAI00], par l'équation suivante :

$$DM[i,j]=\begin{cases} -0.4P_{M(T,N)}[j]+11+17\Delta_z & \text{dB} \quad -3\le \Delta_z < -1 \text{ Bark} \\ \left(0.4P_{M(T,N)}[j]+6\right)\Delta_z & \text{dB} \quad -1\le \Delta_z < 0 \text{ Bark} \\ -17\Delta_z & \text{dB} \quad 0\le \Delta_z < 1 \text{ Bark} \\ -0.15P_{M(T,N)}[j]+\left(0.15P_{M(T,N)}[j]-17\right)\Delta_z & \text{dB} \quad 1\le \Delta_z < 8 \text{ Bark} \end{cases} \tag{B.17}$$

L'indice j est la position en fréquence du masqueur et l'indice i, celle des coefficients spectraux masqués. $DM[i,j]$ est fonction de la puissance en dB du masqueur $P_{M(T,N)}[j]$, qu'il soit tonal ou non-tonal, et de la distance qui sépare le masqueur du masqué $\Delta_z = z[i]-z[j]$ en Barks.

A partir de ces courbes de diffusion du masquage peuvent être calculés les seuils de masquage individuels des masqueurs tonaux TH_T et non-tonaux TH_N tels que

$$\begin{cases} TH_T[i,j]=P_{MT}[j]-0.275\times z[j]+DM[i,j]-6.025 & \text{dB} \\ TH_N[i,j]=P_{MN}[j]-0.175\times z[j]+DM[i,j]-2.025 & \text{dB} \end{cases} \tag{B.18}$$

On constate (*cf.* **Figure B.17**) le pouvoir de masquage supérieur des masqueurs non-tonaux comparés aux masqueurs tonaux comme indiqué par l'équation (B.18).

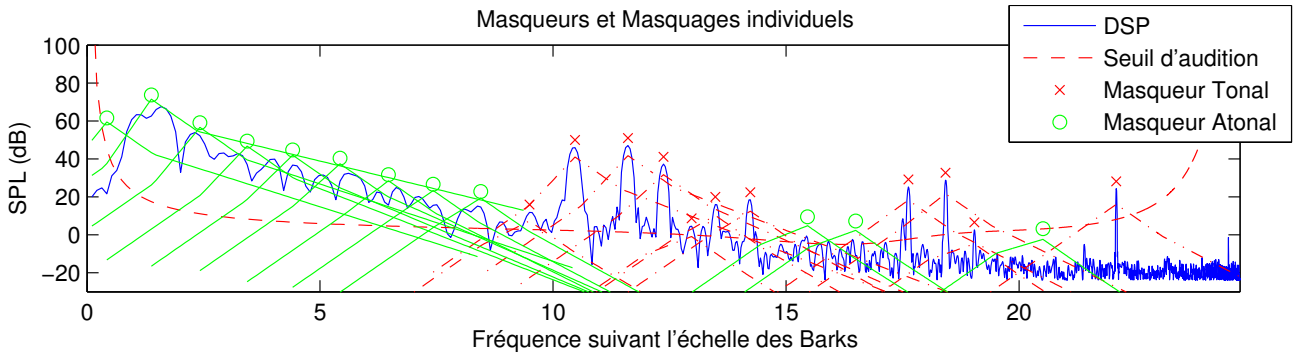

Figure B.17 Superposition du seuil absolu d'audition et des courbes de masquage individuelles relatives aux masqueurs tonaux et non-tonaux extraits de la DSP d'un signal mélange de glockenspiel et de grosse caisse (représentation fréquentielle selon l'échelle des Barks).

B.2.4 Seuil de masquage global et minimum utile pour la quantification

Les seuils de masquage individuels des masqueurs tonals et non-tonaux sont combinés pour calculer le seuil de masquage global. Le modèle psychoacoustique 1 du standard MPEG-1 considère les effets du masquage comme additif et de manière linéaire [PAI00] tel que :

$$TH_G[k] = 10 \times \log_{10}\left(10^{\frac{T[k]}{10}} + \sum_{t=1}^{T_T} 10^{\frac{TH_T[k,t]}{10}} + \sum_{n=1}^{T_N} 10^{\frac{TH_N[k,n]}{10}}\right) \quad \text{dB} \tag{B.19}$$

où $T[k]$ est le seuil d'audition absolu définit à l'équation (B.2), $TH_T[k,t]$ est le seuil de masquage individuel à la fréquence k du masqueur tonal d'indice t pour un nombre total T_T de masqueurs tonaux. $TH_N[k,n]$ est le seuil de masquage individuel à la fréquence k du masqueur non-tonal d'indice n pour un nombre total T_N de masqueurs non-tonaux. Ces seuils de masquages individuels étant définis à l'équation (B.18). Une autre approche proposée par Baumgarte et al. [BAU95] consiste à effectuer une somme non-linéaire des courbes de masquage pour établir le seuil de masquage global. Cette méthode a d'ailleurs été adoptée par le modèle psychoacoustique 2 du standard ISO/IEC MPEG-1 [ISO11172].

Le seuil de masquage global caractérise les modifications du seuil absolu d'audition par addition des puissances de la diffusion basilaire de tout les masqueurs (tonaux et non-tonaux) présents dans la DSP du signal. Le seuil de masquage minimum correspond à la valeur minimale du seuil de masquage global pour chaque sous-bande (b) définie suivant l'échelle des Barks (*cf.* équation (B.4)) tel que :

$$TH_{\min}[b] = \min\left(TH_G[k]\right) \quad \text{dB} \qquad \forall k \in \left\{ j \mid z[j] = b \right\}. \tag{B.20}$$

Ce seuil de masquage minimal TH_{min} est qualifié dans la littérature [PAI00] comme le niveau minimal à partir duquel le bruit de quantification devient juste perceptible *Just Noticiable Distortion* (JND).

Le niveau de bruit de quantification qui peut être introduit dans chaque sous-bande est alors estimé à partir du seuil de masquage minimum et du niveau *SPL* maximum défini pour chaque sous-bande [PAI00] tel que :

$$L_{\max}[b] = \max\left(L[k]\right) \quad \text{dB} \quad \forall k \in \left\{ j \mid z[j] = b \right\}, \tag{B.21}$$

ainsi, le RSM peut être calculé pour chaque sous-bande tel que :

$$RSM[b] = L_{\max}[b] - TH_{\min}[b] \quad \text{dB}. \tag{B.22}$$

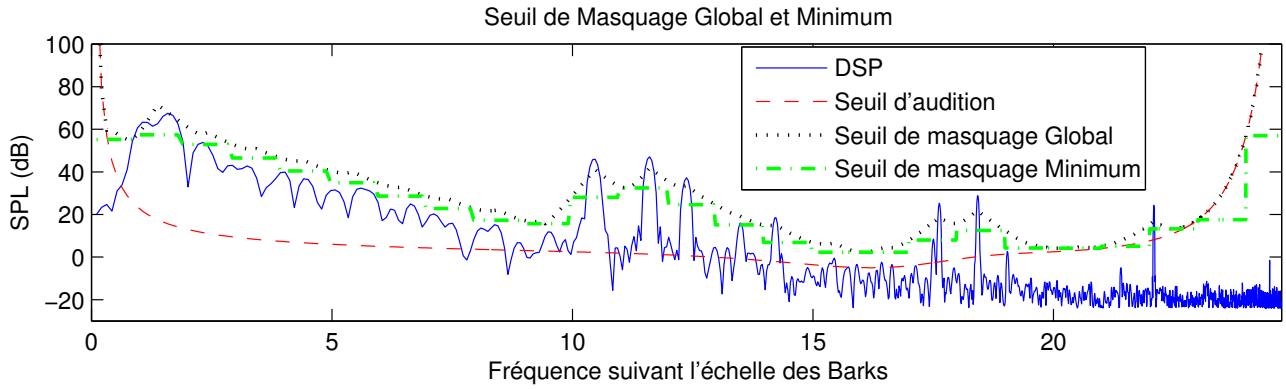

Figure B.18 Superposition du seuil absolu d'audition et des seuils de masquage global et minimum calculés à partir de la DSP d'un signal mélange de glockenspiel et de grosse caisse (représentation fréquentielle selon l'échelle des Barks).

Comme l'illustre la **Figure B.19**, le Rapport Signal à Bruit (RSB), qui dépend du nombre de bits alloués pour la quantification, doit être au moins égal au RSM pour rendre inaudible le bruit de quantification.

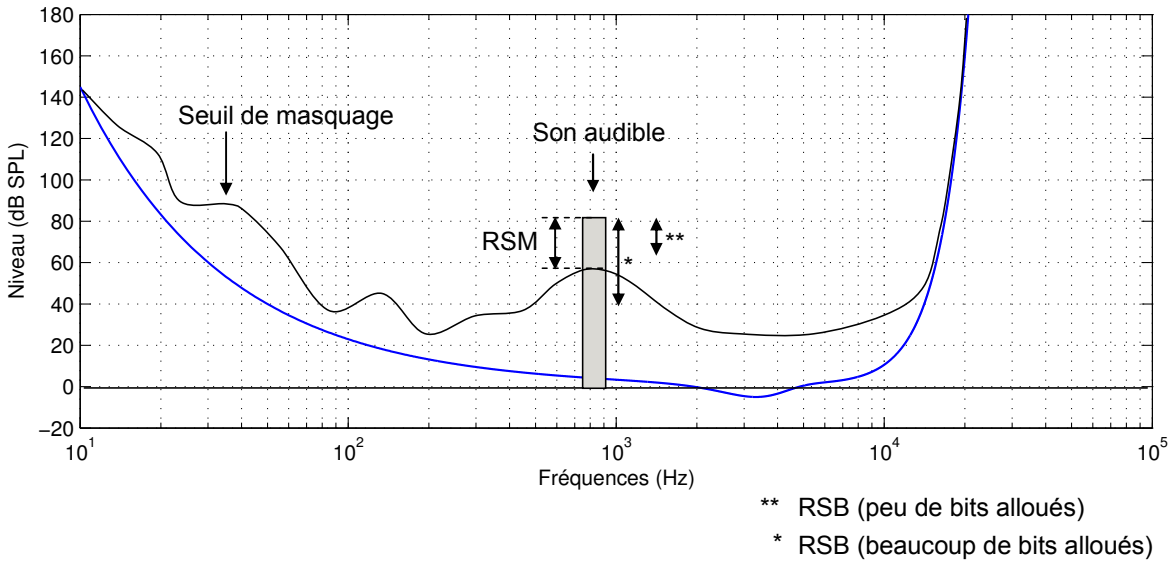

Figure B.19 RSB et RSM. Exemple illustrant les différents Rapport Signal à Bruit (RSB) résultant de l'allocation binaire pour une composante fréquentielle à RSM déterminé.

Johnston a introduit, dans [JOH88], le concept d'entropie perceptuelle pour définir le nombre de bits moyen nécessaire pour la quantification des composantes fréquentielles en sous-bandes sans introduire de dégradations perceptible. Etant donné la DSP du signal analysé, $L[k]$, et l'intensité du seuil de masquage $TH_G[k]$ à chaque fréquence d'indice k=[0,…,N/2-1], l'entropie perceptuelle PE d'une portion de signal est donnée par :

$$PE = \frac{1}{N}\sum_{k=0}^{N/2-1} \max\left\{0, \log_2\left(\sqrt{\frac{L[k]}{TH_G[k]}}\right)\right\} \approx \frac{1}{N}\sum_{k=0}^{N/2-1} \log_2\left(1+\sqrt{\frac{L[k]}{TH_G[k]}}\right). \qquad \text{(B.23)}$$

Ainsi, la mesure d'entropie perceptuelle donne une limite minimale pour le codage audio perceptuel des signaux audio basée sur la transformée fréquentielle du signal et le calcul du seuil de masquage.

C. Compatibilité audio grâce aux techniques *downmix* et *upmix*

La compatibilité entre les différents formats audio spatialisés est rendue possible par des opérations de conversion. Cette approche a été d'abord initiée par les créateurs de contenus (preneurs de son, ingénieurs du son, etc.) de manière à uniformiser les différents formats audio aujourd'hui disponibles. Des solutions techniques évoluées font maintenant parties intégrantes des systèmes audionumériques professionnel et grand public et sont souvent dénommés *décodeurs* par abus de langage.

Le terme *downmix* correspond au procédé qui permet la réduction du nombre de canaux d'un signal multicanal (au format 5.1 ou 7.1 par exemple) pour assurer une compatibilité avec les systèmes stéréophoniques et monophoniques. Typiquement, l'audiophile qui ne possède pas de système de restitution adapté à la reproduction des scènes sonores multicanales peut tout de même disposer du contenu audio converti en un ou deux canaux.

Le terme *upmix* désigne le procédé inverse qui vise à synthétiser un signal multicanal à partir d'un signal monophonique ou stéréophonique. En pratique, une personne qui souhaite utiliser son système de restitution multicanal mais qui ne dispose que d'un contenu audio sur un ou deux canaux peut utiliser ce procédé pour générer artificiellement un signal multicanal adapté à son système de restitution audio.

C.1 Réduction du nombre de canaux par matriçage : *downmix*

L'opération de conversion dénommée *downmix* vise à assurer la compatibilité des signaux audio multicanaux (typiquement au format 5.1) avec les systèmes stéréophoniques en respectant les critères suivants :

- la qualité audio du signal stéréophonique doit être acceptable à l'issue du processus de downmix *i.e.* comparable à celle d'un signal stéréophonique issu d'une prise de son traditionnelle,
- l'énergie du signal stéréo doit être équivalente à l'énergie du signal multicanal original *i.e.* conservation de l'énergie des sources et de l'effet de salle,
- la spatialisation contenue dans le signal multicanal original doit être préservée *i.e.* la position des sources sonores appartenant à l'image frontale doit être identique

La compatibilité avec les systèmes monophoniques peut être établie en respectant seulement ces deux premiers critères. En effet, les informations spatiales peuvent être en partie préservées avec une conversion stéréophonique alors qu'elles ne peuvent qu'être transmises séparément (informations auxiliaires) dans le cas d'une conversion monophonique (*cf.* paragraphe 2.2.3).

En règle générale, la quantité d'informations relatives à l'effet de salle (réverbération notamment) diffère d'un contenu multicanal à un contenu stéréophonique. Cela est dû principalement au fait que, sous l'hypothèse habituellement considérée par les méthodes de *downmix*/*upmix*, l'auditeur bénéficie à la fois d'une scène sonore frontale et de l'effet de salle diffusé principalement par les haut-parleurs arrières. Plus de détails sur la description des scènes sonores multicanales sont donnés au paragraphe 4.1. A l'inverse, avec une écoute stéréophonique, l'effet de salle est perçu comme provenant de la scène frontale. Par conséquent, une simple addition des canaux arrière avec les canaux frontaux résulterait en un signal stéréophonique trop réverbéré alors que l'effet de salle ne constitue pas l'information principale mais juste un indicateur de l'environnement sonore. Un raisonnement similaire peut être mené en considérant l'addition du canal central avec les canaux gauche et droit de façon à ce que l'image stéréophonique résultante ne soit pas trop étroite. Pour assurer un contrôle sur la conversion des signaux multicanaux, les procédés de *downmix* utilisent des coefficients de pondération et éventuellement des modifications de la phase des signaux pour respecter les critères énumérés.

La recommandation de l'UIT, UIT-R BS.775-1 dans [UIT775], préconise des équations d'encodage *i.e.* de matriçage, d'un signal au format 5.0 en un signal stéréophonique tel que :

$$\begin{cases} L_{ITU} = L + \dfrac{1}{\sqrt{2}} \cdot C + \Delta g \cdot Ls \\ R_{ITU} = R + \dfrac{1}{\sqrt{2}} \cdot C + \Delta g \cdot Rs \end{cases} . \qquad \text{(C.1)}$$

Cette recommandation préconise également une équation d'encodage d'un signal au format 5.0 en un signal monophonique tel que :

$$M_{ITU} = \Delta g \left(L + R \right) + C + \frac{1}{2} \left(Ls + Rs \right) . \qquad \text{(C.2)}$$

L'homogénéité des opérations de matriçage définies aux équations (C.1) et (C.2) se vérifie puisque :

$$M_{ITU} = \Delta g \left(L_{ITU} + R_{ITU} \right) . \qquad \text{(C.3)}$$

Le gain $\Delta g = 1/\sqrt{2}$ correspond à une atténuation de -3 dB des canaux arrières et central comparé aux gains attribués aux canaux frontaux. La recommandation de l'ITU reconnaît que ce matriçage n'est pas nécessairement approprié à tous les contenus audio multicanaux et propose la modification des coefficients de matriçage tels que l'atténuation entre les canaux varie entre 0 et -6 dB, soit Δg = [1, …,$1/\sqrt{2}$,… 1/2].

La **Figure C.1** présente d'une part l'effet du matriçage sur les réponses impulsionnelles théoriquement captées par un système multi-microphonique de type coïncident dont les capteurs pointent dans des directions différentes de l'espace relativement au positionnement du système 5.0 (voir [THE01] pour plus de détails sur le système multi-microphonique employé). La réponse impulsionnelle relative au canal gauche reproduit l'ensemble des sections temporelles captées à partir d'une impulsion brève (n°1) émise dans la salle. La réponse impulsionnelle associée au canal central traduit la présence de deux impulsions simultanées (n°1 et n°2) et de leurs réflexions primaires. La réponse impulsionnelle associée au canal arrière gauche traduit la présence des réflexions du son direct et de la réverbération tardive. La **Figure C.1-(a)** présente la réponse impulsionnelle gauche obtenue par *downmix* des réponses impulsionnelles originales associées aux canaux centre, gauche et arrière gauche. L'ordonnancement temporel du son direct, des réflexions (primaires et secondaires) et de la réverbération est conservé après le *downmix* puisqu'aucune différence de temps n'apparaît d'un canal à un autre excepté pour les réflexions du canal arrière gauche qui

peuvent être captées quelques millisecondes avant de l'être pour les canaux frontaux. Le mélange du son direct (impulsions n°1 et n°2) entre les canaux frontaux, et de la réverbération tardive entre le canal gauche et le canal arrière gauche, avec une pondération de -3 dB assure la préservation de l'énergie totale (source et effet de salle) dans le canal résultant.

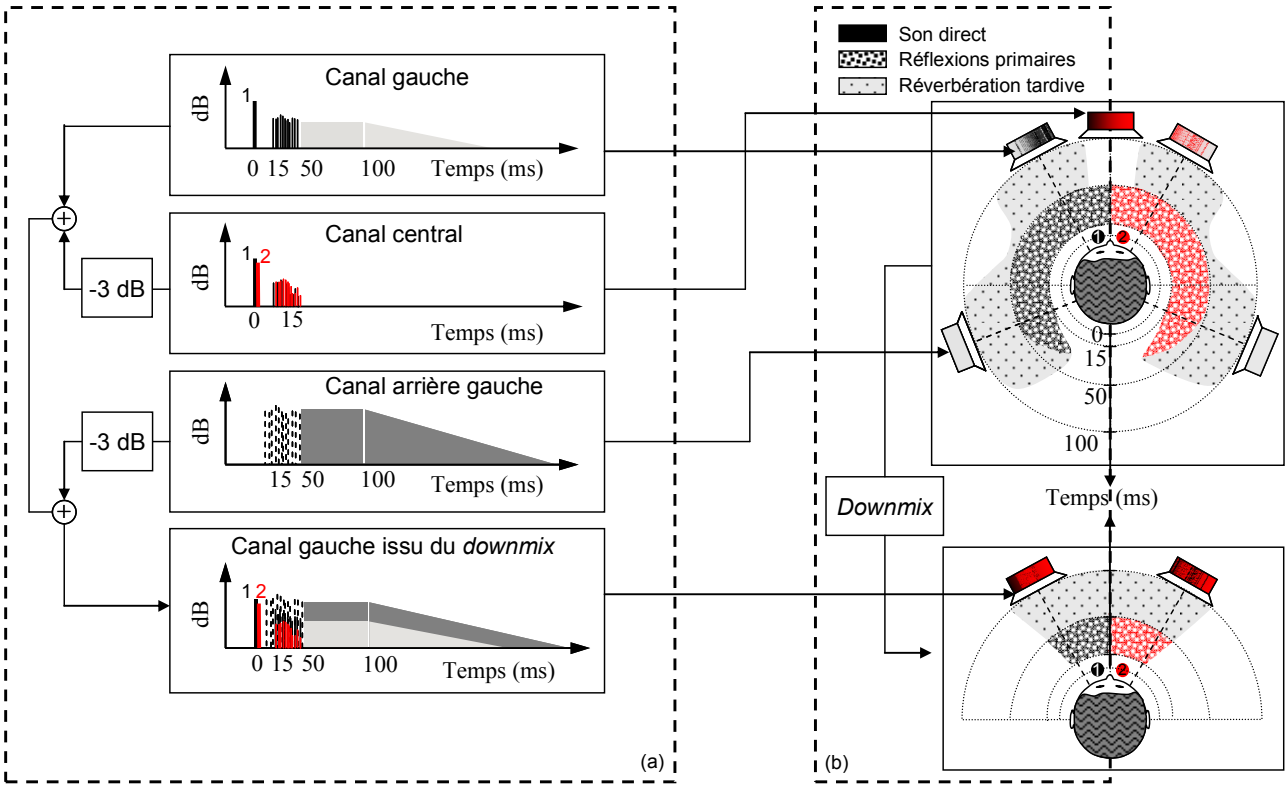

Figure C.1 **Effet du downmix** sur [**(a)** - la réponse impulsionnelle résultante, **(b)** - les sections temporelles du *downmix* stéréo résultant].

D'autre part, la **Figure C.1-(b)** présente l'évolution temporelle des signaux atteignant les oreilles de l'auditeur au centre d'un dispositif 5.0. Chaque canal peut être mis en correspondance avec les réponses impulsionnelles de la **Figure C.1-(a)**. Le son direct (impulsions n°1 et n°2) est donc diffusé par les canaux formant l'image frontale. D'après la loi du premier front d'onde décrit paragraphe 1.1.2, l'auditeur localise les sources à partir du son direct et par suite, comme provenant de deux directions différentes: l'une entre le haut-parleur gauche et le haut-parleur central (impulsions n°1) et l'autre entre le haut-parleur droit et le haut-parleur central (impulsions n°2). La perception des réflexions primaires, diffusées par tous les haut-parleurs, définissent ensuite la largeur et la profondeur des sources. Enfin, la réverbération tardive diffusée par les canaux gauche-droit et arrière gauche-droit procure à l'auditeur la sensation d'enveloppement et le renseigne sur le type et les composants de la salle (degré de réverbération). La comparaison avec le signal stéréophonique issu du *downmix* montre que la position spatiale des sources sonores est conservée. En effet, le son direct relatif aux deux impulsions est présent à la fois dans le canal gauche et le canal droit issu du *downmix*, puisque le canal central contient ces deux impulsions et que ce canal est introduit dans chacun des canaux matriçés (*cf.* équation (C.1)).

Finalement, les critères de conservation de l'énergie, de préservation de la position des sources et de qualité du signal stéréophonique (prise de son équivalente à une prise de son naturelle) issu du *downmix* sont respectés. Seule la perception spatiale du signal stéréophonique issu du downmix diffère de celle procurée par la diffusion du signal 5.0 original. Cette perception spatiale du son étant rendue possible avec les haut-parleurs arrières et la présence de :

- réflexions pour percevoir la largeur et la distance des sources,
- réverbération tardive pour la sensation d'enveloppement.

Outre les équations d'encodage fournies par l'UIT-R, une multitude d'autres jeux de coefficients existent. Citons notamment les travaux de Gerzon, dans [GERZ92b], et de Griesinger, dans [GRI96], qui proposent d'autres alternatives un peu plus sophistiquées dont les jeux de coefficients sont donnés dans [RUM01].

C.2 Conversion d'un signal stéréo en un signal multicanal : *upmix*

De nombreux travaux ont été menés pour synthétiser un signal dit pseudo-stéréophonique à partir d'un contenu audio monophonique [ORB70]-[GERZ92c]. Même si les résultats obtenus sont acoustiquement éloignés d'une prise de son stéréophonique naturelle, l'expérience a été, par la suite, reconduite pour la conversion des signaux stéréophoniques en signaux audio multicanaux typiquement au format 5.0.

La synthèse d'un signal multicanal acoustiquement réaliste à partir d'un signal stéréophonique est envisageable puisque un tel signal est constitué à la fois du contenu audio « basique » *i.e.* les sources sonores, et également des informations spatiales relatives aux positions des sources et à l'effet de salle (informations non-disponibles à partir d'un signal mono). Les objectifs fixés par de telles méthodes d'*upmix* sont de deux types :

- générer un canal central apte à stabiliser l'image frontale constituée des sources sonores *i.e.* positionner correctement les sources en face de l'auditeur,
- extraire l'ambiance ou l'effet de salle contenue dans le signal stéréophonique pour alimenter les canaux *surround.*

Un procédé d'*upmix* dit « aveugle » vise à synthétiser un signal multicanal à partir d'un signal stéréophonique sans aucune connaissance a priori sur sa génération (prise de son naturelle, mixage artificiel ou encore *downmix* d'un signal multicanal). L'idée générale des procédés d'*upmix* aveugle s'insère dans la problématique d'extraction des composantes présentes dans les canaux du signal stéréophonique dénommées composantes :

- primaires ou directionnelles *i.e.* son direct des sources sonores,
- d'ambiance *i.e.* réflexions et réverbération des sources dans l'espace environnant.

L'hypothèse de départ considère que le signal audio multicanal à l'origine du signal stéréophonique (même s'il n'existe pas) a été généré artificiellement ou naturellement tel que la scène frontale soit représentative des sources sonores directionnelles et que la scène arrière corresponde à l'ambiance ou effet de salle (**Figure C.2**). La définition d'une scène sonore multicanale habituellement associée aux contenus audio-visuels est donnée et illustrée au Chapitre 4 - section 4.1.

Nous présentons dans cette partie les techniques employées par deux méthodes représentatives de l'état de l'art en la matière. Irwan et Aarts, dans [IRW02], se basent sur l'Analyse en Composante Principale (ACP) du signal stéréophonique alors qu'Avendano et Jot, dans [AVE02]-[AVE04], proposent un traitement fréquentiel basé sur la cohérence des signaux. Notons d'ailleurs que la méthode basée sur l'ACP réalisée dans le domaine temporel [IRW02] a été récemment étendue à un traitement en sous-bandes de fréquence dans [LI05] dont l'intérêt est discuté au Chapitre 4. L'idée conjointe des deux approches consiste à tenir compte de l'inter-corrélation des canaux du signal stéréophonique soit à partir de la cohérence des canaux estimée dans le domaine fréquentiel (fonction de similarité) dans [AVE04] soit à partir de l'indice de corrélation estimé dans [IRW02]. Ainsi, l'information corrélée est considérée comme relative aux composantes directionnelles et l'information faiblement corrélée entre les canaux est considérée comme relative aux composantes d'ambiance.

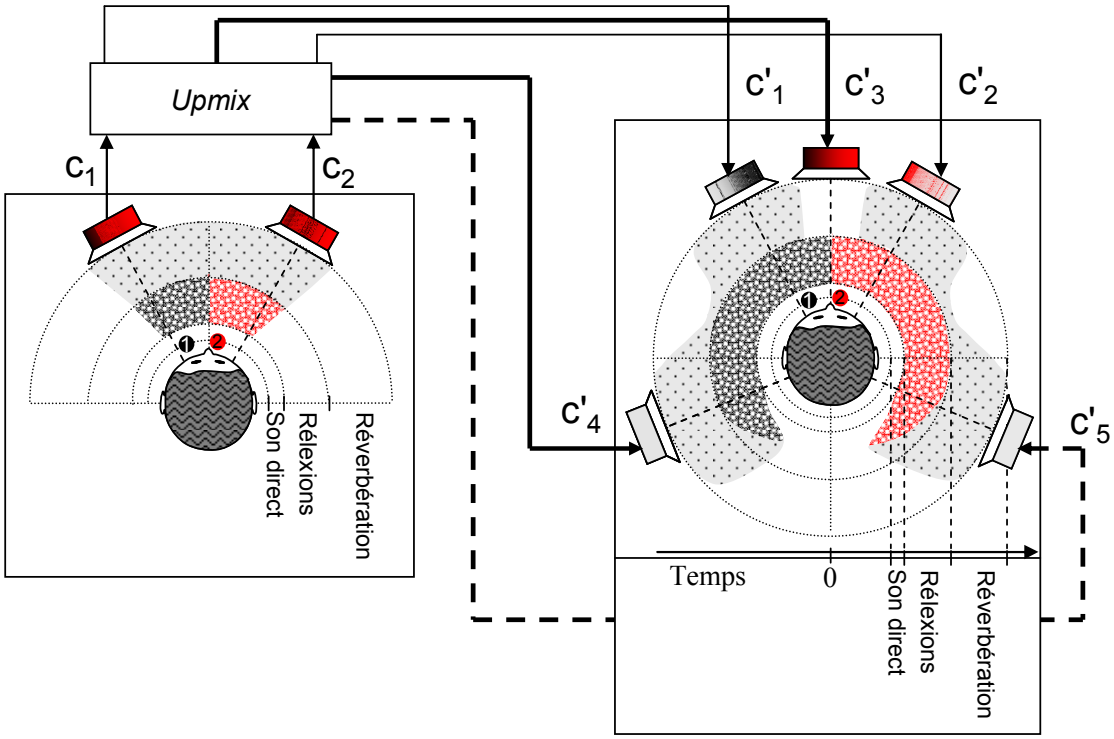

Figure C.2 **Principe de l'*upmix*** : à partir d'un signal stéréo (c_1,c_2), le procédé génère un canal central (c'_3) constitué des composantes directionnelles et deux canaux *surround* (c'_4,c'_5) constitués des composantes ambiances (réflexions et réverbération). Les canaux gauche et droit (c'_1,c'_2) sont dépourvus de certaines composantes extraites.

Outre la séparation des composantes, ces méthodes d'upmix visent à extraire la direction de l'image stéréophonique. Autrement dit, ces algorithmes estiment la position de la source dominante (d'un point de vue énergétique) provenant des haut-parleurs frontaux de façon à pouvoir générer un canal central dont l'énergie des composantes directionnelles coïncide avec la position spatiale des sources.

C.2.1 Renforcement directionnel

C.2.1.1 Approche opérant dans le domaine temporel

Nous avons présenté, dans un contexte de codage audio stéréophonique au paragraphe 2.2.2, comment un matriçage fixe de type somme-différence peut être remplacé par un matriçage adaptatif des canaux qui assure une meilleure concentration de l'énergie.

Dans un contexte d'*upmix* aveugle, l'utilisation de l'ACP est proposée dans [IRW02] de façon à estimer la direction de l'image stéréo au cours du temps en fonction de la covariance des canaux (gauche $c_1[n]$ et droit $c_2[n]$). L'expression de l'angle α est donnée par l'équation (2.4). Les données stéréophoniques peuvent alors être projetées sur la base des vecteurs propres (*cf.* paragraphe 2.2.2).

Le signal dominant d_1, relatif à la plus grande valeur propre, est ensuite utilisé pour générer un canal central constitué des composantes corrélées (sources virtuelles perçues entre les deux haut-parleurs notamment) et dépourvu des informations décorrélées des canaux originaux (réverbération tardive par exemple). Cette approche diffère du décodage *Dolby Pro Logic* (*cf.* paragraphe 2.3.1.3) dans la mesure où l'angle de rotation n'est pas calculé à partir de deux rapports d'énergie entre les signaux issus d'un matriçage de type somme/différence comme indiqué par l'équation (2.27).

Pour obtenir un renforcement directionnel de l'image frontale, la correspondance entre un signal stéréophonique à deux et à trois canaux est réalisée d'une part en doublant la valeur de l'angle estimé (*cf.* **Figure C.3**).

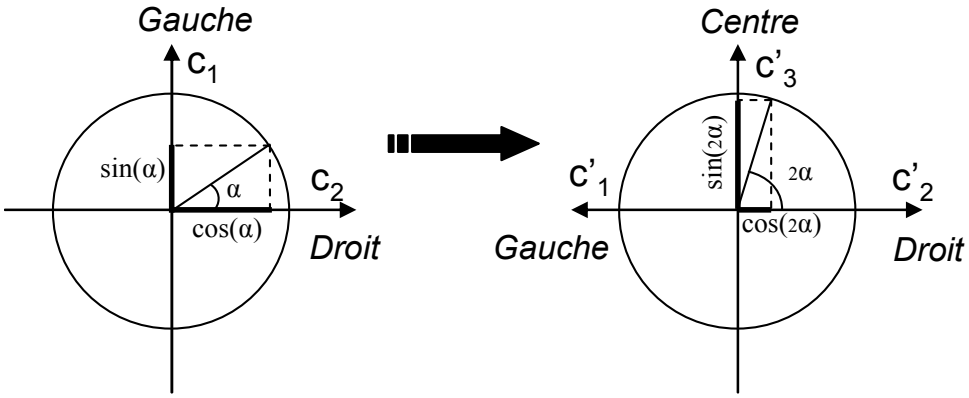

Figure C.3 Renforcement directionnel : deux vers trois canaux en doublant la valeur de l'angle α repérant la direction de l'image stéréophonique.

En outre, cette méthode d'*upmix* (2 vers 3) s'appuie sur la valeur de l'indice de corrélation des canaux originaux défini par :

$$\rho_{c_1c_2}[l] = \frac{r_{12}[l]}{\sqrt{r_{11}[l] \times r_{22}[l]}}. \tag{C.4}$$

Les auteurs de [IRW02] proposent de lever l'ambigüité dans l'estimation de α lorsque les canaux originaux sont faiblement corrélés en limitant les valeurs de l'indice de corrélation tel que :

$$\rho_0[l] = \begin{cases} \rho_{c_1c_2}[l], & \text{si } 0 \le \rho_{c_1c_2}[l] \le 1 \\ 0, & \text{sinon.} \end{cases}, \tag{C.5}$$

pour éviter l'estimation de valeurs négatives de α qui n'ont pas de sens dans l'interprétation donnée par la **Figure C.3**, ainsi $\alpha \in [0;\pi/2]$. De façon à respecter le critère de conservation de l'énergie, un angle β est défini comme représentatif de l'information dit *surround* considérée comme décorrélées en utilisant :

$$\beta[l] = \arcsin\left[1 - \rho_0[l]\right], \beta[l] \in [0;\pi/2]. \tag{C.6}$$

D'après la convention utilisée par les auteurs de [IRW02] présentée à la **Figure C.3**, le signal stéréophonique à trois canaux (c'_1,c'_2,c'_3) a pour expression :

$$\begin{pmatrix} c_1'[l] \\ c_2'[l] \\ c_3'[l] \end{pmatrix} = \begin{pmatrix} g_1[l] & \sin(\alpha[l])\cos(\beta[l]) \\ g_2[l] & \cos(\alpha[l])\cos(\beta[l]) \\ \sin(2\alpha[l]) & 0 \end{pmatrix} \cdot \begin{pmatrix} d_1[l] \\ d_2[l] \end{pmatrix}, \tag{C.7}$$

$$g_1[l]=\begin{cases}-\cos\left(2\alpha[l]\right), & \text{si } \cos\left(2\alpha[l]\right)<0\\ 0, & \text{sinon}\end{cases},$$
$$g_2[l]=\begin{cases}\cos\left(2\alpha[l]\right), & \text{si } \cos\left(2\alpha[l]\right)\geq 0\\ 0, & \text{sinon}\end{cases} \qquad \text{(C.8)}$$

$$\text{avec : } \begin{pmatrix}d_1[l]\\ d_2[l]\end{pmatrix}=\begin{pmatrix}\sin\left(\alpha[l]\right) & \cos\left(\alpha[l]\right)\\ \cos\left(\alpha[l]\right) & -\sin\left(\alpha[l]\right)\end{pmatrix}\cdot\begin{pmatrix}c_1[l]\\ c_2[l]\end{pmatrix}. \qquad \text{(C.9)}$$

C.2.1.2 Approche opérant dans le domaine fréquentiel

L'approche présentée dans [AVE04] repose sur l'estimation de la fonction de similarité des canaux gauche et droit. En notant $F_{C1}[l,k]$ et $F_{C2}[l,k]$ les transformées de Fourier à court terme (TFCT définie en Annexe A.1.1) des canaux, la fonction de similarité est définie telle que :

$$\psi[l,k]=2\frac{\left|F_{C_1}[l,k]F_{C_2}^{*}[l,k]\right|}{\left|F_{C_1}[l,k]\right|^2+\left|F_{C_2}[l,k]\right|^2}, \qquad \text{(C.10)}$$

où * dénotes la conjugaison complexe et k l'indice des fréquences. Sous l'hypothèse que le mélange d'une source, $s[n]$, dans les canaux est réalisé au moyen d'une panoramique d'intensité suivant la loi des tangentes (*cf.* paragraphe 1.2.1.2), alors $c_1[n] = \sin(\theta)\times s[n]$ et $c_2[n] = \cos(\theta)\times s[n]$, d'après [PUL97], et la source virtuelle est perçue à l'azimut θ entre les haut-parleurs écartés de 90°. En remplaçant les termes de TFCT de chaque canal par leur dépendance en θ, la direction de l'image stéréophonique ($\alpha=\theta$) peut finalement être estimée avec :

$$\alpha[l,k]=\begin{cases}\frac{1}{2}\arcsin\left(\psi[l,k]\right), & \text{si } \left|F_L[l,k]\right|<\left|F_R[l,k]\right|\\ \frac{1}{2}\left(\pi-\arcsin\left(\psi[l,k]\right)\right), & \text{sinon}\end{cases}. \qquad \text{(C.11)}$$

D'après la même représentation établie à la **Figure C.3**, l'opération, qui vise à convertir le signal stéréo en trois canaux frontaux, définie trois nouvelles fonctions de pondérations dépendantes du plan temps-fréquence, de la direction de l'image stéréophonique et des gains originellement appliqués aux sources tel que :

$$\begin{pmatrix}F_{C'_1}[l,k]\\ F_{C'_2}[l,k]\\ F_{C'_3}[l,k]\end{pmatrix}=\begin{pmatrix}\frac{g_1[l,k]}{\sin\alpha[l,k]}\\ \frac{g_2[l,k]}{\cos\alpha[l,k]}\\ \frac{\sin\left(2\alpha[l,k]\right)}{\sin\alpha[l,k]+\cos\alpha[l,k]}\end{pmatrix}\cdot\begin{pmatrix}F_{C_1}[l,k]\\ F_{C_2}[l,k]\\ F_{C_1}[l,k]+F_{C_2}[l,k]\end{pmatrix}, \qquad \text{(C.12)}$$

Ces pondérations expriment alors un nouveau panoramique d'intensité opérant dans le plan temps-fréquence. La **Figure C.4** présente une comparaison entre le panoramique original appliqué à la source s pour la stéréophonie à deux canaux, le panoramique par paire suivant

la même loi mais définie pour trois canaux et le panoramique compensé (par le panoramique stéréo déjà appliqués à la source) définie par l'équation (C.12).

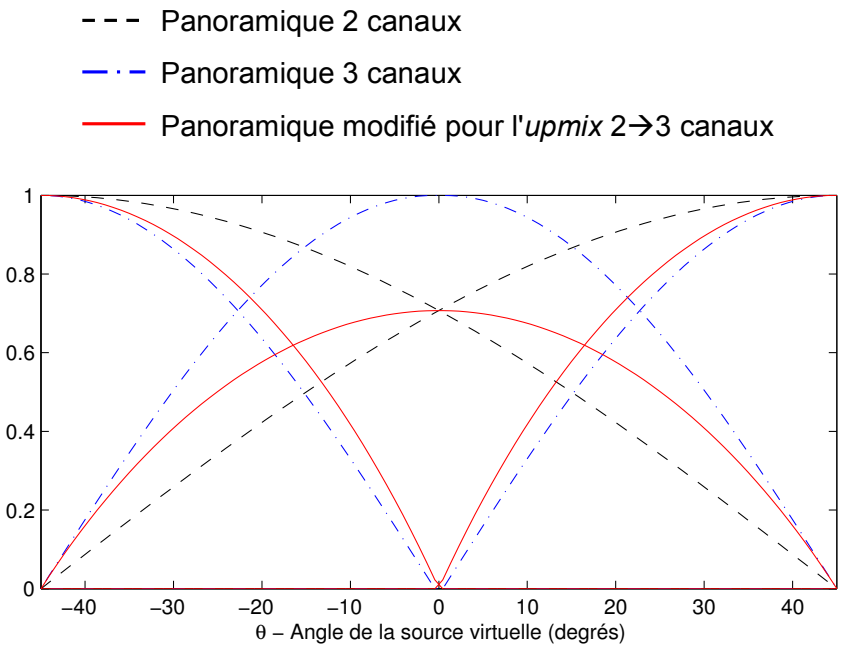

Figure C.4 Panoramique d'intensité (loi des tangentes) pour 2, 3 canaux et modifié pour l'upmix 2→3 canaux.

Cette approche diffère de celle employée dans [IRW02] dans la mesure où les signaux originaux ne sont pas projetés sur la base des vecteurs propres. Par contre, chose qui n'apparaît pas sur l'équation (C.12), cette approche, opérant dans le domaine fréquentiel, ne pondèrent pas directement les TFCT des canaux originaux ou leur mélange mais des composantes fréquentielles issues d'une processus d'extraction dit *unmix* décrit dans [AVE04]. En se basant sur les principes de séparation de sources, notamment ceux décrits dans [JOU00], et de la position des sources dérivée de la similarité donnée par l'équation (C.10), les composantes fréquentielles aux positions spatiales équivalentes sont regroupées (fenêtrage suivant les fréquences) avant d'appliquer le panoramique d'intensité compensé. Après la pondération, une transformation de Fourier à court terme inverse est appliquée pour chaque canal avec la méthode *overlap-add*.

C.2.2 Extraction de l'ambiance

C.2.2.1 Approche opérant dans le domaine temporel

Alors que le procédé de décodage *Dolby Pro Logic* se base sur la différence des canaux pour synthétiser un signal représentatif des informations d'ambiance, l'approche basée sur l'ACP [IRW02] synthétise un signal résiduel constitué des informations décorrélées dans les canaux stéréophoniques. Le signal résiduel d_2, relatif à la plus faible valeur propre, est obtenu par rotations des données stéréophoniques comme indiqué par l'équation (C.9). Pour respecter le critère de conservation d'énergie, le signal d'ambiance est finalement obtenu par :

$$c_4'[n] = \sin(\beta) \times d_2[n]. \qquad \text{(C.13)}$$

Ainsi, l'association des équations (C.7), (C.8) et (C.13) démontrent que l'*upmix* des canaux d_1 et d_2 en quatre canaux (c'_1, c'_2, c'_3, c'_4) conserve l'énergie initiale totale. En effet, le canal d_1 est distribué dans les canaux c'_i (i=1,…,4) au travers de gains dont la somme des carrés vaut 1,

i.e. $g^2_j + \sin^2(2\alpha)=1$ puisque à tout instant un des gains g_i est nul (j=1,2). De la même manière les gains affectés au canal d_2 vérifie la condition, *i.e.* $\cos^2(\beta).(\cos^2(\alpha)+\sin^2(\alpha)) + \sin^2(\beta)=1$.

C.2.2.2 Approche opérant dans le domaine fréquentiel

La méthode d'extraction d'ambiance décrite dans [AVE02] repose sur l'estimation de la fonction de cohérence inter-canal (ICC) à court terme telle que :

$$\mathrm{ICC}[l,k] = \frac{\Phi_{12}[l,k]}{\sqrt{\Phi_{11}[l,k]\Phi_{22}[l,k]}}, \tag{C.14}$$

$$\Phi_{ij}[l,k] = \mu \times \Phi_{ij}[l-1,k] + (1-\mu) \times F_{C_i}[l,k] F^*_{C_j}[l,k], \tag{C.15}$$

où μ est un facteur d'amortissement utilisé pour assurer la continuité de l'estimation au court du temps. Une valeur de la fonction de cohérence proche de 1 signifie que les composantes fréquentielles des signaux analysés sont corrélés et par suite que l'information relative aux sources directionnelles est dominante. A l'inverse une valeur proche de 0 signifie que les composantes fréquentielles analysées sont relatives au champ réverbéré. Le traitement fréquentiel utilisé pour extraire les composantes relatives à l'ambiance sonore consiste à filtrer chaque canal stéréophonique par une fonction dépendante de la cohérence tel que les régions temps-fréquence où la cohérence a une valeur élevée sont fortement atténuées alors que les régions avec une faible cohérence ne sont pas modifiées :

$$\begin{cases} F_{C'_4}[l,k] = f\left(1-\mathrm{ICC}[l,k]\right) \cdot F_{C_1}[l,k] \\ F_{C'_5}[l,k] = f\left(1-\mathrm{ICC}[l,k]\right) \cdot F_{C_2}[l,k] \end{cases}, \tag{C.16}$$

où f est une fonction tangente hyperbolique, par nature non-linéaire, (voir [AVE02] pour plus de détails) qui assure une transition douce pour la séparation des composantes directionnelles et d'ambiance. Après cette séparation fréquentielle des composantes, les signaux d'ambiance $c'_4[n]$ et $c'_5[n]$ sont obtenus par transformation de Fourier (à court terme) inverse avec la méthode *overlap-add*.

C.2.3 Décorrélation par filtrage passe-tout

L'avantage de la méthode d'upmix opérant dans le domaine fréquentiel est de pouvoir générer deux canaux *surround* représentatif de l'ambiance sonore et ainsi d'être capable d'alimenter un système de restitution 5.0. Cependant en pratique, les deux méthodes se rejoignent avec l'utilisation de filtres dit de « décorrélation ». Nous avons présenté au paragraphe (1.1.2.2) le lien entre l'ICC, qui peut être comparé au coefficient de corrélation des canaux stéréophoniques présenté à l'équation (C.4), et la largeur de source ou de scène sonore apparente. Basé sur ce phénomène acoustique, les filtres de décorrélation permettent d'augmenter la perception spatiale du son avec une sensation d'élargissement de la scène sonore. Ce traitement artificiel utilisé pour améliorer la perception de l'espace sonore se base sur les principes de la synthèse de réverbération. Pour réaliser des réverbérateurs artificiels, Schroeder proposa une approche, approfondie plus tard par Moorer [MOO79], fondée sur l'association de filtres en peignes et de filtres passe-tout. D'après [LAR95], les fonctions de transfert du filtre en peigne (*comb filter*) H^{cf} et du filtre passe-tout (*all-pass filter*) H^{apf}, représentés à la **Figure C.5**, sont données respectivement par :

$$H^{cf}(z) = \frac{z^{-P}}{1-az^{-P}} \quad \text{et} \quad H^{apf}(z) = \frac{z^{-P}-a}{1-az^{-P}} \tag{3.17}$$

où a correspond à la distance des pôles et des zéros du cercle unité ou encore à l'atténuation dans le temps de la réponse impulsionnelle des filtres. La valeur a=0,6 réalise un bon compromis entre l'amortissement de la réponse impulsionnelle et la modification de la phase du signal filtré (une valeur plus proche de 1 aura une influence encore plus grande). P désigne à la fois le nombre de pôles (et de zéros) et la distance entre les impulsions de la réponse impulsionnelle (plus P est grand et plus cette distance est grande). En pratique, une valeur de P de l'ordre de 6 à 10 ms (de l'ordre de la centaine d'échantillons) permet de ne pas trop modifier le timbre des signaux filtrés tout en élargissant l'image de l'ambiance perçue. La modification du timbre des signaux audio ou de parole est également appelée coloration (sonorité métallique) et résulte du filtrage en peigne introduit par le bouclage caractéristique de ces filtres présentés à la **Figure C.5**.

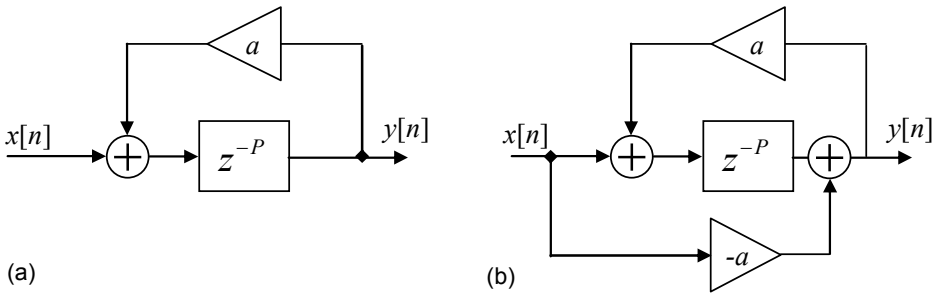

Figure C.5 tirée de [LAR95], **(a) – Filtre en peigne récursif, (b) – Filtre passe-tout.**

La **Figure C.6** présente les réponses en module et en phase d'un filtre en peigne et d'un filtre passe-tout obtenus avec le même paramétrage (a=0,6 et P=10). La réponse en module du filtre passe-tout est plate alors que celle du filtre en peigne va imprimer au signal traité une forte coloration harmonique caractéristique des résonances maximales pour la série de fréquences $\omega_k = 2\pi k / P$ si $a > 0$ ou $\omega_k = (1+2k)\pi / P$ si $a < 0$.

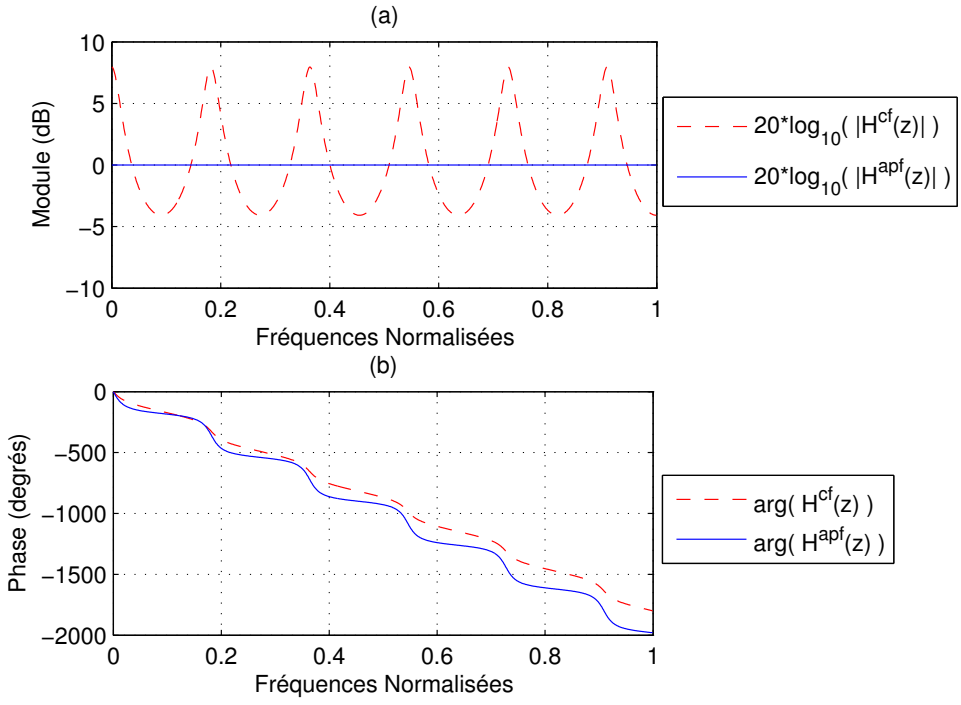

Figure C.6 Réponse en [(a) – module, (b) – phase] d'un filtre en peigne et d'un filtre passe-tout obtenus avec a=0,6 et P=10.

La réponse en phase du filtre passe-tout présente des points d'inflexion ($P/2$ points d'inflexion pour les fréquences positives) plus marqués que ceux du filtre en peigne. Par conséquent, le déphasage introduit par un filtrage passe-tout est plus important et résulte en une meilleure décorrélation des signaux (entre le signal d'entrée et celui de sortie) sans pour autant modifier leurs enveloppes temporelle et fréquentielle.

Comme l'illustre la **Figure C.7**, la perception des signaux aux contenus faiblement corrélés, diffusés en arrière plan, est relative à une image sonore plus large que celle obtenue avec des signaux corrélés d'après le principe de sommation de la localisation (*cf.* 1.1.1.2).

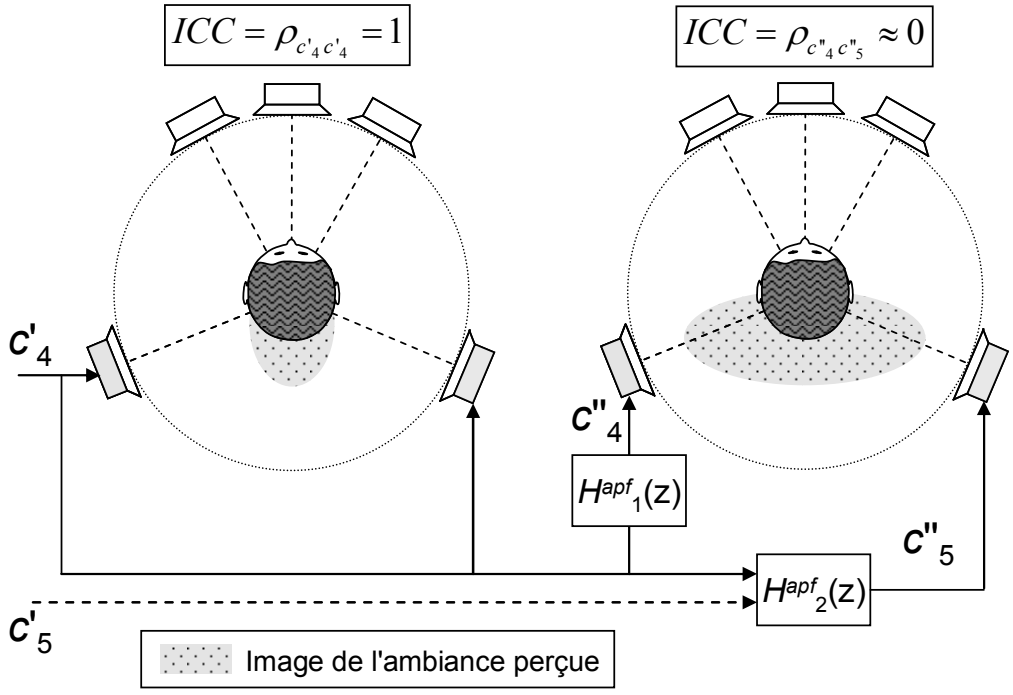

Figure C.7 Filtrage passe-tout décorrélateur $H^{apf}{}_1(z)$ et $H^{apf}{}_2(z)$ pour élargir la zone de l'ambiance perçue et améliorer l'enveloppement.

Chacune des méthodes d'*upmix* en aveugle assure la décorrélation des deux canaux *surround* (c''_4 et c''_5) en utilisant deux filtres passe-tout $H^{apf}{}_1(z)$ et $H^{apf}{}_2(z)$ suivants l'équation (3.17).

C.2.4 Comparaison des méthodes d'*upmix* aveugle

Alors que des tests subjectifs ont été menés pour comparer les performances des méthodes d'*upmix* de signaux stéréophoniques issus d'un matriçage (*cf.* paragraphe 2.3.1.4), la littérature ne présente pas de comparaison des performances des méthodes d'*upmix* plus évoluées opérant en aveugle. D'après nos expérimentations, nous pouvons donner quelques éléments relatifs à la perception des signaux traités par les deux méthodes. L'avantage principal de la méthode d'*upmix* basée sur des techniques fréquentielles est de pouvoir considérer plusieurs sources pour chaque portion de signal temporel. Cependant, les techniques fréquentielles utilisées pour la séparation des composantes introduit des artefacts audibles dus au recouvrement des sources dans le domaine spectral. La méthode d'*upmix* basée sur l'ACP à l'avantage d'introduire moins d'artefacts audibles aux dépens d'une moins bonne séparation des composantes qui résulte en une image sonore plus étroite. Un compromis entre les deux approches consiste à appliquer la méthode basée sur l'ACP mais cette fois-ci dans le domaine fréquentiel comme proposé dans [LI05]. L'intérêt de cette méthode et les performances de la séparation engendrée par une ACP réalisée en sous-bandes de fréquence sont présentés dans le chapitre 4 de ce document.

C.3 Compatibilité entre un signal multicanal et binaural

Il existe également des méthodes de *downmix/upmix* pour assurer la conversion d'un signal multicanal en un signal binaural et inversement. Ainsi, le caractère spatial du signal original peut être complètement conservé dans le cas d'un *downmix* binaural et potentiellement restitué dans le cas de l'*upmix* multicanal sous la contrainte du système de restitution multi-haut-parleurs employé.

L'approche utilisée pour convertir un signal multicanal en un signal binaural repose sur la notion de *haut-parleurs virtuels* pour une écoute au casque. Les contributions de chaque haut-parleur sont traitées par filtrage transaural (*cf.* paragraphe 1.2.1) et ensuite associées pour générer un signal binaural. La méthode présentée dans [JOT99] préconise un encodage ambisonique du signal multicanal qui autorise les mouvements de la tête de l'auditeur lors de la restitution binaurale (voir [DAN00] pour les manipulations du champ sonore: rotations, *focus*). Une évaluation subjective des différentes solutions dites de « *virtual home theater* » aujourd'hui sur le marché pour une reproduction sur casque ou haut-parleurs est proposé dans [LOR04]. Les résultats sont assez proches de ceux présentés au paragraphe 2.3.1.4 dans le sens où aucun système n'arrive à délivrer une qualité audio supérieure à celle du signal stéréo issu du *downmix* (de type passif avec des coefficients fixes au court du temps) du signal multicanal original.

La conversion d'un signal binaural en un signal audio multicanal est proposée par Jakka, dans [JAK05], qui exploite l'ITD estimé en sous-bandes de fréquence pour définir l'azimut ou la direction de l'image sonore perçue. La méthode employée se limite à une reproduction dans le plan horizontal (système 5.0) grâce à la génération d'un signal audio multicanal suivant le principe du panoramique d'intensité de type VBAP (*cf.* paragraphe 1.2.1.2). Le signal binaural est réduit en un signal monophonique (somme normalisée de signaux synchronisés au regard des ITD estimés) et en angles d'azimuts finalement convertis en ILD pour la synthèse finale par VBAP. Bien que les signaux binauraux offre une reproduction audio en trois dimensions (3D), le couple ITD-ILD n'assure pas une reconstruction audio 3D complète (*cf.* paragraphe 1.1), il s'avère même que les confusions avant-arrière dans le plan horizontal posent les limites de la méthode proposée dans [JAK05]. Cependant, ces méthodes apparaissent comme très attrayantes pour assurer la diffusion de signaux audio spatialisés à la fois sur des systèmes multi-haut-parleurs et au casque; dans cette optique les travaux menés par le groupe de travail MPEG vise à intégrer ce type de technologies au sein du standard MPEG *surround* détaillé au paragraphe 2.3.3.

D. Distributions des paramètres utilisés pour le codage basé sur l'ACP

D.1 Base d'apprentissage de signaux stéréo

Nom de l'échantillon / Artiste	Description de la scène sonore (stéréo)	f_S(kHz) / résolution (bits)
Bjork	Voix chantée + sons synthétiques	44100 / 16
Musique Pop	Violons, cuivres, voix chantée (fortes attaques)	48000 / 16
The Corrs	Violon, piano, percussions, guitare, flûte irlandaise	44100 / 16
Fiona Apple	Voix chantée, piano, batterie (cloche)	44100 / 16
Tracy Chapman	Voix chantée, marimba, cloche, percussions	48000 / 16
Sting (downmix 5.1)	Voix chantée, guitares, batterie (forte réverbération)	48000 / 16
Carla Bruni	Voix chantée, guitares (positions latérales), violon	44100 / 16
Dire Strait	Synthétiseur + effets stéréo	48000 / 16
Femi Kuti "Look Around"	Cuivres, synthétiseur, guitares, percussions	44100 / 16
Femi Kuti "Beng"	Saxophone, guitare, batterie, percussions	44100 / 16
James Bond	Voix, bruits de voiture + effets spéciaux	48000 / 16
Jazz	Voix chantée, piano, batterie, saxophone	48000 / 16
Cold Play	Voix chantée, guitares, batterie + effets	44100 / 16
PJ Harvey	Voix chantées (réverbération), piano, guitares, batterie	44100 / 16
J.S. Bach	Concerto (violoncelles)	44100 / 16
A Train	Jazz (saxo, batterie + effet de salle prononcé)	44100 / 16
Jeff Buckley	Voix chantée (réverbération), batterie, guitares	44100 / 16
Mozart	Opéra (voix chantée + orchestre)	44100 / 16
Miles Davis	Cuivres, piano, batterie, contrebasse	44100 / 16
Sweden Radio (downmix 5.1)	Voix en mouvement (fort environnement sonore)	48000 / 16

Tableau D.1 Caractéristiques de la base d'apprentissage stéréo d'une durée totale de 30 minutes.

D.2 Distributions des angles de rotation

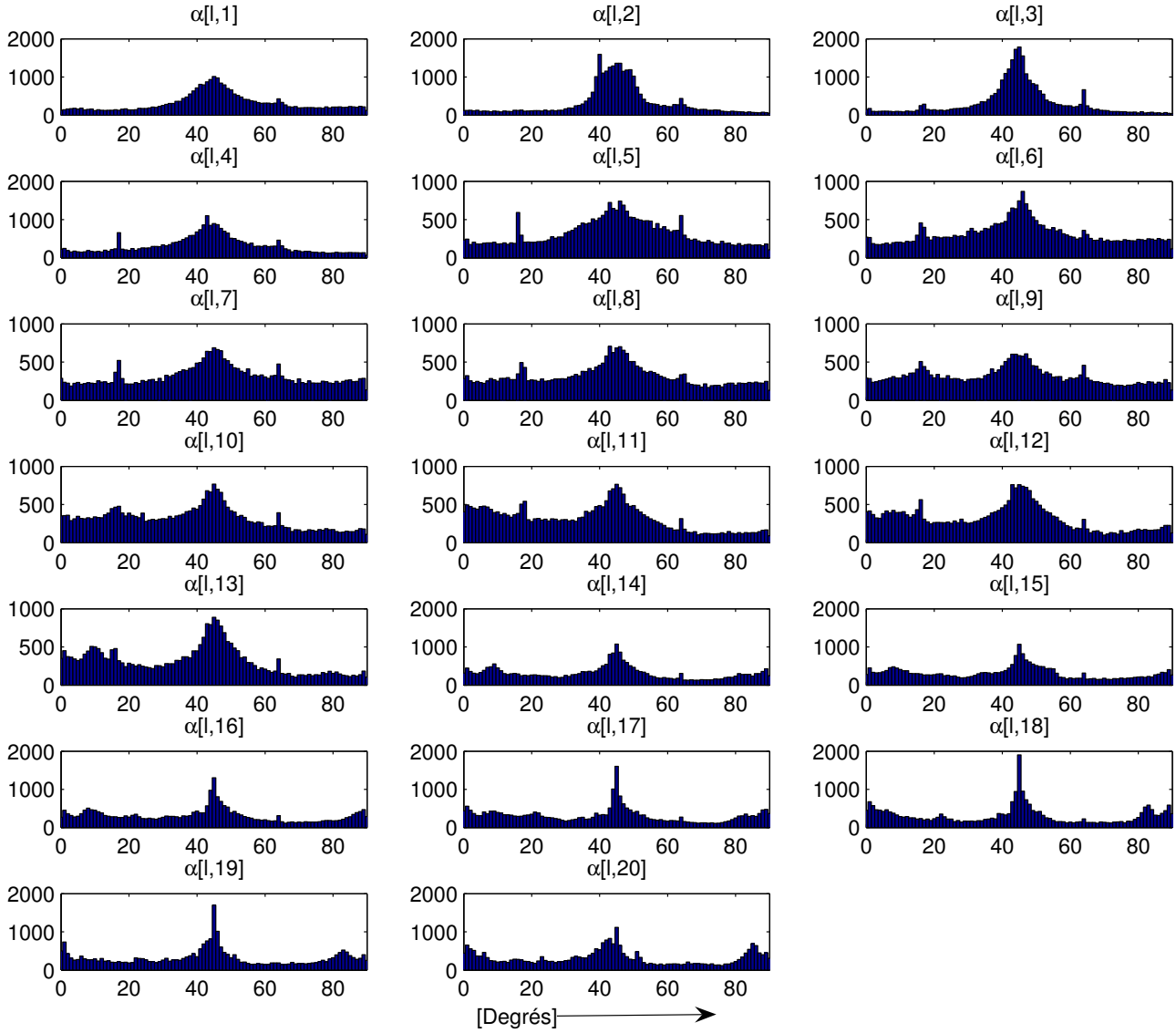

Figure D.1 **Distributions des angles de rotation en sous-bandes** (en degrés) estimés à partir d'un corpus de signaux stéréo, soit 20 signaux d'une durée totale de 30 minutes. La longueur de la fenêtre d'analyse est de 2048 échantillons et le nombre de sous-bandes égal à K_b=20.

A partir de la base d'apprentissage présentée en Annexe D.1, les angles de rotation sont estimés à partir de la covariance des canaux (spectres en sous-bandes) comme indiqué par l'équation (4.41). La conversion en degrés de ces angles de rotation estimés en sous-bandes génère des valeurs qui varient dans l'intervalle [0;90] degrés. La **Figure D.1** présente les distributions des angles de rotation $\alpha[l,b]$ estimés à partir des spectres, d'un large corpus de signaux stéréo, séparés en K_b=20 sous-bandes. Il apparaît assez nettement que plus la fréquence centrale de la sous-bande augmente et plus la distribution des angles estimés s'éloigne d'une distribution gaussienne.

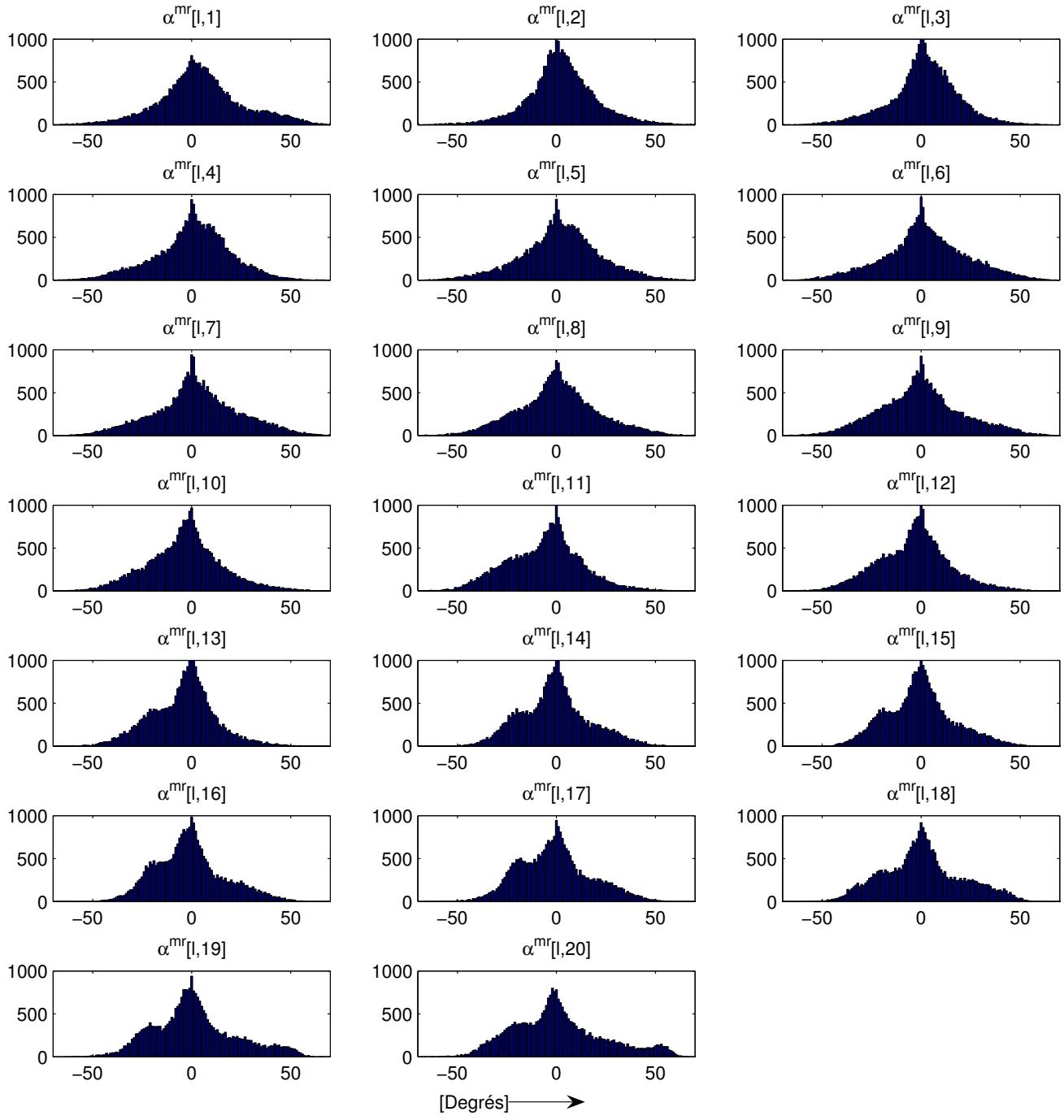

Figure D.2 Distributions en sous-bandes des angles de rotation à moyenne retranchée α_{mr} en degrés. La longueur de la fenêtre d'analyse (indice *l*) du corpus est de N=2048 échantillons et le nombre de sous-bandes K_b=20.

D'après la **Figure D.2** et la **Figure D.3**, les distributions en sous-bandes des angles de rotation différentiels (*cf.* paragraphe 5.1.2.2) sont nettement plus resserrées (piquées) autour de la valeur zéro. Autrement dit, les distributions en sous-bandes des angles différentiels sont nettement moins étalées que les distributions en sous-bandes des angles de rotation à moyenne retranchée et cela d'autant plus que la fréquence centrale des sous-bandes augmente.

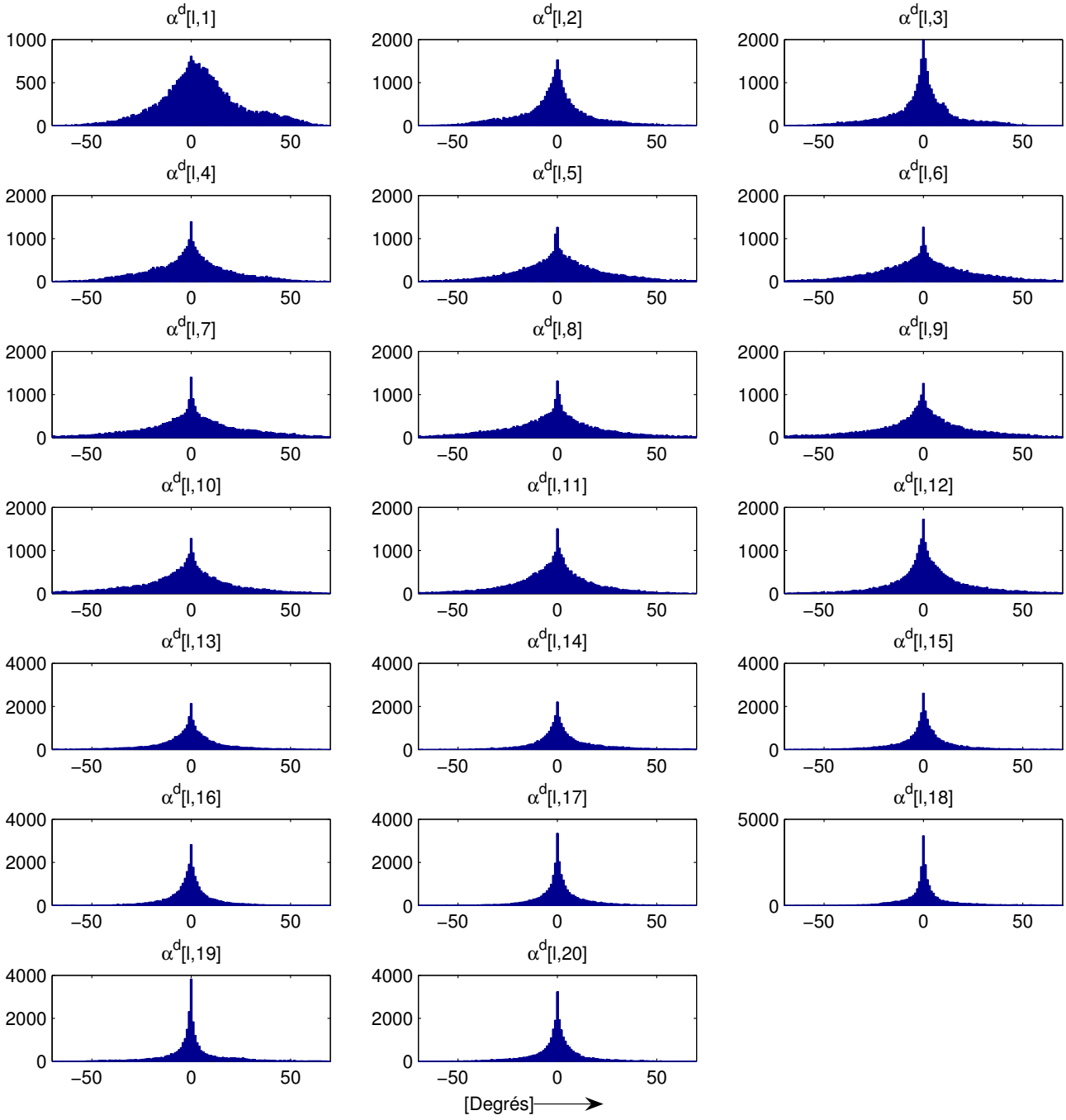

Figure D.3 Distributions en sous-bandes des angles de rotation différentiels α_d en degrés. La longueur de la fenêtre (indice *l*) d'analyse du corpus est de N=2048 échantillons et le nombre de sous-bandes K_b=20.

D.3 Distributions des paramètres énergétiques

Les distributions de l'énergie (différentielle en sous-bandes définie au paragraphe 5.1.2.2) de la composante ambiance (E_{D2}) et du rapport d'énergie entre les composantes issues de l'ACP bidimensionnelle ($RCPA_{12}$ défini au paragraphe 5.1.2.1) sont comparées à partir de la **Figure D.4** et de la **Figure D.5** suite à l'extraction de ces paramètres à partir du même corpus de signaux stéréo (durée de 30 minutes) utilisés en Annexe D.1.

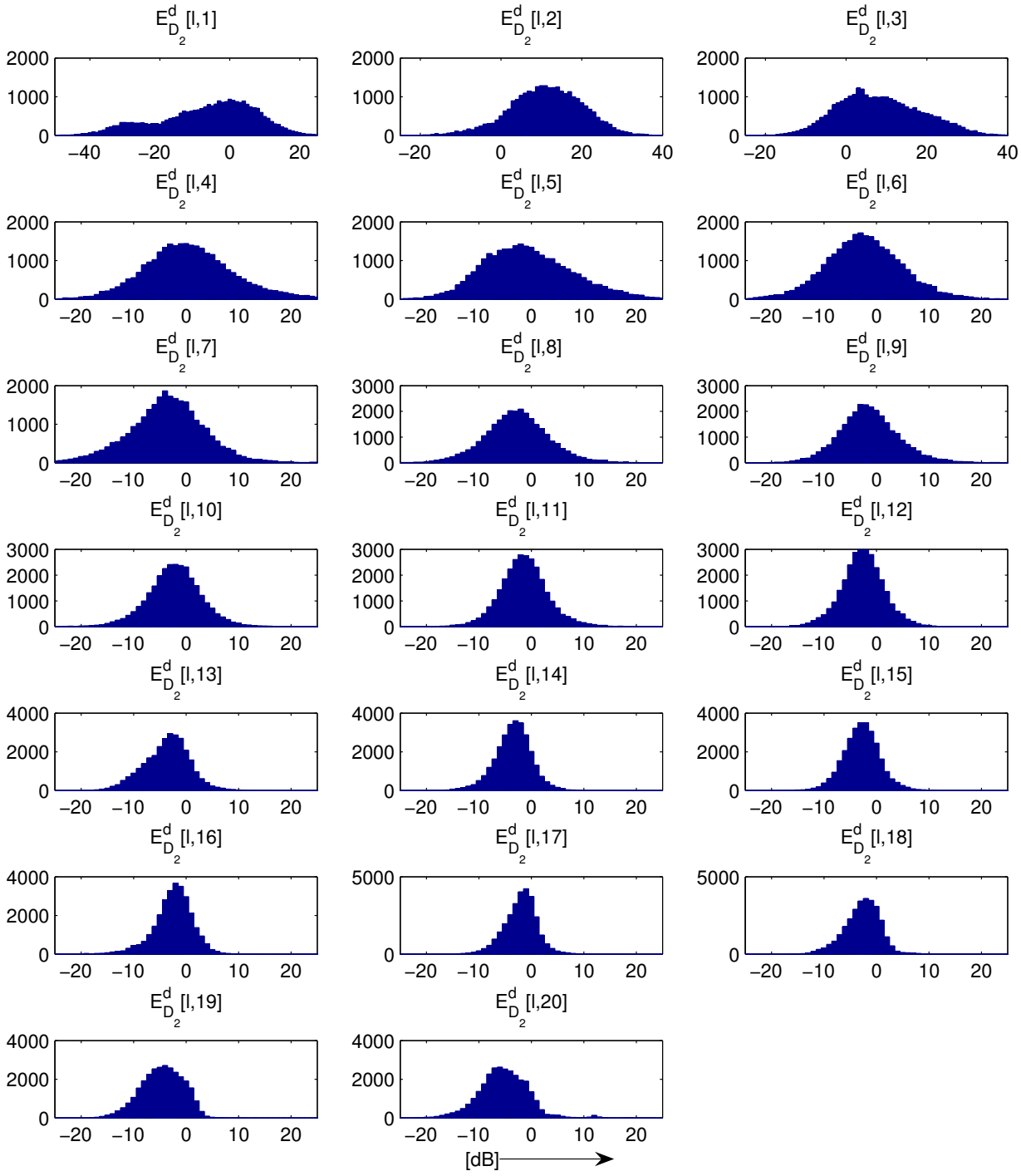

Figure D.4 Distribution (en dB) de l'énergie en sous-bandes de la composante ambiance $E^d_{D2}[l,b]$. La longueur de la fenêtre d'analyse du corpus est de N=2048 échantillons et le nombre de sous-bandes K_b=20.

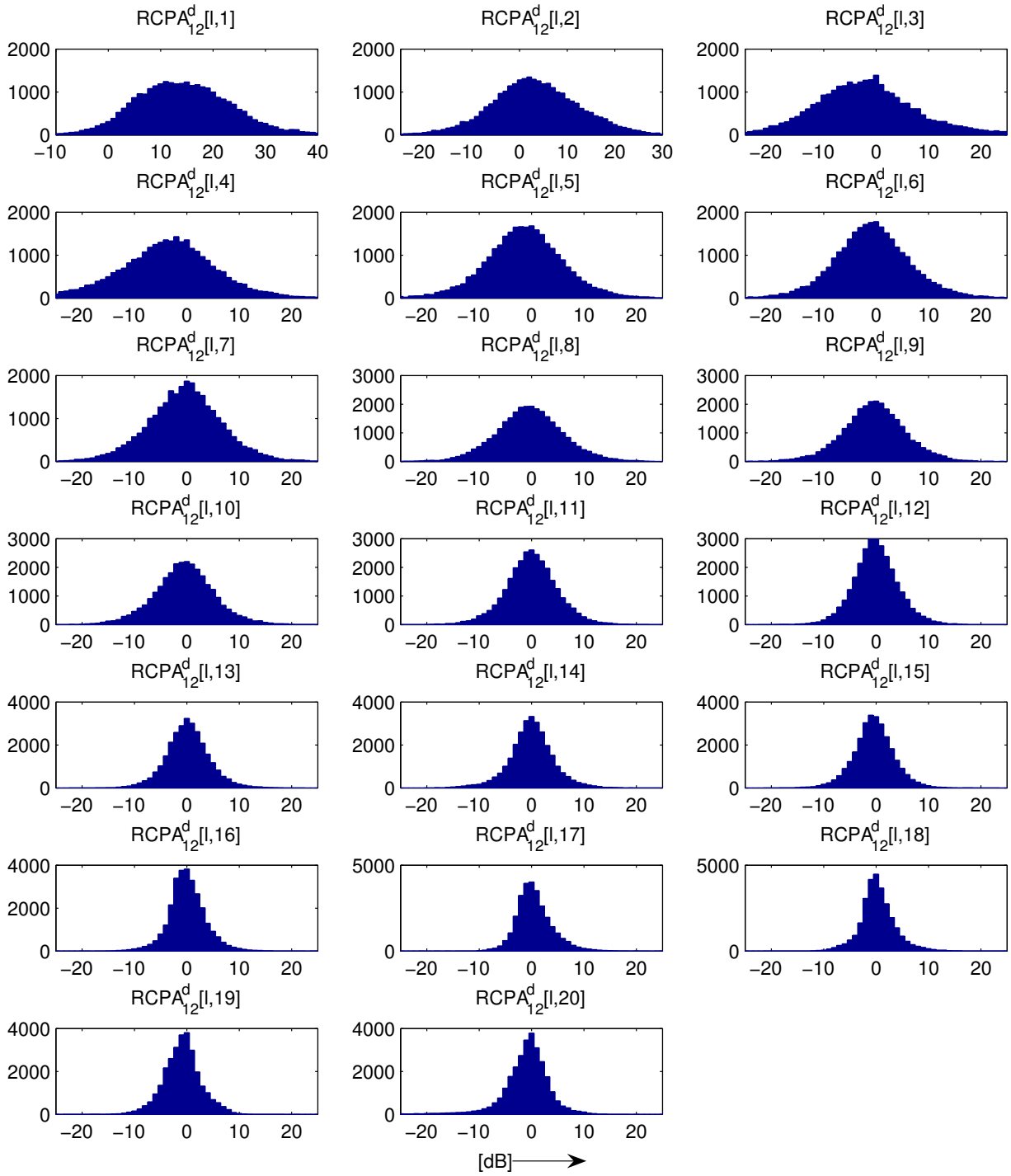

Figure D.5 Distribution (en dB) du RCPA$^{d}_{12}$[*l,b*] en sous-bandes. La longueur de la fenêtre d'analyse du corpus est de N=2048 échantillons et le nombre de sous-bandes K_b=20.

Globalement, les distributions en sous-bandes du $RCPA_{12}$ différentiel apparaissent plus resserrées autour de la valeur zéros (gaussiennes) que celles des énergies en sous-bandes du signal D_2.

REFERENCES BIBLIOGRAPHIQUES

[AES01] **F. Rumsey, D. Griesinger, T. Holman, M. Sawaguchi, G. Steinke, G. Theile et T. Wakatuki**, *AES Technical Document* AESTD1001.1.01-10, *Multichannel surround sound systems and operations*, 2001.

[ASA00] **F. Asano et al.,** "Effect of PCA filter in blind source separation", *Proceedings ICA2000*, pp.57-62, 2000.

[ATS01] **Advanced Television Systems Committee**, "ATSC Standard: Digital Audio Compression (AC-3), Revision A", 2001.

[AVE02] **C. Avendano et J.-M. Jot**, "Ambience Extraction and Synthesis from Stereo Signals for Multichannel Audio Up-mix", Proc. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pages 1957-1960, Orlando, Mai 2002.

[AVE04] **C. Avendano et J.-M. Jot**, "A frequency-Domain Approach to Multichannel Upmix", *J. Audio Eng. Soc.*, 52(7/8):740-749, 2004.

[BAR05] **M. Bartkowiak et T. Żernicki**, "A simple adaptive matrixing scheme for efficient coding of stereo sound", Proc. 13th *European Signal Proc. Conf.*, Sept. 2005, Antalya, Turkey.

[BAS03] **A. Baskind et O. Warusfel**, "Methods for blind computational estimation of perceptual attributes of room acoustics", Proc. AES 22nd *Int. Conf.*, 2003.

[BAU95] **F. Baumgarte, C. Feredikis et H. Fuchs**, "A Non-linear psychoacoustic model applied to the ISO MPEG Layer III Coder", Proc. 99th *AES Convention*, New York, Octobre 1995, preprint 4087.

[BAU02a] **F. Baumgarte et C. Faller**, "Design and evaluation of binaural cue coding schemes," Proc. 113th *AES Convention*, Los Angeles, USA, 2002, preprint 5706.

[BAU02b] **F. Baumgarte et C. Faller**, "Why Binaural Cue Coding is better than Intensity stereo", Proc. 112th *AES Convention*, Munich, 2002, preprint 5575.

[BEL94] **A. Belouchrani et J.-F. Cardoso**, "Maximum likelihood source separation for discrete sources", Proc. *EUSIPCO'94*, pages 768-771, 1994.

[BEL95] **A.J. Bell et T.J. Sejnowski**, "An information-maximization approach to blind source separation and blind deconvolution", *Neural Computation*, 7:1129-1159, 1995.

[BER96] **L. Beranek**, "Concert and Opera Halls: How they sound?", *Acoustical Society of America*, Woodbury, NY, 1996.

[BLA97] **J. Blauert**, *Spatial Hearing: The Psychoacoustics of Human Sound Localization*, MIT Press, Cambridge, USA, 1997.

[BLU31] **A. Blumlein**, "Improvements in and relating to sound transmission, sound recording and sound reproduction systems", *British Patent Specification 394325*, 1931. Reprinted in Stereophonic Techniques, *Aud. Eng. Soc.*, New York, 1986.

[BOI87] **R. Boite et M. Kunt**, *Traitement de la parole*, Presses Polytechniques Romandes, 1ère édition, 1987.

[BOU04] **M. Bouéri et C. Kyirakakis**, "Audio Signal Decorrelation Based on a Critical Band Approach", Proc. 117th *AES convention*, San Francisco, 2004, preprint 6291.

[BOS97] **M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson and Y. Oikawa**, "ISO/IEC MPEG-2 Advanced Audio Coding", *J. Audio Eng. Soc.*, vol. 45, No. 10, 1997.

[BOS02] **M. Bosi et R.E. Goldberg**, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, Dordrecht, 2002.

[BOY02] **R. Boyer**, "Modélisation et Codage de Signaux Audio par Extension du Modèle Sinusoïdal - Représentations Compactes des Signaux à Variations Rapides", *Thèse de l'ENST*, Paris, 2002.

[BOY03] **R. Boyer, S. Essid, K. Abed-Meraim et N. Moreau**, "Modèles sinusoïdaux étendus pour le codage audio", *19ème Colloque sur le Traitement du Signal et des Images*, Paris, Septembre 2003.

[BRA94] **K. Brandenburg et G. Stoll**, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio", *Journal of the Audio Engineering Society*, vol. 42, pp.780-791, Octobre 1994.

[BRE04] **J. Breebaart, S van de Par, A. Kohlrausch et E. Schuijers**, "High-quality parametric spatial audio coding at low bit rates" Proc. 116th *AES Conv.*, Berlin, Mai 2004.

[BRE05a] **J. Breebaart, S. van de Par, A. Kohlrausch et E. Schuijers**, "Parametric Coding of Stereo Audio", *EURASIP Journal on Applied Signal Processing*, 2005:9, 1305-1322.

[BRE05b] **J. Breebaart, J. Herre, C. Faller, J. Roden, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjorling et W. Oomen**, "MPEG spatial audio coding / MPEG surround: overview and current status" Proc. 119th *AES Conv.*, New York, USA, 2005.

[BRI06a] **M. Briand, D. Virette et N. Martin**, "Parametric representation of Multichannel Audio based on Principal Component Analysis", Proc. 120th *AES Conv.*, Paris, 2006, preprint 6813.

[BRI06b] **M. Briand, D. Virette et N. Martin**, "Parametric Coding of Stereo Audio based on Principal Component Analysis", 9th *Int. Conf. on Digital Audio Effects DAFx'06*, Montréal, 2006.

[BRI02] **A.C. den Brinker, E.G.P. Schuijers et A.W.J. Oomen**, "Parametric coding for high-quality audio", Proc. 112th *AES Conv.*, Munich, Mai 2002, preprint 5554.

[BRU99] **D.S. Brungart et W.M. Rabinowitz**, "Auditory localization of nearby sources. Head-related transfer functions", *Journal of the Acoustical Society of America*, vol. 106, pp.1465–1479, 1999.

[BUS06] **S. Busson**, "Individualisation d'indices acoustiques pour la synthèse binaurale", *Thèse de l'Université de la Méditerranée Aix-Marseille II*, 2006.

[CAN00] **G. Canévet**, "Eléments de Psychoacoustique", Laboratoire de Mécanique et d'Acoustique, Janvier 2000.

[CAR97] **J. F. Cardoso**, "Infomax and maximum likelihood for blind source separation". *IEEE Signal Processing Letters*, 4(4):112-114, avril 1997.

[CHE01] **C.I. Cheng et G.H. Wakefield**, "Introduction to head-related transfer functions (hrtfs): representations of hrtfs in time, frequency and space", *J. Audio Eng. Soc.*, 49(4):231-249, 2001.

[CHO71] **J.M. Chowning**, "The simulation of moving sound sources", *J. Audio Eng. Soc.*, vol.19, pp.2-6, 1971.

[COM94] **P. Comon**, "Independant component analysis, A new concept?", *Signal Processing*, 36:287-314, 1994.

[DAN00] **J. Daniel**, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia", *Thèse de l'Université Paris VI*, Juin 2000.

[DAN04] **J. Daniel et S. Moreau**, "Futher study of Sound Field Coding with High Order Ambisonics", Proc. 116th *AES Conv.*, Berlin, 2004.

[DAU00] **L. Daudet**, "Représentations structurelles de signaux audiophoniques – Méthodes hybrides pour des applications à la compression", *Thèse de l'Université de Provence*, 2000.

[DAV95] **M. Davis**, "The AC-3 Multichannel coder", Proc. 95th *AES Conv.*, 1993.

[DOB03] **D. Dobler**, "Real Time Sound Synthesis and transformation of ambients sounds", *Ph.D. Thesis*, University of Central Florida, 2003.

[DIE02] **M. Dietz, L. Liljeryd, K. Kjörling et O. Kunz**, "Spectral Band Replication, a novel approach in audio coding", Proc. 112th *AES Conv.*, Munich, 2002, preprint 5553.

[DUH90] **P. Duhamel et M. Vetterli**, "Fast Fourier Transforms: A tutorial review and a state of the art", *Signal Processing*, 19(4):259-299, 1990.

[DUH01] **T. Duhoo,** "Utilisation de la matrice de rotation des angles d'Euler dans l'étude de problèmes de dimension 3", *Bulletin de l'union des physiciens*, vol.95, Janvier 2001.

[DRE00] **R. Dressler**, "Dolby surround Pro Logic decoder principles of operation", *Technical Report*, Dolby Laboratories, 2000.

[DUR99] **M. Durnerin**, "Une stratégie pour l'interprétation en analyse spectrale. Détection et caractérisation des composantes d'un spectre ", *Thèse de l'INP-Grenoble*, 1999.

[EDL95] **B. Edler**, "Technical description of the MPEG-4 audio coding proposal from University of Hannover and Deutsche Bundespost Telekom", ISO/IEC, JTC1/SC29/WG11 MPEG95/MO414, Oct. 1995.

[EME95] **M. Emerit**, "Simulation binaurale de l'acoustique des salles de concert", *Thèse de l'INP-Grenoble*, CSTB, 1995.

[FAL02] **C. Faller et F. Baumgarte**, "Binaural cue coding: a novel and efficient representation of spatial audio", Proc. *IEEE Int. Conf. Acoustics, Speech, Signal Proc.* (ICASSP'02), vol. 2, pp. 1841-1844, Orlando, USA, Mai 2002.

[FAL04] **C. Faller**, "Parametric Coding of Spatial Audio", *Ph.D. Thesis*, Ecole Polytechnique Fédérale de Lausanne, 2004.

[FAL06a] **C. Faller**, "Parametric Multichannel Audio Coding: Synthesis of Coherence Cues", *IEEE Transactions on Speech and Audio Proc.*: vol.14, pp: 299-310, 2006.

[FAL06b] **C. Faller**, "Parametric Joint-Coding of Audio Sources", Proc. 120th *AES Convention*, Paris, 2006.

[FIE96] **L. Fielder, M. Bosi, G. Davison, M. Davis, C. Todd et S. Vernon**, "AC-2 and AC-3: Low-Complexity Transform-Based Audio Coding", *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist et C. Grewin, Eds., pp. 54-72, AES 1996.

[FUC93] **H. Fuchs**, "Improving Joint Stereo Audio Coding by Adaptive Interchannel Prediction", Proc. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, Oct. 1993.

[GAR97] **W.G. Gardner**, *3-D Audio Using Loudspeakers*. M.I.T. Media Laboratory, Kluwer Academic Publishers, 1997.

[GAR00] **J. Garas**, *Adaptive 3D Sound Sytems*, Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[GER92] **A. Gersho et R.M. Gray**, *Vector quantization and signal compression*, Kluwer, Boston, 1992.

[GER99] **M.A. Gerzon, P.G. Craven, J.R. Stuart, M. Law et J.R. Wilson**, "The MLP Lossless Compression System", 17th *AES Int. Conf.*, Italie, 1999.

[GERZ92a] **M.A. Gerzon**, "Panpot Laws for Multispeaker Stereo", 92nd *AES Conv.*, preprint 3309, 1992.

[GERZ92b] **M.A. Gerzon**, "Compatibility of and conversion between multi-speakers systems", 93rd *AES Conv.*, preprint 3405, 1992.

[GERZ92c] **M.A. Gerzon**, "Signal processing for simulating realistic Stereo images", 93rd *AES Conv.*, preprint 323, 1992.

[GLA90] **B.R. Glasberg et B.C.J. Moore**, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research* 47, pp.103-147, 1990.

[GOL96] **G.H. Golub et C. F. Van Loan**, "Matrix Computations", *The Johns Hopkins University Press*, Third Edition, 1996.

[GOO97] **M.M. Goodwin**, "Adaptive Signal Models: Theory, Algorithms and Audio Applications", *Ph.D. Thesis*, MIT, 1997.

[GRI96] **D. Griesinger**, "Multichannel Matrix Surround Decoders for Two-Eared Listeners", 101st *AES Conv.*, preprint 4402, 1996.

[HAN01] **M. Hans et R. Shafer**, "Lossless compression of digital audio", *IEEE Signal Processing magazine*, pp.21-32, 2001.

[HAY01] **S. Haykin,** *Communication Systems*, 4th edition, Wiley, New York, 2001.

[HER86] **J. Herault et C. Jutten, "**Space or time adaptive signal processing by neural models". Proc. *AIP Conference: Neural Networks for Computing*, pages 206-211, American Institute of Physics, 1986.

[HER92] **J. Herre, K. Brandenburg et E. Eberlein**, "Combined Stereo Coding", 93rd *AES Conv.*, San Francisco, 1992, preprint 3369.

[HER94] **J. Herre, K. Brandenburg and D. Lederer**, "Intensity stereo coding", 96th *AES Conv.*, Amsterdam, 1994, preprint 3799.

[HER04] **J. Herre**, "From Joint Stereo to Spatial Audio Coding – Recent Progress and Standardization", 7th *Int. Conf. on Digital Audio Effects DAFx'04*, Naples, 2004.

[HER05] **J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjorling, E. Schuijers, J. Hilpert and F. Myburg**, "The reference model architecture for MPEG spatial audio coding", Proc. 118th *AES convention*, Barcelona, 2005, preprint 6447.

[HOR00] **C. Hory**, "Mélanges de distributions du χ^2 pour l'interprétation d'une représentation temps fréquence", *Thèse de l'INP-Grenoble*, 2002.

[HOR02] **C. Hory, N. Martin et A. Chehikian**, "Spectrogram segmentation by means of statistical features of non-stationary signal interpretation" *IEEE Transactions on Signal Processing*, vol. 50, No. 12, pp. 2915–2925, Décembre 2002.

[HOT33] **H. Hotelling**, "Analysis of a complex of statistical variables into principal components", *J. Educ. Psychology*, vol. 24, pp. 417-441, 1933.

[HYV97] **A. Hyvärinen et E. Oja**, "A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483-1492, Oct. 1997.

[INSERM] **Site Internet de l'INSERM**, "Promenade autour de la cochlée", INSERM Montpellier, France, http://www.iurc.montp.inserm.fr/cric/audition.

[IRW02] **R. Irwan and R.M. Aarts**, "Two-to-Five Channel Sound Processing", *JAES*, vol. 50, No. 11, Nov. 2002, pp. 914-926.

[ISO11172] **ISO/IEC 11172-3** *Information Technology*, "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 MBit/s, Part 3 Audio", 1993.

[ISO13818] **ISO/IEC 13818-7** *Information Technology*, "Generic Coding of Moving Pictures and Associated Audio, Part 7: Advanced Audio Coding", 1997.

[ISO14496-1] **ISO/IEC 14496-1** *Int. Std.*, "Coding of audio-visual objects – Part 1: Systems (MPEG-4 Systems, 2nd edition)", 1999.

[ISO14496-3] **ISO/IEC 14496-3** *Int. Std.*, "Coding of audio-visual objects – Part 3: Audio (MPEG-4 Audio, 2nd edition)", 2001.

[ISON7138] **ISO/IEC JTC1/SC29/WG11** (MPEG), document N7138, "Report on MPEG Spatial Audio Coding RM0 Listening Tests", 2005.

[ISON8324] **ISO/IEC JTC1/SC29/WG11**, document N8324, *Information Technology* – MPEG Audio Technologies, "Part 1: MPEG surround", 2006.

[JAK05] **J. Jakka**, "Binaural to Multichannel Audio Upmix", *Master's Thesis*, Helsinki University of Technology, 2005.

[JIN04] **G. Jin, A. Corderoy, S. Carlile et A. van Shaik**, "Contrasting monaural and interaural spectral cues for human sound localization", *J. Acoust. Soc. of Am.*, 115(6):3124-3141, 2004.

[JOH88] **J.D. Johnston**, "Estimation of Perceptual Entropy Using Noise Masking Criteria", Proc. *ICASSP*, pp. 2524-2527, Mai 1988.

[JOH92] **J.D. Johnston et A. Ferreira**, "Sum-difference stereo transform coding", *Proc. IEEE Int. Conf. Acoustics, Signal Processing*, vol.2, pp. 569-572, San Francisco, USA, Mars 1992.

[JOT92] **J.-M. Jot**, "Etude et réalisation d'un Spatialisateur de sons par modèles physiques et perceptifs", *Thèse de l'ENST*, Paris, 1992.

[JOT97] **J.-M. Jot, L. Cerveau et O. Warusfel**, "Analysis and synthesis of room reverberation based on statistical time-frequency model", Proc. 103rd *AES Conv.*, preprint 4629, 1997.

[JOT99] **J.-M. Jot, V. Larcher et J.-M. Pernaux**, "A comparative study of 3-D audio encoding and rendering techniques", 16th *AES Int. Conf.*, pages 281-300, Finlande, 1999.

[JOU00] **A. Jourjine, S. Rickard, et O. Yilmaz**, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures", *IEEE Conference on Acoustics, Speech, and Signal Processing,* vol. 5, pages 2985-2988, Juin 2000.

[KAT92] **W.R.T. ten Kate, P.M. Boers, A. Mäkivarita, J. Kuusama, K.E. Christensen et E. Sørensen**, "Matrixing of bit rate reduced audio signals", in Proc. *IEEE Int. Conf. Acoustics, Speech, Signal Proc.* (ICASSP'92), San Francisco, 1992, pp. II 205-208.

[KEN95a] **G.S. Kendall**, "A 3D sound primer: Directional hearing and stereo reproduction", Computer Music Journal, 19(4): 23-46, 1995.

[KEN95b] **G.S. Kendall**, "The decorrelation of audio signals and its impact on spatial imagery", *Computer Music Journal*, 1995.

[KIR97] **O. Kirkeby, P.A. Nelson et H. Hamada**, "The stereo dipole - binaural sound reproduciton using two closely spaced loudspeakers", 102nd *Audio Engineering Society Convention*, Munich, preprint 4463(I6), 1997.

[KOV04] **B. Kövesi, D. Massaloux et A. Sollaud**, “A Scalable Speech and Audio Coding Scheme with Continuous Bitrate Flexibility”, Proc. *IEEE ICASSP*, Mai 2004, vol. 1, pp. 273–276.

[KRA56] **H.P. Kramer et M.V. Mathews**, "A linear coding for transmitting a set of correlated signals", *IRE Trans. Inform. Theory*, vol. 23, n°3, pp. 41-46, Sept. 1956.

[LAB04] **A. Laborie, R. Bruno et S. Montoya,** "High spatial resolution multichannel recording", *Proc.* 116th *AES Conv.*, Berlin, Mai 2004, preprint 6116.

[LAP06] **J. Lapierre et R. Lefebvre**, "On improving parametric stereo audio coding", Proc. 120th *AES Convention*, Paris, 2006.

[LAR95] **J. Laroche**, "Traitement des Signaux Audio-Fréquences", Télécom Paris, 1995.

[LEV98] **S. Levine**, "Audio representations for data compression and compressed domain processing", *Ph.D. Thesis*, Standford University, USA, 1998.

[LI05] **Y. Li and P.F. Driessen**, "An Unsupervised Adaptive Filtering Approach of 2-To-5 Channel Upmix", Proc. 119th *AES convention*, New York, Oct. 2005, preprint 6611.

[LIE02] **T. Liebchen**, "An Introduction to MPEG-4 Audio Lossless Coding", Proc. 113th *AES Conv.*, Oct. 2002, Los Angeles.

[LIE04] **T. Liebchen**, "An Introduction to MPEG-4 Audio Lossless Coding", Proc. *IEEE ICASSP*, Mai 2004, vol. 3, pp. 1012–1015.

[LIT99] **R.Y. Litovsky, H.S. Colburn, W.A Yost et S.J. Guzman**, "The precedence effect", *J. Acoust. Soc. Am.*, 106(4), 1633–1654, 1999.

[LOE48] **M. Loève**, "Fonctions aléatoires de second ordre", in *Processus Stochastiques et Mouvement Brownien*, P. Levy, Ed. Paris, France: Gautiers-Villars, 1948.

[LOR04] **G. Lorho et N. Zacharov**, "Subjective evaluation of Virtual Home Theater sound systems for loudspeakers and headphones", Proc. 116th *AES Conv.*, Berlin, 2004, preprint 6141.

[MAR99] **G. Martin, W. Woszczyk, J. Corey et R. Quesnel**, "Sound source Localization in a Five-Channel Surround sound reproduction system", Proc. 107th *AES Conv.*, preprint 4994, 1999.

[MER93] **D. Mercier**, *Le livre de techniques du son*. Tome III. Fréquences, Paris édition, 1993.

[MIL58] **A.W. Mills**, "On the Minimum Audible Angle", *J. of the Acoust. Soc. of America*, vol. 30, No. 4, pp. 237-246, 1958.

[MIL06] **F. Milloz, J. Huillery et N. Martin**, "Short Time Fourier Transform Probability Distribution for Time-Frequency Segmentation", Proc. *IEEE ICASSP 2006*, III-448.

[MIT04] **N. Mitianoudis et M. Davies**, "Audio signal separation: Solutions and problems", *Int. J. Adapt. Control Signal Process*, 18:299-314, March 2004.

[MOL92] **H. Møller**, "Fundamentals of Binaural Technology", *Applied Acoustics*, N°36, pp171-218.

[MOO79] **J.A. Moorer**, "About this reverberation business", *Computer Music Journal* 3(2):13-18, 1979.

[MOO83] **B.C.J. Moore et B.R. Glasberg,** "Suggested formulae for calculating auditory-filter bandwidths and excitation pattern", *JASA*, vol. 74, pp.750-753, Sept. 1983.

[MOO03] **B.C.J. Moore**, *An Introduction to the Psychology of Hearing*, Academic Press, 5th edition, 2003.

[MOR06] **S. Moreau, J. Daniel et S. Bertet,** "3D sound field recording with higher order ambisonics – Objective measurements and validation of a 4th order spherical microphone", Proc. 120th *AES Conv.*, Paris, Mai 2006.

[NAO95] **I. Naoki, M. Moriya et M. Satoshi**, "High-Quality Audio Coding at less than 64 kbit/s by using Transform-Domain Weighted Interleaved Vector Quantization (TWIN-VQ)", Proc. *IEEE ICCASP'95*, pp. 937-940, 1995.

[OGR05] **P.D. O'Grady, B.A. Pearlmutter et S.T. Rickard**, "Survey of Sparse and Non-Sparse Methods in Source Separation", *International Journal of Imaging Systems and Technology*, 2005.

[OKA98] **T. Okano, L. Beranek, et T. Hidaka**, "Relations among interaural cross-correlation coefficient ($IACC_E$), lateral fraction (LF_E), and apparent source width (asw) in concernt halls", *J. Acoust. Soc. Am.*, 104(1), 255–265, 1998.

[ORB70] **R. Orban**, "A rational technique for synthesizing pseudo-stereo from monophonic sources", *J. Audio Eng. Soc.*, 18(2):157-164, 1970.

[PAI00] **T. Painter et A. Spanias**, "Perceptual Coding of Digital Audio", *Proc. IEEE*, vol. 88, No. 4, April 2000.

[PAN95] **D. Pan**, "A tutorial on MPEG/audio compression", *IEEE Multimedia*, vol. 2, pp.60-74, 1995.

[PEA00] **K. Pearson**, *On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling*, Philosophical Magazine, vol. 50, pp.157-175, 1900.

[PER98] **J.-M. Pernaux, P. Boussard et J.-M. Jot**, "Virtual Sound Source Positioning and Mixing in 5.1 – Implementation of the real-time system Genesis", Proc. *DAFx'98*, Barcelone, 1998.

[PER03] **J.-M. Pernaux**, "Spatialisation du son par les techniques binaurales: Application aux services de télécommunications", *Thèse de l'INP-Grenoble*, 2003.

[PRI86] **J.P. Princen et A.B. Bradley**, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation", *IEEE Trans. ASSP*, Vol. 34, N° 5, Oct. 1986.

[PRI95] **J.P. Princen et J.D. Johnston**, "Audio Coding with Signal Adaptive Filterbanks", *IEEE ICASSP'95*, p.3071-3074, 1995.

[PUL97] **V. Pulkki**, "Virtual sound source positioning using Vector Base Amplitude Panning", *J. Audio Eng. Soc.* 45, 456–466, 1997.

[PUL98] **V. Pulkki, M. Karjalainen et J. Huopaniemi**, "Analysing virtual sound source attributes using a binaural model", Proc. 114th *AES Conv.*, 1998.

[PUL99] **V. Pulkki, M. Karjalainen and J. Huopaniemi**, "Analyzing virtual sound source attributes using a binaural auditory model", *J. of the Audio Eng. Soc.*, 47:203-217, No 4, 1999.

[PUL01a] **V. Pulkki**, "Localization of amplitude-panned sources I: Stereophonic panning", *J. Audio Eng. Soc.* 49(9), 739–752, 2001.

[PUL01b] **V. Pulkki**, "Localization of amplitude-panned sources II: Two- and three-dimensional panning", *J. Audio Eng. Soc.* 49(9), 753–757, 2001.

[PUR97] **H. Purnhagen, B. Edler et C. Feredikis**, "Object-based analysis/synthesis audio coder for very low-bit rates", Proc. 104th *AES Conv.*, Mai 1998, preprint 4747.

[PUR04] **H. Purnhagen, J. Engdegård, J. Rődén, L. Liljeryd**, "Synthetic ambience in parametric stereo coding", Proc. 116th *AES Convention*, Berlin, 2004, preprint 6074.

[RAB78] **L.R. Rabiner et R.W. Schafer**, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[RAO90] **K.R. Rao et P. Yip**, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Boston, MA: Academic Press, 1990.

[RAY07] **L. Rayleigh**, "On our perception of sound direction", *Philos. Mag.* pp.13:214–232. J.W. Strutt, 1907.

[RUM01] **F. Rumsey**, *Spatial Audio*, Focal Press, Music Technology Series, 2001.

[SAX03] **O. Saxod**, "Critères d'identification de signaux sismiques d'avalanches", *Stage de Master*, Laboratoire des Images et des Signaux, INP-Grenoble, 2004.

[SCH70] **B. Scharf**, *Critical Bands*, in Foundations of Modern Auditory Theory, Academic Press, New York, 1970.

[SCH04] **E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard**, "Low complexity parametric stereo coding," Proc. 116th *AES Convention*, Berlin, 2004.

[SOU98] **G.A. Soulodre, T. Grusec, M. Lavoie et L. Thibault**, "Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs", *J. Audio Eng. Soc.*, vol. 46, n°3, pp. 164-177, 1998.

[SPO92] **T. Sporer, K. Brandenburg, et B. Edler**, "The use of multirate filter banks for coding of high quality digital audio", 6th *European Signal Processing Conference (EUSIPCO)*, vol. 1, pp. 211-214, 1992.

[STO00] **G. Stoll et F. Kozamernik**, "EBU listening test on Internet audio codecs", *EBU technical review*, Juin 2000.

[SZC03] **M. Szczerba, F. de Bont, W. Oomen et L. van de Kerkhof**, "Matrixed Multi-Channel Extension for AAC Codec", Proc. 114th *AES Conv.*, Amsterdam, preprint 5796, 2003.

[THE01] **G. Theile**, "Multichannel Natural Music Recording based on Psychoacoustic Principles", extended version Proc.19th *AES Int. Conf.*, 2001.

[TOD94] **C. Todd et al.**, "AC-3: Flexible perceptual coding for audio transmission and storage", *Proc.* 96th *AES Convention*, February 1994, preprint 3796.

[TGP05] **3GPP Technical Specification 26.405 V6.1.0,** *Enhanced aacPlus general audio codec; encoder specification parametric stereo part*, Mars 2005, disponible à l'adresse http://www.3gpp.org.

[UIT] **UIT**, www.itu.int, site web de l'Union Internationale des Télécommunications.

[UIT775] **Recommandation UIT-R BS.775-1**, "Système de son stéréophonique multicanal avec ou sans image associée", 1994.

[UIT800] **Recommandation UIT-T P.800**, "Méthodes d'évaluation subjective de la qualité de transmission", 1996.

[UIT1116] **Recommandation UIT-R BS.1116-1, "**Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio y compris les systèmes sonores multivoies", Question UIT-R 85/10, 1994-1997.

[UIT1534] **Recommandation UIT-R BS.1534-1, "**Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes de codage", Question UIT-R 220/10, 2001-2003.

[VAA04] **R. Väänänen et J. Huopaniemi**, "Advanced AudioBIFS: Virtual Acoustics Modeling in MPEG-4 Scene Description", *IEEE Trans. Multimedia* 6(5): 661-675, 0ct. 2004.

[VAF05] **R. Vafin et W.B. Kleijn**, "Jointly optimal quantization of parameters in sinusoidal audio coding", *IEEE Work. on App. of Signal Proc. to Audio and Acoustics*, New York, 2005.

[VER99] **T.S. Verma**, "A perceptually based audio signal model with application to scalable audio compression", *Ph.D. Thesis*, Standford University, USA, 1999.

[VET95] **M. Vetterli et J. Kovačević**, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[VIL06] **L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen et K. Kjörling**, "MPEG surround: The forthcoming ISO standard for Spatial Audio Coding", Proc 28th *AES Int. Conference*, Piteå, 2006.

[VIN04] **E. Vincent,** "Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux", *Thèse Paris VI*, Université Pierre et Marie Curie, 2004.

[WAA91] **R.G. van der Waal et R.N.J. Veldhuis**, "Subband Coding of Stereophonic Digital Audio Signals", Proc. *IEEE ICASSP'91, Toronto*, 1991, pp. 3601-3604.

[WIL01] **M. Williams et G. Le Dû**, "The Quick Reference Guide to Multichannel Microphone Arrays - Part I: using Cardioid Microphones", Proc. 110th *AES Convention*, preprint 5336, 2001.

[WIN04] **Wing Surround Sound Recommendations Committee**, "Recommendations for Surround Sound Production", *the National Academy of Recording Arts & Sciences*, 2004.

[WOL03] **M. Wolters, K. Kjörling, D. Homm et H. Purnhagen**, "A closer look into MPEG-4 High Efficiency AAC", Proc. 115th *AES Conv.*, New York, 2003.

[YAN03] **D.T. Yang, H. Ai, C. Kyriakakis et C.-C.J. Kuo**, "High-fidelity multichannel audio coding with Karhunen-Loève transform", *IEEE Trans. On Speech and Audio Processing*, vol. 11, pp. 365-380, 2003.

[YAN04] **D.T. Yang, C. Kyriakakis et C.-C. Jay Kuo**, "High Fidelity Multichannel Audio Coding", *EURASIP book series on Signal Processing and Communications*, 2004.

[ZEL77] **R. Zelinski et P. Noll**, "Adaptive transform coding of speech signals", *IEEE Trans. Acoust. Speech, and Signal Processing* ASSP-25, 299–309, 1977.

ABSTRACT

This thesis deals with the extraction of relevant spatial cues for parametric coding of multichannel audio signals. The classical approach used for the parametric coding of stereo or multichannel signals (5.1, 7.1 audio signals, etc.) considers a parametric representation of the multichannel signal based on the auditory localization cues. Associated with a traditional audio coding of the input sum signal, the transmission of these parameters allows the reconstruction of a multichannel signal whose inter-channel cues approximate those of the original multichannel signal. These coding methods offer an effective solution for the data-rate constrained applications since the overall data rate is equivalent to the bit rate of the audio coder (mono or stereo for the compression of a 5.1 audio signal) plus the variable bit rate of the spatial cues. Nevertheless, the subjective audio and spatial quality of the reconstructed audio signal is strongly related to the parametric model which is not completely adapted to all multichannel audio signals.

The first axis of this thesis is looking further into this multichannel audio coding approach in order to improve its performances. Whereas the current parametric coding methods extract the spatial cues with a fixed time-frequency resolution, we suggest to adapt the extraction of the inter-channel parameters to the spectral contents of the audio signals. This approach is based on a time-frequency segmentation process in order to extract the spectral patterns carrying the spatial cues. However, we have not been able to obtain the desirable results based on our experiments which have given difficult interpretations of the signal regions resulting from the segmentation process. We also sought to measure the perceptual degradations introduced by parametric audio coding to refine the parametric representation of the critical time-frequency regions. Our experiments show that the error signal established by the difference between the reconstructed and the original signals (taking into account the instantaneous masking threshold) presents a too important energy distribution on the time-frequency plane to detect some time-frequenciy critical regions.

The second axis of our research relies on a multichannel audio model resulting from the instantaneous mixture of directional sources and ambiences in order to propose an alternative parametric coding method. The time and subband analysis of signals following our model has allowed us to describe the component hierarchisation within the eigenvalues. Then, we have evaluated the performances, in terms of energy concentration, of the Principal Component Analysis (PCA), carried out both in time and in frequency subbands, of two- and three- channel audio signals. To answer the main issue of multichannel compression, we have chosen a parametric approach to carry out the bi- and three-dimensional PCA. We propose a method to extract rotation angles used for the PCA and, a physical interpretation of these angles based on the knowledge of the reproduction sound system. Finally, we use this parametric decomposition of the covariance within a new parametric coding method which relies on the concentration of the dominant sources and the extraction of useful parameters for the audio signal reconstruction. A first subjective listening test has enabled us to define the relevant parameters for the coding method and the second one was carried out to evaluate the performances of our implementation of the parametric stereo coding method. Thanks to the parametric approach used to carry out the three-dimensional PCA, we also describe the extension of the method for parametric coding of 5.0 audio signals.

Keywords: parametric audio coding, upmix, spatial sound, time-frequency analysis and segmentation, Principal Component Analysis, adaptive matrixing.