Académie de Paris
Université Paris 6

THÈSE DE DOCTORAT

École Doctorale de
SCIENCES MÉCANIQUES, ACOUSTIQUE ET ÉLECTRONIQUE

Présentée
par
Alexis Baskind

pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PARIS 6

MODÈLES ET MÉTHODES DE DESCRIPTION SPATIALE DE SCÈNES SONORES

APPLICATION AUX ENREGISTREMENTS BINAURAUX

version provisoire

Rapporteurs et membres du jury

ALAIN DE CHEVEIGNÉ	Examinateur
WILLIAM M. HARTMANN	Rapporteur
JEAN-BAPTISTE LEBLOND	Examinateur
XAVIER MEYNIAL	Examinateur
JEAN-DOMINIQUE POLACK	Directeur de Thèse
GAËL RICHARD	Rapporteur
OLIVIER WARUSFEL	Examinateur

THÈSE

MODÈLES ET MÉTHODES DE DESCRIPTION SPATIALE DE SCÈNES SONORES

APPLICATION AUX ENREGISTREMENTS BINAURAUX

Alexis Baskind

version provisoire

Résumé

L'objet de cette recherche est de proposer des méthodes automatiques de description objective des aspects spatiaux d'une scène sonore enregistrée, sans connaissance préalable ni des sources qui la composent, ni du message sonore diffusé, ni des caractéristiques du lieu d'enregistrement. L'étude se concentre plus spécifiquement sur l'estimation de la direction de la source (supposée unique et stable dans un premier temps) et sur une caractérisation de l'enveloppe de réverbération. Un cadre théorique, inspiré des modèles d'audition spatiale, permet de développer un ensemble homogène de méthodes de détection et d'estimation, basées sur des statistiques non stationnaires d'ordre 2 relatives aux relations entre les voies de l'enregistrement. Les informations obtenues dans chaque bande de fréquences sont ensuite regroupées et interprétées au moyen de descripteurs de plus haut niveau. La pertinence de cette approche est étudiée sur des enregistrements binauraux synthétiques et réels.

Mots-Clefs

Analyse de scènes sonores, réverbération, localisation de source, audition spatiale, description automatique, analyse non stationnaire, égalisation et annulation, enregistrements binauraux

Abstract

This research aims at providing methods for automatically describing objective spatial features of a recorded sound scene, without prior knowledge of the sources, of the sound they emit, or the characteristics of the performance space. The study focuses specifically on estimating the direction of arrival of the sound event (assuming a unique and stable source) and on characterizing the power envelope of the reverberation. A theoretical framework, inspired from spatial hearing models, allows the development of a homogeneous set of methods for detection and estimation. These are based on nonstationary second-order statistics of relations between tracks. Cues obtained in each frequency band are grouped together and interpreted by high-level descriptors. The relevance of this approach is studied on synthetic and real binaural recordings.

Keywords

Sound scene analysis, reverberation, source localization, spatial hearing, automatic description, nonstationary analysis, equalization and cancellation, binaural recordings

Notations et conventions

signal source	signal émis par la source physique avant réverbération
signaux observés	signaux captés par les microphones après réverbération
bicanal	enregistrement sur deux canaux
enregistrement binaural	désigne un enregistrement bicanal effectué sur tête artificielle ou réelle
intercanal	relatif aux relations entre deux canaux d'un enregistrement
interaural	relatif aux relations entre les signaux parvenant aux deux oreilles ou, par extension, aux relations entre les deux canaux d'un enregistrement binaural
latéralisation	écart de la source par rapport au plan médian, repérée par ou la longitude β, en fonction du repère
élévation	altitude de la source, repérée en configuration binaurale par le site φ ou la lattitude δ, en fonction du repère
HRIR	réponse impulsionnelle binaurale mesurée en chambre anéchoïque
HRTF	transformée de Fourier d'une HRIR
TFCT	transformée de Fourier à court-terme
gammatone	banc de filtres approximant les filtres auditifs *roex*
ERB	échelle de filtres en bandes rectangulaires équivalentes
$\rho_{x,y}(\tau)$	corrélation normalisée entre les signaux x et y
α, τ	respectivement le gain et le retard de la méthode de détection et estimation par égalisation et annulation
$D_{\alpha,\tau}$	erreur absolue d'égalisation et annulation
$\varepsilon_{\alpha,\tau}$	erreur normalisée d'égalisation et annulation
ΔT	constante de temps de la fenêtre exponentielle utilisée pour le calcul des corrélations et puissances à court-terme

Table des matières

Introduction **1**

I Modèles de propagation et d'audition spatiale **9**

I Propagation du son dans une salle **11**
1 Modèles déterministes de propagation du son en espace clos 12
1.1 Acoustique géométrique déterministe : modèle par sources images 12
1.2 Théorie modale 14
1.3 Réverbération et filtrage 15
1.4 Discussion 16
2 Modèles stochastiques de réverbération tardive 17
2.1 Notion de champ diffus 17
2.2 Modèles de réverbération diffuse 17
2.3 Discussion 19
3 Modèle composite de propagation 19
3.1 Domaines de validité temps-fréquence 19
3.2 Limites de l'analyse spatiale par cohérence stationnaire 20
3.3 La cohérence à court-terme comme outil d'analyse spatiale de réponses impulsionnelles de salles 22
4 Conclusion 25

II Perception spatiale d'une scène sonore **27**
1 Événement auditif et impression spatiale 28
1.1 Source sonore et événement auditif 28
1.2 L'espace sonore et sa perception 28
2 Localisation anéchoïque de l'événement auditif 32
2.1 Caractéristiques de la perception de la direction 32
2.2 Indices binauraux de perception de la direction 33
2.3 Indices monauraux de perception de la direction 36
2.4 Importance relative des indices monauraux et binauraux dans la perception de la direction 36
2.5 Distance de l'événement auditif en contexte anéchoïque 37
3 Influence de la salle sur la perception de l'événement auditif 37
3.1 L'effet de précédence 38
3.2 Élargissement de l'événement auditif 40
3.3 Perception de la distance de la source en milieu réverbérant 44
4 Perception de l'espace sonore 45
4.1 Perception de la taille de la salle 45
4.2 Perception de l'enveloppement de l'auditeur 45
4.3 Perception de la réverbérance 47
5 Conclusion 47

II Méthodes de détection en milieu réverbérant 49

III Présentation du problème de la détection en milieu réverbérant 51
1 Introduction . . . 52
1.1 Définition du problème à traiter . . . 52
1.2 Détection de source et détection de réverbération . . . 52
2 Exemples de modèles monophoniques . . . 53
2.1 Détection par rupture de stationnarité . . . 53
2.2 Détection d'harmonicité . . . 54
3 Modèles basés sur les similarités entre les canaux . . . 54
3.1 Principe . . . 54
3.2 Modèles auditifs de détection binaurale . . . 55
4 Discussion : détection et estimation . . . 56
5 Conclusion . . . 58

IV Méthode non stationnaire de détection et d'estimation par égalisation et annulation 59
1 Introduction . . . 60
2 Principe de la détection par égalisation et annulation . . . 60
2.1 Rappel : définition de la puissance et de la corrélation . . . 60
2.2 Formalisation du problème . . . 61
2.3 Détection par minimisation de l'erreur absolue . . . 63
2.4 Détection par minimisation de l'erreur normalisée . . . 63
2.5 Erreur absolue ou erreur normalisée ? . . . 64
2.6 Lien avec les méthodes usuelles d'estimation de retard . . . 65
2.7 Discussion . . . 66
3 L'égalisation et annulation pour des signaux non stationnaires . . . 66
3.1 Préliminaire : corrélation et puissance à court-terme . . . 66
3.2 Détection par égalisation et annulation à court-terme . . . 68
3.3 Estimation par minimisation de l'erreur normalisée . . . 69
3.4 Exemple . . . 69
3.5 Aspect computationnel . . . 70
4 Détection en bandes limitées . . . 70
4.1 Principe d'incertitude . . . 72
4.2 Choix du banc de filtres . . . 73
4.3 Retard d'enveloppe . . . 74
4.4 Ambiguïté de phase en bande étroite . . . 76
4.5 Problème de la résolution sur l'estimation du retard en basses fréquences 78
5 Cas particulier de la détection à partir de la transformée de Fourier à court-terme 80
5.1 Approximations de calcul . . . 82
5.2 Minimisation . . . 84
5.3 Lien avec la formule des interférences . . . 84
5.4 Exemple . . . 85
6 Généralisation du principe d'égalisation et annulation à plus de deux canaux . 87
7 Conclusion . . . 88

V Application de la détection par égalisation et annulation aux milieux réverbérants 89
1 Comportement de la détection en milieu réverbérant . . . 90
1.1 Introduction . . . 90
1.2 Évolution de l'indice de détection sur un signal réverbéré . . . 90
1.3 Influence de la durée de la fenêtre d'analyse . . . 93
1.4 Influence du niveau relatif d'énergie réverbérée . . . 97
1.5 Choix du seuil . . . 97
1.6 Exemple . . . 99
1.7 Discussion . . . 100

2 Application aux signaux harmoniques 100
2.1 Introduction 100
2.2 Rehaussement de signaux harmoniques à fréquence fondamentale constante par morceaux 102
2.3 Analyse pitch-synchrone 104
2.4 Exemple 106
3 Conclusion 107

III Méthodes de description spatiale 111

VI Estimation de la direction de la source 113
1 Particularités de l'estimation de direction à partir d'enregistrements binauraux 114
1.1 Spécificité de la prise de son binaurale 114
1.2 Estimation de la direction de provenance à partir d'une représentation temps-espace 115
1.3 Estimation de la direction de provenance par critère de distance 116
1.4 Discussion 118
2 Configuration de la méthode de détection pour des enregistrements binauraux 118
2.1 Organisation structurelle des HRTF 118
2.2 Variation des indices interauraux 121
2.3 Choix du banc de filtres 121
2.4 A propos de la notion de cônes de confusion 126
3 Estimation de la direction par décision sur l'erreur d'égalisation et annulation 128
3.1 Cadre et hypothèses de travail 128
3.2 Principe de l'estimation de la direction 129
3.3 Liens avec les méthodes usuelles d'estimation de la direction de provenance 133
3.4 Prise en compte des retards d'enveloppe 136
3.5 Intégration fréquentielle 136
3.6 Intégration séquentielle 142
3.7 Résolution au sein des cônes de confusion 144
4 Exemple 148
4.1 Contexte 148
4.2 Estimation de fréquence fondamentale 149
4.3 Estimation de la direction 149
4.4 Résultats 150
5 Conclusion 152

VIIDescription de la réverbération 153
1 La réverbération : aspects spatiaux et temporels 154
2 Méthodes d'estimation des durées de réverbération 154
2.1 Estimation des durées de réverbération à partir de mesures 154
2.2 Estimation des durées de réverbération à partir de signaux musicaux . . 157
3 Estimation du temps de réverbération basée sur la détection par égalisation et annulation 159
3.1 Principe 160
3.2 Effet de la troncature 160
3.3 De la difficulté d'estimer la durée de décroissance initiale 164
3.4 Influence de la bande passante 164
3.5 Effet d'un bruit de fond additionnel 165
3.6 Utilisation de moyennes spatiales 168
4 Exemples 168
4.1 Exemple d'analyse sur une scène sonore virtuelle 168
4.2 Exemple d'analyse sur un enregistrement *in situ* 173
5 Conclusion 176
Autre méthode : utilisation de la puissance à court-terme 179

Conclusion **181**

Annexes **185**

A Conventions de notations de coordonnées en prise de son binaurale **187**
1 Définition des plans standard 188
2 Conventions sur les angles 188

B Calculs relatifs à la méthode de détection et estimation par égalisation et annulation **191**
1 Égalisation et annulation en présence de bruit 192
1.1 Formalisation du problème 192
1.2 Détection et estimation lorsque les deux voies du bruit sont décorrélées 192
1.3 Détection et estimation lorsque les deux voies du bruit sont corrélées . . 193
2 Déviation de l'erreur normalisée au voisinage du minimum 194

C Calculs de décroissances sur un modèle localement stationnaire de réverbération tardive **197**
1 Modélisation à bande étroite de la réverbération tardive 198
2 Décroissance intégrée après une impulsion 198
3 Décroissance intégrée après un signal quelconque interrompu 199
4 Intégration sur une décroissance tronquée 200
5 Utilisation de la puissance à court-terme 201

D Publication 1 **203**

E Publication 2 **215**

Bibliographie **223**

Index **229**

Introduction

Introduction

L'ENREGISTREMENT SONORE a une vocation originelle de *témoignage* d'une situation. Depuis son invention il y a deux siècles, il sert ainsi à des fins d'investigation, dans le cadre par exemple d'une démarche scientifique, d'archivage, pour conserver une trace sonore d'un événement, ou de reproduction. L'essor de la bande magnétique au cours du vingtième siècle vient brouiller les pistes, en faisant des technologies d'enregistrement et de reproduction des outils de *création* artistique. L'apparition de la synthèse électronique permet même une création *ex nihilo*, et les procédés de traitement du signal sont depuis longtemps utilisés en production pour créer de toutes pièces des scènes sonores virtuelles, sans nécessaire référence à une situation réelle. La démocratisation des procédés de captation, de traitement et de diffusion du son entraîne un décuplement de l'utilisation de l'enregistrement sonore, et l'apparition des technologies numériques rendent ce dernier plus virtuel que jamais, et surtout reproductible à l'envi.

Dans ce contexte, il est apparu rapidement nécessaire d'organiser cette masse d'informations. Que ce soit à des fins d'archivage, de collections sonores (sonothèques) ou de classement en vue d'une prochaine diffusion, il s'est avéré indispensable de qualifier l'enregistrement d'une situation, par son type (concert, journal télévisé, reportage,...), par la nature et le nombre des sources sonores, sa durée, ou tout autre information pertinente du point de vue de l'auditeur ou de l'usager. Cette tâche d'étiquetage ou de description par le contenu, traditionnellement accomplie "à la main", est de plus en plus confiée à des outils automatiques, avec ou sans supervision humaine. De nouvelles méthodes et de nouveaux standards voient le jour année après année pour nous aider à répertorier les sons enregistrés et à rechercher le plus efficacement possible l'information souhaitée au sein de bases de données de plus en plus importantes.

Parmi les nombreuses tâches d'étiquetage que les techniques modernes sont en mesure d'accomplir, la plupart se concentrent sur une caractérisation, à un plus ou moins haut niveau d'abstraction, des sources ou des sons émis par celles-ci. Qu'il s'agisse de détection, d'estimation (suivi de hauteur ou de partition,...), d'identification ou de reconnaissance (reconnaissance vocale, transcription automatique,...), l'analyse se focalise généralement plus sur les sons que sur leur organisation spatiale ou sur l'espace de diffusion. En ce sens, la réverbération constitue même en général une gêne vis-à-vis de la tâche à accomplir.

Or la caractérisation spatiale d'une scène sonore enregistrée est une information qui peut s'avérer d'une grande utilité :

Toujours dans une optique d'indexation, répertorier des enregistrements, non seulement par une caractérisation des sources sonores, mais également par celle de leurs positions respectives et de l'espace de diffusion, complète les possibilités de recherche au sein la base de données, et autorise par exemple des requêtes du type : "Je recherche les enregistrements de voix chantée pour lesquels le musicien se trouve en face, et le temps de réverbération reste inférieur à 1 seconde".

D'autre part, une caractérisation spatiale d'une scène sonore à partir d'un enregistrement permettrait à l'acousticien d'évaluer une salle en l'absence de mesures. Il y a en effet de nombreuses situations dans lesquelles on souhaiterait prédire la qualité spatiale sans avoir facilement recours à des mesures acoustiques. Ainsi, le mesurage du temps de réverbération en salle pleine est peu envisageable avec les méthodes traditionnelles, si bien qu'il

est fréquent d'avoir recours à des estimations à partir de la mesure en salle vide. De même, il serait utile de pouvoir fournir une caractérisation de la qualité d'un espace acoustique à partir d'un enregistrement dont on ne connaît que peu d'informations sur le lieu de production.

On peut également songer aux applications de rehaussement d'une source donnée par rapport à un milieu réverbérant, ce qu'on appelle la *déréverbération* : quelle que soit la technique employée, il est nécessaire pour effectuer ce genre de tâches de disposer au préalable de connaissances portant au minimum soit sur la position de la source, soit sur la propagation.

De même, d'autres types de traitement nécessitent une caractérisation spatiale de la source et du milieu de diffusion : ainsi, avec l'avènement de la stéréophonie multicanale, il est parfois question de convertir dans ces formats des enregistrements à deux canaux, ce qui suppose de pouvoir y distinguer les sources de la réverbération, de manière à redistribuer une partie de cette dernière sur les canaux d'ambiance.

L'objet de ce travail est de proposer des méthodes de description objective des attributs spatiaux d'une scène sonore enregistrée. Il s'agit donc *in fine* d'estimer, à partir d'un enregistrement effectué hors d'un contexte de mesure, les caractéristiques spatiales des sources (soit la position physique, l'orientation, l'extension spatiale), et de caractériser l'espace de diffusion au travers des aspects les plus pertinents de la réverbération, c'est-à-dire ceux dont l'effet perceptif est le plus notable. Dans un premier temps, l'accent est mis l'estimation de la direction de la source physique et aux statistiques temporelles de la réverbération tardive. On tâche néanmoins de garder à l'esprit ce cadre plus général, et nous nous attacherons à proposer des méthodes qui puissent être étendues à une description spatiale plus complète.

Puisque la description est censée être appliquée à tout type d'enregistrement, on désire que l'analyse soit menée avec un minimum de connaissances et d'hypothèses sur la situation considérée (c'est-à-dire sur la source, le son émis par celle-ci, et sur l'espace de diffusion). Une telle approche, dite *aveugle*, nécessite néanmoins un cadre théorique pour être envisageable, ce qui impose de fait un certain nombre de conditions sur les scènes analysées. Ainsi, en l'absence de toute information sur le signal émis par la source, il est nécessaire de disposer de connaissances minimales *a priori* sur la nature de la propagation. Il suppose donc, premièrement, que l'enregistrement sonore étudié résulte directement de la captation d'une scène sonore réelle. Ceci exclut donc toute manipulation *a posteriori* (montage, mixage), ainsi que la synthèse, sauf lorsque celle-ci a pour vocation de simuler une scène sonore réelle (comme c'est par exemple le cas de la synthèse binaurale, ou de la réverbération artificielle par convolution). On considère de plus que l'espace de diffusion est clos, et que la réverbération qui s'y développe tende *in fine* vers un champ idéalement diffus. On écarte donc le cas d'échos flottants persistants, d'espaces clos aux dimensions trop disparates comme des couloirs, de volumes couplés, ou encore d'espaces semi-ouverts. Finalement, le système de captation de l'onde sonore est supposé connu avec le maximum de détails. Cette connaissance peut être structurelle (position des microphones, directivité, réponse en fréquence, géométrie d'un éventuel obstacle entre eux) ou systémique (réponse du système de microphones, vu comme une "boite noire", à des stimuli provenant de toutes les directions envisageables).

Bien que l'on souhaite que les méthodes proposées dans cette étude soient suffisamment générales pour être applicables à tout type de configuration de microphones (du moment que celui-ci est connu au sens des hypothèses mentionnées ci-dessus), on envisagera principalement leur application à des enregistrements *binauraux*. Le terme d'"enregistrement binaural" est un abus de langage assez courant désignant un enregistrement effectué grâce à deux microphones miniatures placés dans les conduits auditifs soit d'un sujet humain, soit d'une tête artificielle. On choisit de se focaliser sur ce type d'enregistrement pour plusieurs raisons.

Premièrement, un enregistrement binaural est à deux canaux uniquement, ce qui ne simplifie pas la tâche (plus le nombre de canaux est restreint, plus grande est l'incertitude spatiale), mais facilite en revanche beaucoup sa conceptualisation. En partant de l'hypothèse

que le cas de plusieurs canaux ne diffère pas fondamentalement dans le principe que celui de deux canaux, nous proposons des modèles et des méthodes de description, que nous étudierons en profondeur, appliquées à un système relativement simple, mais généralisables à des configurations plus complexes.

De plus, ce type d'enregistrement est, par rapport à d'autres systèmes de captation à deux canaux, riche en informations susceptibles de conduire à une description spatiale. En effet, il induit à la fois des *différences de temps* et d'*intensité* entre les deux canaux, ces différences dépendant étroitement de la direction de la source physique et, en champ proche, de sa distance. De nombreux systèmes de captation, et en particulier ceux à deux canaux (couples X-Y, M-S) induisent uniquement des différences d'intensité entre les deux canaux, alors que d'autres, comme les couples AB de temps, n'induisent que des différences de temps. Par rapport à des systèmes de temps et intensité, comme les couples AB / ORTF, il offre l'avantage (ou l'inconvénient, suivant le point de vue selon lequel on se place) d'introduire des différences de temps et d'intensité dépendant plus étroitement de la fréquence, cette dépendance étant liée à la géométrie du buste.

Pour finir, on peut souligner l'intérêt croissant pour les techniques binaurales de captation et de synthèse, cet intérêt étant favorisé par la démocratisation des outils d'enregistrement et des moyens de traitement.

Une différence fondamentale entre les configurations usuelles de captations et une configuration binaurale est que cette dernière est intrinsèquement *individuelle* : du point de vue tant de la diffusion que de l'analyse, elle ne prend tout son sens que vis-à-vis de la morphologie de la personne ayant effectué l'enregistrement.

Une première approche envisagée (Baskind et Warusfel, 2002)[1] pour résoudre le problème posé ci-dessus consiste à se rapprocher le plus possible d'une situation de mesure, par exemple en tentant d'estimer par déconvolution aveugle les réponses impulsionnelles associées au trajet source-récepteurs, pour pouvoir appliquer ces mêmes indices. Cependant, cette option se heurte en pratique à deux limitations :

D'une part, le paradigme inhérent à la notion de réponse impulsionnelle de salle, qui suppose entre autres que la propagation est linéaire et à temps invariant, est d'une validité discutable dans le cas d'enregistrements quelconques (et en particulier lorsque ceux-ci sont de longue durée), car la source est malgré tout sujette à de petits mouvements, et le milieu de propagation peut varier ; ainsi, la structure microscopique de la réponse impulsionnelle est susceptible d'être bouleversée dans le détail au cours du temps, bien que les caractéristiques de l'enveloppe soient inchangées. La solution qui a été proposée pour minimiser l'effet de ces fluctuations consiste à ne s'intéresser qu'aux tous premiers instants de l'effet de salle.

D'autre part, une mesure se veut une description exhaustive d'une situation donnée (dans la limite du paradigme mentionné ci-dessus), alors qu'un enregistrement peut être très incomplet en termes d'information, notamment au niveau spectral.

Cette approche a en outre un défaut majeur, qu'il s'est avéré difficile de pallier : en l'absence de toute connaissance sur le signal source, il est impossible de juger grâce à la méthode proposée de la *qualité* de l'estimation de la réponse impulsionnelle. Il a donc été décidé de l'écarter, au profit de méthodes liées à un cadre théorique plus complet, ce qui permet à la fois d'effectuer l'estimation des indices objectifs à la base de la description, et d'avoir accès à l'erreur vis-à-vis de ce cadre sous-jacent.

La première partie de ce document, par la présentation des fondements théoriques de la propagation et de la perception spatiale, vise à dégager la nature des informations à rechercher.

Le chapitre I traite ainsi des modèles de propagation acoustique usuels en espace clos, et souligne la nécessité d'une double approche, à la fois déterministe et stochastique, pour appréhender la totalité des phénomènes de la réverbération. Il aborde également le problème de la transition entre les zones de validité de ces différents modèles, à travers l'évolution temporelle de la cohérence entre deux points de l'espace. Cette notion de cohérence s'avère primordiale, puisqu'elle permet de distinguer les tous premiers instants de l'effet de salle,

[1] Cette publication est reproduite en annexe D.

pour lesquels les signaux reçus ne diffèrent en première approximation que par des gains et des retards, de la réverbération tardive.

Le chapitre II, en rappelant les principales causes et phénomènes de l'audition spatiale, vise à isoler quelles informations sont les plus pertinentes. On s'intéresse notamment aux indices permettant à l'auditeur de juger de la position d'une source, à l'influence des réflexions dans la salle sur ce jugement, et à la formation d'une impression auditive sur l'espace de diffusion. Il y est entre autres rappelé qu'au fur et à mesure du déroulement de la réverbération, l'auditeur est de moins en moins attentif à la distribution de chacune des réflexions, se basant finalement sur l'enveloppe énergétique uniquement pour qualifier la réverbération tardive.

Puisque nous disposons à ce stade d'une meilleure connaissance de la nature des informations à rechercher, il s'agit de proposer des méthodes permettant de repérer les moments auxquels ces informations sont les plus saillantes. Cette tâche de *détection* est étudiée au cours de la seconde partie.

Le chapitre III définit plus précisément le problème de la détection, et le subdivise en deux tâches, qui sont la détection des plages de temps pour lesquelles la source est active, et la détection des plages de temps pour lesquelles la réverbération peut être étudiée librement. Cette détection nécessite inévitablement un modèle de signaux, portant soit sur la nature des signaux émis par les sources, soit sur la nature de la propagation. Après un bref rappel de quelques méthodes monophoniques de détection portant sur des modèles de signaux sources, on se concentre sur les méthodes portant sur les relations entre les signaux reçus en plusieurs points de l'espace. Cette approche nécessite que l'enregistrement comporte deux ou plusieurs voies, et l'on suppose cette hypothèse acquise pour la suite.

Au chapitre IV est étudiée une méthode de détection consistant à repérer les plages de temps pour lesquelles les signaux reçus en deux voies de l'enregistrement ne diffèrent que par des gains et des retards. Cette méthode, dite par *égalisation et annulation*, permet simultanément d'estimer ces gains et retards, et donc de fournir des indices très utiles pour juger ultérieurement de la direction de la source. Après en avoir présenté les principes fondamentaux, nous étudions le couplage de cette méthode avec une analyse en bandes étroites, de manière à pouvoir en ajuster la résolution fréquentielle, et ainsi s'adapter au plus grand nombre de situations envisageables. Appliqué à deux canaux uniquement, le principe d'égalisation et annulation ne permet de considérer qu'une seule source active à chaque instant. Nous introduisons brièvement à cet effet les principes d'une généralisation de cette méthode à des enregistrements sur plus de deux canaux, qui permet de gérer le cas de sources actives simultanément.

La méthode est jusqu'ici présentée de manière très théorique. Le chapitre V se consacre à son application pratique au cas de signaux réverbérés. Il permet de définir plus précisément la détection de source et de réverbération et étudie l'influence conjointe des paramètres de l'analyse et du niveau de réverbération. Finalement, est proposée une méthode complète de détection dans le cas particulier de signaux harmoniques.

La troisième partie propose finalement des méthodes de description spatiale reposant étroitement sur le résultat de la détection.

Tout d'abord, est présentée au chapitre VI une méthode d'estimation de la direction de la source. Celle-ci s'inspire en partie de techniques basées sur la connaissance des retards et des différences de niveau dans chaque bande de fréquences, mais s'en démarque, en rappelant qu'elles permettent difficilement de tenir compte des ambiguïtés et erreurs intrinsèques à la méthode d'estimation de ces indices. À partir de ce constat, il est proposé d'intégrer l'estimation de la direction au sein de la méthode de détection par égalisation et annulation. C'est à ce stade qu'intervient réellement la configuration de prise de son : en effet, puisque par hypothèse on connaît la réponse du système de microphones à des stimuli provenant du plus grand nombre possible de directions, la méthode de décision consiste à juger laquelle de ces directions est la plus probable vis-à-vis de la situation présente. Dans le cas d'une configuration binaurale, cette connaissance du système repose sur les fonctions de transferts de tête (HRTF). Si la source est supposée immobile, il est possible de gagner en précision spatiale

au cours du temps, en complétant petit-à-petit la connaissance de la situation. La méthode de décision proposée ici est appliquée à des exemples simples, ainsi qu'à un enregistrement binaural *in situ*.

Le dernier chapitre s'attache à la caractérisation temporelle de l'enveloppe de réverbération. On y rappelle avant tout les méthodes usuelles d'estimation des durées de réverbération, tant à partir de mesures de réponses impulsionnelles qu'à partir de signaux musicaux. Ceci permet de dégager le principe d'intégration rétrograde en tant que représentation privilégiée pour cet effet. Nous nous attachons alors à l'étude de son couplage à la méthode de détection de réverbération proposée pour une estimation automatique du temps de réverbération à partir d'enregistrements quelconques, en insistant sur les spécificités de ce type d'analyse par rapport à celle de réponses impulsionnelles. Deux exemples d'enregistrements sont analysés, l'un résultant de la simulation d'une scène sonore par convolution, et l'autre étant un enregistrement en situation réelle.

Première partie

Modèles de propagation et d'audition spatiale

I

Propagation du son dans une salle

L'OBJET DE CE CHAPITRE est, en rappelant brièvement les bases de la propagation du son en espace clos, de mettre en évidence les notions susceptibles d'aider à en décrire les aspects spatiaux. En particulier, au travers de la présentation des modèles déterministes (section 1) et stochastiques (section 2) usuels, on s'attache à souligner l'évolution de la cohérence spatiale et temporelle, qui reflète la transition au cours de l'effet de salle d'une onde cohérente jusqu'à un champ diffus (section 3).

1 MODÈLES DÉTERMINISTES DE PROPAGATION DU SON EN ESPACE CLOS

1.1 Acoustique géométrique déterministe : modèle par sources images

Ce modèle, qui est directement issu de la loi de la réflexion du son, vise à se ramener à une situation similaire à celle d'un champ libre (qui fournit une solution très simple à l'équation des ondes), en remplaçant chaque réflexion sur une paroi par la source image correspondante. Ceci suppose d'une part que chaque réflexion soit spéculaire, et donc que la paroi puisse être considérée comme parfaitement plane et lisse au regard des longueurs d'onde considérées, de sorte que diffraction et diffusion puissent être négligées, et que les parois courbes puissent être considérées comme localement planes. De plus, il est fréquent d'assimiler la source à une source ponctuelle (ce qui suppose que le point d'écoute soit à une distance suffisante), de telle sorte que les ondes émises par la source principale et chacune des sources images soient considérées comme sphériques.

En supposant valides ces hypothèses, on est en mesure de substituer à une représentation spatio-temporelle de l'effet de salle une vision purement spatiale : le temps que met l'onde directe et chacune des réflexions pour atteindre un point d'écoute correspond à la distance entre ce point et (respectivement) la source principale ou la source image correspondante dans un espace tridimensionnel contenant la salle proprement dite et la source principale, ainsi que toutes les images de cette salle et de cette source, issues d'une ou de multiples symétries par rapport aux parois (Kuttruff, 2000).

Le cas d'école de la salle parallélipédique (figure I.1) permet d'illustrer simplement ce principe, car les salles images forment un pavage régulier de l'espace. On peut émettre une première remarque à partir de cet exemple simple quant à l'évolution de la nature de l'effet de salle au cours du temps : alors que pour les premières réflexions (c'est-à-dire pour des sources images à de courtes distances du point d'écoute sur la figure I.1), les positions de la source et du point d'écoute sont d'une grande importance, au fur et à mesure que le temps passe, le nombre de réflexions augmente, et celles-ci se répartissent de manière relativement uniforme dans une couronne de rayon $R = ct$ autour du centre de la salle. Au regard des distances considérées, les positions de la source et du point d'écoute n'ont plus aucune influence sur les statistiques de cette répartition. Ainsi, il est par exemple possible de fournir une estimation de la **densité temporelle de réflexions spéculaires**, indépendante des positions de la source et du point d'écoute :

$$\frac{dN_{ref}}{dt} \simeq \frac{4.\pi.c^3}{V}.t^2 \ (s^{-1})$$

Cette formule, appliquée à des géométries plus complexes, pose les limites du modèle par sources images. Pour s'en convaincre, on peut déterminer un ordre de grandeur du nombre d'échos sur toute la réponse ; prenons le cas d'une salle de taille moyenne, de volume $V = 250\,m^3$ et de temps de réverbération $TR_{60} = 1\,s$. Le nombre d'échos entre $t = 0$ et $t = TR_{60}$ vaut :

$$N_{ref}(t = TR_{60}) \simeq \frac{4.\pi.c^3}{V}.\frac{TR_{60}^3}{3}$$

Ce qui donne $N_{ref}(TR_{60}) \simeq 600000$! Ce nombre ne donne qu'une idée de la complexité de l'approche déterministe sur l'ensemble de la réponse, et permet d'en souligner les limites, tant du point de vue du déterminisme du modèle, que du traitement informatique des données : en effet, si certaines statistiques de la distribution des réflexions tardives sont indépendantes des positions de la source et du point d'écoute, la date d'arrivée de chaque réflexion y est au contraire extrêmement sensible. De toute manière, l'audition n'est pas en mesure de percevoir l'effet individuel d'un grand nombre de réflexions, même cohérentes, lorsque celles-ci sont trop rapprochées les unes des autres (voir chapitre suivant), si bien qu'une approche déterministe, même si elle était possible, n'apporterait pas réellement d'information pertinente supplémentaire.

Il est possible de quantifier la rapidité de la transition vers ce régime non déterministe : on définit par le **temps de mélange** l'ordre de grandeur de la durée à partir de laquelle l'audition n'est plus sensible à l'effet individuel de chaque réflexion (Polack, 1988). On peut

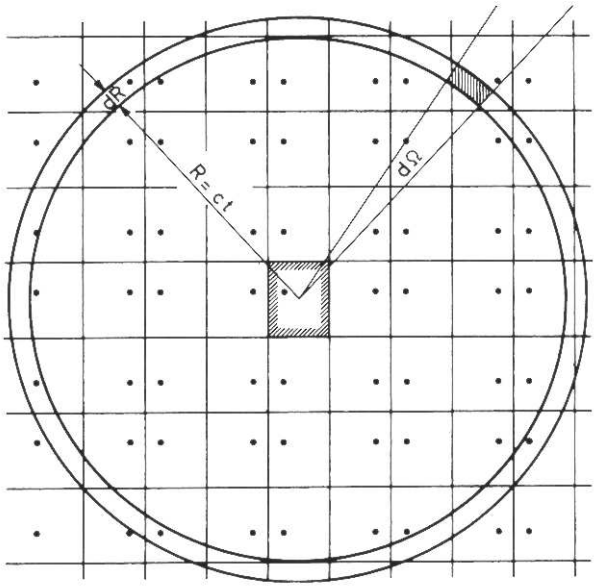


FIG. I.1 – **Méthode des sources images pour une salle parallélipédique (représentation bidimensionnelle)** (repris de Kuttruff (2000))

estimer ce temps de mélange en considérant que 10 échos dans un intervalle de temps Δt correspondant au seuil d'écho de l'audition suffisent à créer un mélange perceptivement convaincant quel que soit le point d'écoute, si la salle est ergodique. On trouve alors :

$$t_{mix} \simeq \sqrt{\frac{10V}{4\pi c^3 \Delta t}}\,(s)$$

Pour un seuil d'écho de l'ordre de 20 millisecondes, qui est une valeur cohérente avec les résultats d'expériences sur l'effet de précédence, on trouve $\sqrt{\frac{10}{4\pi c^3 \Delta t}} \simeq 10^{-3}$, si bien que la formule se ramène à l'expression simplifiée :

$$t_{mix} \simeq \sqrt{V}\,(ms)$$

Quant à la distribution énergétique des réflexions, celle-ci dépend étroitement des propriétés mécaniques des parois. Dans le cas théorique où celles-ci sont non absorbantes, l'énergie reste constante dans la salle, ce qui peut être confirmé en observant que si le nombre de réflexions croît proportionnellement à t^2, leur énergie décroît en $\frac{1}{t^2}$, puisque la longueur du trajet parcouru est proportionnelle à t. Dans le cas général où les parois sont absorbantes, celles-ci sont usuellement supposées à réaction locale, de sorte que l'impédance soit indépendante de l'angle d'incidence. Dans ces conditions, le coefficient de réflexion est une fonction du cosinus de cet angle. Cependant, une approximation courante en acoustique architecturale consiste à considérer un coefficient de réflexion en intensité "équivalent" indépendant de l'angle d'incidence égal à $|R|^2 = 1 - \alpha$, α étant le coefficient d'absorption équivalent de la paroi. Ces considérations permettent de déduire les formulations du temps de réverbération établies par Sabine et par Eyring (Kuttruff, 2000).

Limites du modèle, incohérence spatiale et temporelle

Cependant, ces considérations ont une portée limitée par la pertinence des hypothèses de travail ; il a déjà été noté que le modèle par sources images n'est valide que dans la mesure ou les réflexions sont parfaitement spéculaires. Or cette hypothèse, si elle fait sens pour les premières réflexions, est de plus en plus fausse au fur et à mesure de la propagation :

petit-à-petit, la proportion d'énergie diffusée et diffractée sur les parois est de plus en plus importante, si bien que les ondes perdent leur cohérence spatiale et temporelle[1]. En pratique, cela se traduit à la fois par une dispersion temporelle des ondes successives (incohérence temporelle), et par une chute de la corrélation et de la cohérence entre deux points d'écoute (incohérence spatiale)[2].

Ceci est une limite supplémentaire à l'approche géométrique déterministe. Les notions d'incohérence temporelle et spatiale sont traditionnellement d'une grande importance en acoustique architecturale, et sont d'un intérêt majeur pour cette étude : la cohérence spatiale permet de juger du stade de la réverbération à un instant donné à partir de deux points d'écoute, sans nécessiter la connaissance des signaux émis par la source, et la cohérence temporelle indique quels sont les plages de temps les plus significatives pour la caractérisation de l'onde directe.

1.2 Théorie modale

La connaissance préalable des modes propres, qui sont les solutions de l'équation des ondes sans second membre (en tenant compte cette fois-ci des conditions aux limites propres à la salle), permet d'effectuer une expansion modale approchée de la fonction de Green caractérisant la propagation entre deux points, et permettant ainsi de déduire la pression en un point quelle que soit l'excitation (Lyon, 1984; Tohyama et al., 1991) :

$$\begin{aligned} H\left(\vec{x}_s,\vec{x}_o,\omega\right) &\simeq \sum_m \frac{\psi_m\left(\vec{x}_s\right).\psi_m\left(\vec{x}_o\right)}{\omega^2-\omega_m^2-2j\omega\alpha_m} \\ &\simeq \sum_m \frac{\psi_m\left(\vec{x}_s\right).\psi_m\left(\vec{x}_o\right)}{2\omega_m}\left(\frac{1}{\omega-\omega_m-j\alpha_m}-\frac{1}{\omega+\omega_m-j\alpha_m}\right) \end{aligned} \tag{I.1}$$

En supposant que les fonctions modales soient réelles, on peut déduire de cette équation une formulation temporelle caractérisant la propagation :

$$h\left(\vec{x}_s,\vec{x}_o,t\right)=\sum_m \frac{1}{\omega_m}.\psi_m\left(\vec{x}_s\right).\psi_m\left(\vec{x}_o\right).sin\left(\omega_m t\right).e^{-\alpha_m t}u(t)$$

Dans ces expressions, $u(t)$ représente l'échelon de Heaviside, $\vec{x}_s$ et $\vec{x}_o$ sont respectivement les positions de la source et du récepteur, ψ_m désigne la fonction propre associée au m-ième mode de la salle, ω_m la pulsation propre de ce mode, et α_m la constante de décroissance correspondante, supposée négligeable devant ω_m, et liée au temps de réverbération à cette fréquence par :

$$\alpha_m=\frac{6.91}{T_r\left(\omega_m\right)}$$

Les pôles de chaque terme sont $\Omega_m^{\pm}=\pm\omega_m+j\alpha_m$, et représentent les valeurs propres solutions de l'équation d'onde sans second membre. Dans le plan des fréquences complexes, ces pôles sont situés sur la **ligne polaire**, qui dans le cas d'un temps de réverbération constant est la droite $y=\alpha$ parallèle à l'axe des réels, et dans le cas général la courbe $y=\alpha(\omega)$, sachant que $\alpha(\omega)$ est considéré comme lentement variant (Tohyama et al., 1991; Tohyama, 1997).

Densité modale, recouvrement modal

Jusqu'à présent, il n'a pas été fait mention de la distribution des modes propres sur l'axe des fréquences. A partir du cas d'une salle parallélipédique, il est possible de fournir une formulation approchée de la **densité modale**, c'est-à-dire du rapport du nombre de modes propres à la largeur de l'intervalle fréquentiel considéré, pour une fréquence donnée :

$$\frac{dN_{mod}}{df}\left(f\right)\simeq\frac{4\pi V}{c^3}.f^2\ \left(Hz^{-1}\right)$$

[1] Les notions de corrélation et de cohérence seront précisées en section suivante, ainsi qu'au chapitre suivant

[2] Cette distinction entre incohérence spatiale et temporelle est somme toute assez artificielle, ces deux effets étant dûs aux mêmes causes physiques.

Tout comme la formule exprimant la densité temporelle de réflexions, cette formule est appliquée à toute type géométrie.

La densité modale augmentant avec la fréquence et les modes ayant une largeur de bande non nulle, ces derniers vont interférer les uns avec les autres, et d'autant plus lorsque la fréquence augmente. Pour quantifier cet aspect, on définit le **recouvrement modal** comme étant le nombre de modes excités à une fréquence donnée :

$$M(f) = B_m . \frac{dN_{mod}}{dt}(f)$$

B_m est la bande passante à -3 dB du m-ième mode, c'est-à-dire $B_m = \frac{\delta_m}{\pi}\ (Hz)$. Donc :

$$M(f) = \frac{4.V.6.91}{c^3.T_r(f)}.f^2$$

Tout comme le modèle déterministe par sources images perd de sa pertinence lorsque la densité de réflexions est trop élevée, le modèle modal n'est plus valide lorsque la densité modale est trop élevée. Pour la plupart des géométries de salles, il est en effet impossible de prédire le comportement de la fonction de Green lorsque le recouvrement modal est trop important. La **fréquence de Schroeder**, qui donne un ordre de grandeur de la fréquence à partir de laquelle la théorie modale n'est plus valide, correspond à un recouvrement modal de 3, c'est-à-dire, si le temps de réverbération est indépendant de la fréquence :

$$\begin{aligned} M(f) = 3 \quad &\Leftrightarrow f_s^2 = \tfrac{3.c^3}{4*6.91}\tfrac{T_r}{V} \\ &\Leftrightarrow f_s \simeq 2000\sqrt{\tfrac{T_r}{V}}\ (Hz) \end{aligned}$$

La cohérence entre deux points d'écoute est là encore fonction du domaine considéré : alors qu'un mode isolé (ce qui est parfois le cas en basses fréquences) offre une cohérence parfaite en champ stationnaire, la superposition de plusieurs modes distincts, au fur et à mesure que la fréquence augmente, est source d'interférences qui la font chuter...et de même pour la corrélation, à ceci près que celle-ci dépend de la bande passante des signaux considérés.

1.3 Réverbération et filtrage

Il est très courant de considérer la propagation acoustique comme un filtrage linéaire, c'est-à-dire un processus linéaire à temps invariant. C'est en effet dans ce cadre qu'il est envisageable de caractériser celle-ci au moyen de réponses impulsionnelles mesurées. L'essor récent des techniques de synthèse de réverbération par convolution ont également contribué à propager ce point de vue.

Cependant, les cas pour lesquels il s'agit d'une description fidèle de la réalité sont en fait rares : si l'hypothèse de linéarité de la propagation dans une salle, vues les longueurs d'onde considérées, est très raisonnable, celle de l'invariance temporelle est plus douteuse dans une situation réelle. En effet, elle sous-entend entre autres que la source et le récepteur considérés soient parfaitement immobiles, ce qui est très rare, et ce n'est même pas toujours le cas en situation de mesure. Ainsi, les mesures de réponses impulsionnelles binaurales de salle, par exemple, ne peuvent être effectuées au moyen de séquences de Golay, car cette méthode est extrèmement sensible à la moindre modification du canal acoustique. Le fait que les mouvements de tête, fussent-ils minimes, suffisent à empêcher ces mesures, indique clairement que le canal ne peut être considéré à temps invariant.

De plus, les variations des propriétés acoustiques de l'air en fonction de la température, de l'hygrométrie, ou de la présence de turbulences, notamment, influent sur la propagation de telle sorte que celle-ci ne puisse plus être considérée comme à temps invariant même lorsque les transducteurs sont immobiles. Ainsi, Blesser (2001) rappelle qu'un changement de température de 1°C à température ambiante, ou une variation d'hygrométrie de 1%, suffit à faire varier la célérité du son d'environ 0,2%. Si des variations de cet ordre sont négligeables vis-à-vis des réflexions précoces, elles sont susceptibles de bouleverser le détail de la

réverbération tardive, et en particulier dans les hautes fréquences, même si l'enveloppe, qui reste l'aspect pertinent d'un point de vue perceptif, demeure inchangée.

Ces considérations forment un obstacle conséquent à l'une des approches envisagées dans le cadre de cette étude, visant à estimer "la" réponse impulsionnelle caractérisant la propagation pour ensuite appliquer des méthodes traditionnelles de description de l'effet de salle (voir l'annexe D). La solution proposée alors consiste à ne s'intéresser qu'au début de la réponse, et à appliquer une fenêtre exponentielle permettant de minimiser l'influence de la réverbération tardive.

Un problème supplémentaire inhérent aux méthodes par déconvolution est le fait que même dans le cas où l'hypothèse d'une propagation invariante dans le temps est valable (ce qui est par exemple le cas d'enregistrements soumis à des réverbérations synthétiques par convolution), la réponse impulsionnelle correspondante est à phase mixte dans la majorité des situations. En effet, les travaux de Tohyama (1997) et Tohyama et Lyon (1989) ont montré à partir du modèle modal que, si quand la source et le récepteur sont à la même position la réponse impulsionnelle est à phase minimale, lorsque ceux-ci s'éloignent l'un de l'autre, la proportion de zéros à phase maximale augmente d'autant plus vite que le temps de réverbération est important. Ainsi, la réponse impulsionnelle ne peut être qu'en partie estimée lorsque le rapport d'énergie directe sur l'énergie réverbérée est trop défavorable. Ceci dit, la théorie (très brièvement rappelée ci-dessus) stipule que les pôles de la fonction de transfert, qui sont intégralement contenus dans la partie à phase minimale, ont une partie réelle inversement proportionnelle au temps de réverbération, de sorte qu'il est envisageable d'estimer celui-ci sans nécessiter la connaissance du résidu passe-tout. Ainsi, Polack (1988) constate que l'enveloppe de l'autocorrélation d'une réponse impulsionnelle de salle décroît à la même vitesse que l'enveloppe du reflectogramme (aspect notamment utilisé ultérieurement par Hansen (1995)). Or l'autocorrélation d'un signal contient la même information que sa partie à phase minimale dans une décomposition phase minimale/passe-tout, car l'information de phase est supprimée.

1.4 Discussion

En théorie, les deux modèles physiques déterministes (modèle modal et modèle par sources images) sont équivalents, et chacun peut suffire à décrire complètement la propagation. Cette équivalence est d'ailleurs démontrable sur des cas très simples : ainsi, Allen et Berkley (1979) rappellent qu'une équivalence formelle peut être établie dans le cas d'une salle parallélipédique entre les expressions théoriques de la réponse impulsionnelle (obtenue par la méthode des sources images) et de la fonction de transfert (obtenue par expansion modale). La méthode des orbites, proposée par Polack (2001), est une autre méthode permettant de lier le formalisme modal au modèle par sources images : en étudiant les trajectoires fermées, ou "orbites", il est possible d'estimer la fonction de Green moyenne de la salle, et donc la densité modale.

Cependant, en dehors de ce type de cas d'école, les deux modèles ne permettent plus de caractériser totalement la propagation en espace clos. En effet, la régularité des sources images dans une salle parrallélipédique parfaite est brisée dès lors que l'on s'écarte de ce modèle. Une analogie simple est le cas d'une salle présentant deux miroirs parallèles : si les premières images de la salle sont placées régulièrement sur un axe perpendiculaire aux miroirs, les défauts de géométrie et de parallélisme de ces derniers font diverger les salles images d'ordre élevé d'une manière difficilement prévisible.

Puisque malgré l'emploi simultané de ces deux modèles déterministes, il subsiste une importante zone d'ombre, correspondant aux hautes et moyennes fréquences et pour la partie tardive de l'effet de salle, il est nécessaire de les compléter par une description de type stochastique.

2 MODÈLES STOCHASTIQUES DE RÉVERBÉRATION TARDIVE

2.1 Notion de champ diffus

Le concept de champ diffus est d'une importance particulière dans cette étude, puisqu'il sert à préciser la limite inférieure de la corrélation (ou cohérence) entre deux points d'écoute, qui sera le principal indice pour distinguer les instants pour lesquels l'onde directe est prédominante de ceux pour lesquels seule la réverbération est présente.

Un champ diffus est un champ uniforme et isotrope, c'est-à-dire pour lequel l'énergie incidente en chacun des points est répartie de manière égale selon toutes les directions et à toutes les positions ; de plus, on considère les phases distribuées de façon aléatoire, de sorte que les ondes incidentes en un point soient indépendantes d'une direction à l'autre et que leurs énergies puissent être ajoutées. Un champ diffus défini de la sorte est une vue de l'esprit : il n'existe pas de champ sonore parfaitement diffus, même s'il est possible de s'en approcher dans certains cas comme celui des chambres réverbérantes.

Il est également fréquent de considérer les situations pour lesquelles le champ diffus est stationnaire et ergodique dans le domainte temporel, bien que ces deux notions ne participent pas de la définition au sens strict. Dans ce cadre, il est possible de calculer la cohérence et la corrélation théoriques entre les signaux de pression en deux points du champ : ainsi, Cook et al. (1955) proposent un calcul de la corrélation $\rho_{\omega_0}(\tau = 0)$ en bande infiniment étroite et pour un retard nul (ce qui équivaut à la cohérence $\Phi(\omega_0)$ [3]) basé sur une décomposition en ondes planes du champ diffus, et trouvent :

$$\rho_{\omega_0}(\tau = 0) = \Phi(\omega_0) = \frac{sin\left(\frac{\omega_0.d}{c}\right)}{\left(\frac{\omega_0.d}{c}\right)}$$

...d étant la distance en mètres entre les points considérés. La généralisation à des retards non nuls et à des bandes de largeur non infinitésimale donne (Jacobsen et Roisin, 2000) :

$$\rho_{\omega_0}(\tau) = \frac{sin\left(\frac{\omega_0.d}{c}\right)}{\left(\frac{\omega_0.d}{c}\right)}.cos(\omega_0\tau).\frac{sin\left(\frac{\Delta\omega.\tau}{2}\right)}{\left(\frac{\Delta\omega.\tau}{2}\right)}$$

Dans cette expression, ω_0 et $\Delta\omega$ désignent respectivement la pulsation centrale et la largeur de la bande considérée. Il est important de rappeler que ce calcul est mené pour les pressions acoustiques, et ne peut donc être vérifié expérimentalement qu'avec des microphones omnidirectionnels. Dans le cas de microphones directionnels (orientés identiquement), la formulation est différente, mais la tendance globale est identique : en bande étroite, les signaux sont parfaitement corrélés en très basses fréquences ; lorsque la fréquence considérée augmente, la corrélation chute et atteint zéro pour une certaine fréquence f_c dépendant de l'écartement entre les points, puis fluctue autour de zéro avec une amplitude de plus en plus faible. Pour des signaux de pression, cette fréquence du premier passage à zéro de la cohérence vaut

$$f_c = \frac{c}{2d}$$

Ainsi, même un champ idéalement diffus ne permet pas une décorrélation totale des signaux.

2.2 Modèles de réverbération diffuse

Il a déjà été mentionné qu'il n'existe pas de champ sonore parfaitement diffus, même s'il est parfois possible de s'en approcher, par exemple en chambre réverbérante. En effet, dans

[3] On rappelle que la **cohérence** est définie par le rapport de l'interspectre par la racine du produit des autospectres ; la **corrélation** (au sens déterministe) est la moyenne du produit temporel de deux signaux après application d'un retard sur l'un d'entre eux ; la **corrélation normalisée** est le rapport de la corrélation par le produit des puissances ; finalement, le **coefficient de corrélation** est la valeur maximale de la corrélation normalisée. En bande étroite, on peut montrer que le carré du module de la cohérence est égal à la valeur de l'enveloppe de la corrélation normalisée pour un retard nul (Jacobsen et Roisin, 2000) ou, de manière équivalente, que la valeur de la corrélation normalisée pour un retard nul est égale à la partie réelle de la cohérence. Ces notions de corrélation et de cohérences seront précisées et développées au chapitre IV.

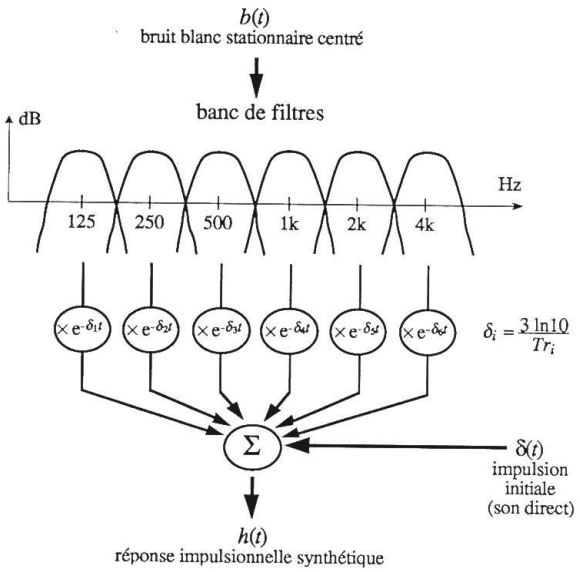


FIG. I.2 – **Technique de synthèse de réverbération proposée par Moorer (1979)** (repris de Jot (1992))

un champ réverbérant, si l'absorption des parois n'est pas trop importante, la multiplication au cours du temps du nombre de réflexions et de leur direction d'incidence permet de tendre vers un champ diffus. La vitesse à laquelle se forme ce champ diffus est directement liée à la géométrie des parois : plus celles-ci sont diffusantes (au sens de la loi de Lambert), plus l'approximation d'un champ diffus sera rapidement pertinente.

Contrairement aux hypothèses de travail ci-dessus, un champ diffus obtenu de la sorte au cours de la réverbération est dans le cas général non stationnaire dans le domaine temporel, et donc non ergodique, puisque l'énergie présente dans la salle décroît avec le temps dès lors que la source fait silence. Cela dit il est courant d'employer l'hypothèse de **locale stationnarité** à l'égard de tels champs : celle-ci consiste à considérer que si l'on considère une plage de temps très courte, on peut toujours considérer le champ comme stationnaire et ergodique, et donc que le signal de pression sous la forme d'un produit d'un signal stochastique stationnaire par une enveloppe déterministe à variations lentes.

C'est cette hypothèse qui permet de proposer un modèle de signaux de réverbération diffuse distinguant le processus stochastique de la décroissance exponentielle de l'énergie. Ainsi, Moorer (1979) propose une synthèse de réverbération basée sur plusieurs processus aléatoires indépendants dans des bandes de fréquences distinctes, multipliés par des enveloppes exponentielles dont la vitesse de décroissance dépend de la bande considérée, ce afin de modéliser la dépendance fréquentielle du temps de réverbération (voir figure I.2). Quelques années plus tard, Polack (1988) emploie un modèle simplifié apparenté à celui de Moorer, qui consiste à modéliser la réverbération tardive de la réponse impulsionnelle par une équation du type :

$$h(t) = n(t).e^{-\alpha.t}$$

Dans cette équation, $n(t)$ désigne un processus aléatoire (par exemple un bruit blanc en supposant le spectre diffus et le temps de réverbération indépendants de la fréquence), et $\alpha = \frac{6.91}{T_r}$ est le coefficient de décroissance de l'enveloppe. Ce modèle est valable dans la mesure où le temps de réverbération est indépendant de la fréquence, donc par exemple en bande étroite. De plus il ne concerne encore une fois que la fin de la réponse.

Dans une étude préalable (Baskind, 1999; Baskind et Polack, 2000) a été proposée une généralisation du modèle de Moorer, en considérant la réponse impulsionnelle comme une

superposition aléatoire de modes exponentiellement décroissants :

$$h(t) = \begin{cases} 0 & si\, t < 0 \\ \int_{-\infty}^{+\infty} N(f) . e^{j2\pi ft} . e^{-\alpha(f).t} & si\, t \geq 0 \end{cases} \quad \text{(I.2)}$$

En fréquence, ce modèle s'exprime sous la forme :

$$H(f) = \int_{-\infty}^{+\infty} N(f_0) . \frac{1}{\alpha(f_0) + j2\pi(f - f_0)} . df_0 \quad \text{(I.3)}$$

...$N(f)$ étant un processus a priori aléatoire non blanc, son spectre étant égal au spectre diffus résultant de la superposition dense de modes résonants. Il s'agit d'un modèle initialement stochastique, mais il peut aisément être mis en relation avec l'expansion modale de la fonction de Green proposée en section 1.2 : en effet, la mise en relation des équations I.1 et I.3 permet d'écrire :

$$N(f_0) = \frac{j}{4\pi} \sum_m \frac{\psi_m(\vec{x}_s) . \psi_m(\vec{x}_o)}{f_m} . [\delta(f_0 - f_m) + \delta(f_0 + f_m)]$$

La fonction $N(f)$ est discrète, et représente donc les amplitudes successives des modes aux fréquences propres correspondantes. Pour un recouvrement modal élevé, elle devient quasi-continue. Ce modèle peut donc aisément être mis en relation, de par sa définition même, tant avec les modèles temporels (géométriques) que fréquentiels (modaux).

2.3 Discussion

Alors que les modèles physiques présentés en section précédente sont des modèles exocentriques, c'est-à-dire aptes à décrire en théorie tout point de la salle sans se focaliser sur un en particulier, les différents modèles de réverbération diffuse sont **monophoniques** : ils ne tiennent pas compte de la dépendance des informations reçues en deux ou plusieurs points d'un champ réverbérant, et ne permettent donc pas d'émettre une quelconque hypothèse quant aux relations entre ces derniers.

De plus, il est nécessaire, pour pouvoir appréhender les aspects spatiaux de l'effet de salle, de considérer ces modèles de réverbération tardive dans une perspective globale, aux côtés des deux modèles physiques déterministes.

3 MODÈLE COMPOSITE DE PROPAGATION

3.1 Domaines de validité temps-fréquence

Puisqu'il est impossible de fournir un modèle unifié de propagation en espace clos qui soit valable à tout instant et pour toutes les fréquences, il s'agit de considérer alternativement l'une des trois classes de modèles envisageables, en fonction du domaine temps-fréquence considéré (Jot et al., 1997) :

- En basses fréquences (c'est-à-dire pour des fréquences inférieures à une limite de l'ordre de la fréquence de Schroeder), le formalisme modal est valide. Les informations entre plusieurs points d'écoute sont cohérentes.
- Pendant un laps de temps relativement court (inférieur à une durée de l'ordre du temps de mélange) après l'arrivée de l'onde directe, le modèle spéculaire est applicable. La cohérence des informations entre plusieurs points d'écoute dépend du nombre de réflexions simultanées, de leur direction d'incidence, et de la résolution temporelle considérée : elle est maximale à l'arrivée de l'onde directe, puis décroît au fur et à mesure de l'arrivée de nouvelles réflexions et de leur diffusion.
- Pour la majeure partie du déroulement de la réverbération (moyennes et hautes fréquences, durées supérieures au temps de mélange), aucun des deux modèles n'est valable, et il est alors nécessaire de considérer la réverbération comme un processus stochastique.

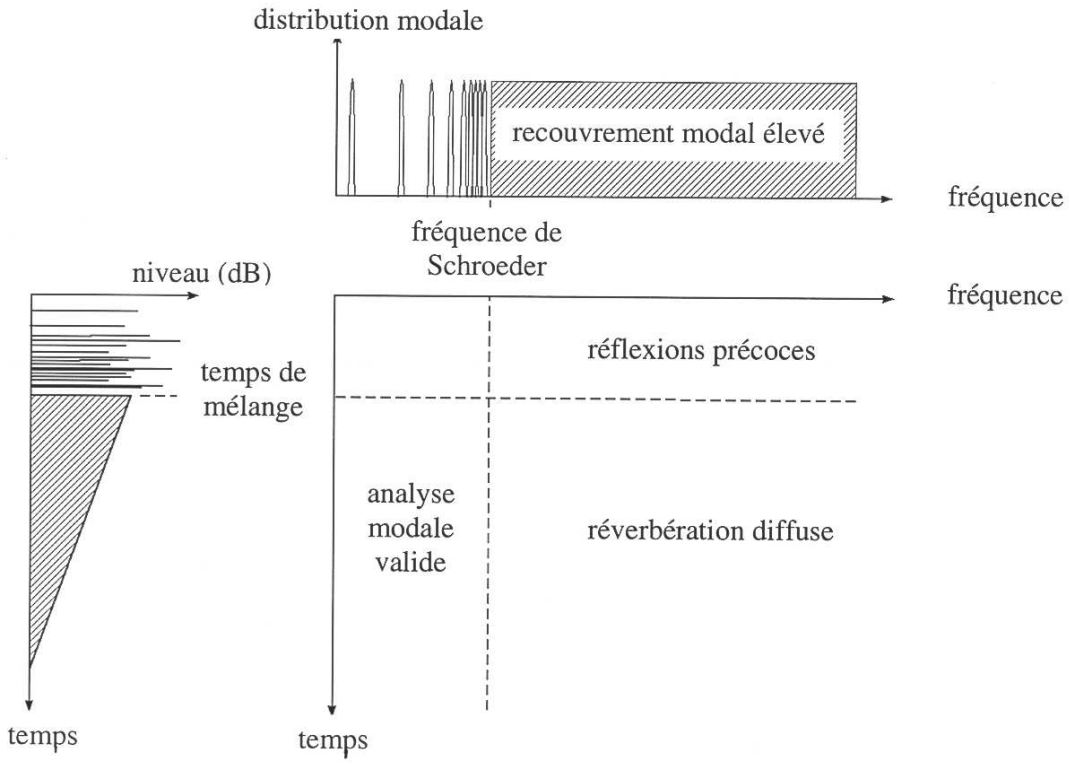


FIG. I.3 – **Représentation schématique des domaines de validité des modèles de propagation en espace clos**

La figure I.3 résume ce principe. Il s'agit d'un point de vue schématique, car en pratique la transition en temps et en fréquence d'un modèle déterministe vers un modèle stochastique n'est pas brusque, mais progressive, et dépend de plusieurs facteurs, tant liés à l'analyse (résolution en temps et en fréquence) qu'aux propriétés géométriques et physiques de la salle. Ainsi, la vitesse à laquelle on bascule d'un modèle spéculaire à un modèle diffus dépend de la diffusion des parois, et donc de la fréquence, puisque la quantité d'énergie diffusée sur une paroi donnée est fonction de la comparaison entre l'ordre de grandeur de ses irrégularités et la longueur d'onde. Cet aspect n'est pas considéré dans la définition du temps de mélange, qui se base sur des considérations perceptives dans le cadre de salles faiblement diffusantes.

3.2 Limites de l'analyse spatiale par cohérence stationnaire

Puisqu'une réverbération ne peut être considérée comme diffuse dans sa totalité, les techniques stationnaires d'estimation de la cohérence, par exemple utilisant la méthode du périodogramme ou du corrélogramme, fournissent des résultats qui ne peuvent pas être comparés avec la formulation théorique de la corrélation en bande étroite proposée par Cook et al. (1955).

On peut se rendre compte de ce phénomène sur la figure I.4 : la salle est un volume vide à section carrée d'environ $650\,m^3$, aux parois réfléchissantes, de temps de réverbération de 3 secondes en basses fréquences et de 1 seconde à 10 kHz. L'effet de salle est représenté par une réponse impulsionnelle binaurale mesurée sur un sujet humain se trouvant à 90 cm de la source, par déconvolution à partir d'un sinus glissant, à une fréquence d'échantillonnage de 44100 Hz. Le temps de mélange vaut 25 millisecondes, et la fréquence de Schroeder 136 Hz.

La première figure indique la cohérence intercanale calculée par la méthode du périodogramme moyenné de Welch utilisant des fenêtres de Hanning de 1024 points non recou-

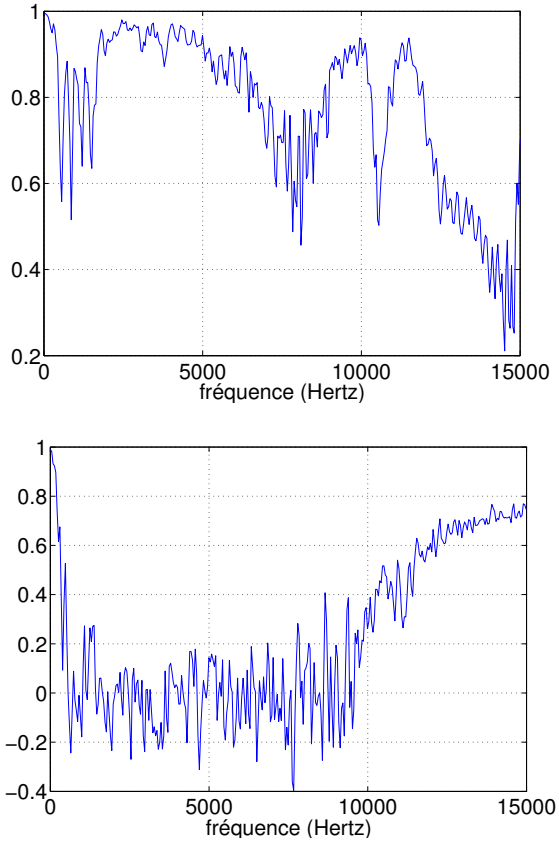


FIG. I.4 – **Limites de l'analyse spatiale par cohérence stationnaire** : La réponse impulsionnelle analysée est une réponse binaurale. La figure supérieure représente la cohérence stationnaire calculée sur la résultante de la convolution de cette réponse par une séquence de bruit blanc de 23 secondes (seule la partie stationnaire est considérée). La figure inférieure représente la cohérence calculée de la même manière, mais la réponse impulsionnelle a ici été tronquée de 20000 échantillons (0,45 secondes) au début avant convolution.

vrantes, appliquée à la résultante de la convolution des deux canaux de cette réponse par une séquence de bruit blanc de 23 secondes, seule la partie stationnaire étant considérée. La présence du son direct et des réflexions précoces dans le signal force la cohérence à rester à un niveau globalement élevé (supérieur à 0,5) à toutes les fréquences ; on note bien le lobe principal en basses fréquences comme le prévoit la théorie, mais à celui-ci s'ajoutent d'autres zones de forte cohérence, comme par exemple entre 2 kHz et 6 kHz, alors que la cohérence théorique en champ diffus est quasiment nulle à ces fréquences.

En revanche, si l'on s'affranchit de la partie précoce de la réponse impulsionnelle, on rejoint bien plus fidèlement la théorie. La partie inférieure de la figure I.4 représente la cohérence calculée de la même manière, mais la réponse impulsionnelle a au préalable été tronquée de ses 20000 premiers échantillons, ce qui représente 0,45 secondes. On distingue très nettement le premier lobe, coupant à environ 500 Hz ; au-delà, la cohérence reste inférieure à 0,25 jusqu'à 9 kHz, puis remonte progressivement en très hautes fréquences jusqu'à environ 0,7. Cette recorrélation des signaux en hautes fréquences s'explique par la présence de bruit de mesure : en effet, celui-ci correspond principalement au bruit acoustique, ainsi qu'au bruit de fond électronique et électromagnétique ; or, puisque les deux canaux ont été enregistrés simultanément, ce dernier est très similaire d'une voie à l'autre. Cette résurgence du bruit de fond (et donc de la cohérence) intervient principalement dans les hautes fréquences puisque le temps de réverbération y est relativement faible, si bien que le rapport signal à bruit est défavorable. L'utilisation d'une transmission symétrique du signal électrique, peu

courante en mesure de salles, devrait pouvoir réduire largement le bruit électromagnétique, et donc de fait la cohérence entre les deux voies du bruit de fond.

Puisque la prise de son est binaurale, on ne respecte pas en théorie les hypothèses à la base du calcul de cohérence en champ diffus proposé par Cook et al (1955), et dont le résultat est indiqué en section 2.1 : en effet, celui-ci suppose que les signaux considérés soient de purs signaux de pression, captés par exemple par des microphones omnidirectionnels. Or la prise de son binaurale est par essence directionnelle, et c'est même cet aspect qui permet une localisation de la source sur des critères spectraux ou de différences interaurales d'intensité (voir chapitres II et VI). Cependant, en basses fréquences, les retards interauraux sont de l'ordre de $\frac{3a}{c}.sin(\theta)$ dans le plan horizontal (θ étant l'azimut et a le rayon de la tête pour un modèle sphérique) (Kuhn, 1977), et la réponse en fréquence relative à chacun des microphones est quasiment indépendante de la direction de provenance, si bien que la configuration binaurale est équivalente à un couple de microphones omnidirectionnels distants d'une trentaine de centimètres. Cette distance est cohérente avec la fréquence de coupure d'environ 500 Hz mentionnée ci-dessus.

On parvient donc à la conclusion que si le champ réverbéré tend progressivement vers un champ diffus, la présence de réflexions spéculaires empêche de révéler celui-ci. L'étude de la cohérence, qui est une forte source d'information vis-à-vis de la distribution spatiale de l'effet de salle, dans des salles autres que des chambres réverbérantes nécessite l'emploi de méthodes plus adaptées que les méthodes stationnaires.

3.3 La cohérence à court-terme comme outil d'analyse spatiale de réponses impulsionnelles de salles

Le caractère non stationnaire d'une réponse de salle incite à employer une définition de la cohérence elle aussi non stationnaire, ce qui est le cas de la **cohérence à court-terme**. La cohérence à court-terme a été notamment employée par Allen et al. (1977), puis par Avendano et Jot (2002) comme outil de détection sur des signaux musicaux, et elle est utilisée de la sorte dans cette étude, notamment au chapitre V. Or son emploi en tant qu'outil de description de l'effet de salle à partir de réponses impulsionnelles est également tout-à-fait justifiable. Ainsi, Blauert (1997) rappelle un principe énoncé par Danilenko, selon lequel la corrélation à court-terme entre les deux voies d'une réponse de salle reste élevée tant que des réflexions spéculaires émergent, puis décroît graduellement au fur et à mesure que l'on tend vers un champ diffus (figure I.5). L'application de la cohérence à court-terme à une réponse impulsionnelle bicanale est une généralisation de ce principe à une analyse en bandes étroites.

La section 5 du chapitre IV présente plus en détail cet outil, et le met en relation avec la corrélation à court-terme ; on se contente d'indiquer ici la définition de la cohérence à court-terme, qui est similaire à celle du périodogramme de Welch, si ce n'est que les contributions locales aux interspectres et autospectres sont ici pondérées par une fenêtre d'oubli, qui est typiquement une fonction exponentiellement décroissante du temps :

Soient $X[p,k]$ et $Y[p,k]$ les transformées de Fourier discrètes à court-terme des signaux (discrets) considérés $x[n]$ et $y[n]$ (p désigne la plage temporelle considérée, et k l'indice fréquentiel) ; on définit l'**interspectre à court-terme** $S_{XY}[p,k]$ et les **autospectres à court-terme** $S_{XX}[p,k]$ et $S_{YY}[p,k]$ à partir des transformées de Fourier à court-terme selon :

$$S_{XY}[p,k] = \sum_{q=-\infty}^{p} Y^*[q,k].X\,[q,k]\,.e^{-\frac{p-q}{\Delta T.F_e}}$$

$$S_{XX}[p,k] = \sum_{q=-\infty}^{p} |X[q,k]|^2\,.e^{-\frac{p-q}{\Delta T.F_e}}$$

$$S_{YY}[p,k] = \sum_{q=-\infty}^{p} |Y[q,k]|^2\,.e^{-\frac{p-q}{\Delta T.F_e}}$$

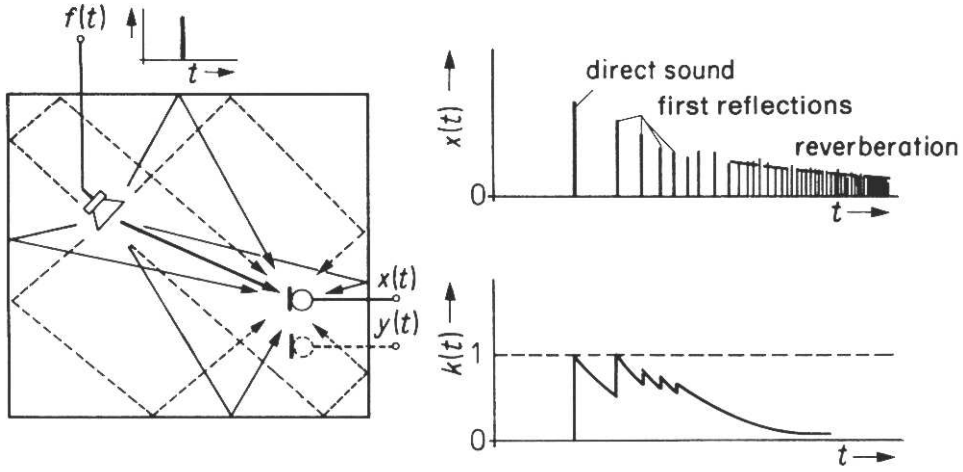


FIG. I.5 – **Évolution de la corrélation entre deux microphones plongés dans un champ réverbérant** (repris de Blauert (1997))

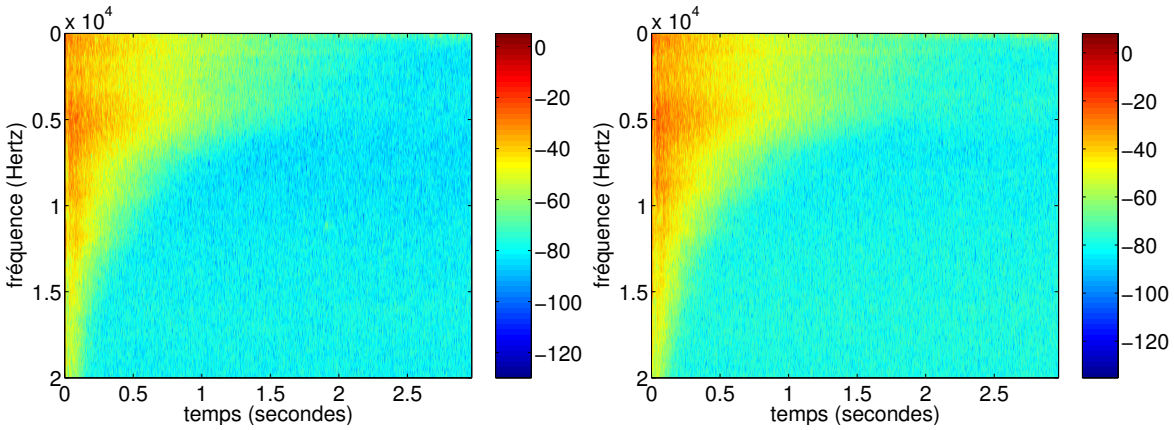
Dans ces expressions, ΔT désigne la constante de temps de la fenêtre d'oubli (en secondes), et F_e est la fréquence d'échantillonnage des signaux. La cohérence à court-terme vaut alors simplement :

$$\Phi_{XY}[p,k] = \frac{S_{XY}[p,k]}{\sqrt{S_{XX}[p,k].S_{YY}[p,k]}}$$

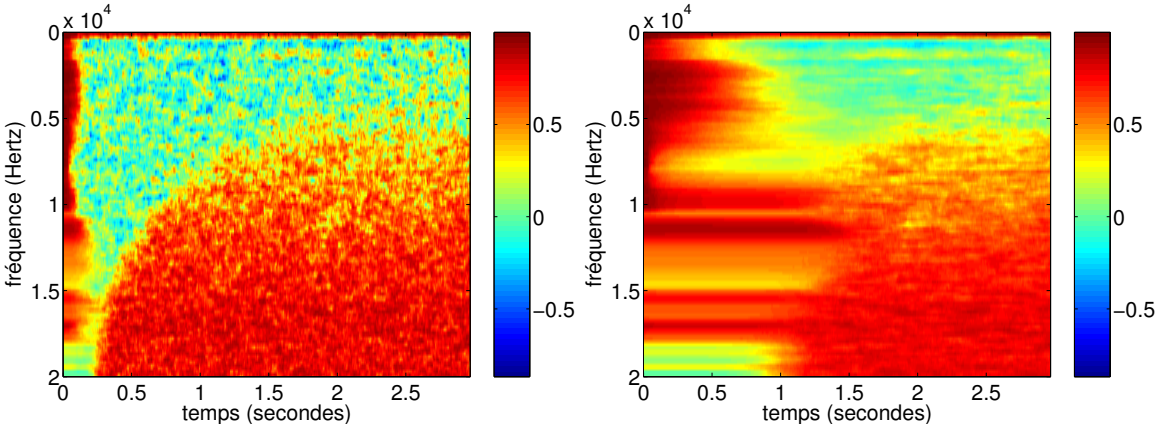
Les paramètres de la cohérence à court-terme sont d'une part les paramètres de la transformation de Fourier à court-terme (c'est-à-dire la forme et la longueur de la fenêtre d'analyse spectrale et le taux de recouvrement) à la base du calcul, et d'autre part la constante de temps τ de la fenêtre d'oubli. La valeur minimale de cette constante de temps est déterminée par la longueur de la fenêtre d'analyse temps-fréquence : en effet, cette dernière doit être nécessairement beaucoup plus courte que la fenêtre d'oubli servant à calculer les spectres, pour que cette dernière joue convenablement son rôle de moyenne (en d'autres termes, pour que la variance de l'estimation soit suffisamment faible). Cependant, la longueur de la fenêtre d'oubli conditionnant la résolution temporelle de l'analyse (pour la partie précoce) et la validité de l'hypothèse de stationnarité locale (pour la partie diffuse), il est délicat de conférer une valeur importante à τ.

La figure I.6 permet de représenter un exemple de calcul par cohérence à court-terme : les deux figures supérieures représentent les spectrogrammes issus de la transformation de Fourier à court-terme des deux voies de la réponse impulsionnelle binaurale étudiée en section 3.2, pour une fenêtre de Hanning de 6 millisecondes, avec un recouvrement de 75%. Les deux figures intermédiaires représentent la cohérence à court-terme correspondantes, pour deux valeurs de la constante de temps (20 et 100 millisecondes). Dans les deux cas, on distingue plusieurs zones, représentées schématiquement sur la figure inférieure. Dans le cas de la constante de temps longue, la cohérence estimée est plus stable dans la partie diffuse, et non sujette à des fluctuations rapides comme dans le cas où la constante de temps vaut 20 ms. Ceci la rend plus lisible, et en accord avec la théorie de Cook et al. (1955). La figure I.7 permet justement d'observer l'évolution temporelle de la cohérence à court-terme dans le cas d'une fenêtre d'oubli longue. En revanche, cet avantage a une contrepartie, qui est la perte de la résolution temporelle au début de la réponse impulsionnelle : alors que le cas où la constante de temps vaut 20 ms indique que la partie fortement cohérente reste confinée dans les premières 100 millisecondes de la réponse impulsionnelle (le temps de mélange vaut 25 ms), la transition est beaucoup plus lente dans le cas où la fenêtre d'oubli est longue, et n'intervient pas avant une seconde dans les hautes fréquences.

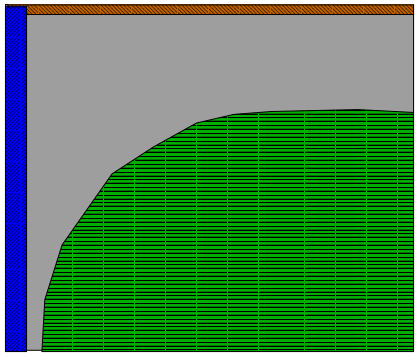
Quoi qu'il en soit, la cohérence à court-terme, bien que non suffisante pour quantifier



Spectrogramme calculé par FFT avec une fenêtre de Hanning de 256 points (6 ms) avec un recouvrement de 75%, pour les voies gauche et droite.



Partie réelle de la cohérence à court-terme, pour des constantes de temps $\Delta T = 20\,ms$ (gauche) et $\Delta T = 100\,ms$ (droite)



Représentation schématique de la cohérence à court-terme : la zone en bleu représente la partie précoce (spéculaire) de l'effet de salle ; la zone en marron représente le domaine de forte cohérence du champ diffus en basses fréquences ; la zone en vert correspond au domaine temps-fréquence pour lequel le champ diffus est noyé dans du bruit fortement cohérent d'une voie à l'autre ; finalement, la zone grise représente le domaine où le champ peut-être considéré diffus, incohérent d'une voie à l'autre, et prépondérant par rapport au bruit de fond.

FIG. I.6 – **La cohérence à court-terme comme outil d'analyse spatiale de la réverbération** : représentations temps-fréquence en énergie et en cohérence (partie réelle) d'une réponse impulsionnelle binaurale de salle.

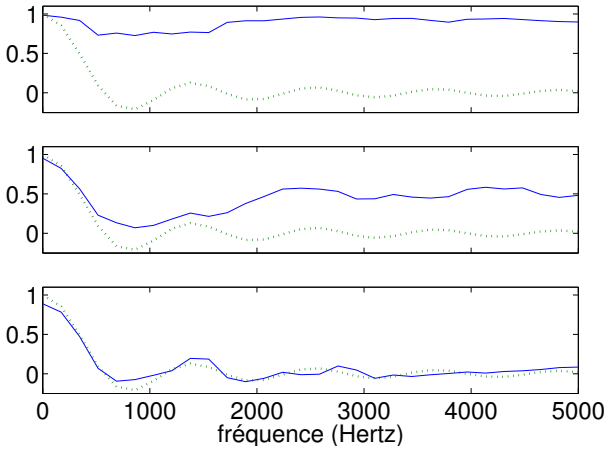


FIG. I.7 – **Évolution temporelle de la partie réelle de la cohérence à court-terme** ($\tau = 100\,ms$). Figure supérieure : cohérence à l'instant $t = 140\,ms$; figure intermédiaire : $t = 720\,ms$; figure inférieure : $t = 1,59\ s$

les aspects spatiaux de la scène sonore, permet de les qualifier. Lorsque la cohérence dans une bande de fréquences est égale à 1, les signaux dans cette bande ne diffèrent que par un gain et un retard, alors que plus elle est faible, plus ils sont décorrélés. La superposition des réflexions successives à l'onde directe est traduite par une perte de cohérence au cours du temps, et il est donc possible grâce à cet outil de représenter la transition progressive vers un champ diffus.

4 CONCLUSION

Ce chapitre, en rappelant les bases théoriques des modèles déterministes et stochastiques de propagation du son en espace clos et leurs domaines de validités, a permis de dégager quelques unes des notions fondamentales vis-à-vis du problème qui nous concerne : d'une part, l'étude des aspects spatiaux à partir d'un nombre restreint de points d'observation nécessite de s'intéresser aux **relations intercanales**, car dans une approche monophonique, une majeure partie de l'information spatiale est perdue ; d'autre part, il est nécessaire d'appréhender l'effet de salle comme un phénomène **non stationnaire**. Les techniques stationnaires comme la cohérence (au sens stationnaire) ne sont pas adaptées, car elles ne permettent pas de distinguer les instants où l'onde directe est prépondérante de ceux où le champ réverbéré est prépondérant ; or, c'est uniquement sur la base de cette distinction que l'on est en mesure de décrire précisément les différents aspects spatiaux de la scène sonore.

La recherche de techniques non stationnaires d'estimation des statistiques du deuxième ordre a conduit à l'utilisation de la **cohérence à court-terme** pour une caractérisation spatiale d'un effet de salle à partir de réponses impulsionnelles. De plus il est rappelé (et ce point sera plus amplement développé au chapitre IV) que la cohérence peut être considérée comme un cas particulier de la corrélation normalisée dans le cas de bandes étroites. Ces notions de corrélation et de cohérence non stationnaires sont à la base de la majorité des outils développés dans le cadre de cette étude.

II

Perception spatiale d'une scène sonore

LA PERCEPTION D'UNE SCÈNE SONORE commence par l'identification des sources, puis leur caractérisation sous de nombreux aspects, dont spatiaux, c'est-à-dire leurs positions et largeurs apparentes. A ceci vient s'ajouter la perception de l'environnement acoustique, qui, si l'on se limite aux espaces clos, véhicule les informations sur sa taille, ainsi que sur les sensations assez liées d'enveloppement et de réverbérance. La perception de la salle fusionne rarement avec celle de la source, créant ainsi deux types de flux auditifs (ou plus dans le cas de plusieurs sources). Griesinger (1997) pense que ces deux modes de perception sont traités par deux processus physiologiques distincts, mais que ce soit le cas ou non, le principe de séparation perceptive entre source et salle reste consensuel, aussi bien dans la communauté scientifique que parmi les spécialistes en écoute critique. Ce chapitre adopte ce point de vue, tout d'abord en l'explicitant un peu plus longuement (section 1), puis en détaillant la perception anéchoïque de la position de la source (section 2) et l'influence de la salle sur les attributs spatiaux de l'événement auditif (section 3), puis la perception de la salle, qui dépend de la position de la source et du message véhiculé (section 4). Cette subdivision, même si elle propose en apparence une organisation proche d'une description chronologique de l'effet de salle, s'en distingue dans le principe, et permet d'éviter de séparer la perception spatiale de la salle de celle de la source au moyen de bornes temporelles arbitraires.

1 ÉVÉNEMENT AUDITIF ET IMPRESSION SPATIALE

1.1 Source sonore et événement auditif

L'étude de la perception humaine nécessite au préalable de faire une séparation nette entre le monde physique et le monde perceptif. Blauert (1997) distingue ainsi **l'événement sonore**, c'est-à-dire l'onde de pression créée par la source physique, de l'**événement auditif**, c'est-à-dire la représentation interne à l'espace perceptif, qui lui est associé. Cette distinction n'est pas simplement formelle, mais repose sur le fait que le monde perceptif diffère du monde physique sur bien des points : par exemple, en vertu de l'effet de précédence, deux ou plusieurs événements sonores peuvent créer un unique événement auditif, dont la position peut coïncider avec une source sonore ou non (c'est le cas par exemple de la stéréophonie) ; de même, un événement sonore peut créer deux événements auditifs (ce qui est plus rare en écoute naturelle). Il y a également de très nombreuses dissimilitudes spatiales entre ces deux mondes : la position et la largeur d'un événement auditif ne correspondent pas forcément à celles de la source physique qui en est la cause, surtout en l'absence d'informations visuelles, qui aident à résoudre certaines confusions. Cette distinction est fondamentale, car il est nécessaire de prendre de la distance par rapport au point de vue selon lequel la perception, et en particulier l'audition, est un capteur permettant de décrire avec plus ou moins d'erreur une situation avérée.

1.2 L'espace sonore et sa perception

Parallèlement à cet événement auditif, se développe au cours de l'écoute une représentation interne de l'espace sonore contenant la ou les sources. Cet espace sonore peut correspondre à un environnement acoustique (clos ou non) dans lequel se trouve l'auditeur, soit à un espace virtuel issu d'un travail de production et reproduit sur haut-parleurs. Cela dit, la majeure partie des recherches s'est portée sur l'audition dans les salles de concerts, ou dans un contexte anéchoïque.

L'histoire des recherches sur la perception des salles (Kahle, 1995) est marquée par une compréhension de plus en plus complète des différents phénomènes perceptifs et de leurs causes physiques et physiologiques, à laquelle il a été nécessaire d'associer un vocabulaire sans cesse plus riche et plus précis. La question du lexique est cruciale, car ses lacunes, les confusions ou multiples interprétations d'un même terme sont autant de freins à la structuration de l'écoute critique, ainsi qu'à son étude subjective et objective. Cette remarque concerne tous les aspects de la perception de l'espace sonore, et donc en particulier les aspects spatiaux, rassemblés sous le terme générique d'**impression spatiale**.

L'impression spatiale

Historiquement, le premier aspect "spatial" de la perception d'une salle à avoir été étudié est la **réverbérance**. Sabine (1922) s'y intéresse à cause de son effet sur la clarté du discours, et la considère comme l'un des critères majeurs vis-à-vis de l'appréciation de la salle[1]. Quarante ans plus tard, Beranek (1962) introduit parmi ses 18 attributs permettant de qualifier l'acoustique d'une salle les notions spatiales d'**intimité** (*intimacy*), de **balance** (dont la balance spatiale) et de **mélange** (*blend*). L'intimité, ou présence, symbolise la sensation d'être au sein d'un espace clos, et qui plus est au sein d'un espace clos de petites dimensions, et peut donc être relié à la notion de **taille apparente de la salle**.

Mais aucun de ces deux travaux, ni ceux de leurs contemporains, ne mentionne les aspects géométriques des percepts liés à l'écoute spatiale. Barron (1971), poursuivant les travaux de Marshall et de de Keet, étudie de manière approfondie l'**élargissement apparent de la source**, qu'il attribue aux rôles conjoints des réflexions précoces et de l'effet de précédence. Cet élargissement de l'événement auditif est désigné sous le terme d'**impression**

[1] A noter que la réverbérance est donc ici considérée sous son aspect temporel (durée de la réverbération), et non spatial (sensation d'enveloppement par le champ diffus)

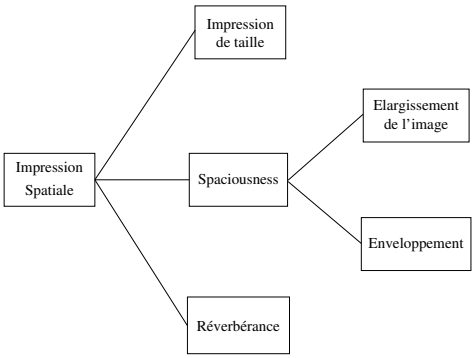


FIG. II.1 – **Les différents aspects de l'impression spatiale** (d'après Potter (1993))

spatiale qui, contrairement à l'acception actuelle, désignait alors au sens de Barron cet effet très précisément. Ce point de vue est remis en question ultérieurement par Kuhn (1977), puis par Blauert et Lindemann (1986a), ces derniers définissant l'impression spatiale comme "le concept de type et de taille d'un espace réel ou simulé auquel l'auditeur arrive spontanément lorsqu'il est baigné dans un champ sonore approprié". Cette définition, qui n'est peut-être pas entièrement satisfaisante puisqu'elle ne mentionne que la salle et non pas la ou les sources, inclut de fait la réverbérance et la taille apparente de la salle, les auteurs préférant le terme de "spaciousness"[2] pour désigner la largeur apparente de la source. Les travaux, encore plus récents, de Bradley et Soulodre (1995a) ont permis de formaliser la notion d'**enveloppement de l'auditeur** (*listener envelopment*, ou LEV), qui désigne la sensation d'immersion dans un champ réverbérant. Il a paru naturel aux auteurs d'intégrer cet aspect au vocable de "spaciousness" , au même titre que l'élargissement apparent de la source.

De ces différents travaux découle la taxonomie majoritairement admise aujourd'hui au sein de la communauté scientifique (voir par exemple Blauert (1997) ou Potter (1993)), dont une représentation graphique, issue de la thèse de Jan Potter, est donnée figure II.1 . On peut cependant regretter que la localisation de la source ne soit pas intégrée à cette classification : en effet, la propagation dans la salle a un tel effet sur la perception de la direction (décalage de l'image, élargissement, voire délocalisation) et de la distance qu'il peut paraître naturel d'intégrer la localisation à l'impression spatiale. Cette exclusion résulte de la distinction faite au sein de la communauté scientifique entre l'écoute en situation anéchoïque et l'écoute en espace clos. Blauert rappelle dans son ouvrage "Spatial Hearing" (1997, introduction) qu' "il n'y a pas d'écoute non spatiale". Un corollaire serait qu'il n'y a pas de lieu, clos ou non, qui ne soit potentiellement vecteur d'impression spatiale.

On pourrait également ajouter que cette classification découle principalement d'expériences basées sur le paradigme simplifié d'une source unique et statique, et est donc difficilement applicable à des scènes complexes.

La qualité spatiale selon Rumsey

La taxonomie de la qualité spatiale proposée récemment par Rumsey (2002), beaucoup plus complète, apporte une réponse à ces deux dernières remarques. Le but affiché est de développer un vocabulaire permettant de décrire au mieux des scènes sonores complexes, d'où la nécessité de prendre en compte le cas de plusieurs sources. Chaque attribut est donc considéré comme relatif à l'espace sonore, à une source prise individuellement, ou bien à l'ensemble de sources. La distinction entre attributs individuels des sources et attributs de l'ensemble a un sens, car même si les deux sont corrélés, il n'y a pas nécessairement équivalence. Par exemple, la largeur individuelle d'une source peut ne pas être définissable si

[2]intraduisible, se réfère à l'extension spatiale d'un objet. Le terme de "spatialité" (équivalent de l'anglais "spatiality"), parfois employé, n'est pas adéquat, puisqu'il désigne de manière plus basique le fait qu'un objet possède ou non des attributs spatiaux.

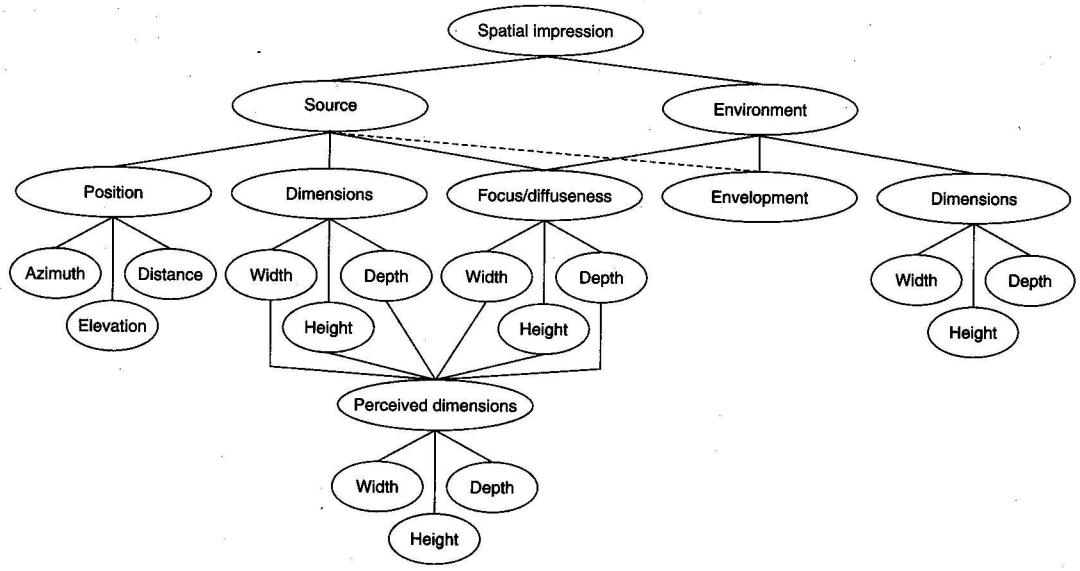


FIG. II.2 – **La qualité spatiale selon Mason et Rumsey** (repris de Rumsey (2001))

celle-ci est fortement mélangée à ses voisines, ou la largeur de l'ensemble peut être supérieure à la largeur des sources qui le composent. La classification tient compte de notions dynamiques particulièrement pertinentes dans le cas de scènes reproduites, telle la stabilité des images.

Rumsey tâche également de résoudre certaines confusions possibles dans la terminologie usuelle. Ainsi, il relève une ambiguïté sémantique courante, propre au vocable d' "attributs spatiaux" : s'agit-il des *attributs spatiaux*, c'est-à-dire géométriques, ou bien des *attributs de l'espace*, c'est-à-dire relatifs à l'espace acoustique dans lequel les sources sont baignées ? Ces deux catégories ne sont pas totalement disjointes, puisque les attributs spatiaux rassemblent les attributs spatiaux liés à la source et ceux liés à l'espace de diffusion, ce qui entretient un peu plus la confusion. Rumsey opte clairement pour le premier point de vue, excluant ainsi tous les critères non géométriques, comme la réverbérance ou la chaleur, de la notion de qualité spatiale. Ce point de vue pourrait néanmoins être discuté, car la réverbérance résulte d'un phénomène à la fois spatial (enveloppement par le champ diffus) et temporel.

La terminologie proposée par Rumsey, et que l'on retrouve entre autres dans les travaux de Mason (voir figure II.2), est la suivante :

- **attributs de largeur** : largeur individuelle d'une source, largeur de l'ensemble, largeur de l'environnement, largeur de la scène
- **attributs de distance et de profondeur** : distance et profondeur individuelles d'une source, distance et profondeur de l'ensemble, profondeur de l'environnement, profondeur de la scène
- **attributs d'immersion** : enveloppement individuel par une source, enveloppement par l'ensemble, enveloppement par l'environnement, présence
- **attributs divers** : biais latéral de la scène, biais avant-arrière de la scène, stabilité et précision de localisation des sources, stabilité de la scène, homogénéité de la scène

Curieusement, il n'est pas fait mention dans cette classification de la position des sources, alors que s'y trouve la notion de précision de localisation (*source focus*), ainsi que les attributs dynamiques correspondants (stabilité). Il peut être également intéressant de noter la nuance faite entre "enveloppement par la salle" et "enveloppement par la source". Cette subdivision du critère d'enveloppement, qui se base sur la distinction perceptive entre source et salle, rejoint le point de vue de Griesinger développé ci-dessous.

L'exhaustivité de cette taxonomie contraste fortement avec le vocabulaire de description

spatiale adopté usuellement par les chercheurs et les acousticiens, en complétant ce dernier (notion de profondeur de source[3], de biais/stabilité de la localisation,...), et en l'adaptant au cas de scènes complexes. Ceci peut rendre délicat son utilisation dans un cas pratique, tant la nuance entre certains critères est parfois subtile, et rend encore plus ardue toute tentative de modélisation objective. Cependant, cette classification, très proche du vocabulaire d'expertise employé par les ingénieurs du son, paraît beaucoup mieux adapté à une situation réelle et à des scènes auditives complexes que ne l'est la classification usuelle[4].

L'impression spatiale selon Griesinger

Griesinger (1997) définit l'impression spatiale comme "l'intensité de la sensation de se trouver dans un espace clos". Au sens de Rumsey, cette définition désignerait plus une "impression d'espace" qu'une "impression spatiale". Griesinger exclut donc la largeur apparente de l'impression d'espace, et s'oppose à l'utilisation actuelle du vocable de "spaciousness", qui selon lui devrait désigner le caractère spacieux ou non d'une salle, et non pas l'extension spatiale des événements auditifs. Il considère donc l'impression d'espace comme l'ensemble des indices perceptifs permettant de se sentir entouré par des murs, ce qui en fait une notion plus restreinte et plus précise que son acception usuelle.

Son point de vue sur la perception spatiale se base sur l'hypothèse selon laquelle l'analyse d'une scène auditive par le cerveau traite parallèlement deux types de flux auditifs : des flux de premier plan (*foreground streams*), liés aux sources sonores, et un flux d'arrière-plan (*background stream*), lié à la salle. Tous deux sont porteurs d'impression spatiale, mais ne sont pas perceptibles simultanément, le flux d'arrière-plan étant masqué par les sources lorsque celles-ci émettent du son. Notre attention se focalise tour à tour sur la source, puis sur la salle lorsque la source fait silence, et on comprend alors que l'impression spatiale varie largement en fonction du message sonore et du type de salle. Il distingue ainsi trois formes d'impression spatiale, deux étant liées à la source, et la troisième à la salle :

- **l'impression spatiale continue** (*continuous spatial impression*, ou CSI) : il s'agit de la seule forme d'impression spatiale pour des sources entretenues, puisque la réverbération est alors masquée. Cette impression spatiale correspond à un champ sonore totalement enveloppant avec ou sans une source localisable avec précision. Griesinger l'attribue à une fluctuation continue des indices interauraux, due à l'interférence entre l'onde directe et les réflexions précoces, et renforcée par des modulations d'amplitude ou de fréquence de la source (vibrato). Un exemple de cette impression spatiale est la sensation partielle ou totale de délocalisation lorsque lors d'un concert, un chanteur tient une note pendant plusieurs secondes. Cette notion est à rapprocher de celle d' "enveloppement par la source" proposée par Rumsey.
- **l'impression spatiale précoce** (*early spatial impression*, ou ESI) : l'ESI est l'impression spatiale permettant de juger de la présence des murs dans une très petite salle, qui ne peut être enveloppante puisque la réverbération n'y est pas assez longue. L'ESI s'observe pour des sources non continues, et accompagne l'élargissement apparent de la source, mais en s'en distinguant dans le principe, puisqu'il s'agit d'une impression liée à la sensation de présence de la salle. Griesinger attribue la CSI et l'ESI aux réflexions latérales précoces.
- **l'impression spatiale d'arrière-plan** (*background spatial impression*, ou BSI) : cette forme d'impression spatiale n'est perceptible qu'aux moments ou la source fait silence, après un certain laps de temps nécessaire pour révéler la perception de la salle. Il s'agit des seuls instants auxquels on est sensible à la réverbération tardive si celle-ci est audible (la durée de la réverbération doit être suffisante, ce qui n'est généralement pas le cas des petites salles). Cette forme d'impression spatiale est à rapprocher des notions d'enveloppement et de réverbérance aux sens traditionnels des termes.

Cette théorie, qui se base sur des notions de **masquage** et de ségrégation de flux, et adopte un

[3]Blauert et Lindemann (1986a) font néanmoins mention d'une sensation d'augmentation de profondeur de source dues aux composantes basses fréquences des réflexions latérales (voir section 3.2)

[4]On retrouve ici le dilemme fréquent entre le besoin de décrire des scènes complexes réalistes et la capacité à pouvoir en contrôler la totalité des paramètres afin de les reproduire en laboratoire.

point de vue assez original et délicat à saisir, d'autant plus qu'elle utilise des termes usuels en leur conférant un sens différent. Cependant, le contraste qu'elle offre avec le point de vue usuel est riche d'enseignements. L'un de ses principaux apports est de nous permettre de prendre de la distance par rapport aux mesures objectives d'une salle, en replaçant la perception spatiale en tant que phénomène non pas absolu, mais relatif au contexte et au message sonore.

2 LOCALISATION ANÉCHOÏQUE DE L'ÉVÉNEMENT AUDITIF

La localisation, dépend étroitement de l'espace sonore (réel ou virtuel), des caractéristiques géométriques des sources, et du son qui est diffusé. Cependant, dans un souci de simplification préliminaire, on se limite ici à la localisation d'une source physique dans un contexte anéchoïque, ce qui écarte le cas de la stéréophonie ou celui des espaces réverbérants.

2.1 Caractéristiques de la perception de la direction

Cette section vise à présenter quelques résultats sur les propriétés de la localisation, sans vocation d'exhaustivité. Ces résultats concernent des sons très simples comme des bruits de largeurs de bande diverses ou des sons purs, et la plupart du temps dans un contexte de laboratoire (écoute dichotique au casque ou naturelle en chambre anéchoïque) pour simuler au mieux les conditions de champ libre. La généralisation à des situations plus réalistes est donc à faire avec prudence.

Deux caractéristiques usuelles permettant de juger de la faculté de localisation sont les biais et flou de localisation. Le **biais de localisation** est la différence entre la direction de la source sonore et la direction moyenne de l'événement auditif. Le **flou de localisation** (*localisation blur*) ou angle minimum audible (*minimum audible angle*) est défini comme le seuil différentiel de perception angulaire, c'est-à-dire la déviation angulaire pour laquelle 50% des sujets perçoivent un déplacement de l'événement auditif (Mills, 1958; Blauert, 1997, sections 1.3.1 et 2.1).

Localisation dans le plan horizontal

Les performances de la localisation dans le plan horizontal sont, pour résumer, maximales à l'avant, et minimales sur les côtés. Ainsi, pour des séquences de bruit blanc de 100ms (Blauert, 1997, section 2.1), le biais de localisation est nul en face, et de l'ordre de 10° sur le côté (une source sonore à 80° est perçue à 90°). Pour les mêmes stimuli, le flou de localisation varie de $\pm 3{,}5°$ en face à $\pm 10°$ sur les côtés. Pour des sons purs ou des bruits à bande étroite, on observe la même dépendance azimutale, avec de plus une forte dépendance à la fréquence centrale : par exemple (Mills, 1958) pour une source frontale, le flou est minimal ($\pm 1°$) dans les médiums (autour de 800 Hz), faible ($\pm 2°$) entre 3000 Hz et 6000 Hz, et maximal ($\pm 3°$) entre 1500 et 3000 Hz, ainsi qu'entre 6000 et 10000 Hz ; pour une source latérale (à 75°), il est de toute façon important (de l'ordre de $\pm 10°$ au minimum), mais diverge entre 1000 Hz et 3000 Hz, et entre 5000 Hz et 10000 Hz, la localisation étant impossible. Cette perte de la faculté de localisation des sons purs hautes fréquences trouve un corollaire dans la localisation des bruits haute fréquence à bande étroite. Ainsi, les résultats de Musicant et Butler (1985) indiquent que la localisation de tels signaux ne dépend pas de la position de la source physique, mais uniquement de la fréquence centrale.

Aucun résultat sur la localisation n'a pu être trouvé sur les basses fréquences (en dessous de 250 Hz), qui rende compte entre autres de la perte de la précision de localisation dans les sub-basses (en dessous de 100 Hz). En particulier, il est difficile de savoir si cet effet existe également en contexte anéchoïque, ne serait-ce que parce qu'il est rare de trouver des salles qui soient anéchoïques à des fréquences aussi basses. Cependant, il est probable que ce soit le cas, et que la nature même des basses fréquences en soient à l'origine : en effet, la différence de phase maximale (c'est-à-dire celle correspondant à un retard interaural de

l'ordre de 0,7 ms) diminue proportionnellement à la fréquence, et il est vraisemblable que le système auditif soit limité par sa résolution en très basses fréquences.

Localisation dans le plan médian

La localisation dans le plan médian est beaucoup plus sensible à la nature de la source, et de manière générale plus imprécise que dans le plan horizontal. Lorsque la source est une voix continue et familière, le biais est nul en face, et maximal (de l'ordre de 15°) lorsque la source est à la verticale. Le flou de localisation augmente lui aussi avec l'élévation. Dans la direction frontale, il est estimé à 9° pour une voix continue et familière par Damaske et Wagener, à 17° pour une voix continue et non familière par Blauert, et à 4° pour du bruit blanc par Wettschurek. La localisation d'un bruit à bande étroite dans le plan médian ne dépend plus de la position de la source, mais uniquement de la fréquence centrale [5].

Ainsi, la précision de la localisation, et son adéquation avec la position de la source physique, varie en fonction de la direction : elle est maximale en face, puis diminue au fur et à mesure que l'on s'approche des côtés ou du dessus.

Confusions avant-arrière

En l'absence d'indices visuels et si la tête est immobile, de nombreuses études ont pu mettre en évidence des confusions avant-arrière avec des sons purs ou bien des bruits filtrés en bande étroite (Rayleigh, 1876, 1907; Stevens et Newman, 1936).

Rôle des attaques

On peut se demander si la saillance de l'attaque a une influence sur la précision de la localisation : la localisation est-elle meilleure dans un contexte anéchoïque si le temps de montée d'un son est court ? Une expérience de Perrot (1969), étudiant l'influence de la durée de montée sur le seuil différentiel de perception angulaire, semble indiquer que non, alors qu'une autre menée par Rakerd et Hartmann (1986), utilisant une autre méthode de test, indique un effet significatif, mais faible, de la durée de l'attaque. Deux études successives de Rakerd et Hartmann (1985; 1986) montrent que la localisation est toujours possible en contexte anéchoïque pour des sons purs très longs, et pour ainsi dire sans attaque, puisque le temps de montée est de 7 secondes. Néanmoins, ces deux études sont menées sur des sinusoïdes pures, et pour un nombre très limité de fréquences, et il est difficile de généraliser ces résultats à tous types de signaux.

Quoi qu'il en soit, la présence d'une attaque n'est pas nécessaire à la localisation dans un contexte anéchoïque. En revanche, les attaques ont un rôle primordial en milieu bruité ou réverbérant, car elles fournissent des indices interauraux non ambigus.

2.2 Indices binauraux de perception de la direction

La localisation binaurale tire parti des dissemblances entre les sons perçus par les deux oreilles, qui sont dues principalement au rôle d'obstacle joué par la tête, qui oblige les ondes à parcourir un chemin supplémentaire pour atteindre l'oreille contralatérale.

Différences interaurales d'intensité

Historiquement, la première explication à notre faculté de perception de la direction de l'événement sonore est avancée par Lord Rayleigh (1876), qui souligne le rôle de la perception de la différence d'intensité entre les deux oreilles due à l'effet d'ombre de la tête. Cependant, il conclut l'article en indiquant que cette explication est incomplète, car cette différence est très faible pour des longueurs d'onde grandes par rapport aux dimensions de la tête (ce qui

[5] Tous ces résultats sont tirés de l'ouvrage "Spatial Hearing" de Jens Blauert (1997, section 2.1)

correspond à des fréquences inférieures à environ 500 Hz), alors que la localisation est toujours possible. Une modélisation sphérique de la tête, ainsi que les mesures (Feddersen et al., 1957), viennent confirmer ce constat.

Différences interaurales de phase

Rayleigh introduit plus tard la notion de différences interaurales de phase (Rayleigh, 1907) et indique leur rôle dans la localisation des sons en basses fréquences (inférieures à 1.6 kHz), hypothèse confirmée ultérieurement par de nombreux travaux (Stevens et Newman, 1936; Mills, 1958). En effet, d'une part le système auditif est incapable de détecter des différences de phase à hautes fréquences[6], et d'autre part, lorsque la longueur d'onde du signal est inférieure ou égale à la différence de marche maximale (c'est-à-dire pour les fréquences supérieures ou égales à environ 1.5kHz), il y a une ambiguïté intrinsèque car une différence de phase donnée correspond à deux ou plusieurs azimuts possibles (Stevens et Newman, 1936). Il existe une autre ambiguïté, valable elle à toutes les fréquences, sur le calcul de la différence interaurale de phase (Blauert, 1997, section 2.4.1) : suivant que l'on choisisse l'onde gauche ou droite comme référence, on obtient deux positions différentes. Cette ambiguïté est partiellement résolue par le système auditif qui privilégie la direction la plus frontale, mais le problème subsiste lorsque la différence de phase avoisine les 180°, ce qui n'est possible que pour des fréquences supérieures à environ 500 Hz. En effet, plusieurs expériences, citées par Blauert (1997, section 2.4.1), attestent l'existence d'ambiguïtés de localisation pour des signaux dichotiques à bande étroite, ambiguïté augmentant avec la fréquence centrale, et dépendant du type de signal : elle est importante dès 600 Hz pour des sinusoïdes, alors qu'elle reste faible jusqu'à quelques 1200 Hz pour des sinusoïdes modulées en fréquence. Cette ambiguïté peut être partiellement résolue grâce aux mouvements de tête.

Retards interauraux d'enveloppe

Cependant, une localisation intracrânienne est possible pour des signaux courts de fréquence de porteuse supérieure à 1.6 kHz, et soumis uniquement à un retard (pas de différences de niveau). Ceci amène à la notion de retard interaural d'enveloppe (Leakey et al., 1958; Henning, 1974) : à ces fréquences, il peut y avoir fusion (et donc latéralisation de l'événement auditif) de deux stimuli présentés de manière dichotique, ayant pour porteuses des bruits filtrés passe-haut et décorrélés (Schubert et Wernick, 1969).

Il semble néanmoins que le rôle des retards interauraux d'enveloppe en hautes fréquences soit faible par rapport à celui des différences d'intensité, et surtout des différences de phase en basses fréquences (Trahiotis et Bernstein, 1986; Macpherson et Middlebrooks, 2002).

Théorie unifiée de la localisation binaurale

La théorie unifiée de la localisation binaurale (issue entre autres des travaux de Rayleigh) par estimation des différences interaurales de phase ou d'intensité selon le domaine fréquentiel est appelée théorie "duplex" de la localisation. Elle fait autorité en la matière depuis plus d'un siècle, et a subi peu de modifications, le principal ajout étant dû à la découverte de notre sensibilité aux retards interauraux d'enveloppe. Le point de vue actuel est donc que la localisation binaurale repose sur ces deux types d'indices, les différences d'intensité et les différences de temps (retards de phase en basses fréquences, et retards d'enveloppe ou de groupe en hautes fréquences). Le poids respectif de ces indices, qui est typiquement estimé à partir d'expériences de localisation varie en fonction de la fréquence Wightman et Kistler (1992); Macpherson et Middlebrooks (2002) :

– les différences interaurales d'intensité (*interaural level differences*, ou ILD) sont prises en compte à toutes les fréquences, mais leur importance relative diminue en basses fréquences (en dessous de 2kHz).

[6] Il y a deux raisons à cela : d'une part, les cellules ciliées internes ne codent plus la structure temporelle fine au delà de 1.5 kHz environ, et codent uniquement l'enveloppe ; d'autre part, les neurones ne sont pas capables de se caler sur la phase de signaux à ces fréquences.

– les retards de phase (*interaural phase delays*, ou IPD) ne sont perceptibles (et donc pris en compte) que pour des fréquences inférieures à environ 1,6 kHz.
– les retards d'enveloppe (*interaural enveloppe delays*, ou IED), appelés également par abus de langage retards de groupe (*interaural group delays*, ou IGD[7]), sont pris en compte pour des signaux de fréquences supérieures ou égales à 1,6 kHz, mais leur poids vis-à-vis de la localisation est moindre par rapport aux autres indices, et varie en fonction des individus.

Intégration fréquentielle

L'application de ces résultats à des sons plus complexes nécessite de s'intéresser à la manière dont le système auditif traite des stimuli à large bande : sont-ils analysés tels quels, ou bien décomposés en signaux à bande étroite traités séparément ? La seconde hypothèse semble la plus plausible, au vu des résultats d'expériences psychoacoustiques, faisant appel le plus souvent à une écoute dichotique.

Ainsi, l'expérience de Toole et Sayers (1965) consistait à étudier la latéralisation d'un stimulus dichotique (train d'impulsions) soumis à un retard interaural important (supérieur à 2ms) et variable. Il s'avère que ce stimulus est subdivisé en plusieurs événements auditifs latéralisés à différentes positions, chaque événement auditif étant lié à une ou plusieurs harmoniques du signal. La position de chaque événement auditif est conforme à ce que l'on pouvait prédire en fonction de la différence de phase à la fréquence correspondante. Les signaux sont soumis à des retards interauraux non naturels, si bien que l'expérience ne correspond pas à une situation réaliste, mais elle met en évidence un traitement séparé des différentes composantes harmoniques. De même, Perrot et al (1970), étudiant la latéralisation de sons purs dichotiques de fréquences différentes à gauche et à droite, montrent que les deux signaux ne sont pas fusionnés en un seul événement auditif si leurs fréquences sont trop différentes. Ceci pourrait indiquer que l'analyse est effectuée en bande étroite, les deux signaux ne fusionnant pas lorsque leurs bandes critiques ne se recouvrent pas.

Il semble que la durée de l'attaque joue un grand rôle dans la fusion d'événements auditifs (groupement simultané). Ainsi, la fusion de signaux dichotiques basée sur un simple retard d'enveloppe dépend étroitement de la rapidité de l'attaque de cette enveloppe, et de sa durée (Perrot et al., 1970).

La fonction de transfert interaurale (quotient des HRTF ipsilatérale et contralatérale) contient la totalité des indices interauraux pour une position donnée de la source. Cependant, le gain et la phase de ces fonctions de transfert varie beaucoup en fonction de la fréquence, et l'on peut se demander si l'audition est sensible à tous ces détails. Plusieurs expériences, dont celles menées récemment par Macpherson et Middlebrooks (2002) semblent indiquer que ce n'est pas le cas, et que la localisation binaurale travaille par intégration de différences de niveau et de temps bande critique par bande critique, plutôt que par une analyse en bande étroite des spectres interauraux de phase et d'intensité.

Physiologie de la localisation binaurale

Les hypothèses mentionnées précédemment sur l'intégration fréquentielle sont largement confirmées par nos connaissances actuelles sur la physiologie du système auditif : l'analyse tonopique effectuée par la cochlée, qui vise à effectuer un traitement séparé des fréquences, est effective à tous les niveaux de l'audition, des noyaux du tronc cérébral jusqu'au cortex auditif. De plus, cette analyse a une résolution fréquentielle limitée par celle de la cochlée, ce qui appuie l'hypothèse selon laquelle tous les détails fins du spectre interaural ne sont pas nécessaires à la localisation binaurale.

Quant au traitement des différences interaurales effectué par les neurones, plusieurs théories ont été formulées. Les mécanismes physiologiques de la localisation pourraient être ainsi basés :

[7] les retards interauraux de phase et de groupe sont rassemblés sous le terme de "différences interaurales de temps" (*interaural time differences*, ou ITD)

- Sur des différences de temps uniquement (sachant que dans ce modèle, les différences de niveau sont transformées en différences de temps)
- Sur des différences de niveau uniquement (sachant que dans ce modèle, les différences de temps sont transformées en différences de niveau)
- Sur des différences de temps et de niveau, traitées séparément

Au vu de découvertes plus récentes sur la physiologie, la dernière hypothèse paraît la plus réaliste (Blauert, 1997). Néanmoins, les deux types de traitement sont physiologiquement plausibles.

2.3 Indices monauraux de perception de la direction

Le phénomène de localisation monaurale dans le plan horizontal a pu être également mis en évidence dès le 19e siècle par Rayleigh (1882) sur un sujet sourd d'une oreille, qui pouvait localiser non sans erreurs des voix, mais était incapable de localiser des sons purs.

Si l'on considère les informations ne venant que d'une seule oreille, le seul indice permettant de juger de la localisation est le filtrage subi par l'onde au travers des multiples réflexions sur le crâne, le torse, et surtout le pavillon (Gardner et Gardner, 1973; Gardner, 1973; Roffler et Butler, 1968a), ce qui nécessite que le signal contienne des hautes fréquences (vues les dimensions du pavillon). L'utilisation de cette information implique deux conditions : d'une part, il est nécessaire que la source soit familière à l'auditeur, pour qu'il soit capable d'en évaluer les modifications spectrales dues à ces réflexions ; d'autre part, il faut que sa bande passante soit suffisamment large, car en bande étroite, il devient impossible de distinguer un filtrage dépendant de la fréquence d'un simple gain.

En effet, plusieurs expériences psychophysiques confirment que la localisation monaurale de l'événement auditif ne correspond à la direction de la source que si les stimuli sont à large bande, et s'ils contiennent des hautes fréquences, c'est-à-dire au dessus de 4 kHz (Roffler et Butler, 1968a). Dans ce cas, le flou de localisation diminue au fur et à mesure que la source devient familière (Blauert, 1997). En bande étroite, la localisation dans le plan médian ne dépend plus de la position effective de la source, mais uniquement de la fréquence centrale de sa bande passante (Roffler et Butler, 1968b; Blauert, 1997, section 2.1).

2.4 Importance relative des indices monauraux et binauraux dans la perception de la direction

La localisation monaurale joue un grand rôle dans la localisation, car elle permet à une personne ayant une audition normale d'aider à résoudre nombre des confusions avant-arrière de l'écoute binaurale, notamment dans le plan médian où les différences interaurales sont quasi-nulles[8]. De manière plus générale, elle permet de lever l'ambiguïté de la localisation au sein d'un même **cône de confusion**. Les cônes de confusion sont les surfaces pour lesquelles les différences interaurales de temps et d'intensité sont constantes. En assimilant en première approximation la tête à une sphère rigide et en se plaçant en champ lointain, les cônes de confusion sont de véritables cônes (d'où leur nom) ayant pour axe commun l'axe interaural. Mais en réalité, la présence de dissymétries ne permet plus de définir des surfaces où les indices interauraux sont rigoureusement constants à toutes les fréquences. Cela dit, en acceptant une certaine tolérance de variation, il est toujours possible de définir des zones spatiales pour lesquels les indices interauraux sont peu variables.

On peut se demander quel est la contribution respective des indices binauraux (de temps et d'intensité) et monauraux (spectraux) sur la localisation dans un même cône de confusion. Au sujet des indices monauraux, la plupart des recherches s'accordent sur le fait que les informations spectrales (monaurales) présentes au niveau de l'oreille ipsilatérale sont les indices primordiaux pour la détermination du **site** (appelé également **élévation**) de l'événement auditif, et donc la résolution des confusions avant/arrière. Quant à l'oreille contralatérale, des expériences perceptives menées par Morimoto (2001) semblent indiquer que l'influence

[8] Certains chercheurs soutiennent néanmoins que les dissymétries de la tête, et notamment du pavillon, sont des indices utiles pour une localisation dans le plan médian selon des critères interauraux. (Searle et al., 1976)

des indices spectraux de l'oreille contralatérale sur la perception de l'élévation diminue au fur et à mesure que la source s'éloigne du plan médian. En revanche, les indices interauraux apportent rarement une information suffisamment fiable pour l'estimation de l'élévation.

La participation d'indices monauraux à la localisation en azimut chez les gens ayant une audition normale, à l'instar des personnes sourdes d'une oreille, n'est pas encore clairement prouvée, sachant que certaines recherches soutiennent cette hypothèse (Musicant et Butler, 1985), alors que d'autres la récusent (Macpherson et Middlebrooks, 2002). Le système auditif est de toute façon capable de localisation azimutale avec un oreille sourde, mais le flou de localisation est bien plus important qu'en écoute binaurale. Le fait que ce flou de localisation diminue avec l'expérience semble indiquer que ce mécanisme est au moins partiellement inhibé en écoute binaurale.

2.5 Distance de l'événement auditif en contexte anéchoïque

La perception anéchoïque de la distance dépend de trois causes physiques, répertoriées par Pellegrini (2001) :

D'une part, l'intensité physique varie avec la distance, selon une loi dépendant étroitement de la géométrie de la source. Si celle-ci est suffisamment lointaine, le champ sonore obéît à l'approximation en ondes sphériques dont la courbure locale (c'est-à-dire comparée à la distance entre les deux oreilles) peut être négligée, si bien que l'intensité baisse de 6 dB pour chaque doublement de distance (loi en 1/r).

En champ proche (moins de quelques mètres), cette approximation n'est plus valable, et la courbure des ondes entraîne une modification du spectre dépendant très sensiblement de la distance. Notamment, les basses fréquences peuvent être amplifiées en vertu de l'effet de proximité pour une source s'apparentant à une source dipolaire ; de plus, les différences interaurales de temps et d'intensité dépendent elles aussi, de fait, de la distance. Ce phénomène de parallaxe acoustique est maximal lorsque la source est dans l'axe interaural, et nul dans le plan médian.

Pour des grandes distances (supérieures à 15 m), les distorsions linéaires dues au trajet aérien doivent être prises en compte, et le timbre de la source en est altéré, notamment en hautes fréquences qui sont les plus atténuées avec la distance.

Le seul indice binaural n'étant effectif qu'en champ proche, il est impossible de juger efficacement de la distance d'une source non familière dans un contexte anéchoïque (Coleman, 1962). Ainsi, pour une source non familière, il est impossible de savoir si une variation de sonie est due à une variation d'intensité de la source physique ou bien à une modification de sa distance, alors qu'une connaissance *a priori* de la source nous donne un indice supplémentaire basé sur son timbre : la distorsion hautes fréquences caractéristique des cuivres dans les *forte* ("cuivrage"), ou bien la distinction entre de la voix chuchotée, parlée ou criée, nous permettent de juger de l'intensité de la source physique. La perception anéchoïque de la distance en champ lointain est donc **relative** au contexte.

3 INFLUENCE DE LA SALLE SUR LA PERCEPTION DE L'ÉVÉNEMENT AUDITIF

Tout environnement acoustique, et en particulier une salle, a une très forte influence sur la perception spatiale de la source. Les réflexions successives dues à la propagation dans une salle ne modifient généralement pas la direction moyenne de l'événement auditif, mais ont tendance à l'élargir et donc à rendre sa localisation plus imprécise (section 3.2). En revanche, la perception anéchoïque de la distance est complétée, voire supplantée par un autre mécanisme auditif dépendant des réflexions précoces et du niveau de la réverbération tardive (section 3.3).

Cette section et la suivante traitent de l'influence de l'effet de salle respectivement sur la perception de la source et celle de la salle. Pour faire la distinction entre ces deux percepts, il est nécessaire avant tout de rappeler le principe de l'**effet de précédence** (appelé également **effet d'antériorité**) et son rôle vis-à-vis de la séparation subjective entre source et salle (section 3.1).

3.1 L'effet de précédence

Le rôle premier de l'effet de précédence semble être de faciliter la perception d'une source et de sa localisation en milieu réverbérant, en inhibant au mieux la perception des réflexions au moyen d'un mécanisme de fusion temporelle et spatiale. Il s'agit d'un phénomène complexe de **masquage** mettant en jeu des mécanismes monauraux et binauraux, et dont le comportement dépend très largement du type de son diffusé (enveloppe temps-fréquence notamment) ainsi que des paramètres physiques des réflexions, soit leurs gains et retards relativement à l'onde directe. De fait, la plupart des nombreuses études sur le sujet depuis sa découverte sont basées sur le cadre simplifié d'un son direct avec une unique réflexion, en chambre anéchoïque ou bien en restitution sur casque (simulation binaurale, ou positionnement des sources par ITD et ILD).

Le terme "effet de précédence" désigne en fait plusieurs phénomènes perceptifs distincts, mais liés par le fait qu'ils se réfèrent à la manière dont le système auditif traite un ou plusieurs échos. Par "écho", on entend une source secondaire (par exemple une réflexion dans une salle) retardée par rapport à la source principale, dite primaire, et fortement ou totalement corrélée à celle-ci. Le type de phénomène perçu dépend principalement du retard et du niveau relatif de la source secondaire.

Localisation par sommation

Lorsque le retard est inférieur à un seuil de l'ordre de 1 ms[9], la source principale et la source secondaire créent un unique événement auditif, appelé "source fantôme", dont la localisation dépend de leurs deux positions : pour un retard nul et à niveau égal, l'événement auditif se situe généralement à mi-chemin des deux sources. Sinon, il se déplace vers la source la plus forte et/ou la plus précoce. Cet effet de "localisation par sommation" (*summing localization*), qui est à la base de l'écoute stéréophonique bicanale, a ainsi été longuement étudié pour des sources frontales placées symétriquement par rapport au plan médian, mais il est généralisable à de nombreuses configurations de sources, avec toutefois moins d'efficacité : par exemple, deux sources physiques écartées de 60°, qui est l'écart normalisé en écoute stéréophonique bicanale, mais situées horizontalement et sur le coté (de part et d'autre du plan interaural), créent bien une source fantôme si l'on tente d'y placer une source monophonique au moyen d'un panoramique d'intensité, mais sa localisation est beaucoup plus sensible aux différences de niveau, si bien qu'elle a tendance à "sauter" d'un haut-parleur à l'autre même pour des faibles différences de niveau (Theile et Plenge, 1977). Ceci explique la difficulté de positionner une source sur les côtés dans le cadre d'une écoute stéréophonie multicanale, comme en tétraphonie ou en 5.1.

Blauert (1997, section 4.4.2) fait mention d'un phénomène intéressant : le timbre de la source fantôme n'est quasiment pas altéré par une modification du retard de la source secondaire, contrairement à ce que l'on pourrait supposer étant donnée la présence de filtrages en peigne au niveau des oreilles. En revanche, la coloration est audible en écoute monaurale, ce qui laisse supposer l'existence d'un mécanisme de compensation binaurale de tels effets de timbre, à rapprocher aux principes de détection binaurale ("cocktail-party") et de suppression binaurale de réverbération : le système auditif est capable de compenser ou de tenir compte de certaines différences binaurales pour améliorer la perception de la source.

Loi du premier front d'onde

Dès lors que le retard de la source secondaire dépasse environ 1 ms, et en deçà d'un seuil dépendant de son niveau (la relation entre le niveau maximal et le retard est dite "seuil d'écho"), la localisation de l'événement auditif se confond avec la position de la source primaire, et la source secondaire n'est pas perçue distinctement. On parle de dominance de localisation (*localization dominance*), ou de "loi du premier front d'onde" (*law of the first wavefront*). Néanmoins, sa perception est altérée, et plus particulièrement sa sonie, son timbre

[9] En fait ce retard est variable en fonction du niveau du signal, et peut même atteindre plusieurs dizaines de millisecondes si le niveau relatif de la réflexion est suffisant Barron (1971).

et sa largeur apparente, pour peu que son niveau soit supérieur au seuil de détectabilité, dit "seuil masqué". L'augmentation de la sonie est l'effet le plus saillant, effet qui résulte de l'intégration de l'énergie de la source secondaire à l'événement auditif. Cet effet est utilisé en sonorisation pour renforcer le son sans le délocaliser au moyen de hauts-parleurs de renforcement en façade ou en salle. Contrairement au cas précédent, le **timbre** de la source est dans ce cas sensiblement altéré par la présence d'une réflexion, en particulier lorsque le retard avoisine les 20 ms (Barron, 1971). Cette modification du timbre, qui est due au filtrage en peigne provoqué par l'interaction entre les sources primaire et secondaire, est malgré tout toujours plus faible qu'en écoute monaurale. Le seul effet spatial qui soit perceptible est l'**élargissement de l'événement auditif**, qui correspond à une plus grande incertitude de localisation : l'événement auditif est toujours globalement situé au niveau de la source primaire, mais la source secondaire perturbe cette localisation qui perd en précision.

Perception de l'écho

Si le niveau de la source secondaire dépasse le seuil d'écho, un écho est perçu en tant qu'événement auditif distinct. La précision de localisation de cet écho est faible pour des retards courts, et s'améliore lorsque le retard augmente (Litovsky et al., 1999). De plus, et à l'image de la précision de localisation du son direct, elle dépend de l'azimut, et est ainsi sensiblement meilleure à l'avant que sur les côtés (Litovsky et Macmillan, 1994). Le seuil d'écho dépend largement du type de signal (par exemple, plus le signal est court, plus le seuil est faible), de l'angle entre les sources précoce et secondaire, ainsi que du niveau relatif de la réflexion (plus le niveau relatif est faible, plus le seuil est important). De plus, la notion même de seuil d'écho est relativement floue et subjective, et les données issues d'expériences psychophysiques varient ainsi grandement suivant le critère imposé à l'auditeur ("écho à peine audible", "un son est audible au niveau du haut-parleur secondaire", "l'écho et le son direct ont la même sonie", "écho gênant", etc...) (Blauert, 1997; Litovsky et al., 1999).

Inhibition du premier son

Lorsque le niveau de la source secondaire est suffisamment fort par rapport à celui de la source primaire (niveau relatif de l'ordre de 20 dB à 40 dB), un phénomène inverse à la loi du premier front d'onde peut parfois se produire : la source primaire n'est plus audible, et le seul événement auditif est localisé au niveau de la source secondaire. Cet effet de masquage rétroactif est appelé "inhibition du premier son" (*inhibition of the primary sound*) . Un exemple concret est l'inhibition du son direct dans le cadre d'un champ fortement réverbérant (Blauert, 1997, section 3.3).

Causes physiologiques de l'effet de précédence

Si l'on présente le premier son à une oreille uniquement, et l'écho à l'autre oreille uniquement (présentation dichotique), la loi du premier front d'onde est toujours valable (Blauert, 1997). Ceci semble indiquer qu'un mécanisme d'inhibition contralatérale en est au moins partiellement responsable. De fait, l'effet de précédence est généralement considéré comme un phénomène binaural. Cependant, cette hypothèse est incomplète, et ne permet pas par exemple d'expliquer l'effet de précédence dans le cas où l'onde directe et/ou l'onde réfléchie sont dans le plan médian. Plusieurs études ont ainsi mis en évidence l'existence d'un phénomène monaural de précédence (Blauert, 1971; Rakerd et al., 1997; Litovsky et al., 1997), qui correspond selon toute logique au phénomène bien connu de masquage temporel. Cet effet, qui serait dû à un mécanisme d'inhibition ipsilatérale, présente des valeurs de seuil du même ordre que celles de son équivalent binaural.

Généralisation à plus de deux sources

Ces remarques sur l'effet de précédence peuvent globalement être généralisées au cas de plusieurs échos, et c'est cette généralisation qui permet d'appréhender la perception d'une

salle de concert.

Peu d'études ont été menées sur la localisation par sommation dans le cas de plusieurs échos, mais les résultats de quelques expériences (Blauert, 1997) inclinent à penser que le phénomène est identique à celui observé pour un écho unique, c'est-à-dire que pour des retards faibles entre plusieurs sources, une seule source est perçue, dont la position dépend de celles des sources physiques. Il est probable qu'un éventuel déplacement de l'événement auditif soit principalement conditionné par les premières sources images, et que le "pouvoir délocalisant" d'une source s'amenuise plus celle-ci est tardive.

Concernant la loi du premier front d'onde, le principe est identique à celui enoncé pour une seule réflexion, la différence primordiale étant la modification des seuils : lorsqu'un ou plusieurs échos sont présents entre le son direct et l'écho cible, tous les seuils relatifs à l'effet de précédence (seuil masqué, seuil d'écho, seuil d'égale sonie...) sont repoussés plus tardivement (Barron et Marshall, 1981; Litovsky et al., 1999). Cet effet est appelé "accumulation de la précédence" (*buildup of precedence*) (Litovsky et al., 1999).

Pour finir, il a déjà été mentionné que l'inhibition du son direct dans le cas d'un milieu très réverbérant peut être considérée comme une généralisation de l' "inhibition du premier son".

Malgré la similarité des effets observés, la modification des seuils laisse penser que les données expérimentales obtenues pour un écho unique quant à la position et à la largeur de l'événement auditif en fonction du retard et du niveau relatif peuvent difficilement être utilisées à titre prédictif dans le cas d'échos multiples.

3.2 Élargissement de l'événement auditif

L'élargissement dû à la distribution précoce de l'énergie

Les expériences psychophysiques sur l'effet de précédence ont mis en évidence un phénomène d'élargissement de l'événement auditif lorsqu'une réflexion est perceptible (par rapport au cas de la source primaire présentée seule) mais non audible distinctement. Plusieurs études se sont intéressées plus spécifiquement à cet aspect, et notamment à sa manifestation dans le cas de la propagation en espace clos. A la suite de l'hypothèse de Marshall selon laquelle les réflexions latérales précoces avaient une importance subjective sur la perception des salles de concert, Barron (1971) étudie l'effet d'une réflexion latérale unique sur ce qu'il appelle l' "impression spatiale", et qui sera défini plus tard comme la **largeur apparente de la source** (*apparent source width, auditory source width*, ou encore ASW). Ces résultats sont synthétisés par Barron en un schéma reproduit en figure II.3. Dix ans plus tard, Barron et Marshall publient ensemble (Barron et Marshall, 1981) des résultats plus complets basés sur une simulation simple de salle avec deux réflexions latérales symétriques et une réverbération synthétique. Ces travaux ont montré que cet effet d'élargissement de l'événement auditif dépendait principalement de la direction d'incidence et du niveau relatif des réflexions précoces, ainsi que du niveau global du son et de son spectre. De plus, les réflexions provenant d'angles supérieurs à 90° (entre 90° et 180°) "n'étaient pas perçues comme provenant de l'arrière, mais produisaient juste une sensation agréable d'enveloppement" (sous-entendu en plus de l'effet d'élargissement).

Ainsi, ces résultats indiquent que les réflexions venant du plan médian (en face, au dessus ou derrière) provoquent des changements de sonie ou de coloration, mais ne créent pas de sensation d'élargissement, alors que plus les réflexions sont proches de l'axe interaural, plus cette sensation est forte. En revanche, il semble que le retard d'une réflexion unique n'ait que très peu d'incidence sur l'impression d'espace (Barron, 1971), si cette réflexion se situe en dessous du seuil d'écho, ce qui correspond à un retard maximal de l'ordre de 80 ms. La même expérience avec deux réflexions placées à $\pm 40°$ (Barron et Marshall, 1981) conduit à une conclusion identique.

Plusieurs expériences indiquent que l'impression d'espace dépend également du niveau sonore global lorsque le signal est large bande, si bien que la largeur apparente ne peut être considérée comme un phénomène linéaire. Ainsi, Cremer (1989) souligne la différence qualitative entre l'impression spatiale créée par des sons faibles et forts : "J'ai observé dans

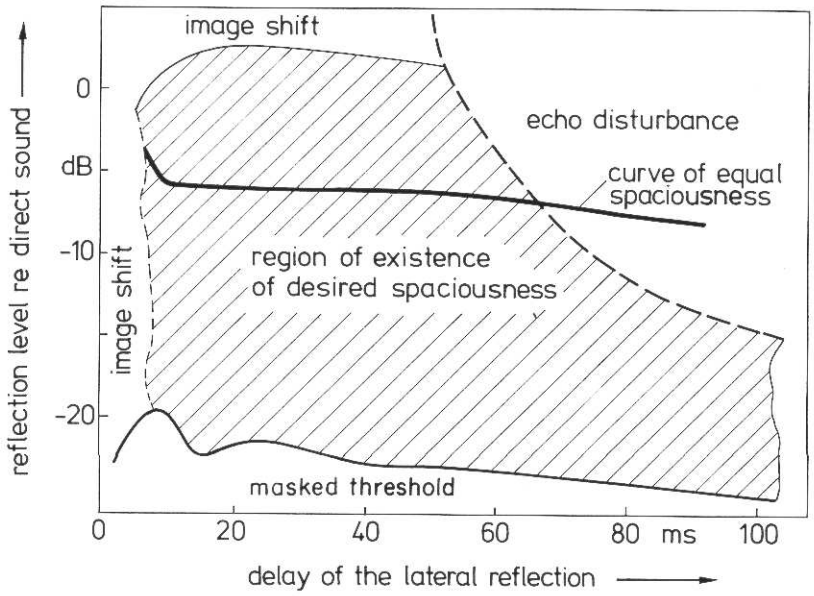


FIG. II.3 – **Etude par Barron (1971) de l'effet d'une réflexion latérale (azimut de 40°) sur l'impression spatiale** en fonction de son délai et de son niveau relatif. On note le déplacement de l'image auditive pour des faibles délais, quel que soit le niveau, les seuil masqué, le seuil d'écho, et la "région d'existence d'une impression d'espace adéquate", qui correspond à la zone pour laquelle on note un élargissement de l'événement auditif (repris de Blauert (1997)).

la salle de la Philarmonie de Berlin que, d'une part, lorsque la source joue *piano*, elle semble concentrée sur la scène ; la direction apparente du son coïncide avec l'extension spatiale des sources physiques. Mais en *fortissimo*, nous prenons conscience de la salle à cause du son réfléchi par les murs et le plafond : pour ainsi dire, les murs eux-même semblent être en train de jouer.". Deux propositions d'explications ont été avancées pour expliquer ce phénomène : Cremer (1989) en attribue la responsabilité à la non-linéarité de la courbe (logarithmique) liant l'intensité sonore à la sonie, alors que Wettschurek (cité par Potter (1993, section 2.5)), ainsi que Blauert (1997) se basent sur la dépendance fréquentielle des courbes d'isosonie, et sur le fait que les basses fréquences ont une plus forte vertu spatialisante que les autres fréquences du spectre audible (voir paragraphe suivant). Cette influence prépondérante que l'intensité sonore exerce en espace clos à la fois sur la sonie, sur la largeur apparente de la source, et sur la présence de la salle, permet au musicien de renforcer considérablement les nuances de son jeu, et spécialement lorsque celui-ci est habitué à la salle.

Rôle des basses fréquences

Un effet qui n'a pas encore été mentionné est le rôle primordial du contenu spectral, et en particulier des basses fréquences, sur l'impression d'espace. Beaucoup d'auteurs soutiennent cette hypothèse (Blauert, 1997; Blauert et Lindemann, 1986a; Barron et Marshall, 1981; Morimoto et Maekawa, 1988). En fait, celles-ci semblent avoir une telle influence sur l'extension spatiale de la source que la source est complètement délocalisée lorsqu'elles sont présentées seules, si bien qu'elles suscitent ainsi une sensation s'apparentant plus à de l'enveloppement qu'à un élargissement.

Ainsi, Barron et Marshall (1981) indiquent que les basses fréquences « donnent une impression subjective d'être entouré (*surrounded*) par le son, une sensation que l'on peut désigner par le mot "enveloppement" ». Cette remarque est à lier aux résultats de Blauert et

Lindemann (1986a) qui distinguent les effets perceptifs des hautes et basses fréquences : selon eux, alors que les composantes hautes fréquences des réflexions latérales sont responsables de la sensation de largeur (extension angulaire de l'événement auditif), les composantes basses fréquences provoquent une sensation de profondeur (*"front-back extension"*), et donc d'enveloppement de l'auditeur. Cela dit, l'effet d'élargissement dû aux hautes fréquences, dès lors que celles-ci sont présentes dans la distribution précoce de l'énergie, prédomine sur l'effet d'enveloppement dû aux basses fréquences. Les résultats d'expériences menées par Morimoto et Maekawa (1988) semblent indiquer que même dans ce dernier cas, les basses fréquences ont un effet très significatif sur l'élargissement apparent de la source. Cet effet, qui est à mettre en relation avec la perte de précision de localisation dans les très basses fréquences, semble indiquer que la présence de réflexions précoces étend cette zone d'incertitude jusqu'à quelques centaines de Hertz, si l'on se fie à ces résultats, ainsi qu'à ceux issus de tests informels de Barron et Marshall (1981).

Néanmoins, cette hypothèse est remise en question par Potter (1993, chapitre 4), qui suggère de considérer non seulement le pouvoir spatialisant de chaque bande de fréquences, mais également sa saillance : étudiant le seuil de détectabilité d'un changement de niveau ou de corrélation interaurale sur une bande d'octave à fréquence centrale variable, il montre que celui-ci est maximal pour 500 Hz, et décroît en basses et hautes fréquences. Il en conclut que les basses fréquences ont un fort pouvoir spatialisant lorsqu'elles sont présentées seules, mais que celui-ci diminue si les signaux sont à large bande. La contradiction entre ces résultats et ceux des études précédemment mentionnées (Barron et Marshall, 1981; Blauert et Lindemann, 1986a; Morimoto et Maekawa, 1988) pourrait être expliquée par les différences de procédures expérimentales : en effet, d'une part les stimuli sont dichotiques, alors que les autres études utilisent toutes une simulation de salle basée sur un champ sonore en chambre anéchoïque ; d'autre part, le test de Potter est une expérience de discrimination, alors que les autres se basent sur un jugement absolu de l'impression spatiale.

Si l'on adhère au point de vue majoritaire selon lequel les basses fréquences ont un rôle primordial vis-à-vis de l'impression spatiale, on peut se demander de quelle nature est la sensation l'enveloppement suscité par les basses fréquences. Le commentaire de Barron et Marshall cité ci-dessus laisse à penser qu'elle s'apparente plus à un enveloppement de l'auditeur par l'événement auditif devenu extrêmement large qu'à un enveloppement par la réverbération. On retrouve ici la distinction proposée par Rumsey (2002) entre "enveloppement par la source" et "enveloppement par la salle". On pourrait objecter à cette distinction la remarque de Cramer citée ci-dessus ("...les murs eux-même semblent être en train de jouer."), mais tout laisse à croire que son impression se rapportait en fait à la perception des réflexions tardives, puisque la sensation d'enveloppement par la réverbération que celles-ci suscitent dépend elle aussi du niveau sonore global (voir section 4.2). L'hypothèse de Wettschurek et Blauert sur l'influence du niveau global sur la largeur apparente prend ici tout son sens : puisque le seuil d'audibilité des basses fréquences est relativement élevé comparé à celui des fréquences moyennes, leur contribution à l'élargissement de la source n'est significative qu'à partir d'un niveau global suffisant.

Influence de la distribution d'énergie tardive sur la perception de la largeur de l'événement auditif

Beaucoup d'expériences sur l'impression spatiale menées jusqu'à la fin des années 1980 a consisté à étudier l'influence de la distribution spatiale de l'énergie précoce uniquement sur l'élargissement de l'événement auditif. Barron (1971) s'est intéressé au rôle de la réverbération sur l'impression spatiale, mais de manière anecdotique et non systématique. En fait, Bradley et Soulodre (1995a), puis Bradley et al. (2000), ont mis en évidence un effet significatif de la proportion d'énergie tardive (plus de 80ms après le son direct) sur la perception de cette largeur : l'augmentation de cette proportion provoque non seulement une augmentation de la sensation d'enveloppement de l'auditeur par la salle (voir section 4.2), mais également une perte de sensibilité de la largeur de l'événement auditif à la fraction d'énergie latérale précoce. Il s'agit donc d'un phénomène de masquage spatial rétroactif.

L'élargissement vu comme une incertitude de localisation

L'idée, assez intuitive, selon laquelle l'élargissement de l'événement auditif correspond littéralement à une perte de précision de sa localisation, a curieusement peu été étudiée de manière systématique. En effet, alors qu'un nombre conséquent de recherches ont été menées sur la localisation et sa résolution en contexte anéchoïque, peu de travaux concernent la localisation dans les espaces clos. On peut néanmoins citer les recherches de Hartmann (1983), qui semblent indiquer que la localisation dans une salle est perturbée par les réflexions latérales précoces et au contraire renforcée par les réflexions frontales. En particulier, les sons purs stationnaires sont quasiment impossibles à localiser en présence de réverbération, et spécialement dans les basses fréquences. Plus récemment, un test de jugement de la qualité spatiale menée par Berg et Rumsey (2001) indique entre autres que la localisation et la largeur apparente sont anticorrélés. Le lien entre faculté de latéralisation (ou, pour généraliser, de localisation) et élargissement apparent de la source a également été mis en exergue par Potter (1993, section 6) : la chute de la cohérence interaurale, lorsque celle-ci reste a un niveau raisonnable, a un effet faible mais significatif sur l'angle minimal audible, et donc sur la précision de localisation. Lorsque la corrélation est vraiment très faible, la localisation devient très difficile, voire impossible.

Cette perte de précision de la localisation peut s'expliquer grâce à l'influence des réflexions précoces : même si l'effet de précédence résout une grande part de ces ambiguïtés, il ne permet pas de s'affranchir totalement de la perception de ces réflexions. Celles-ci sont donc sources d'incohérence vis-à-vis des indices interauraux de localisation, incohérence qui peut être de deux types (Blauert, 1997, section 4.5.1) :

- **Incohérence spectrale**
  Dans le cas de sons continus, la présence d'une réflexion, par effet de peigne, fait fluctuer les différences interaurales de temps et d'intensité en fonction de la fréquence. La localisation, qui se base sur l'intégration de ces indices sur la totalité des bandes fréquentielles utiles, pourrait être gênée par cette "incohérence spectrale", ce qui serait une cause de l'élargissement. Ceci explique entre autres pourquoi les sons purs sont beaucoup moins facilement localisables en espace clos qu'en champ libre (Rakerd et Hartmann, 1985, 1986).
- **Incohérence temporelle**
  Au cours du message sonore, l'arrivée successive des réflexions fait fluctuer les indices interauraux en fonction du temps. Si le système auditif était capable de suivre toutes ces changements, on percevrait une source ponctuelle se déplaçant constamment et rapidement (la position qu'aurait cet événement auditif est appelé "pseudo-angle" par Griesinger (1997)) ; en pratique, cette fluctuation est trop rapide pour l'audition, d'où l'élargissement de l'événement auditif.

L'influence de ces fluctuations sur la perception est intimement liée à la nature de la source. Si celle-ci est continue, la double incohérence due aux réflexions, associée à celle provoquée par la source si celle-ci est modulée en amplitude, ou *a fortiori* en fréquence, la délocalisent partiellement ou complètement. Blauert, puis Griesinger, se sont intéressés aux effets distincts provoqués par une fluctuation des différences de temps ou d'intensité : une fluctuation d'ILD uniquement donne la sensation d'une source ponctuelle dans un champ enveloppant, alors qu'une fluctuation d'ITD uniquement délocalise complètement la source, ne laissant qu'un champ enveloppant l'auditeur.

Si la source n'est plus continue mais intermittente, les réflexions précoces dégradent les indices interauraux relativement peu ambigus fournis par les attaques. En effet, si l'attaque du signal n'a pas d'influence sur la faculté de localisation en champ libre (voir la section 2), elle est primordiale en champ diffus (Rakerd et Hartmann, 1985, 1986). L'expérience quotidienne nous montre par exemple qu'un instrument possédant des attaques précises, comme une guitare ou une trompette, est plus facilement localisable qu'un instrument avec des attaques plus douces. L'hypothèse corollaire, soutenue notamment par Griesinger (1997), est que la largeur apparente d'une source dépend de sa nature.

L'effet d'élargissement se retrouve dans des expériences dichotiques dans lesquels les signaux gauche et droite sont des bruits partiellement incohérents (Blauert et Lindemann,

1986b; Potter, 1993; Kendall, 1995). Cet effet, qui est à la base de la technique de pseudo-stéréophonie, participe des mêmes causes que celui dû à la présence de réflexions latérales précoces en écoute naturelle. Ceci permet d'introduire la **corrélation interaurale précoce** (*early interaural cross corrélation*, ou $IACC_e$, ou encore $IACC_0^{80}$) comme mesure objective de la largeur apparente de la source. Cette mesure, qui peut être effectuée en pleine-bande ou sur des bandes d'octaves jugées particulièrement importantes (typiquement 0.5 -, 1.0 - et 2.0 kHz), est effectivement bien corrélée au jugement de largeur apparente (Barron, 1971). Néanmoins, elle n'est pas pertinente en basses fréquences, puisqu'alors qu'elle fournit systématiquement un résultat proche de 1 (ce qui correspondrait à un source jugée ponctuelle), la largeur apparente est au contraire jugée maximale (Morimoto et Maekawa, 1988; Potter, 1993; Griesinger, 1999).

Une autre mesure objective de l'impression spatiale, la **fraction d'énergie latérale** (notée LF ou LF_0^{80}), a été proposée ultérieurement par Barron et Marshall (1981). Elle généralise le lien mentionné précédemment entre le jugement d'impression spatiale due à une réflexion unique et la projection de l'énergie de cette réflexion sur l'axe interaural. Il y a beaucoup de similitudes de principe entre la fraction d'énergie latérale LF et l'incohérence (1-IACC) (Barron et Marshall, 1981; Griesinger, 1999), et Barron (1983) montre l'équivalence, sous certaines conditions, certes assez restrictives[10], entre ces deux mesures.

Ces deux mesures fixent la limite temporelle de l'intégration de l'énergie précoce à 80 ms. Cette limite provient d'une expérience avec réflexion unique et avec un type de signal particulier (Barron, 1971). Or on sait que la présence de plusieurs réflexions repousse tous les retards de l'effet de précédence ; de plus, il semble que le jugement d'impression spatiale dépende de la nature de la source. Cette limite est donc arbitraire, car elle ne tient pas compte du contexte. De même, aucune de ces deux mesures ne tient compte du niveau sonore, ni de la distribution tardive de l'énergie.

3.3 Perception de la distance de la source en milieu réverbérant

Il a été mentionné en section 2.5 que si l'on écarte le cas d'une source très proche, la perception de la distance en milieu anéchoïque n'est envisageable que si l'on dispose d'un minimum de connaissances sur la source. La présence de réverbération rend cette perception plus objective, ce que l'on peut expliquer au moyen de considérations simples sur la propagation en espace clos.

Avant tout, on peut rappeler que la réverbération tardive tend en théorie vers un champ diffus, et donc que son énergie est indépendante de la position de la source. Puisque l'énergie relative à l'onde directe dépend elle fortement de la distance à la source, on en vient naturellement à considérer le rapport d'énergie directe sur l'énergie réverbérée comme un indice permettant au système auditif de juger de la distance de la source sonore. Cette hypothèse a été de nombreuses fois confirmée en pratique (Zahorik, 2002), et l'on considère généralement que la proportion d'énergie directe dans la réverbération est l'indice prédominant dans le jugement de distance à la source.

Cependant, il n'est pas tenu compte dans cette hypothèse de la nécessaire distinction entre intensité directe objective et intensité perçue : en effet, puisque les réflexions précoces intégrées au son direct en vertu de l'effet de précédence ont pour effet entre autres d'accroître l'intensité perçue de ce dernier, il paraît raisonnable d'en tenir compte dans la définition d'un indice objectif de distance basé sur l'énergie (Kahle, 1995). De même, et pour généraliser, on peut se demander si la distribution spatiale et temporelle des réflexions précoces ne joue pas un rôle dans la perception de la distance : dans la majorité des situations, plus la source est lointaine, plus la direction de provenance des premières réflexions se rapproche de celle du son direct, et plus leur retard relatif est faible. Pellegrini (2001) a récemment proposé une expérience pour vérifier cette hypothèse, grâce à un processeur d'acoustique virtuelle utilisant un modèle simple de salle, qui permet de jouer entre autres sur la direction d'incidence et le temps d'arrivée de chacune des trois premières réflexions non frontales. La

[10] Le son direct est supposé frontal, la distribution des réflexions précoces symétrique par rapport au plan médian, et le signal source est supposé de courte durée et basses fréquences, de manière à ce que la tête puisse être considérée comme transparente d'un point de vue acoustique, et donc approchée par deux microphones omnidirectionnels.

validation perceptive de ce simulateur semble confirmer la pertinence de cette hypothèse. Cependant, il est difficile de conclure sur l'influence séparée des trois paramètres réglant les premières réflexions (direction d'incidence, temps d'arrivée et niveau) sur la distance perçue, puisque le simulateur règle les trois en même temps sur des considérations géométriques.

Contrairement à l'intensité sonore et au timbre de la source, le niveau relatif de la réverbération et les caractéristiques des premières réflexions constituent des indices absolus pour la perception de la distance, c'est-à-dire qu'ils permettent de juger de la distance d'une source non familière, sans nécessiter de connaissances *a priori* sur son intensité ou son timbre.

4 PERCEPTION DE L'ESPACE SONORE

Au fur et à mesure de la propagation, les réflexions qui parviennent à l'auditeur, de plus en plus nombreuses, ne sont plus intégrées à l'événement auditif lié à la source, mais contribuent à la formation d'un événement auditif distinct lié à la salle, ou plus généralement, à l'espace sonore dans lequel sont baignées les sources. Cette impression de l'espace sonore se manifeste à la fois par une sensation de présence de la salle et de ses dimensions (section 4.1), et par la perception d'un champ sonore enveloppant, perception tant spatiale (section 4.2) que temporelle (section 4.3).

4.1 Perception de la taille de la salle

A partir de considérations géométriques simples, on peut constater que le temps d'arrivée des réflexions spéculaires dépend étroitement de la taille de la salle. L'équation de Kuttruff (2000), selon laquelle la densité temporelle de réflexions à un instant donné est inversement proportionnelle au volume de la salle, permet d'appuyer ce constat. On en vient naturellement à l'hypothèse duale dans le monde perceptif, selon laquelle la perception de la taille de la salle est liée principalement au temps d'arrivée des réflexions précoces. Ainsi, le **temps d'arrivée de la première réflexion** (*Initial Time Delay Gap*, ou ITDG) a été proposé par Beranek (1962; 1996) comme un indice objectif de la sensation d'intimité. Beranek considère que ce temps d'arrivée doit être au plus de 20 millisecondes pour obtenir une bonne sensation d'intimité. Plus récemment, Pellegrini (2001) a proposé dans son simulateur d'acoustique virtuelle un rendu de la taille basé sur huit réflexions spéculaires précoces, sachant que les trois premières permettent également de régler la distance apparente (voir section 3.3), alors que les cinq dernières ne dépendent que du volume de la salle.

Néanmoins, cette corrélation entre ITDG et intimité est remise en question par Barron (1993; 1971), qui remarque que puisque l'ITDG est minimal pour les sièges loin de la scène, la sensation d'intimité devrait être plus importante que pour les sièges près de la scène, alors que c'est généralement l'inverse qui est constaté.

4.2 Perception de l'enveloppement de l'auditeur

L'enveloppement de l'auditeur (*listener envelopment*, ou LEV) peut être défini comme l'impression subjective d'être immergé dans un champ réverbérant. La distinction entre largeur de l'événement auditif et sensation d'enveloppement est assez récente, et a été clairement formalisée par Bradley et Soulodre (1995a), et la distinction entre enveloppement par la source et enveloppement par la salle est encore plus récente (Griesinger, 1997; Rumsey, 2002).

Il y a beaucoup de similitudes entre les causes et manifestations des sensations d'enveloppement et d'élargissement de la source : d'une part, s'il s'avère que ces deux notions sont bien distinctes d'un point de vue de perceptif, elles ne sont pas complètement indépendantes, l'augmentation de la proportion d'énergie tardive entraînant à la fois une augmentation de l'enveloppement et une perte de sensibilité à la largeur apparente de la source, comme cela a été déjà mentionné dans la section 3.2 ; d'autre part, tout comme la largeur apparente dépend de la distribution spatiale de l'énergie précoce, l'enveloppement dépend de la distribution spatiale de l'énergie tardive : les réflexions latérales ont un plus fort pouvoir de

décorrélation interaurale que les réflexions quasi-frontales, et l'enveloppement en est d'autant augmenté. L'influence de réflexions venant de l'arrière n'a que peu été étudié. Si l'on se fie aux remarques informelles de Barron et Marshall (1981), les réflexions précoces provenant de l'arrière ont déjà un pouvoir enveloppant, et l'on peut vraisemblablement penser que les réflexions tardives venant de l'arrière ont au moins le même effet. Selon Morimoto, cité par Griesinger (1999), les réflexions venant de l'arrière ont un pouvoir enveloppant plus fort que les réflexions frontales, ce qui l'a amené à proposer une mesure basée sur le niveau relatif d'énergie entre avant et arrière.

Comme la largeur apparente de la source, la sensation d'enveloppement est elle aussi affectée par le niveau sonore global (Bradley et Soulodre, 1995b; Bradley et al., 2000). Il est vraisemblable que les explications qui ont été avancées vis-à-vis de l'influence du niveau global sur la largeur apparente de source soient également valables vis-à-vis de la sensation d'enveloppement.

Influence de la distribution d'énergie précoce sur l'enveloppement de l'auditeur

De même que la distribution d'énergie tardive à un rôle inhibiteur sur la largeur apparente de source, Bradley et al. (2000) ont montré que la distribution d'énergie précoce joue sur la sensation d'enveloppement : pour un niveau sonore constant, l'enveloppement diminue lorsque la proportion d'énergie précoce (mesurée grâce au critère de clarté C_{80}) augmente. Ainsi, il s'avère que les distributions précoce et tardive de l'énergie ont des influences contraires sur la largeur apparente et l'enveloppement. La conception d'une salle, qui nécessite un équilibre adéquat entre les deux sensations en fonction de l'esthétique souhaitée, nécessite donc de gérer la balance entre l'énergie précoce et l'énergie tardive.

L'enveloppement comme résultant de l'incohérence interaurale tardive

Il a été mentionné (section 3.2) que dès lors que l'on présente aux oreilles deux signaux faiblement incohérents, l'événement auditif correspondant avait tendance à s'élargir. Si l'incohérence est plus prononcée, la localisation devient de plus en plus difficile et l'auditeur se sent peu à peu complètement enveloppé (puis, lorsque les signaux sont fortement décorrélés, il n'y a plus de fusion, et l'auditeur perçoit deux événements auditifs distincts au niveau des écouteurs). Or dans une salle réelle, la corrélation interaurale à court-terme est importante à l'instant d'arrivée de l'onde directe, puis décroît au fur et à mesure que de nouvelles réflexions parviennent à l'auditeur, provoquant ainsi un élargissement de la source, puis une sensation d'enveloppement lorsqu'elle passe en dessous d'un certain seuil, sensation d'enveloppement qui n'est pas liée perceptivement à la source, puisque l'incohérence ainsi que le retard par rapport au son direct ne permettent plus à l'effet de précédence d'intégrer les réflexions tardives à l'événement auditif créé par le son direct.

Mesures objectives de l'enveloppement

Le parallèle entre les causes d'élargissement et celles de l'enveloppement, et notamment l'influence de la direction d'incidence sur l'enveloppement, ont amené Bradley et Soulodre (1995a; 1995b) à la présentation d'un nouvel indice objectif dérivé de la fraction d'énergie latérale précoce, appelé le **niveau relatif d'énergie latérale tardive** (noté LG_{80}^{∞}) ; cet indice, comparé à des variantes ainsi qu'à la **corrélation interaurale tardive** (notée $IACC_l$, ou $IACC_{80}^{\infty}$), offre la meilleure corrélation avec le jugement d'enveloppement sur les résultats de leurs expériences, en une relation quasi-linéaire.

La remarque concernant la pertinence des mesures objectives de la largeur apparente sont toujours valables ici : l'indice LG n'est pas suffisant pour caractériser objectivement la sensation d'enveloppement : il ne tient compte ni du niveau sonore, ni du rôle inhibiteur des réflexions précoces, et lui aussi utilise la borne temporelle arbitraire de 80 ms. Néanmoins, il est nécessaire pour que la mesure soit répétable que sa définition soit la plus stricte possible, et ces différentes mesures, bien que simplificatrices, ont prouvé leur efficacité à caractériser une salle.

4.3 Perception de la réverbérance

En se référant à la définition de Blauert et Lindemann (1986a), la réverbérance est la "perception de la rémanence temporelle des événements sonores". Cette définition incite à établir un lien entre la sensation de réverbérance et la notion objective de durée de réverbération.

Il est nécessaire avant tout de rappeler que la notion de durée de réverbération n'est pas définie de manière unique. Si la théorie, rappelée au chapitre I, prévoit que la réverbération tardive résulte d'une superposition dense de modes présentant une décroissance exponentielle, en pratique son enveloppe est rarement exponentielle sur toute sa durée : ainsi, la présence d'éventuels volumes couplés dans la salle (ce qui est fréquent dans une salle de concert comprenant des balcons ou une cage de scène profonde, par exemple) induit la superposition de plusieurs décroissances exponentielles, dont l'importance relative varie avec la position par rapport à chacun des volumes. Même si l'on laisse de côté cette question des doubles décroissances, il faut distinguer une nouvelle fois la réverbération tardive, qui participe d'un régime diffus, de la partie précoce de l'effet de salle, pour laquelle les réflexions spéculaires peuvent sensiblement modifier l'allure de l'enveloppe de la réverbération. Il existe ainsi deux classes de définitions de la durée de réverbération (AFNOR, 2000) :

- le **temps de réverbération** (TR) caractérise la réverbération tardive : il correspond à la durée que met l'énergie dans la salle pour décroître de 60 dB en régime diffus après excitation par une source stationnaire. Pour les raisons évoquées ci-dessus, il est nécessaire de ne pas considérer la première partie de la décroissance, si bien que la référence temporelle pour mesurer cette durée a été arbitrairement fixée au moment où l'énergie à déjà décru de 5 dB depuis l'extinction de la source. En pratique, la dynamique de la mesure permet rarement d'estimer la durée,nommée T_{60}, nécessaire à l'énergie pour décroître de -5 dB à -65 dB, et l'on a recours à des définitions plus souples, qui consistent à estimer la durée de décroissance de l'énergie de -5 dB à -35 dB (T_{30}), voire de -5 dB à -25 dB (T_{20}), sachant que ces durées sont ramenées à une décroissance de 60 dB par multiplication par deux ou trois, selon le cas.
- la **durée de réverbération initiale** (*Early Decay Time*, ou EDT) caractérise la décroissance précoce, participant à la fois des réflexions spéculaires que des premiers instants de la réverbération diffuse. Elle correspond à la durée nécessaire à l'énergie pour décroître de 0 dB à -10 dB, 0 dB correspondant au niveau lors de l'extinction de la source.

Le temps de réverbération est depuis très longtemps considéré comme un indice primordial vis-à-vis de la sensation de réverbérance. Néanmoins, Schroeder indique que la durée de réverbération initiale était mieux corrélée avec cette dernière. Ceci s'explique simplement par le fait que l'on a finalement peu l'occasion d'entendre la décroissance totale de la salle dans une situation réelle, celle-ci étant masquée par l'arrivée de nouveaux événements. Cela dit, le temps de réverbération est loin d'être une mesure non pertinente vis-a-vis de la qualité d'une salle : en effet, la comparaison avec la durée de réverbération initiale donne une indication du caractère diffus ou non de la salle, ainsi que sur la densité de réflexions précoces, et donc entre autres sur la clarté. Certaines salles modernes ont été justement conçues en jouant fortement sur ce degré de liberté entre temps de décroissance précoce et tardive de la réverbération, dans un sens comme dans un autre. Par exemple, maintenir en même temps un EDT relativement bas et un TR assez long (deux secondes ou plus) permet d'allier une importante clarté à la sensation d'une réverbération présente et riche (Barron, 1993; Beranek, 1996).

5 CONCLUSION

Le point de vue adopté ici sur les différents aspects de l'audition spatiale en espace clos, global mais non exhaustif, permet de dégager deux grandes catégories : d'une part, la perception spatiale de la source, c'est-à-dire principalement celle, plus ou moins précise en fonction du contexte, de sa position ; d'autre part, la caractérisation perceptive de la réverbération, aussi bien spatiale que temporelle.

Dans l'optique d'une description **automatique** des aspects spatiaux d'une scène sonore, ceci permet de préciser la nature des informations à rechercher au sein d'un enregistrement

puisque, rappelons-le, on ne s'intéresse qu'aux informations qui font sens pour un auditeur. Le fait que l'on se consacre plus particulièrement aux enregistrements binauraux favorise encore un peu plus cette mise en relation, mais les principes qui sont présentées ici sont pour la plupart suffisamment généraux pour être transposables à la plupart des configurations d'écoute ; en revanche, la généralisation des méthodes de description proposées par la suite à d'autres configurations n'est pas forcément immédiate : par exemple, il serait nécessaire de tenir compte du rôle majeur qu'occupe l'effet de localisation par sommation en écoute stéréophonique pour en déduire l'organisation spatiale automatiquement, ce qui sort du cadre fixé pour cette étude, puisque dans ce cas, il ne s'agit plus uniquement d'extraire des informations objectives de la scène, mais de les interpréter au moyen d'un modèle auditif.

Il s'agit maintenant de proposer un ensemble de méthodes, directement inspirées par les notions évoquées dans ce chapitre et le précédent, visant à décrire deux aspects objectifs de la scène sonore pertinents d'un point de vue perceptif, c'est-à-dire la position physique de la source, et la caractérisation de l'enveloppe de réverbération.

Deuxième partie

Méthodes de détection en milieu réverbérant

III

Présentation du problème de la détection en milieu réverbérant

Comme cela a été mentionné en introduction de ce document, la première opération à effectuer est une détection des plages de temps les plus à même de fournir les informations recherchées. Avant de développer le formalisme sous-jacent aux méthodes de détection proposées dans cette étude, il est nécessaire de définir précisément la tâche de détection à effectuer (section 1). Toute détection reposant sur un modèle de signaux, on propose en sections 2 et 3 quelques modèles susceptibles de mener à bien cette opération. Finalement, le choix de la méthode employée par la suite est justifié en section 4.

1 INTRODUCTION

1.1 Définition du problème à traiter

La terme de détection peut être appliqué à de nombreux problèmes très variés, désignant à chaque fois une tâche différente. Il est donc nécessaire de préciser sa signification pour le problème spécifique soulevé par cette étude.

Une tâche de détection, dans son sens général, consiste à repérer les instants auxquels le **signal utile** (c'est-à-dire les informations qui font sens vis-à-vis du problème à traiter) émerge du **bruit** (c'est-à-dire le reste de l'information disponible). Si l'on applique ce principe au problème présent, on peut définir la détection comme la tâche consistant à repérer dans l'enregistrement les données permettant d'inférer des informations sur les aspects spatiaux de la scène sonore.

Or le chapitre II a permis de dégager deux classes d'attributs spatiaux permettant de qualifier une scène sonore : ceux relatifs à la **source**, qui consistent principalement en des informations de **position**, et ceux relatifs à la **réverbération**, caractérisant l'organisation **spatiale** du champ réverbéré (enveloppement), ainsi que, si l'on élargit la notion d'attributs spatiaux, son organisation **temporelle** (durée de l'enveloppe de réverbération). Le problème de détection est donc double :

- la **détection de source** consiste à rechercher les plages de temps auxquels on peut au mieux **localiser la source**
- la **détection de réverbération** consiste à rechercher les plages de temps sur lesquelles on peut le mieux **décrire la réverbération**.

1.2 Détection de source et détection de réverbération

En première approche, les problèmes de la détection de source et de réverbération peuvent être considérés comme duaux :

D'une part, les moments auxquels la source est la mieux localisable sont ceux pour lesquels le bruit (c'est-à-dire principalement la réverbération des événements précédents si la source est unique) est le plus faible possible, et la source en activité, de telle sorte que les indices de localisation soient non ambigus. La détection de source se ramène donc en fait à une détection de la **présence** d'une source active[1].

D'autre part, les moments auxquels la réverbération est la plus aisée à décrire (en l'absence d'un traitement supplémentaire de type déconvolutif) sont ceux où la source fait silence, car dans le cas contraire, la réverbération est au moins partiellement masquée par l'activité de la source. La détection de réverbération se ramène donc à une détection de l'**absence** d'activité de la source.

En fait, il est nécessaire de nuancer quelque peu cette dichotomie, car elle considère que le champ acoustique n'est dû qu'à une source localisable et à la réverbération dont elle est directement la cause. Ceci exclut notamment :

- le bruit de fond électronique et électromagnétique : le bruit de fond électronique ne gêne l'estimation de la décroissance uniquement sur des dynamiques importantes, comme dans le cas par exemple d'une réponse impulsionnelle. En revanche, le bruit de fond d'origine électromagnétique est plus gênant, puisqu'il est fréquent, et en particulier lorsque les liaisons électriques des microphones aux préamplificateurs sont très proches les unes des autres, qu'il se retrouve quasiment à l'identique sur toutes les voies des signaux observés.
- les sources acoustiques non localisables, c'est-à-dire le bruit de fond acoustique ambiant, ou bien des sources très lointaines, qui ne sont pas localisables car ayant subi trop de réflexions. Cependant, si les supports spectraux de la réverbération et de ces sources de bruit sont disjoints, ces dernières ne gênent pas l'estimation, car l'ensemble de l'analyse qui est présentée dans cette étude est effectuée en bandes étroites.

[1] On verra par la suite que durant l'intervalle de temps pendant lequel la source est active, la qualité de la détection évolue, étant meilleure pendant l'attaque que pendant l'entretien

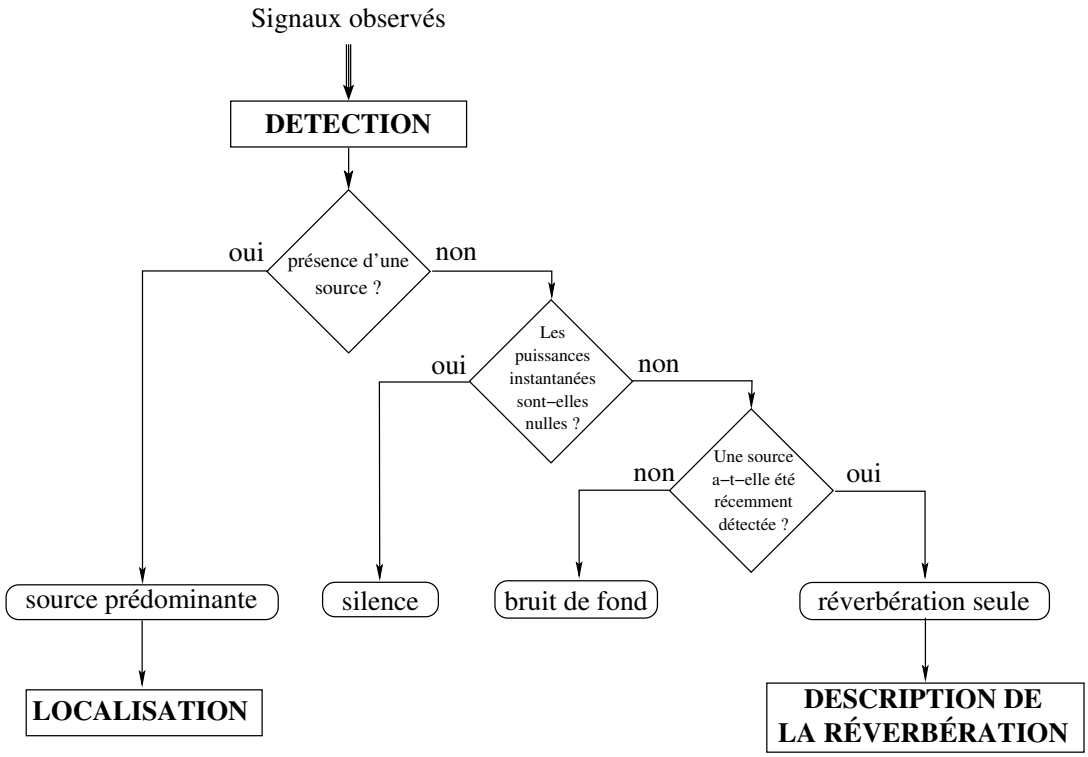


FIG. III.1 – **Schéma de principe de la détection bicanale en milieu réverbérant**

La figure III.1 illustre le principe évoqué ci-dessus, et présente ainsi l'algorithme de décision basé sur une détection de signal qui sera utilisé dans cette étude.

La détection de signal est un cas particulier de décision, et repose donc, du moins dans une tâche de détection automatique comme c'est la cas ici[2], sur un **modèle de signaux**. L'unique alternative consiste dans ce cas à estimer si les signaux observés contiennent ou non un signal qui corresponde à ce modèle, et si oui, en quelle proportion. Plusieurs modèles de signaux sont envisageables, qu'ils soient monophoniques (section 2) ou non (section 3).

2 EXEMPLES DE MODÈLES MONOPHONIQUES

On rappelle ici deux classes de méthodes de détection, qui ont l'avantage d'être monophoniques, mais qui impliquent des hypothèses quant aux signaux sources qui leur sont propres.

2.1 Détection par rupture de stationnarité

Si l'on suppose que le signal présente des attaques marquées, il est possible de se baser sur cette **rupture de stationnarité** pour détecter le début du signal. Une approche courante consiste à étudier les variations de la puissance moyenne à court-terme du signal en bande étroite (Martin, 1995; Rodet et Jaillet, 2001). Une autre méthode est de tenter de prédire le devenir du signal, et de baser la détection sur les instants où l'erreur de prédiction devient importante. La prédiction peut se baser par exemple sur le fait qu'un signal stationnaire peut se décomposer en une somme de sinusoïdes (Thornburg et Gouyon, 2000), ou bien en utilisant un filtrage prédictif adaptatif, qui tente d'estimer les valeurs actuelles du signal observé à partir des échantillons plus anciens (Supper et al., 2003).

[2]Lorsque la théorie de la détection est employée en psychologie, le modèle sous-jacent est la plupart du temps un modèle probabiliste de réponse interne à un stimulus sensoriel.

Ces méthodes sont soumises à une limitation intrinsèque, car étant donnée la définition même du modèle de signaux, elles ne peuvent permettre une description spatiale de signaux stationnaires, car elles ne visent pas la détection de la **présence** d'une source active, mais la détection des **attaques**, ou tout autre évènement de courte durée . On peut ainsi repérer les moments où la source **devient** active, mais il est difficile d'envisager *a contrario* le repérage des moments où devient inactive, car même si la source s'éteignait brusquement sans résonance propre, la réverbération empêche que cette extinction soit de courte durée. Elles ne permettent donc pas de juger si, entre deux attaques, une source est présente ou non.

2.2 Détection d'harmonicité

En supposant maintenant que la source est harmonique et que le bruit de fond est inharmonique, la détection de la source peut se baser sur une évaluation de l'harmonicité (ou périodicité), c'est-à-dire du rapport d'énergie entre la partie harmonique et la partie inharmonique du signal.

Il existe de très nombreuses manières d'estimer l'harmonicité d'un signal et, si cela a un sens, sa période fondamentale. Cheveigné et Kawahara (2002) ont récemment proposé une méthode de détection d'harmonicité et d'estimation de fréquence fondamentale basée sur un principe de retard et annulation : considérant un signal $x(t)$, on forme la différence

$$d_\Delta(t) = x(t) - x(t - \Delta)$$

, fonction d'un retard Δ variable, et l'on cherche celui qui minimise la puissance de la différence.

Ce principe très simple permet d'estimer l'harmonicité et la fréquence fondamentale de signaux harmoniques monodiques avec une très bonne précision. Néanmoins, l'application à des signaux harmoniques réverbérés pose un problème de taille : en effet, la réverbération associée à une source harmonique monodique, bien que non harmonique au sens strict (voir ci-dessous), peut être considérée comme un deuxième flux quasi-harmonique, si bien que le signal total (c'est-à-dire le mélange de la source non réverbérée et de la réverbération) n'est plus monodique mais polyphonique. Une solution consiste à appliquer un modèle de détection d'harmonicité double, par exemple en mettant en cascade deux détecteurs (Cheveigné, 1993). L'hypothèse d'une source unique et monodique dans un champ réverbéré permet de poser des contraintes supplémentaires visant à stabiliser la détection (Baskind et Cheveigné, 2003)[3].

Contrairement au cas précédent, la distinction entre les moments où l'onde directe est prépondérante (source active) et ceux où seule la réverbération est présente (source inactive) est envisageable, en se basant sur des critères d'inharmonicité : en effet, Wu et Wang (2003) indiquent que la réverbération corrompt l'harmonicité d'un signal, et qu'il est même envisageable d'estimer le temps de réverbération à partir de l'évaluation de cette inharmonicité. Sur cette base, on peut envisager de pouvoir détecter la réverbération. Cela dit, l'hypothèse d'harmonicité de la source est cruciale dans ce cas, et on peut se demander si un tel système serait toujours robuste vis-à-vis de sources elles-mêmes partiellement inharmoniques.

3 MODÈLES BASÉS SUR LES SIMILARITÉS ENTRE LES CANAUX

3.1 Principe

Le modèle sous-jacent à ces méthodes n'est pas exactement un modèle de signaux, mais un modèle de relations entre les voies des signaux observés. En l'absence de toute connaissance concernant les signaux émis par la source, la principale caractéristique qui permette de distinguer l'onde directe du champ réverbéré est la cohérence spatiale de la première, alors que, comme cela a été mentionné au chapitre I, le champ réverbéré tend dans l'idéal vers un champ diffus, qui ne possède plus de cohérence spatiale, excepté en basses fréquences.

[3]Cette publication est reproduite en annexe E.

Cette cohérence spatiale se traduit par une **similarité** entre les signaux reçus par les microphones, c'est-à-dire que l'on considère dans des cas simples que ces signaux ne diffèrent que par des retards et des gains. Cette similarité peut être constatée au moyen de nombreuses techniques. On ne citera ici que celles basées sur des statistiques d'ordre 2 :

- d'une part, et pour chacune des paires de signaux possibles, la **corrélation** est maximale pour le retard égal au retard entre les deux signaux concernés.
- d'autre part, la **puissance de la somme** des signaux **resynchronisés** est elle aussi maximale. Il s'agit des techniques par **retard et addition** (*delay and add*).
- pour finir, la **puissance de la différence** des signaux **resynchronisés** et **égalisés** est minimale, voire nulle en l'absence de bruit additionnel. Il s'agit des techniques par **égalisation et annulation** (*equalization and cancellation*).

Puisque la valeur du maximum (ou respectivement, du minimum) donne une indication de la qualité du modèle, la recherche de ce maximum (respectivement de ce minimum) par balayage de tous les retards (respectivement de tous les retards et gains) envisageables permet de juger si une source obéissant à l'hypothèse de similarité est présente ou non.

Ce principe est à la base de beaucoup de techniques de détection par antennes de microphones (notamment à travers la technique par retard et addition), ainsi que de la plupart des modèles auditifs de détection binaurale. Le formalisme des méthodes de détection par antennes étant peu adapté au cas d'une configuration à deux microphones, on présente brièvement ces derniers, qui bien que destinés à des fins de modélisation et non de description automatique, ont permis de développer des notions qui sont d'une grande utilité dans cette étude.

3.2 Modèles auditifs de détection binaurale

Modèles de détection binaurale par corrélation

Historiquement, le premier modèle de cette classe est le **modèle de coïncidence** proposé par Jeffress (1948) : il s'agit d'un modèle de détection et de latéralisation sous forme d'un réseau de neurones utilisant des éléments physiologiquement plausibles, c'est-à-dire principalement des lignes à retards et des neurones de type excitation-excitation ("neurones EE"), dont la sortie ne s'active que lorsque les deux entrées sont simultanément activées. En assimilant le fonctionnement de ces neurones EE à une simple multiplication, on obtient une implémentation numérique sous forme de corrélations à court-terme (voir figure III.2). La détection correspond en ce cas à une activation du motif de coïncidence, qui en l'absence de signaux cohérents à gauche et à droite, est théoriquement nul. Le retard pour lequel l'activation est maximale correspond à la différence de temps entre les signaux détectés.

Modèles de détection binaurale par égalisation et annulation

Le principe d'interaction binaurale par égalisation et annulation (*Equalization and Cancellation Model*, ou modèle E-C) a été proposé par Durlach (1963) pour modéliser les expériences mettant en évidence la supériorité de la détection binaurale sur la détection monaurale (effet cocktail-party), expériences conduisant à la notion de **Seuil de masquage binaural** (*Binaural Masking Level Difference*, ou BMLD), et à mettre en relation avec l'effet *cocktail-party* : dans ces expériences, un signal cible stationnaire est masqué par un signal parasite, les deux étant présentés de manière dichotique, avec des relations de phase différentes. Durlach suppose que le système auditif cherche une transformation qui puisse annuler le signal masquant, en deux opérations : tout d'abord, en multipliant et retardant le signal à une oreille pour que le signal masquant soit identique à gauche et à droite (égalisation), puis en soustrayant les deux (annulation). Si le signal cible n'a pas les mêmes relations de phase que le signal masquant, il n'est lui pas annulé par ce processus, mais juste filtré. Il souligne également le parallèle entre ce principe et les méthodes de pointage par antennes : annuler le signal masquant revient à pointer le zéro de la directivité de l'antenne vers la direction de la source parasite.

Ce modèle a été plus tard étendu par Culling et Summerfield (1995) pour modéliser le regroupement simultané sur plusieurs bandes. Le coeur de ce modèle modifié (*modified EC*

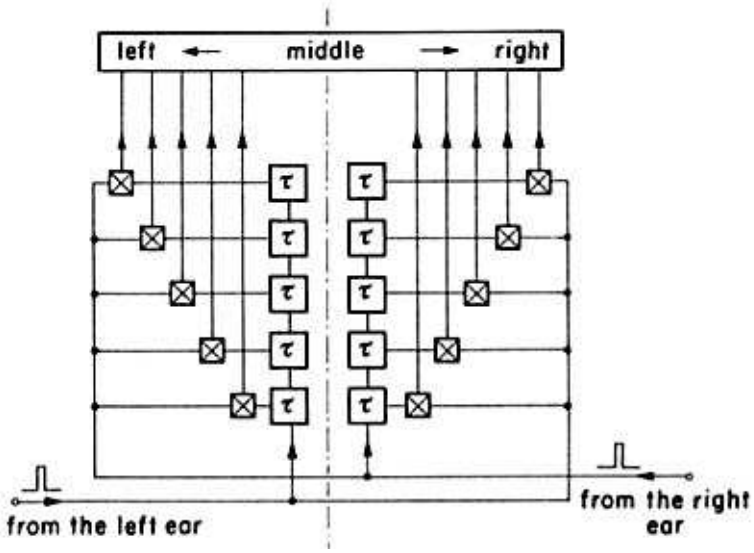


FIG. III.2 – **Le modèle de coïncidence de Jeffress** (repris de Blauert (1997))

model) consiste à appliquer le modèle de Durlach, bande critique par bande critique, à partir de signaux issus d'un modèle de cochlée.

Il s'agit selon Durlach, et contrairement au modèle de Jeffress, non pas de proposer une explication physiologique plausible de l'audition binaurale, mais plutôt de considérer celle-ci comme une "boite noire", et de fournir un outil prédictif des phénomènes de détection binaurale. D'autre part, ce modèle n'a pas pour vocation initiale de s'intéresser aux aspects spatiaux de la source, mais uniquement à sa détectabilité. Néanmoins, Durlach indique que, considéré sous un angle différent, il peut également expliquer notre sensibilité aux différences interaurales, la phase d'égalisation nécessitant leur estimation.

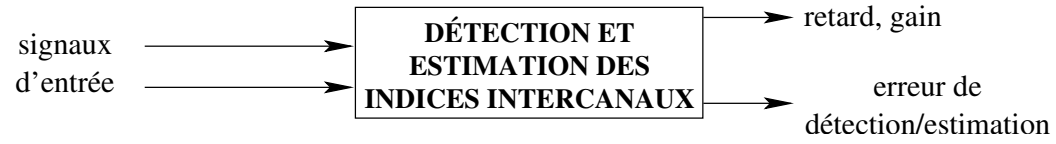
Liens entre les deux classes de modèles

Ces deux modèles, bien que différant dans le principe et dans les buts, ont en fait un fonctionnement très proche : ainsi, remplacer dans le réseau de Jeffress, comme le propose Breebaart (2001), les cellules excitation-excitation (EE) par des cellules excitation-inhibition (EI), qui sont inhibées lorsque les deux entrées sont activées, revient à proposer un modèle par égalisation et annulation avec un gain fixé à 1. A l'inverse, la puissance du signal issu de l'égalisation et annulation peut se décomposer, comme le montre Cheveigné (1998) dans le cas monophonique, et comme on le rappellera au chapitre IV, en une somme pondérée des puissances de chacun des signaux et de la corrélation interaurale.

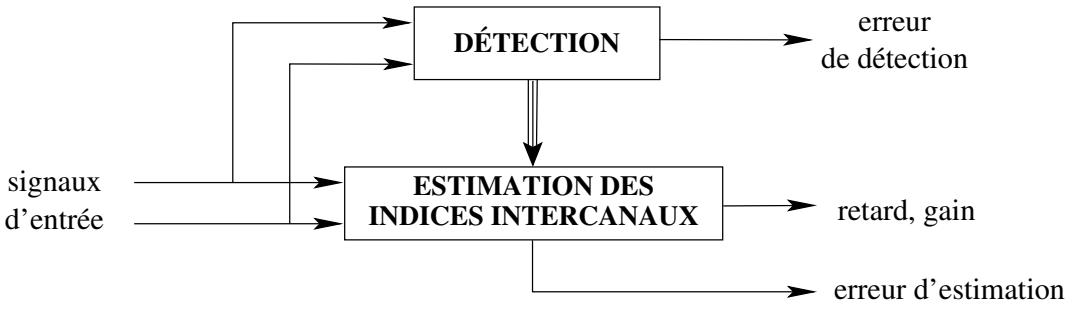
4 DISCUSSION : DÉTECTION ET ESTIMATION

Il est utile de rappeler que, puisque toute tâche de détection consiste à repérer dans du bruit un signal répondant au modèle sous-jacent à la méthode, elle permet simultanément d'estimer les paramètres de ce modèle : ainsi, en détectant si un signal contient une composante harmonique à un instant donné on a en même temps accès à la valeur de la période de cette composante, si elle existe. De même, juger par méthode de corrélation si deux signaux contiennent chacun des composantes égales à un retard près permet d'accéder à la connaissance de ce retard, qui est celui qui maximise la corrélation.

Or, en l'absence de toute connaissance sur le signal émis par une source, les indices les plus pertinents pour estimer la position physique de cette source, sont justement les différences de temps et de niveau entre les signaux captés en deux points de l'espace.



(a) la détection et l'estimation des indices de localisation participent du même modèle



(b) la détection et l'estimation des indices de localisation sont distincts

FIG. III.3 – Schémas de principe, lorsque la détection et l'estimation des indices de localisation participent du même modèle (cas de la détection par similarité), et distincts (cas notamment de la détection par harmonicité ou rupture de stationnarité)

Ainsi, les modèles monophoniques ont plusieurs avantages, dont justement celui de pouvoir travailler sur des signaux monophoniques, mais ne permettent pas d'effectuer la détection et l'estimation des indices de localisation en une seule opération, comme c'est le cas pour les modèles par similarité, puisque le modèle sous-jacent à ce type de détection n'est pas un modèle **spatial** mais un modèle **temporel**, se basant sur le comportement non-stationnaire (dans le cas de la détection par rupture de stationnarité), ou au contraire le comportement stationnaire (dans le cas de la détection d'harmonicité), du signal source.

Le corollaire de la remarque précédente est que puisque, dans le cas des modèles monophoniques, la détection et l'estimation des indices de localisation sont nécessairement confiées à des méthodes distinctes, il devient difficile de juger de l'**erreur** d'estimation de ces indices. La détection repose bien dans ces méthodes sur l'évaluation d'une erreur, mais celle-ci (erreur de prédiction ou inharmonicité) ne peut plus être directement reliée à une description spatiale.

Cet aspect est illustré en figure III.3 : dans le cas où la détection et l'estimation des indices de localisation participent du même modèle, l'erreur de détection permet également de quantifier de l'erreur d'estimation des indices de localisation, alors que dans le cas où ces deux tâches sont confiées à des méthodes distinctes, l'erreur de détection ne peut servir à juger de la qualité de l'estimation, et il est nécessaire de trouver une autre méthode pour cela.

Pour cette raison, et également puisque les méthodes monophoniques supposent le respect d'hypothèses fortes sur la nature des signaux sources, on choisit de se concentrer sur les méthodes de détection par similarité inter-canaux. Parmi les trois classes présentées, on se concentre sur les méthodes **par égalisation et annulation**, puisque celles-ci font varier non plus un, mais deux paramètres, qui sont le retard et le gain. La détection permet donc simultanément de connaître le retard et le gain optimaux à l'instant considéré.

5 CONCLUSION

Après avoir défini la nature du problème de détection dans le cadre de cette étude, on a présenté dans ce chapitre quelques unes des méthodes envisageables, et opté pour l'une des grandes classes de méthodes, qui consiste à étudier les éventuelles similarités entre les voies des signaux observés pour juger de la présence ou non d'un champ sonore cohérent, et donc, par extrapolation, d'une source sonore active dans la scène. Plus précisément, parmi les trois types de méthodes de détection par similarité basées sur des statistiques d'ordre 2, on a retenu le principe d'égalisation et annulation, qui, sorti de son contexte originel de modèle auditif, s'avère plus complet que les autres, puisqu'il permet d'estimer non seulement le retard de propagation entre les deux points d'écoute, mais également l'atténuation.

Ce principe d'égalisation et annulation, qui est ici présenté de manière très grossière, nécessite maintenant d'être explicité de manière bien plus complète pour envisager son application à des signaux complexes.

IV

Méthode non stationnaire de détection et d'estimation par égalisation et annulation

Ce chapitre propose une méthode non stationnaire de détection, couplée à une estimation des paramètres physiques permettant par la suite de juger de la direction de la source. On commence en section 2 par en ébaucher le principe sur une analyse à long-terme et en pleine bande. Puis, en soulignant les limites de ce type d'analyse sur des signaux non-stationnaires, on présente en section 3 une déclinaison à court-terme de cette méthode. Un couplage avec une analyse préalable par banc de filtres, étudié en section 4, permet de proposer un outil d'analyse à résolution temporelle et fréquentielle ajustable en fonction du problème à traiter. Dans le même ordre d'idées, on démontre en section 5 que puisque la transformation de Fourier à court-terme peut être considérée comme un banc de filtres, les méthodes élaborées s'y appliquent également, et gagnent en rapidité de calcul au prix d'une approximation dont la pertinence dépend de la largeur de chaque bande. Finalement, en section 6 est ébauchée une possible généralisation à des configurations de prise de son sur plus de deux canaux.

1 INTRODUCTION

La méthode présentée ici reprend l'idée énoncée par Durlach (1963) pour son modèle d'audition ; cependant, il faut noter deux différences majeures :

D'une part, le principe d'égalisation et annulation est ici sorti de son contexte de modèle auditif. Il ne s'agit pas de modéliser le mécanisme de détection binaurale pour prédire son comportement, mais de proposer des méthodes de traitement de signal pour effectuer une tâche automatique de détection.

D'autre part, cette idée est en quelque sorte renversée dans la méthode proposée ici : en effet, le modèle Durlach vise à expliquer le phénomène de démasquage binaural par **annulation du signal masquant**. Ici, on cherche au contraire à **annuler une hypothétique source cible**, pour juger *a posteriori* de sa présence ou non. Si l'on reprend l'analogie de Durlach avec le traitement d'antennes, on ne cherche pas à pointer le zéro de directivité de l'antenne dans la direction du signal masquant pour réveler le signal masqué, mais à étudier s'il existe une orientation de l'antenne pour laquelle la source cible est complètement annulée.

Cette dernière mise au point est loin d'être purement formelle, puisqu'elle révèle que les paradigmes sous-jacents au modèle de Durlach et à la présente étude diffèrent dans le principe : en effet, dans les expériences de détection binaurale, le signal masquant et le signal masqué sont généralement relativement cohérents d'une voie à l'autre, la distinction s'opérant sur la mise hors phase de l'un ou de l'autre ; ici le signal cible est mélangé à un "bruit" correspondant principalement au champ réverbéré, qui est, si l'on rappelle les notions évoquées dans le chapitre I, décorrélé d'une voie à l'autre, excepté dans les très basses fréquences.

Les méthodes présentées ici utilisent en partie le formalisme et les méthodes de détection proposés par Cheveigné et Kawahara (2002) pour l'estimation monophonique de fréquence fondamentale, en l'adaptant au problème de la détection bicanale. Cependant, l'analogie entre les deux problèmes trouve ses limites dans deux constats :

- on considère ici que les signaux observés diffèrent non seulement en temps (une voie est en retard sur l'autre), mais également en niveau. Il est donc nécessaire de compléter le formalisme, en adjoignant, dans la phase d'égalisation, un gain au retard de synchronisation.
- les différences de temps et de niveau dépendent de la fréquence dans la majorité des situations. Pour tenir compte de cet aspect, la méthode d'égalisation et annulation est étendue au cas d'une analyse en bandes étroites, dans les sections 4 et 5.

2 PRINCIPE DE LA DÉTECTION PAR ÉGALISATION ET ANNULATION

2.1 Rappel : définition de la puissance et de la corrélation

Signaux d'énergie infinie

On définit, pour des signaux d'énergie infinie, la puissance d'un signal $x[n]$ donné sous la forme :

$$P_x = \lim_{N \to \infty} \left\{ \frac{1}{2N+1} \sum_{n=-N}^{N} x^2[n] \right\}$$

On définit de même l'intercorrélation de deux signaux $x[n]$ et $y[n]$ par :

$$C_{xy}[\tau] = \lim_{N \to \infty} \left\{ \frac{1}{2N+1} \sum_{n=-N}^{N} y[n].x[n+\tau] \right\}$$

L'intercorrélation normalisée est définie par :

$$\rho_{xy}[\tau] = \frac{C_{xy}[\tau]}{\sqrt{P_x.P_y}}$$

Signaux d'énergie finie

Si les signaux sont d'énergie finie, on a recours préférentiellement à l'énergie au lieu de la puissance :

$$E_x = \sum_{n=-\infty}^{+\infty} x^2[n]$$

... sachant qu'en pratique, les bornes de la sommation se restreignent au support temporel du signal. L'intercorrélation s'écrit alors :

$$C_{xy}[\tau] = \sum_{n=-\infty}^{+\infty} y[n].x[n+\tau]$$

... et l'intercorrélation normalisée :

$$\rho_{xy}[\tau] = \frac{C_{xy}[\tau]}{\sqrt{E_x.E_y}}$$

Pour la suite, on suppose que les signaux sont **d'énergie finie**, mais les calculs sont identiques dans le cas contraire.

2.2 Formalisation du problème

On se place tout d'abord volontairement dans un cadre très simplifié : on considère le cas d'un champ de pression émis par une source supposée ponctuelle en situation anéchoïque, capté par deux microphones de directivités indépendantes de la fréquence, et sans obstacle entre eux (figure IV.1). Sous ces hypothèses, et en désignant par $s[n]$ le signal source échantillonné, les signaux observés peuvent s'écrire sous la forme à temps discret :

$$\begin{cases} x\,[n] = A_x.s\,[n-\Delta_x] + b_x\,[n] \\ y\,[n] = A_y.s\,[n-\Delta_y] + b_y\,[n] \end{cases} \quad \text{(IV.1)}$$

Dans cette écriture, les gains $A_x$et A_y incluent les gains de la partie électrique, l'atténuation acoustique due aux trajets, et l'atténuation due à la directivité des microphones et à la position angulaire de la source par rapport à chacun d'entre eux. Les retards $\Delta_x$et Δ_y incluent les retards électriques et le retard acoustique dû aux trajets entre source et récepteurs. Les bruits $b_x\,[n]$ et $b_y\,[n]$ correspondent au bruit de fond acoustique (qui inclut notamment la réverbération des flux précédents) et électrique, qui est en principe décorrélé du signal source. Quel que soit le signal $s[n]$, il est toujours possible de se ramener par dilatation et translation temporelle au cas où $A_x = 1$ et $\Delta_x = 0$, si bien que le système IV.2 peut se réécrire sous la forme :

$$\begin{cases} x\,[n] = s\,[n] + b_x\,[n] \\ y\,[n] = A.s\,[n-\Delta] + b_y\,[n] \end{cases} \quad \text{(IV.2)}$$

Il est possible de s'affranchir du signal source en écrivant :

$$y\,[n] = A.x\,[n-\Delta] + (b_y\,[n] - A.b_x\,[n-\Delta]) \quad \text{(IV.3)}$$

Ainsi, si l'on néglige dans un premier temps le bruit de fond, les signaux électriques obtenus ne diffèrent que d'un retard et d'un gain, qui sont fonction de la différence de marche entre les trajets acoustiques et de la position relative de la source vis-à-vis de chacun des microphones.

$$y\,[n] = A.x\,[n-\Delta] \ \ \forall n \quad \text{(IV.4)}$$

ce qui équivaut simplement à

$$y\,[n] - A.x\,[n-\Delta] = 0 \ \forall n \quad \text{(IV.5)}$$

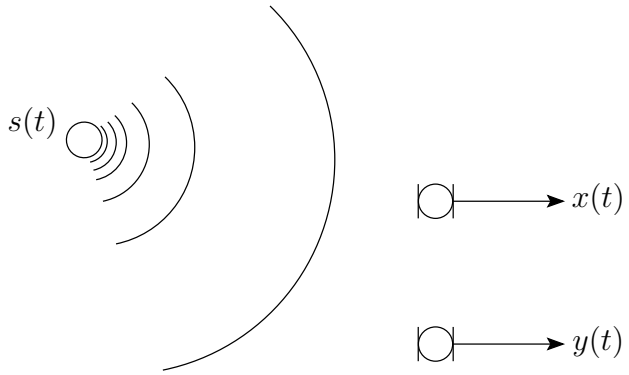


FIG. IV.1 – **Propagation idéale en champ libre d'une onde sonore** issue d'une source à laquelle correspond le signal $s(t)$ (par exemple, s'il s'agit d'un haut-parleur, le signal électrique qui l'alimente, ou le signal issu de la captation de la source acoustique en champ proche), et captée par deux microphones de directivité indépendante de la fréquence.

Erreur absolue

En règle générale, les gain A et le retard Δ sont inconnus, et il est nécessaire de les estimer : on définit pour ce faire l'**erreur absolue** d'égalisation et annulation, comme étant l'énergie de la différence après application d'un gain et d'un retard sur $x[n]$:

$$D_{\alpha,\tau} = \sum_{n=-\infty}^{+\infty} \left(y\left[n\right] - \alpha.x\left[n+\tau\right]\right)^2 \tag{IV.6}$$

L'erreur absolue s'exprime simplement grâce aux énergies de $x[n]$ et $y[n]$ et à leur intercorrélation :

$$D_{\alpha,\tau} = E_y + \alpha^2 E_x - 2\alpha C_{xy}[\tau]$$

Cette formulation rend son calcul bien plus aisé, puisqu'il suffit de connaître à l'avance les énergies et l'intercorrélation pour toutes les valeurs du retard τ envisageables pour calculer l'erreur absolue pour toute valeur de α et de τ.

Son minimum, toujours en l'absence de bruit, est obtenu pour le couple $(\alpha, \tau) = (A, -\Delta)$, et vaut simplement :

$$D_{A,-\Delta} = 0$$

Donc, si ce minimum est unique, la recherche systématique des minima de l'erreur absolue sur toutes les valeurs du couple (α, τ) envisageables permet en théorie d'estimer la différence de temps Δ et la différence d'intensité A qui correspondent au retard de propagation entre les deux microphones de l'onde générée par la source.

Cependant, c'est en présence de bruit que la notion de détection prend tout son sens : il est inutile de détecter un signal obéissant au modèle gain-retard, si l'on sait que seul lui est éventuellement présent dans les signaux observés[1] ! Si les signaux cibles sont mélangés à du bruit, le minimum de l'erreur n'est plus nul, mais fonction du rapport signal sur bruit. Ce résultat est montré en section 1 de l'annexe B. On y montre également qu'en fonction du niveau de corrélation des deux voies du bruit, l'estimation est plus ou moins biaisée.

Une manière d'effectuer la détection consiste donc à juger, en fonction de la valeur du minimum de l'erreur absolue, si les signaux observés permettent d'attester ou non de la présence d'une source physique active se propagant en champ libre.

[1] Plus exactement, la détection se ramène dans ce cas simplement à savoir si la puissance des signaux observés est nulle ou non.

2.3 Détection par minimisation de l'erreur absolue

La minimisation s'effectue nécessairement en deux étapes : d'une part, par annulation de la dérivée partielle par rapport à α, puis par recherche systématique en fonction de τ au sein des minima possibles (tous solutions de la minimisation par rapport à α) du minimum global. Traduit mathématiquement, cela signifie que si le couple (α_1, τ_1) est un minimum de l'erreur absolue, il obéit à :

$$\begin{cases} \left.\frac{\partial D_{\alpha,\tau}}{\partial \alpha}\right|_{\alpha_1,\tau_1} = 0 \\ \tau_1 = arg \left\{ \underset{\tau}{min} \ \left(D_{\alpha_1,\tau}\right) \right\} \end{cases}$$

... ce qui équivaut à :

$$\begin{cases} \tau_1 = arg \left\{ \underset{\tau}{max} \ \left(C_{xy}\left(\tau\right)\right) \right\} \\ \alpha_1 = \frac{C_{xy}(\tau_1)}{E_x} \end{cases} \tag{IV.7}$$

Si l'on écarte pour l'instant le cas de signaux sources harmoniques, il n'existe qu'un seul maximum à la fonction d'intercorrélation $C_{xy}\left(\tau\right)$. On a donc montré que le couple (α_1, τ_1) est un minimum unique de l'erreur absolue. Ce minimum vaut, après calculs :

$$D_{min} = D_{\alpha_1,\tau_1} = E_y.\left(1 - \rho_{xy}^2\right) \tag{IV.8}$$

Dans cette expression, ρ_{xy} désigne le coefficient d'intercorrélation, défini comme le maximum de l'intercorrélation normalisée :

$$\rho_{xy} = \underset{\tau}{max} \ \{\rho_{xy}[\tau]\}$$

2.4 Détection par minimisation de l'erreur normalisée

Comme toute tâche de détection, la décision repose ici sur la comparaison de l'erreur par rapport à un **seuil**. Un problème est que puisque le minimum est proportionnel à la puissance des signaux observés, il faudrait pour définir ce seuil connaître ces puissances. Pour dépasser cette limitation, on à recours à une **normalisation** de l'erreur, qui permet de s'assurer que celle-ci reste confinée dans des limites connues à l'avance, et indépendantes du contexte.

Erreur normalisée

Sur le modèle de l'erreur absolue, on forme la fonction :

$$S_{\alpha,\tau} = \sum_{n=-\infty}^{+\infty} \left(y\left[n\right] + \alpha.x\left[n+\tau\right]\right)^2$$

On définit l'**erreur normalisée** par :

$$\varepsilon_{\alpha,\tau} = \frac{D_{\alpha,\tau}}{D_{\alpha,\tau} + S_{\alpha,\tau}}$$

Puisque les fonctions $S_{\alpha,\tau}$ et $D_{\alpha,\tau}$ sont positives ou nulles quelles que soient les valeurs de α et de τ, l'erreur normalisée est **toujours comprise entre 0 et 1**.

D'autre part, la fonction $S_{\alpha,\tau}$ s'exprime elle aussi comme combinaison linéaire des énergies et de l'intercorrélation :

$$S_{\alpha,\tau} = E_y + \alpha^2 E_x + 2\alpha C_{xy}[\tau]$$

Donc l'erreur normalisée s'écrit :

$$\varepsilon_{\alpha,\tau} = \frac{1}{2} - \frac{\alpha C_{xy}[\tau]}{E_y + \alpha^2 E_x} \tag{IV.9}$$

Ainsi, si les signaux $x[n]$ et $y[n]$ sont corrélés de manière positive, l'erreur normalisée est comprise entre 0 et $\frac{1}{2}$. Si le signaux $x[n]$ et $y[n]$ sont anticorrélés, l'erreur normalisée est comprise entre $\frac{1}{2}$ et 1.

Minimisation de l'erreur normalisée

Si le couple (α_2, τ_2) est un minimum de l'erreur normalisée, il obéit à :

$$\begin{cases} \left.\frac{\partial \varepsilon_{\alpha,\tau}}{\partial \alpha}\right|_{\alpha_2,\tau_2} = 0 \\ \tau_2 = arg \left\{ \begin{array}{c} min \\ \tau \end{array} \ (\varepsilon_{\alpha_2,\tau}) \right\} \end{cases}$$

Soit, après calculs :

$$\begin{cases} \tau_2 = arg \left\{ \begin{array}{c} max \\ \tau \end{array} \ (C_{xy}(\tau)) \right\} \\ \alpha_2 = sgn\left(C_{xy}(\tau_2)\right).\sqrt{\frac{E_y}{E_x}} \end{cases} \quad \text{(IV.10)}$$

...où $sgn(.)$ désigne la fonction signe. On trouve donc que les estimations τ_1 et τ_2 des retards sont identiques pour les deux méthodes. Si l'on suppose que le coefficient de corrélation est positif, le gain optimal vaut simplement :

$$\alpha_2 = \sqrt{\frac{E_y}{E_x}}$$

On retrouve ici la méthode usuelle et intuitive d'estimation de gain par simple rapport de puissances, méthode fréquemment utilisée en estimation d'indices interauraux à partir de réponses impulsionnelles en pleine bande ou en bande étroite (Tollin, 1998; Martin, 1995; Daniel, 2000; Larcher, 2001).

D'autre part, l'erreur normalisée minimale vaut :

$$\varepsilon_{\alpha_2,\tau_2} = \frac{1}{2}.\left(1 - \left|\rho_{xy}(\tau_2)\right|\right) \quad \text{(IV.11)}$$

2.5 Erreur absolue ou erreur normalisée ?

Les deux méthodes d'estimation qui viennent d'être proposées ne fournissent donc pas les mêmes estimations du gain A. En fait, on peut observer d'après les équations IV.7 et IV.10 que les valeurs du gain optimales α_1 et α_2 trouvées par les deux méthodes sont liées par la relation suivante :

$$\alpha_1 = \left|\rho_{xy}\right|.\alpha_2$$

Ces deux méthodes de minimisation sont donc équivalentes si les signaux x et y sont effectivement égaux à un retard et un gain près (puisque dans ce cas $|\rho_{xy}| = 1$), mais ont des comportements différents en présence de bruit : en effet, dans le cas limite où le signal est complètement noyé dans le bruit et non détectable, la première méthode fournit un gain nul, contrairement à la seconde. En fait, on montre en section 1 de l'annexe B que les deux méthodes sont biaisées en présence de bruit, *a fortiori* lorsque celui-ci est corrélé d'une voie à l'autre.

On opte donc pour la méthode par minimisation de l'erreur normalisée, car comme cela a déjà été mentionné, elle convient mieux au problème de la détection, puisqu'elle fournit un indice, l'erreur normalisée minimale, qui est indépendant de la puissance des signaux observés. On appelle cet indice **indice de détection.** Cet indice, noté par la suite ε_{min}, est compris entre 0 et $\frac{1}{2}$:

- Si $\varepsilon_{min} = 0$, alors les signaux ne contiennent pas de bruit additionnel. Le modèle gain-retard est une description fidèle de la réalité, car $y[n] = \alpha_2.x\left[n - \tau_2\right] \forall n$
- Si $\varepsilon_{min} = \frac{1}{2}$, les signaux $x[n]$ et $y[n]$ sont totalement décorrélés, et les signaux observés ne contiennent aucune composante qui réponde au modèle.

2.6 Lien avec les méthodes usuelles d'estimation de retard

Équivalence avec l'estimation du retard basée sur l'intercorrélation

Comme indiqué au chapitre III, beaucoup de modèles auditifs de localisation ou de latéralisation utilisent la corrélation interaurale, et tâchent de la maximiser (ou sa valeur absolue) pour trouver le retard optimal. Cette méthode est également couramment utilisée dans de nombreuses applications ne relevant pas forcément de la modélisation (Larcher, 2001; Martin, 1995), et parfois critiquée, notamment pour son incapacité à fournir une indication sur la précision de l'estimation (Daniel, 2000).

Or il s'avère d'après les calculs ci-dessus que la méthode d'estimation du retard par maximisation de la corrélation intercanale trouve un cadre théorique grâce au modèle gain-retard : que l'on minimise l'erreur absolue ou l'erreur normalisée, le retard optimal est également celui qui maximise la valeur absolue du coefficient de corrélation. Il n'y a pas toutefois équivalence totale avec la méthode visant à maximiser le coefficient de corrélation (sans valeur absolue) : en effet, puisqu'il n'y a *a priori* pas de contrainte de signe sur le gain α, le modèle gain-retard gère les signaux anti-corrélés en fournissant dans ce cas un gain négatif. Cela dit, dans ce contexte précis de localisation de signaux naturels, le cas de gains négatifs, qui correspond à un hors-phase, n'a pas de sens. Pour cette raison, et également pour éviter une partie des ambiguïtés de phase en bande étroite, la minimisation est effectuée avec la contrainte supplémentaire $\alpha > 0$. Dans ce cas, le retard trouvé par cette méthode est égal à celui trouvé par maximisation de la corrélation intercanale.

Lien avec les méthodes de régression linéaire sur la phase

La relation entre spectre et corrélation permet de déduire une formulation fréquentielle de l'erreur : soient $S_{xx}(\nu)$ et $S_{yy}(\nu)$ les **autospectres** de $x[n]$ et $y[n]$, et $S_{xy}(\nu)$ leur **interspectre**. La relation de Wiener-Khintchine permet de les lier respectivement aux autocorrélations et à l'intercorrélation par simple transformée de Fourier. Ainsi, l'interspectre vaut :

$$S_{xy}(\nu) = \sum_{\tau=-\infty}^{+\infty} C_{xy}[\tau].e^{-j2\pi\nu\tau}$$

L'erreur absolue peut être écrite à partir de l'équation IV.17 :

$$D_{\alpha,\tau} = \int_{-\frac{1}{2}}^{+\frac{1}{2}} S_{yy}(\nu).d\nu + \alpha^2.\int_{-\frac{1}{2}}^{+\frac{1}{2}} S_{xx}(\nu).d\nu - 2.\alpha.\int_{-\frac{1}{2}}^{+\frac{1}{2}} S_{xy}(\nu).e^{j.2\pi\nu\tau}.d\nu$$

Puisque l'interspectre $S_{xy}(\nu)$ est à symétrie hermitienne, il vient :

$$D_{\alpha,\tau} = 2.\int_{0}^{+\frac{1}{2}} S_{yy}(\nu).d\nu + 2.\alpha^2.\int_{0}^{+\frac{1}{2}} S_{xx}(\nu).d\nu - 2.\alpha.\int_{0}^{+\frac{1}{2}} |S_{xy}(\nu)|\,.cos\left(.2\pi\nu\tau + \angle S_{xy}(\nu)\right).d\nu \tag{IV.12}$$

La minimisation par rapport au retard τ (qui peut être menée ici de manière analytique) donne donc :

$$\left.\frac{\partial D_{\alpha,\tau}}{\partial\tau}\right|_{\alpha,\tau_{min}} = 0 \Longleftrightarrow \int_{0}^{+\frac{1}{2}} |S_{xy}(\nu)|\,.2\pi\nu.sin\left(.2\pi\nu\tau_{min} + \angle S_{xy}(\nu)\right).d\nu = 0$$

Si l'on suppose maintenant que la quantité $(.2\pi\nu\tau_{min} + \angle S_{xy}(\nu))$, qui est en fait l'erreur de phase, est faible, on peut effectuer un développement limité au premier ordre du sinus[2], qui permet d'obtenir la formulation suivante du retard optimal :

$$\tau_{min} = -\frac{\int_{0}^{+\frac{1}{2}} |S_{xy}(\nu)|\,.2\pi\nu.\angle S_{xy}(\nu).d\nu}{\int_{0}^{+\frac{1}{2}} |S_{xy}(\nu)|\,.\left(2\pi\nu\right)^2.d\nu}$$

[2]Effectuer le développement limité au premier ordre nécessite que la fonction à approcher soit continue et dérivable, ce qui suppose ici que l'on travaille avec la phase déroulée.

Cette formulation rappelle l'équation donnant le retard estimé par régression linéaire de la phase déroulée de l'interspectre, l'ordonnée à l'origine étant nulle :

$$\tau_{regression} = -\frac{\int_{-\frac{1}{2}}^{+\frac{1}{2}} 2\pi\nu.\angle S_{xy}(\nu).d\nu}{\int_{-\frac{1}{2}}^{+\frac{1}{2}} (2\pi\nu)^2 .d\nu}$$

Ainsi, à supposer que la phase soit suffisamment linéaire, l'estimation du paramètre de retard du modèle par égalisation et annulation, et donc la méthode d'estimation de retard par maximisation de la corrélation, équivaut à une **régression linéaire sur la phase déroulée de l'interspectre pondérée par son module**.

2.7 Discussion

On vient de présenter l'application du principe de détection et d'estimation par égalisation et annulation à un cas idéal, qui est celui de signaux d'énergie finie captés par des transducteurs omnidirectionnels sans obstacle entre eux. La tâche de détection se ramène à comparer un indice indépendant du temps, qui est l'indice de détection à un seuil, ce dernier étant défini de manière à distinguer le cas du bruit seul de celui où un signal cohérent d'une voie à l'autre est présent. On montre également que des méthodes couramment employées pour estimer les différences de temps et d'intensité, qui sont respectivement la maximisation de l'intercorrélation et le rapport des puissances, trouvent un cadre théorique cohérent, qui permet de juger, grâce à l'indice de détection, de la **qualité** de l'estimation.

Cependant, il est impossible dans l'état actuel d'appliquer cette méthode à des situations plus complexes : en effet, dans le cas d'enregistrements musicaux par exemple, les signaux émis par les sources sont **non stationnaires**, et puisque c'est également le cas de la réverbération, le rapport signal sur bruit change constamment pendant le déroulement de l'enregistrement. Même si l'on suppose la source physique et les récepteurs fixes (et donc les différences de temps et de niveau indépendantes du temps), il est à craindre qu'une analyse à long-terme comme celle présentée ci-dessus ne puisse fonctionner, la rapport signal sur bruit global (c'est-à-dire calculé sur toute la durée des signaux) étant trop défavorable.

On s'attache donc ci-dessous à proposer une extension du principe permettant d'effectuer le même type d'analyse **à court-terme**, de manière à pouvoir se concentrer sur les instants les plus significatifs au cours du temps.

3 L'ÉGALISATION ET ANNULATION POUR DES SIGNAUX NON STATIONNAIRES

3.1 Préliminaire : corrélation et puissance à court-terme

On définit la **corrélation à court-terme**[3], à l'instant n et pour le retard τ (tous deux en échantillons), notée $C_{xy}[n,m]$, entre deux signaux $x[n]$ et $y[n]$, comme suit :

$$C_{xy}[n,\tau] = \sum_{p=-\infty}^{+\infty} x[p+\tau].y[p].w[n-p] \qquad \text{(IV.13)}$$

...$w[n]$ étant la fenêtre d'intégration, qui est supposée positive ou nulle, et de somme $\sum_{q=-\infty}^{+\infty} w[q]$ égale à 1 . A noter que si la fenêtre est à support temporel infini, constante et égale à 1, on retrouve la définition de la corrélation temporelle à long-terme de signaux d'énergie infinie, soit :

$$C^0_{xy}[\tau] = \lim_{N \to +\infty} \left\{ \frac{1}{2N+1}. \sum_{p=-N}^{+N} x[p+\tau].y[p] \right\}$$

[3]Cette représentation est aussi appelée *correlogram* par Slaney et Lyon (1993), *correlatogram* par Blauert (1997), ou fonction de corrélation locale (*local correlation function*) par Polack (1984)

Puisque la fenêtre est positive ou nulle et de somme finie, elle est par conséquent nulle aux limites :

$$\lim_{q \to +\infty} \{w[q]\} = \lim_{q \to -\infty} \{w[q]\} = 0$$

A partir de cette formulation de la corrélation, on définit la **puissance à court-terme** [4], notée $P_x[n]$, d'un signal $x[n]$ par son autocorrélation à court-terme pour un retard nul :

$$P_x[n] = C_{xx}[n, 0] = \sum_{p=-\infty}^{+\infty} x^2[p].w[n-p] \tag{IV.14}$$

Normalisation de la corrélation à court-terme

La normalisation de la corrélation à court-terme s'effectue, comme dans le cas de la corrélation stationnaire, grâce à l'inégalité de Schwartz. Celle-ci nous indique dans ce cas que :

$$|C_{xy}[n, \tau]| \leq \sqrt{P_x[n+\tau].P_y[n]}$$

On peut alors définir la **corrélation normalisée à court-terme**, notée $\rho_{xy}[n, \tau]$:

$$\rho_{xy}[n, \tau] = \frac{C_{xy}[n, \tau]}{\sqrt{P_x[n+\tau].P_y[n]}}$$

Par construction, on est assuré que cette fonction est à valeurs entre -1 et +1.

On peut noter que plusieurs auteurs utilisent une définition légèrement différente de la corrélation normalisée, qui consiste à remplacer le facteur $P_x[n+\tau]$ par $P_x[n]$. Dans ce cas on ne peut plus affirmer qu'en toutes circonstances la valeur absolue de la corrélation normalisée est inférieure à 1. Néanmoins, si la fenêtre d'analyse est suffisamment longue par rapport aux retards maximaux envisagés, $P_x[n+\tau] \simeq P_x[n]$, si bien que les définitions sont équivalentes.

Choix de la fenêtre

La fenêtre d'intégration joue un grand rôle dans le résultat. Sans chercher à passer en revue tous les types de fenêtre possibles, on peut dégager trois notions primordiales :

- **causalité** : la fenêtre peut être, entre autres, centrée, causale ou anticausale, sachant que tous trois sont équivalents à un décalage temporel près lorsque la fenêtre est à support temporel fini. Les fenêtres causales (c'est-à-dire celles à valeurs sur $\{0...+\infty\}$), en plus d'être incontournables dans le cas d'une analyse en temps réel, assurent comme leur nom l'indique que l'information analysée ne soit pas postérieure à l'instant considérée, et leur utilisation est donc particulièrement pertinente dans une tâche de détection nécessitant une bonne précision temporelle, et donc ici.
- **forme de la fenêtre** : les définitions des corrélations et puissances à court-terme autorisent toute forme de fenêtre, de durée finie ou infinie, pourvu qu'elle soit de somme finie. Cependant, le choix de la fenêtre n'est pas sans conséquences, notamment car il conditionne la précision de la détection dans le domaine temporel : en particulier, si la fenêtre est causale et discontinue en 0, comme c'est le cas par exemple pour des fenêtres rectangulaires, et *a fortiori* exponentielles, le détecteur réagit beaucoup plus rapidement à un nouvel évènement que si la fenêtre est continue. De nombreux auteurs utilisent une fenêtre rectangulaire, entre autres pour sa simplicité de mise en oeuvre .
- **longueur de la fenêtre** : la longueur de la fenêtre est d'une importance capitale vis-à-vis de la stabilité et de la pertinence de la détection : en effet, si celle-ci est trop courte par rapport à l'échelle temporelle des fluctuations du signal (ou des fluctuations de son enveloppe s'il s'agit d'un signal à bande étroite), elle n'assure plus son rôle de moyenne, car les quantités estimées (c'est-à-dire les puissances et corrélations non normalisées)

[4] Cette définition correspond à celle de la **modulation** donnée par Polack (1984)

sont trop fluctuantes. Cela dit, le cas inverse n'est pas non plus souhaitable, car si la fenêtre est trop longue, la résolution temporelle sera réduite, si bien que les événements de trop courte durée seront impossibles à détecter, alors que les événements plus durables pourront être détectés avec un grand retard, et leurs effets masqueront d'autres événements postérieurs d'énergie moindre.

Le choix s'est porté ici sur une fenêtre exponentielle causale, c'est-à-dire une fenêtre du type :

$$w[n] = \begin{cases} \left(1 - e^{-\frac{T_e}{\Delta T}}\right).e^{-\frac{n.T_e}{\Delta T}} & si\ n \geq 0 \\ 0 & si\ n < 0 \end{cases} \tag{IV.15}$$

...où T_e est la période d'échantillonnage, et ΔT est la constante de temps de la fenêtre. Le facteur $\left(1 - e^{-\frac{T_e}{\Delta T}}\right)$ est un facteur de normalisation permettant d'assurer que la somme soit égale à 1. Cette fenêtre répond aux exigences mentionnées ci-dessus, et offre deux avantages qui lui sont propres : d'une part, la discontinuité en 0 permet de détecter plus efficacement les transitoires que d'autres fenêtres de largeur équivalente mais de forme plus douce (du type des fenêtres gaussiennes ou de Hanning) ; d'autre part, elles peuvent être implémentées au moyen d'un filtrage autorégressif du premier ordre, et permettent donc un calcul rapide des corrélations à court-terme. Quant à la valeur de la constante de temps, il a pu être vérifié en pratique qu'il est souhaitable que celle-ci soit au moins égale à 10 fois l'inverse de la bande passante des signaux considérés.

Ce choix d'une fenêtre exponentielle n'est malgré tout pas incontournable, et on peut mettre en évidence en pratique que l'analyse se comporte de manière assez similaire vis-à-vis des attaques du signal avec d'autres formes de fenêtres présentant une discontinuité en 0, comme les fenêtres rectangulaires. La différence principale entre fenêtre exponentielle et fenêtre rectangulaire est le comportement lors des extinctions du signal.

3.2 Détection par égalisation et annulation à court-terme

Muni de ces définitions de la corrélation et de la puissance à court-terme, on est en mesure d'étendre le principe exposé en section 2 à une analyse à court-terme. Les calculs restent identiques, si ce n'est que toutes les quantités dépendent désormais du temps, repéré par l'indice n.

Erreur absolue

L' erreur absolue est maintenant définie par :

$$D_{\alpha,\tau}[n] = \sum_{p=-\infty}^{+\infty} \left(y[p] - \alpha.x[p+\tau]\right)^2 .w[n-p] \tag{IV.16}$$

Si l'on exprime cette erreur en fonction des puissances et corrélations à court-terme liées aux signaux $x[n]$ et $y[n]$, il vient :

$$D_{\alpha,\tau}[n] = P_y[n] + \alpha^2.P_x[n+\tau] - 2\alpha.C_{xy}[n,\tau] \tag{IV.17}$$

Erreur normalisée

La fonction somme vaut :

$$S_{\alpha,\tau}[n] = \sum_{p=-\infty}^{+\infty} \left(y[p] + \alpha.x[p+\tau]\right)^2 .w[n-p] \tag{IV.18}$$

La somme des deux énergies donne :

$$D_{\alpha,\tau}[n] + S_{\alpha,\tau}[n] = 2.\left[P_y[n] + \alpha^2.P_x[n+\tau]\right] \tag{IV.19}$$

On désigne donc par l'erreur normalisée du modèle la fonction suivante :

$$\varepsilon_{\alpha,\tau}[n] = \frac{D_{\alpha,\tau}[n]}{D_{\alpha,\tau}[n] + S_{\alpha,\tau}[n]} \tag{IV.20}$$

$$\varepsilon_{\alpha,\tau}[n] = \frac{1}{2}\left(1 - 2\alpha.\frac{C_{xy}[n,\tau]}{P_y[n] + \alpha^2.P_x[n+\tau]}\right) \tag{IV.21}$$

3.3 Estimation par minimisation de l'erreur normalisée

Dans un formalisme à court-terme, si, à l'instant n, le couple $(\alpha_2[n], \tau_2[n])$ est un minimum de l'erreur normalisée, il obéit à :

$$\begin{cases} \left.\frac{\partial\varepsilon_{\alpha,\tau}[n]}{\partial\alpha}\right|_{\alpha_2[n],\tau_2[n]} = 0 \\ \tau_2[n] = arg\left\{\begin{array}{c} min \\ \tau \end{array}\ (\varepsilon_{\alpha_2,\tau}[n])\right\} \end{cases} \tag{IV.22}$$

La première étape consiste à s'affranchir du gain α. La minimisation de la première équation permet d'exprimer, pour tout retard τ, le gain qui minimise l'erreur normalisée :

$$\frac{\partial\varepsilon_{\alpha,\tau}[n]}{\partial\alpha} = 0 \Leftrightarrow \alpha_{min2(\tau)}[n] = sgn\left(C_{xy}[n,\tau]\right).\sqrt{\frac{P_y[n]}{P_x[n+\tau]}} \tag{IV.23}$$

L'erreur normalisée vaut alors :

$$D_{\alpha_{min2(\tau)}[n],\tau}[n] = 2.P_y[n].\left(1 - |\rho_{xy}[n,\tau]|\right) \tag{IV.24}$$

La minimisation de cette erreur normalisée par rapport à τ (qui équivaut à la maximisation de l'intercorrélation), en vertu de la seconde équation du système IV.22, fournit les valeurs optimales $\alpha_2[n]$ et $\tau_2[n]$ du gain et du retard à l'instant considéré. Soit, pour résumer (et en supposant le coefficient de corrélation positif) :

$$\begin{cases} \tau_2 = arg\left\{\begin{array}{c} max \\ \tau \end{array}\ \left(C_{xy}\left(\tau\right)\right)\right\} \\ \alpha_2 = \sqrt{\frac{E_y}{E_x}} \end{cases} \tag{IV.25}$$

L'indice de détection, qui est, on le rappelle, défini comme étant égal à l'erreur normalisée minimale vaut donc :

$$\varepsilon_{min}[n] = \varepsilon_{\alpha_2[n],\tau_2[n]}[n] = \frac{1}{2}.\left(1 - |\rho_{xy}\left[n\right]|\right) \tag{IV.26}$$

... $\rho_{xy}\left[n\right]$ désignant le coefficient d'intercorrélation à l'instant n.

- Si $\varepsilon_{min}[n] = 0$, alors le modèle est une description fidèle de la réalité dans l'intervalle de temps considéré : $y[p] = \alpha_2.x[p+\tau] \forall p \in \{n...(n+W-1)\}$
- Si $\varepsilon_{min}[n] = \frac{1}{2}$, les signaux x et y sont décorrélés dans la fenêtre temporelle concernée.

3.4 Exemple

La figure IV.2 illustre graphiquement un exemple de détection dans un cas simple : le signal source est un bruit rose (obtenu par filtrage en $\frac{1}{\sqrt{f}}$ d'une séquence de bruit blanc) identique sur les deux voies à un retard et un gain près. Ce signal est noyé dans du bruit blanc dont les deux voies sont indépendantes mais de même puissance ; le rapport signal sur bruit mesuré sur la totalité des signaux est de 15,3 dB à gauche, et de 9,3 dB à droite, mais varie autour de cette moyenne, puisque la puissance instantanée d'un bruit rose est bien plus fluctuante que celle d'un bruit blanc. La fenêtre employée pour le calcul des puissances et corrélations est une fenêtre exponentielle causale, de constante de temps égale à 10 ms.

La théorie suppose le calcul d'une représentation tri-dimensionnelle $\varepsilon_{\alpha,\tau}[n]$ de l'erreur normalisée du modèle gain-retard, fonction du gain α, du retard τ et du temps n. Cependant, le recours à l'expression analytique IV.23, en supposant le coefficient de corrélation positif, permet de ne calculer en pratique que la représentation bi-dimensionnelle $\varepsilon_{\alpha_{min2(\tau)}[n],\tau}[n]$. La minimisation de cette représentation par rapport au retard τ permet de calculer à chaque instant l'indice de détection $\varepsilon_{min}[n]$, le retard optimal $\tau_{min}[n]$, ainsi que le gain optimal $\alpha_{min}[n]$, encore une fois par recours à l'équation IV.23.

Sur la figure IV.2, on peut observer d'une part que le signal est détecté dans le bruit avec une très bonne précision temporelle, l'erreur normalisée $\varepsilon_{min}[n]$ chutant rapidement de 0,5 à 0, dès lors que le signal est présent, aspect que l'on peut constater sur la représentation l'erreur normalisée représentée sur l'étape **D**, qui présente un minimum stable et précisément repéré sur l'axe des retards. Le retard optimal $\tau_2[n]$ est aléatoire en dehors de la zone utile, et, dès lors que le signal à été détecté, se fixe de manière stable à la valeur exacte du retard, soit ici -0,9 ms (l'erreur maximale d'estimation est inférieure à 0.1%). Contrairement à la théorie, le gain optimal $\alpha_2[n]$ n'est pas exactement nul en dehors de la zone utile.

En revanche, dès que le signal est présent, le gain se stabilise autour de la valeur exacte, soit 0,5, avec toutefois plus de fluctuations que pour le retard (l'erreur maximale est d'environ 6% dans ce cas), que l'on peut expliquer aisément grâce au biais dû au bruit, mentionné en section 2 et calculé en annexe B : puisque le signal est un bruit rose, sa puissance à court-terme , et donc le rapport signal sur bruit, présentent d'importantes fluctuations, qui se répercutent dans l'estimation du gain. D'autre part, le fait que le gain ne soit pas nul en l'absence de signal ne pose pas de réel problème, car la connaissance de l'erreur nous indique de toute façon que les valeurs du retard et du gain sont non significatives dans ces zones temporelles.

3.5 Aspect computationnel

Le calcul de l'erreur du modèle et sa minimisation seraient en théorie des opérations très coûteuses[5], si d'une part cette erreur n'était pas une fonction analytique du gain, et d'autre part si l'on ne pouvait avoir recours à une expression sous forme de corrélations et de puissances. Ceci permet de n'avoir jamais à calculer en pratique l'erreur comme représentation tri-dimensionnelle du gain, du retard et du temps, bien que cette représentation soit implicite. Néanmoins, le calcul des corrélations à court-terme est lui-même très coûteux, et il est nécessaire de disposer de méthodes de calcul optimisées. De ce point de vue, l'utilisation de fenêtres de type rectangulaire ou, comme ici, exponentielle, est un atout indéniable.

Un autre aspect qui mérite d'être mentionné est que le fait que les signaux soient échantillonnés empêche de connaître la corrélation pour des retards non multiples de la période d'échantillonnage. Ceci est particulièrement gênant dans le cas d'une détection de type binaural, où les retards sont généralement confinés dans la plage $[-0,7\,ms; 0,7\,ms]$, ce qui ne représente qu'une soixantaine de possibilités. En écartant d'emblée le suréchantillonnage qui est très coûteux, on peut néanmoins utiliser une méthode d'estimation de retards non entiers très répandue, qui est l'interpolation parabolique : en interpolant les corrélations et puissances autour du retard entier optimal, on peut fournir une estimation correcte de retard non entier.

4 DÉTECTION EN BANDES LIMITÉES

Dans beaucoup de situations, le modèle de détection par égalisation et annulation proposé en section 3 n'est pas directement applicable : en effet, il a déjà été mentionné que, puisque la directivité de tout microphone varie en fonction de la fréquence, et/ou à cause d'un éventuel obstacle placé entre eux, les différences de temps et d'intensité dépendent de la fréquence, et c'est justement le cas dans une prise de son binaurale. Dans ce cas, le modèle gain-retard n'est pas adéquat, et l'erreur sera importante. De plus, puisque les signaux sources n'occupent pas forcément toute le spectre audible, une estimation en pleine bande

[5] Ce qui est d'autant plus vrai lorsque l'analyse est effectuée en bandes étroites (voir section 4)

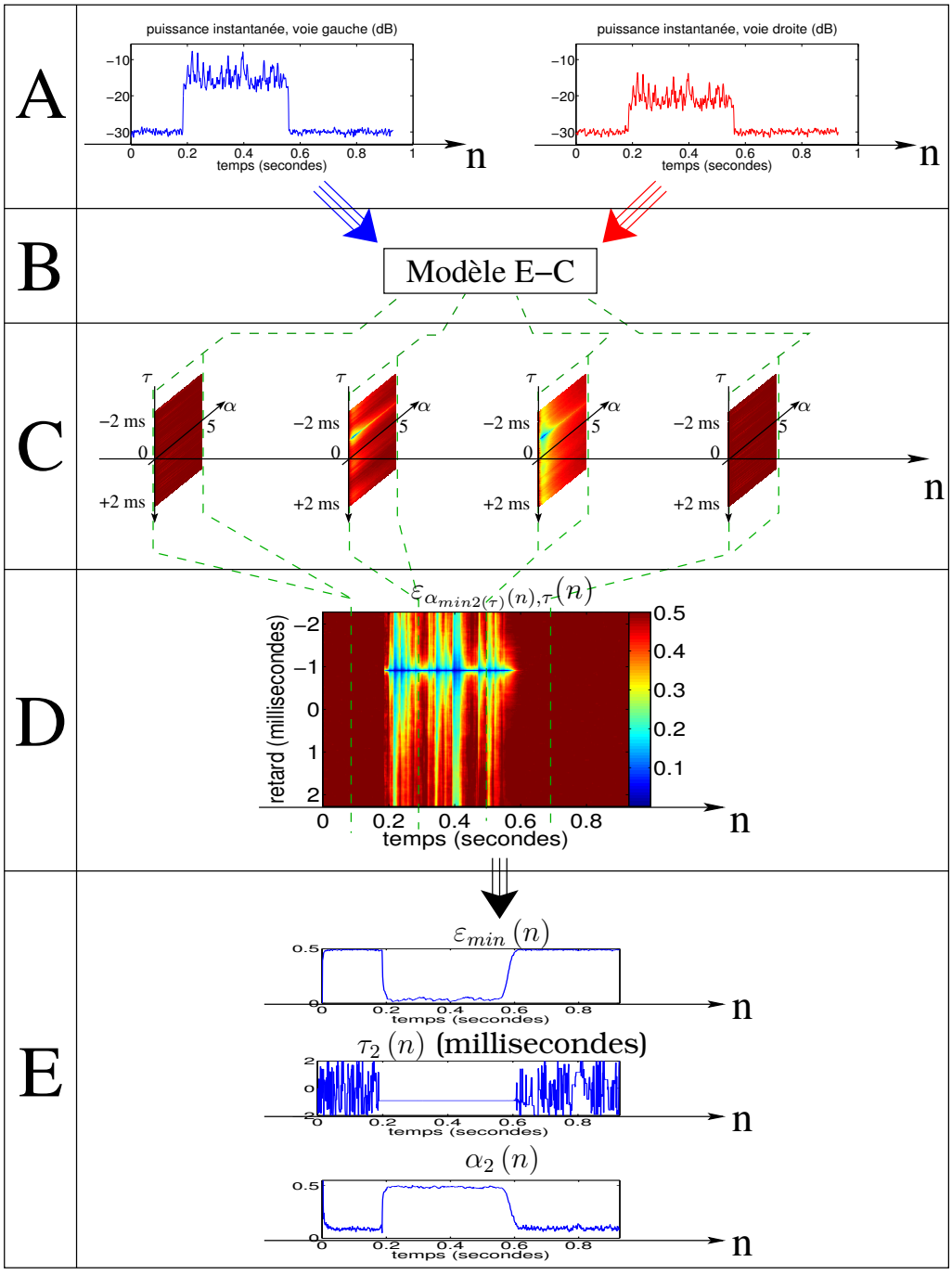


FIG. IV.2 – **Exemple de détection par égalisation et annulation** : le signal source est une séquence de bruit rose d'une durée de 380 ms, identique sur les deux voies à un retard (0,9 ms) et un gain (0,5) près, et auquel se superpose une séquence de bruit blanc par voie (les puissances à court-terme sont affichées en **A**). Le modèle gain-retard (étape **B**) permet de calculer à chaque instant n un motif d'erreur d'égalisation et annulation fonction d'un gain α et d'un retard τ variables (étape **C**). La connaissance de la relation entre gain optimal $\alpha_{min2(\tau)}[n]$ et retard permet de réduire cette représentation à une représentation temps-retard $\varepsilon_{\alpha_{min2(\tau)}[n],\tau}[n]$(étape **D**). Enfin, la recherche à chaque instant du retard optimal $\tau_2[n]$ permet de calculer l'indice de détection (c'est-à-dire l'erreur minimale) et les paramètres optimaux du modèle (étape **E**).

ne fournira pas le même résultat en fonction de la bande passante du signal. Cette instabilité de la détection et de l'estimation rend difficile toute tentative de localisation.

Cela dit, cette dépendance fréquentielle est en fait une riche source d'informations, car dans l'hypothèse où l'on connaît avec une précision suffisante le comportement de la fonction de transfert interaurale en fonction de la direction de provenance de la source, elle permet de préciser un peu mieux encore l'estimation de la position, notamment en permettant une meilleure résolution de la localisation au sein du cône de confusion. Mais il est nécessaire pour pouvoir en tenir compte d'effectuer l'estimation bande de fréquences par bande de fréquences. Un tel traitement permet d'une part de fournir des estimations plus stables (puisque l'erreur du modèle diminue avec la bande passante), et d'autre part de tenir implicitement compte du rapport signal sur bruit dans chaque bande : en effet, si une bande de fréquences ne contient pas ou peu d'énergie relative au signal, mais uniquement du bruit de fond ou de la réverbération, l'erreur dans cette bande sera importante, et l'estimation ne sera donc pas retenue.

4.1 Principe d'incertitude

L'extension du modèle gain-retard au cas de bandes limitées issues d'une analyse par transformée de Fourier à court-terme ou d'un banc de filtres se heurte à la limitation intrinsèque de cette dernière qui est sa résolution temps-fréquence. En effet, toute analyse temps-fréquence ou temps-échelle est soumise au principe d'incertitude d'Heisenberg-Gabor, en vertu duquel la résolution temporelle d'une fenêtre $h(t)$ et la résolution fréquentielle de sa transformée de Fourier $H(f)$ sont soumises à l'inégalité suivante, quels que soient l' instant t_0 et la fréquence f_0 considérés :

$$\Delta t.\Delta f \geq \frac{1}{\pi} \tag{IV.27}$$

...où Δt et Δf désignent les résolutions en temps et en fréquence de la fenêtre $h(t)$, définies à partir des moments (temporel et spectral) centrés au deuxième ordre de la fenêtre :

$$\Delta t^2 = 2.\frac{1}{\|H\|_2} \|(t-t_c).h(t)\|_2 = 2.\frac{\int_{-\infty}^{+\infty} (t-t_c)^2 .h^2(t).dt}{\int_{-\infty}^{+\infty} h^2(t).dt}$$

$$\Delta f^2 = 2.\frac{1}{\|H\|_2} \|(f-f_c).H(f)\|_2 = 2.\frac{\int_{0}^{+\infty} (f-f_c)^2 .|H(f)|^2 .df}{\int_{0}^{+\infty} |H(f)|^2 .df}$$

...$\|H\|_2$ désignant l'énergie de la fenêtre, et t_c et f_c ses centroïdes temporel et fréquentiel, définis par :

$$t_c = \frac{\int_{-\infty}^{+\infty} t.h^2(t).dt}{\int_{-\infty}^{+\infty} h^2(t).dt}$$

et

$$f_{c=}\frac{\int_{0}^{+\infty} f.|H(f)|^2 .df}{\int_{0}^{+\infty} |H(f)|^2 .df}$$

L'égalité pour l'équation IV.27, qui donne un compromis optimal entre résolution temporelle et fréquentielle, est obtenue pour une fenêtre gaussienne, appelée dans ce cas fenêtre de Gabor.

Cette question de la résolution temps-fréquence concerne directement le problème abordé ici, car elle indique que la résolution temporelle de la détection dépend directement de la résolution fréquentielle. Or, dans le cas d'un milieu réverbérant, la détection, et donc l'estimation, ne pourront être effectuées avec une bonne précision **que pendant le laps de temps entre l'arrivée de l'onde directe et celle de la première réflexion**, comme on le verra au chapitre VI. Or ce laps de temps est très court (typiquement de l'ordre d'une dizaine de millisecondes), et il sera difficile d'effectuer une détection de bonne qualité si la résolution temporelle est trop mauvaise.

Il est donc nécessaire de faire un compromis entre résolution temporelle et résolution fréquentielle, en jouant sur la forme et la longueur de la réponse impulsionnelle du filtre pour une bande donnée : si l'on opte pour un filtre court par rapport aux retards envisagés et à la résolution temporelle recherchée, on obtient une bonne résolution temporelle, mais on limite la résolution fréquentielle. Au contraire, si l'on opte pour un filtre long, on perd en précision temporelle, mais on gagne en précision fréquentielle.

4.2 Choix du banc de filtres

Le choix du banc de filtres dépend donc du problème à traiter. Les caractéristiques principales d'un filtre sont sa résolution temporelle Δt, sa résolution fréquentielle Δf (toutes deux définies ci-dessus), ainsi que sa résolution en phase, cette dernière étant déterminée par le **facteur de qualité**. Il existe plusieurs définitions du facteur de qualité (la plus courante est définie à partir de la bande passante à -3 dB), celle employée dans cette étude est la suivante :

$$Q = \frac{\Delta f}{f_c}$$

...les quantités Δf et f_c correspondant aux définitions ci-dessus.

Sans entrer dans le détail (la question du choix du banc de filtres pour la détection bicanale sera abordée plus spécifiquement au chapitre VI), on peut néanmoins mentionner certaines des catégories les plus courantes de bancs de filtres :

Bancs de filtres uniformes On peut classer dans cette catégorie la transformée de Fourier (discrète) à court-terme (TFCT), la transformée en cosinus discrète à court-terme et sa version modifiée (*Modified Discrete Cosine Transform*, ou MDCT). Ces filtres sont, par définition, à largeur de bande constante, et le facteur de qualité est donc variable, faible en basses fréquences, et élevé en hautes fréquences, si bien que si la résolution temporelle du banc est constante, sa résolution en phase dépend de la fréquence. Ils ont pour principaux avantages d'être implémentables au moyen d'algorithmes rapides de type FFT, et de permettre une reconstruction parfaite du signal (sous certaines conditions). En revanche, puisque les bancs de filtres uniformes donnent la même importance à toutes les fréquences, et puisque les indices les plus pertinents pour la localisation de signaux naturels se situent entre quelques centaines de hertz et environ 6 kHz, ce type de filtres peut avoir tendance à privilégier les hautes fréquences au détriment des basses fréquences, si l'on ne prend pas les précautions nécessaires, par exemple en diminuant le poids relatif des canaux hautes fréquences. De toute manière, il est nécessaire pour obtenir une bonne résolution fréquentielle en basses fréquences d'avoir un grand nombre de filtres au total, ce qui limite l'intérêt de disposer de méthodes de calcul rapides.

Bancs de filtres temps-échelle On peut classer dans cette catégorie la transformée en ondelettes discrète, ainsi, entre autres, que la plupart des bancs de filtres en tiers d'octave ou en octave. Les représentations temps-échelle sont définies au moyen de la dilatation ou la contraction temporelle d'un filtre donné. Dans ce cas le facteur de qualité est constant, si bien que la résolution temporelle est mauvaise en basses fréquences, et bonne en hautes fréquences (et inversement pour la résolution fréquentielle). Tous les filtres temps-échelle ne permettent pas forcément une reconstruction parfaite, mais cette question est secondaire vis-à-vis de cette étude. En revanche, le fait que la densité de filtres et leur résolution soit plus élevée en basses fréquences qu'en hautes fréquences est d'un grand intérêt, conformément aux remarques mentionnées ci-dessus.

Filtres auditifs Il s'agit de filtres hybrides (c'est-à-dire ni uniformes, ni temps-échelle) visant à modéliser l'analyse tonotopique effectuée par le système auditif. Cette modélisation peut être effectuée selon deux angles d'approches :

Certains filtres modélisent la physiologie de la cochlée, et plus particulièrement son comportement hydromécanique. Ils comprennent généralement une partie linéaire (filtrage par

les oreilles externes et moyenne, réponse passive de la cochlée sous forme d'une batterie de filtres passe-bas résonnants) et une partie non-linéaire, constituée d'une part par la modélisation de la transduction de la vibration de la membrane basilaire en décharges nerveuses effectuée par les cellules ciliées internes, et d'autre part par la prise en compte du rôle des cellules ciliées externes, entre autres dans la dépendance de la raideur de la membrane basilaire au niveau sonore. C'est par exemple le cas du modèles de filtrage auditif proposé par Lyon, et implémenté sous forme numérique par Slaney (1988).

L'autre grande classe de filtres auditifs résulte d'un point de vue psychophysique de l'audition, et a pour but de représenter sous forme de filtrage linéaire la notion de masquage fréquentiel. Diverses implémentations ont été proposées, utilisant soit l'échelle des **bandes critiques** de Zwicker, soit celle des **bandes rectangulaires équivalentes** (*Equivalent Rectangular Bandwidth*s, ou ERB). Les deux échelles sont assez proches, et notamment dans les deux cas, la largeur de bande augmente avec la fréquences en moyennes et hautes fréquences. La différence intervient surtout en basses fréquences (en-deçà de 1 kHz) : les bandes critiques sont alors quasiment constantes, alors que les bandes rectangulaires équivalentes sont faiblement croissantes avec la fréquence. La question de la forme des filtres est d'importance, la largeur des bandes rectangulaires équivalentes n'étant pas suffisante pour les caractériser. La détermination de la forme du filtre est issue de tests psychoacoustiques, par exemple en étudiant le masquage d'un son pur par un bruit dont une bande a été coupée ("*notched noise masker*") (Hartmann, 1996). Ainsi, Patterson et al. (1982) proposent les filtres exponentiels arrondis (*rounded exponential filter*, ou **roex**) comme approximation du filtrage auditifs.

Bien que destinés premièrement à des fins de modélisation de l'audition, ces filtres, et en particulier ceux issus de la deuxième catégorie, qui sont généralement linéaires, peuvent être envisagés dans des situations plus pratiques, et on verra que leurs caractéristiques propres, notamment la loi reliant le facteur de qualité à la fréquence centrale, permettent d'envisager leur utilisation dans un contexte plus général. Le principal défaut de ce type de filtrage est son coût de calcu l élevé, et particulièrement pour les modèles cochléaires. Certaines implémentations sont néanmoins plus légères, au prix de quelques approximations. Ainsi, les filtres dits *gammatone* (Patterson et al., 1991; Slaney, 1993) sont une implémentation des filtres *roex* sous forme d'un filtrage ARMA du quatrième ordre, et sont donc peu coûteux, et de plus causaux.

Aspect computationnel

Si l'on applique les méthodes de détection par égalisation et annulation sur les signaux directement issus du banc de filtres, on décuple la quantité de mémoire nécessaire pour les données, ainsi que le coût de calcul de détection, par rapport à la situation de pleine bande. Or il est possible en partie (en fonction du banc de filtres) de ce surcroît de charge computationnelle si les signaux sont sous-échantillonnés par décimation en sortie du banc de filtres. Cependant, le facteur de sous-échantillonnage maximal admissible pour ne pas perdre d'informations étant directement fonction de la largeur de la bande considérée, il faut opérer une décimation différente dans chaque bande lorsque le banc de filtres est non uniforme.

4.3 Retard d'enveloppe

Puisque les signaux considérés sont à bande limitée, il est possible de considérer séparément le **signal porteur** (typiquement une sinusoïde) de l'**enveloppe**. Jusqu'ici, on s'est attaché à l'estimation du retard intercanal dans l'hypothèse où celui-ci était indépendant de la fréquence. Dans ce cas, la phase est linéaire, et l'estimation du retard est sans ambiguïté. Cependant, si la transmission est plus complexe, la phase ne peut plus être considérée comme linéaire. Dans ce cas, et si l'on se place en bandes limitées, il est utile de faire la distinction entre **retard de porteuse** et **retard d'enveloppe**. Le retard estimé par maximisation de la corrélation des signaux correspond alors au retard entre les porteuses des signaux. Or le **retard d'enveloppe** est une quantité tout aussi importante que le retard de phase en

transmission de l'information en bandes étroites, et ne pas en tenir compte revient à négliger une grande partie de l'information contenue dans la phase. D'ailleurs, la plupart des modèles de détection binaurale, par souci de cohérence avec la physiologie du système auditif, proposent conjointement à l'estimation des retards interauraux de phase une estimation des retards interauraux d'enveloppe. Ces notions de retard de porteuse et de retard d'enveloppe sont également à mettre en relation avec celles, respectivement, de **retard de phase et de retard de groupe** , en analyse de Fourier.

L'estimation du retard d'enveloppe grâce au modèle gain-retard suppose au préalable une estimation de l'enveloppe des signaux analysés. Cette **détection d'enveloppe** peut être menée de plusieurs façons :

Détection de l'enveloppe par transformation de Hilbert

La transformée de Hilbert permet un calcul rigoureux de l'enveloppe d'un signal à bande étroite, lorsque sa bande passante ne contient pas la fréquence nulle. En effet, un tel signal peut-être écrit sous la forme d'une modulation d'amplitude :

$$x(t) = e_x(t).cos\left(2\pi f_0 t + \varphi\right)$$

Dans cette notation, $e_x(t)$ représente l'enveloppe du signal, ou modulation, et f_0 est la fréquence de la porteuse. Si la bande passante du signal est suffisamment étroite pour ne pas contenir la fréquence nulle (c'est-à-dire si le module $|X(f)|$ de sa transformée de Fourier est à valeurs sur $\left[-f_0 - \frac{B}{2}, -f_0 + \frac{B}{2}\right] \cup \left[f_0 - \frac{B}{2}, f_0 + \frac{B}{2}\right]$, B étant la largeur de bande, avec $f_0 \gg \frac{B}{2}$), alors le signal analytique correspondant, calculé par transformation de Hilbert, s'écrit :

$$\widetilde{x}(t) = e_x(t).exp\left(j\left(2\pi f_0 t + \varphi\right)\right)$$

Si l'on suppose de plus que $e_x(t)$ est à valeurs positives, alors :

$$|\widetilde{x}(t)| = e_x(t)$$

Détection de l'enveloppe par redressement et filtrage passe-bas

Le principe général de ces méthodes de détection est de ramener le signal en bande de base, par l'intermédiaire d'une transformation simple qui entre autres crée une composante continue sur la porteuse, suivie d'un filtrage passe-bas dont le but est de ne conserver que cette composante continue. Les transformations de ce type les plus courantes sont la mise au carré, et le redressement simple ou double alternance, ce dernier visant respectivement à annuler ou inverser le signe du signal lorsque la porteuse est à valeurs négatives.

Pour que cette transformation permette de détecter l'enveloppe avec précision, il est nécessaire que la fréquence de coupure du filtre soit dans l'intervalle $\left[\frac{B}{2}, f_0 - \frac{B}{2}\right]$. Ceci sous-entend bien sûr encore une fois que cette bande passante soit suffisamment étroite, c'est-à-dire que $f_0 \gg \frac{B}{2}$. De plus, il faut que le filtre soit un passe-bas idéal. Si ce n'est pas le cas, la sortie du détecteur pourra contenir une partie de l'information autour de la fréquence de la porteuse, et sera donc un mélange de l'enveloppe idéale et du signal initial.

Ces techniques trouvent un équivalent dans la physiologie de l'audition : en effet, les stéréocils, qui sont à la base de la transduction mécano-électrique effectuée au sein des cellules ciliées internes, ne permettent l'ouverture des canaux ioniques qu'une alternance sur deux, si bien que les modèles les plus simples de cellules ciliées sont constitués d'un redressement simple alternance suivi d'un filtrage passe-bas du premier ordre à une fréquence de l'ordre de 800 Hz (Blauert et Cobben, 1978). Des modèles plus sophistiqués, comme celui de Schroeder et Hall (1974) ou celui de Meddis (1986) permettent, notamment grâce à une prise en compte de l'évolution de la quantité de neurotransmetteurs, de tenir compte d'autres phénomènes, comme la sensibilité aux attaques[6]. Ces modèles sont cohérents avec la sensibilité

[6]Il est de plus nécessaire de rappeler que tous ces modèles ne représentent pas la valeur instantanée de l'intensité électrique dans la fibre nerveuse, mais soit la **probabilité** de décharges nerveuses à tout instant, soit le potentiel de membrane

de l'audition aux retards de phase et d'enveloppe : en effet, aux basses fréquences, l'enveloppe n'est pas détectée, car la porteuse est de fréquence trop basse, si bien que les retards estimés sont des retards de phase. En hautes fréquences (c'est-à-dire pour des fréquences au moins deux fois supérieures à la fréquence de coupure du détecteur, soit entre 1,5 kHz et 2 kHz), seule l'enveloppe est présente, la porteuse n'est plus présente à la sortie du détecteur, et les retards estimés sont des retards d'enveloppe. Aux fréquences intermédiaires, la sortie du détecteur est un mélange de l'enveloppe et du signal modulé redressé, ce qui assure une transition douce des retards de phase vers les retards de groupe.

4.4 Ambiguïté de phase en bande étroite

L'équation IV.12, appliquée à des signaux à bande étroite, permet de mettre en évidence un problème inhérent à toutes les méthodes d'estimation de retard par corrélation, qui est l'ambiguïté de phase en bande étroite. Pour s'en convaincre, on peut rappeler que si les signaux $x[n]$ et $y[n]$ sont à bande très étroite, alors c'est aussi le cas pour la corrélation $C_{xy}[n,\tau]$. Ainsi, dans la notation de l'équation IV.12, on peut considérer que le facteur $cos\left(2\pi\nu\tau + \angle S_{xy}(n,\nu)\right)$est quasiment constant sur la bande considérée. En définissant l'erreur de phase $\varepsilon_{ph}(n,\nu)$ par

$$\varepsilon_{ph}(n,\nu) = \angle S_{xy}\left(n,\nu_c\right) + 2\pi\nu\Delta$$

, il vient donc :

$$\begin{aligned} D_{\alpha,\tau}[n] \simeq \quad & 2.\int_0^{+\frac{1}{2}} S_{yy}(n,\nu).d\nu + 2.\alpha^2.\int_0^{+\frac{1}{2}} S_{xx}(n,\nu).d\nu \\ & -2.\alpha.cos\left(2\pi\nu_c\left(\tau-\Delta\right) + \varepsilon_{ph}(n,\nu)\right)\int_0^{+\frac{1}{2}} \left|S_{xy}(n,\nu)\right|.d\nu \end{aligned}$$

...ν_c étant la fréquence centrale de la bande. On voit directement apparaître le problème, qui est que dans ce cas limite, l'erreur est minimale pour toutes les valeurs de τ égales à Δ **modulo la période** $\frac{1}{\nu_c}$. La figure IV.3 montre un exemple pratique de cette ambiguïté sur un cas simple : les signaux sources sont deux bruits blancs identiques à un retard près de 4,5 ms, traités par un filtre basse-bande de Butterworth d'ordre 2, de fréquence centrale fixe $f_0 = 2\,kHz$ et de largeur de bande variable. On peut observer que pour un faible facteur de qualité, le retard est repérable sans ambiguïté, avec une résolution quantifiable par la largeur temporelle du lobe principal, qui est de l'ordre de $\frac{1}{2.f_0}$ (cette valeur correspond à un filtre parfaitement rectangulaire). Lorsque le facteur de qualité augmente, les lobes secondaires prennent de plus en plus d'importance, et bien que le lobe principal ne s'élargisse pas, il devient difficile de discriminer quel maximum local correspond au retard initial.

En fait cette ambiguïté se présente également lorsque les signaux considérés sont harmoniques, ou plus généralement s'ils sont constitués de partiels aux mêmes fréquences sur les deux canaux. En ce cas, la corrélation est périodique par rapport à l'axe des retards, et l'erreur sera donc définie à la période fondamentale près. Il est également important de noter que cette ambiguïté ne peut exister que pour des fréquences supérieures à un seuil critique valant $\frac{1}{2.\tau_{max}}$, τ_{max} étant le retard maximal envisageable entre les deux signaux sources. En effet, en-deçà de ce seuil, l'intervalle de temps considéré ne contient pas une période entière, si bien qu'un seul maximum de la corrélation est visible. Cette ambiguïté existe en écoute binaurale, comme cela est indiqué à la section 2.2 du chapitre II : dans ce cas, le retard maximal varie de 0,7 millisecondes à 1 milliseconde en fonction des individus, et la fréquence minimale à partir de laquelle l'ambiguïté est possible varie donc entre 500 et 700 Hz. Cette fréquence de coupure dépend de l'azimut de la source : ce seuil minimal correspond à une source latérale (et implique une difficulté de discrimination gauche-droite de signaux à bande étroite à ces fréquences lorsque ceux-ci sont fortement latéralisés), alors que pour une source physique frontale le seuil est deux fois plus important, si bien qu'une ambiguïté n'est envisageable que pour des fréquences supérieures à 1kHz.

Il ne s'agit pas à proprement parler d'une imprécision due à la méthode d'estimation employée, puisque cette ambiguïté est intrinsèque au signal lui-même, et à sa nature quasi-harmonique. Pour tenter de lever cette ambiguïté d'estimation, il est nécessaire de disposer

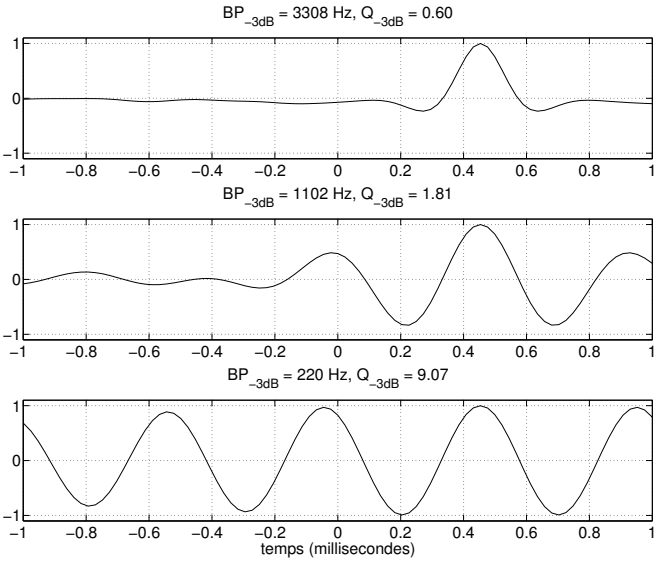


FIG. IV.3 – Intercorrélation de deux bruits blancs identiques à un retard près de 4,5 ms, traités par un filtre passe-bande (filtre de Butterworth d'ordre 2) de fréquence centrale $f_0 = 2\,kHz$ et de largeur variable, indiquée par la bande passante BP_{-3dB} et le facteur de qualité à -3 dB $Q_{-3dB} = \frac{f_0}{BP_{-3dB}}$

d'une part de la corrélation sur plusieurs bandes de fréquences, et d'autre part d'un **modèle** de signaux, lié à la spécificité de la propagation physique entre la source et la configuration de prise de son considérée, qui permette de lier entre elles les estimations des retards et gains sur toutes les bandes considérées. Ainsi, si l'on considère le cas spécifique d'une prise de son effectuée au moyen de deux microphones omnidirectionnels distants idéaux, sans obstacle entre eux et en l'absence de réverbération, le retard est indépendant de la fréquence. Ce cas est illustré par la figure IV.4 : l'intercorrélation présente un ou plusieurs maxima par bande, dont les positions sur l'axe des retards varient avec la fréquence centrale du filtre. Quel que soit le banc de filtres employé, l'un des maxima cependant est commun à toutes les bandes, et correspond au retard entre les deux signaux. Une sommation (ou, de manière équivalente, une moyenne) de l'intercorrélation normalisée sur toutes les bandes permet de le faire clairement apparaître.

Cette idée de sommation du motif de corrélation (appelée également *summary correlation* ou *summary correlogram*) est notamment utilisée au sein de modèles monauraux d'analyse computationnelle de scènes auditives (Brown et Cooke, 1994; Slaney et Lyon, 1993; Ellis, 1996) ; elle y sert non seulement à estimer la période fondamentale d'un signal harmonique (la période étant le pendant monophonique du retard intercanal), mais également de faire une analyse de plus haut niveau, comme une reconnaissance phonétique sur des signaux de parole. Elle peut être appliquée au cas de modèles binauraux de latéralisation (Pulkki, 1999; Schauer et Paschke, 1999), mais non sans difficultés, car l'hypothèse d'un retard indépendant de la fréquence n'est plus valable dans le cas d'une écoute binaurale, et en particulier lorsque la source est fortement latéralisée. On peut se rendre compte de cette difficulté grâce à la figure IV.5 : le bruit blanc utilisé dans le cas précédent est cette fois-ci filtré par un couple de HRTF mesurées sur une tête artificielle KEMAR par Gardner et Martin (1994), dans le plan horizontal et pour un azimut de 60°. Puisque le maximum de corrélation se déplace d'une bande à l'autre, et puisque de plus celui-ci n'est même plus égal à 1, le maximum de la somme sur toutes les bandes n'est plus aussi clairement défini : si l'on est malgré tout encore en mesure de l'estimer, il ne peut servir que d'indication permettant de sélectionner dans chaque bande le retard parmi toutes les possibilités. Pullki propose d'effectuer

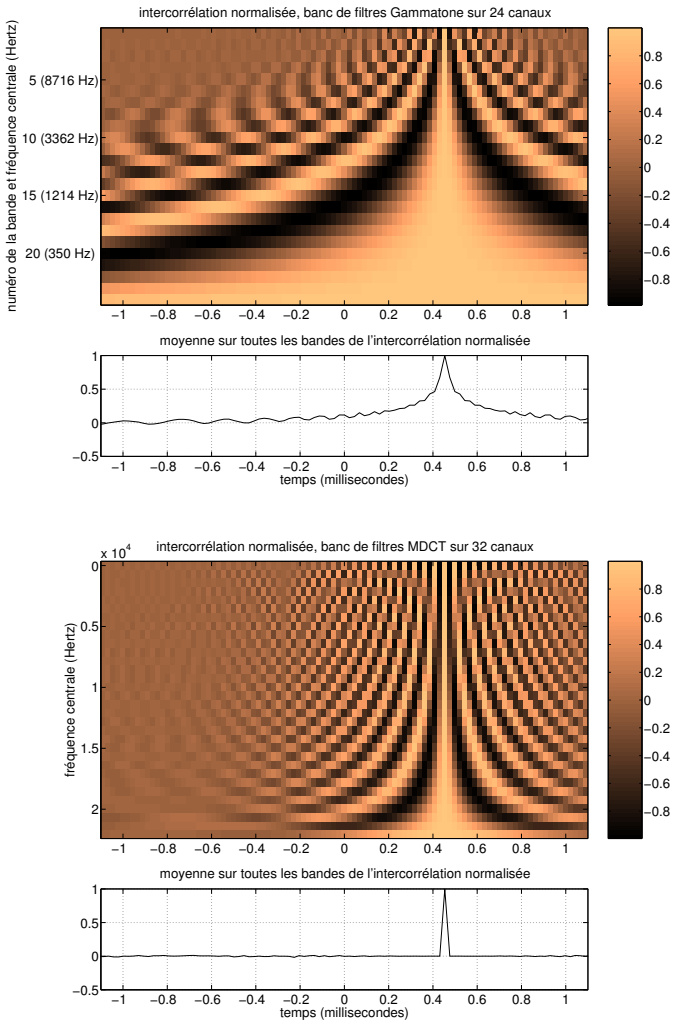


FIG. IV.4 – **Ambiguïté de phase en bande étroite (1)** : Motif l'intercorrélation de deux bruits blancs à 44100 Hz, identiques à un retard près traités par un banc de filtres Gammatone à 24 canaux (haut) ou MDCT à 32 canaux (bas), et moyenne sur toutes les bandes de fréquences.

cette sommation uniquement pour les bandes de fréquence centrale inférieure à 1,5 kHz afin d'éviter l'ambiguïté, mais même à ces fréquences, le maximum de l'autocorrélation présente des variations qui sont loin d'être négligeables (variations principalement dues à la mauvaise résolution temporelle en basses fréquences, aspect qui sera explicité en section 4.5), comme nous le montre par exemple la figure IV.6. Effectuer la somme entraîne irrémédiablement une perte d'information.

4.5 Problème de la résolution sur l'estimation du retard en basses fréquences

La remarque de la section précédente, selon laquelle la largeur du lobe principal de l'autocorrélation est inversement proportionnelle à la fréquence centrale du filtre considéré, amène à une autre source d'imprécision, liée cette fois-ci à la **résolution temporelle** de l'estimation. Ce problème, qui a déjà été mentionné à la section 2.6 à partir d'une formulation fréquentielle, est intimement lié à la présence de bruit additionnel : en effet, en l'absence de bruit, le maximum de la corrélation est indépendant de la fréquence. Cependant, si la résolution temporelle est faible, ce qui est le cas en basses fréquences, la corrélation est quasiment

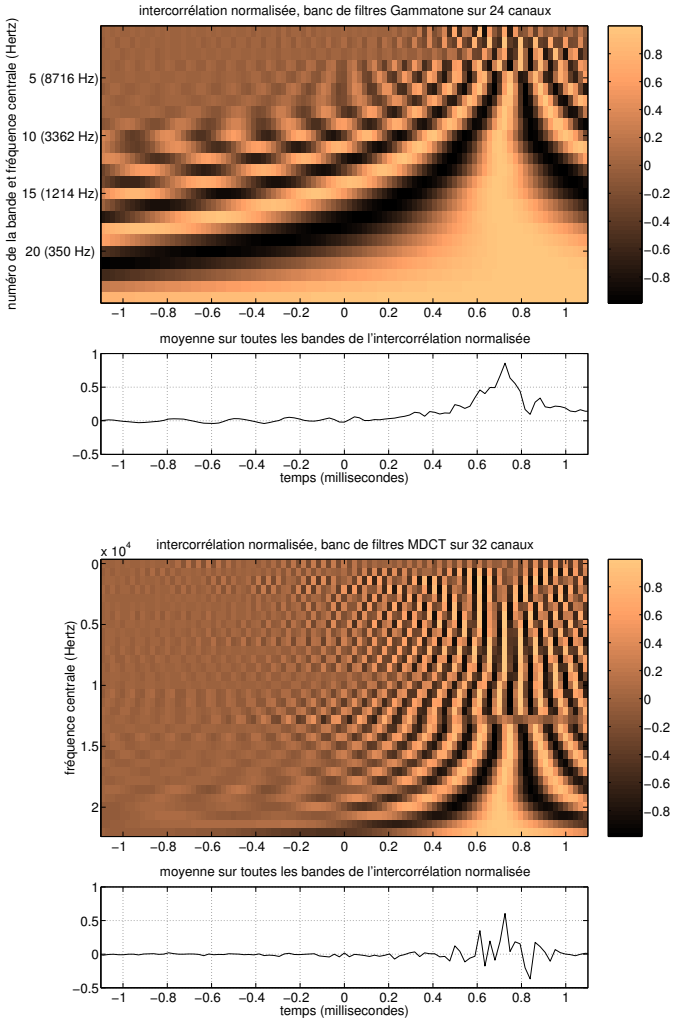


FIG. IV.5 – **Ambiguïté de phase en bande étroite (2) :**Motif l'intercorrélation d'un signal bicanal issu de la convolution d'un bruit blanc par un jeu de HRTF (KEMAR) à 60° d'azimut et 0° de site traités par un banc de filtres Gammatone à 24 canaux (haut) ou MDCT à 32 canaux (bas), et moyenne sur toutes les bandes de fréquences.

constante sur le domaine de retards considérés (typiquement $[-1ms; 1ms]$ dans le cas binaural), et la détection du maximum est extrêmement sensible à la présence de perturbations.

La figure IV.7 illustre ce propos. Le signal source considéré est toujours un bruit blanc échantillonné à 44,1 kHz, et retardé d'une voie à l'autre. Un bruit blanc est ajouté sur chacune des voies (les deux bruits blancs parasites sont indépendants entre eux), le rapport signal sur bruit étant de 40 dB. Pour les fréquences centrales élevées (500 Hz et 2kHz), l'estimation est robuste malgré le bruit, alors que pour une fréquence centrale de 50 Hz, l'estimation est complètement faussée.

La nature de la perturbation est la corrélation des deux voies du bruit, et la corrélation entre le bruit et le signal. En effet, bien que les deux voies du bruit soient sensées être décorrélées entre elles et décorrélées du signal, le fait que l'intercorrélation soit calculée sur une durée finie (ici 370 ms) entraîne inévitablement une surestimation en basses fréquences. Le corollaire est que **la résolution de l'estimation du retard dépend de la longueur de la fenêtre utilisée**. Dans le cas d'une salle, ce problème est encore plus présent, puisque le champ diffus (qui est la principale source de bruit) est par nature corrélé en basses fréquences. Si l'on ajoute à ce constat le fait que dans une situation réelle, les signaux diffèrent

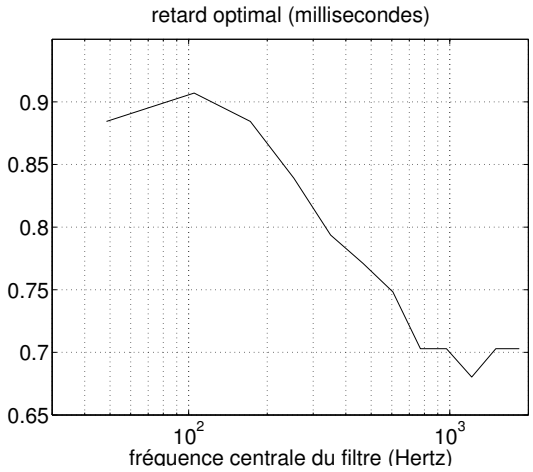


FIG. IV.6 – **Dépendance fréquentielle de l'estimation du retard interaural en basses fréquences** : Retard optimal par bandes de fréquence pour un signal bicanal issu de la convolution d'un bruit blanc par un jeu de HRTF (KEMAR) à 60° d'azimut dans le plan horizontal traités par un banc de filtres Gammatone à 24 canaux

peu en intensité en basses fréquences, puisque la différence de marche y est faible, on arrive à la conclusion que le modèle par égalisation et annulation ne permet pas d'estimer correctement les gains et retards intercanaux en basses fréquences dès lors que le rapport signal sur bruit est défavorable.

5 CAS PARTICULIER DE LA DÉTECTION À PARTIR DE LA TRANSFORMÉE DE FOURIER À COURT-TERME

Le cas de la **transformée de Fourier à court-terme**, ou **transformation de Gabor discrète**, mérite d'être traité séparément des autres bancs de filtres, ne serait-ce que parce que les filtres utilisés sont ici complexes. En effet, il est alors nécessaire de modifier légèrement la définition de la fonction différence pour qu'elle continue à fournir des valeurs réelles et positives. D'autre part, il est nécessaire de déterminer si l'on utilise pour le calcul de la transformée de Fourier à court-terme la convention passe-bande ou la convention passe-bas, la formulation n'étant pas tout-à-fait identique dans ce dernier cas à cause du filtrage hétérodyne à la sortie du banc de filtres (Crochiere et Rabiner, 1983). On peut rappeler ces deux écritures de la transformée à court-terme d'un signal $x[n]$:

– en convention passe-bande :

$$\tilde{X}_R[p,k] = \sum_{n=0}^{N-1} g[n].x[pR+n].e^{-j2\pi\frac{k}{N}n} \tag{IV.28}$$

– en convention passe-bas :

$$X_R[p,k] = e^{-j2\pi\frac{k}{N}pR}.\tilde{X}_R[p,k] \tag{IV.29}$$

$$X_R[p,k] = \sum_{n=pR}^{pR+N-1} g[n-pR].x[n].e^{-j2\pi\frac{k}{N}n} \tag{IV.30}$$

Dans ces deux définitions, $g[n]$ est la fenêtre d'analyse, à valeurs sur $\{0...(N-1)\}$, R est le pas d'analyse, p est l'indice temporel d'analyse, et k est l'indice de la bande de fréquences considérée. L'emploi d'un pas d'analyse R supérieur à 1 correspond à un sous-échantillonnage des signaux issus de la transformation de Fourier à court-terme, qui ont une largeur de bande très inférieure aux signaux d'origine. Il est malgré tout nécessaire, pour éviter la perte d'information par repliement spectral, de ne pas donner au pas d'analyse une valeur trop

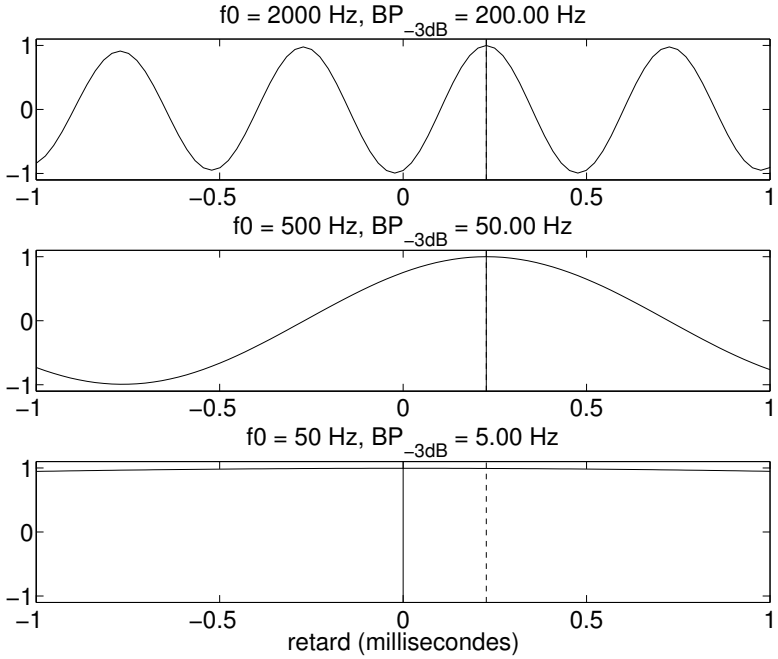


FIG. IV.7 – **Problème de la résolution sur l'estimation du retard en basses fréquences :** Intercorrélation de deux bruits blancs identiques à un retard près de 2,27 ms, avec un bruit additionnel (rapport signal à bruit de 40 dB), traités par un filtre passe-bande (filtre de Butterworth d'ordre 2) de fréquence centrale f_0 variable et de facteur de qualité à -3dB $Q_{-3dB} = 10$. Pour chaque cas, la ligne verticale continue indique le maximum estimé, et la ligne verticale interrompue indique le maximum théorique.

importante. On peut montrer que ce pas doit être strictement inférieur à la longueur de la fenêtre[7] pour que la reconstruction soit parfaite.

Il est utile de noter qu'il s'agit de deux définitions légèrement différentes du même concept mathématique. La convention passe-bas, bien que peu usitée en pratique, est souvent jugée plus intuitive, et est très souvent utilisée en calcul formel. La convention passe-bande est de très loin la plus employée en pratique, notamment puisqu'en évitant le filtrage hétérodyne, on se ramène à une succession de transformées de Fourier, que l'on peut calculer par exemple par méthode FFT. Néanmoins, les deux conventions sont totalement équivalentes en termes d'information disponible. Dans le cas présent, on commence le calcul à partir de la convention passe-bande, puisqu'il s'agit d'un banc de filtres linéaires sans modulation hétérodyne, ce qui est le cadre utilisé dans la section 4.

Soient $\tilde{X}_R[p,k]$ et $\tilde{Y}_R[p,k]$ les transformées de Fourier à court-terme en convention passe-bande des signaux $x[n]$ et $y[n]$ calculées selon la définition IV.28. L'hypothèse inhérente au modèle d'égalisation-annulation, menant en large bande au modèle gain-retard présenté en début de chapitre :

$$y[n] = \alpha.x[n+\tau]$$

... se traduit dans ce formalisme par :

$$\tilde{Y}_R[p,k] = \alpha.\tilde{X}_R\left[p + \frac{\tau}{R}, k\right] \tag{IV.31}$$

Le gain α et le retard τ dépendent ici de la bande k considérée, bien que ce ne soit pas explicite dans la notation afin de l'alléger. Cette formule fait appel à une valeur de $\tilde{X}[p,k]$ qui n'a pas été forcément calculée, si le pas d'analyse R n'est pas égal à 1. On laisse pour l'instant de côté cet aspect, et l'on choisit temporairement $R = 1$, si bien que les transformées de

[7] inférieur ou égal dans le cas d'une fenêtre rectangulaire

Fourier à court-terme sont les sorties directes des bancs de filtres, sans décimation. On peut former, sur le modèle de la définition IV.16, la fonction différence pour des filtres complexes :

$$D_{\alpha,\tau}[p,k] = \sum_{q=-\infty}^{+\infty} \left|\tilde{Y}_1[q,k] - \alpha.\tilde{X}_1\left[q+\tau,k\right]\right|^2 .w[p-q] \tag{IV.32}$$

...$\tilde{X}_1[q,k]$et $\tilde{Y}_1[q,k]$désignant les transformées de Fourier à court-terme avec $R=1$. En développant, il vient :

$$D_{\alpha,\tau}[p,k] = P_{\tilde{Y}_1}[p,k] + \alpha^2.P_{\tilde{X}_1}\left[p+\frac{\tau}{R},k\right] - 2\alpha.\left|C_{\tilde{X}_1,\tilde{Y}_1}[p,\tau,k]\right|.cos\left(\angle C_{\tilde{X}_1,\tilde{Y}_1}[p,\tau,k]\right) \tag{IV.33}$$

Les puissances par bande sont définies par :

$$P_{\tilde{X}_1}[p,k] = \sum_{q=-\infty}^{+\infty} \left|\tilde{X}_1[q,k]\right|^2 .w[p-q]$$

et

$$P_{\tilde{Y}_1}[p,k] = \sum_{q=-\infty}^{+\infty} \left|\tilde{Y}_1[q,k]\right|^2 .w[p-q]$$

, et l'intercorrélation pour la bande de fréquences k vaut :

$$C_{\tilde{X}_1,\tilde{Y}_1}[p,\tau,k] = \sum_{q=-\infty}^{+\infty} \tilde{Y}_1^*[q,k].\tilde{X}_1\left[q+\tau,k\right].w[p-q]$$

La relation IV.29 permet d'exprimer les corrélations et puissances à court-terme en fonction des transformées de Fourier en convention passe-bas :

$$P_{\tilde{X}_1}[p,k] = P_{X_1}[p,k]$$

$$P_{\tilde{Y}_1}[p,k] = P_{Y_1}[p,k]$$

$$C_{\tilde{X}_1,\tilde{Y}_1}[p,\tau,k] = e^{j2\pi\frac{k}{N}\tau}.C_{X_1,Y_1}[p,\tau,k]$$

...$P_{X_1}[p,k]$, $P_{Y_1}[p,k]$ et $C_{X_1,Y_1}[p,\tau,k]$ désignant respectivement les puissances et l'intercorrélation des transformées de Fourier en convention passe-bas. On peut donc exprimer la fonction différence sous la forme :

$$D_{\alpha,\tau}[p,k] = P_{Y_1}[p,k] + \alpha^2.P_{X_1}\left[p+\frac{\tau}{R},k\right] - 2\alpha.\left|C_{X_1,Y_1}[p,\tau,k]\right|.cos\left(\angle C_{X_1,Y_1}[p,\tau,k] + 2\pi\frac{k}{N}\tau\right) \tag{IV.34}$$

5.1 Approximations de calcul

Le calcul de la fonction de différence proposé en IV.33 ou en IV.34 nécessite en théorie la connaissance à chaque instant de la transformée à court-terme des deux signaux, ce qui, si l'on se place dans le cas d'un analyse en bande limitées, est peu envisageable d'un point de vue purement computationnel, tant à cause du coût de calcul que de la mémoire nécessaire. Néanmoins, on peut recourir à deux approximations pour optimiser le calcul.

Approximation pour des faibles retards

Si la longueur de la fenêtre est importante par rapport aux retards maximaux envisagés (de l'ordre de 1 ms dans le cas d'enregistrements de type binauraux), alors la transformée de Fourier à court-terme en convention passe-bas est quasiment invariante par translation temporelle. Pour s'en convaincre, on peut remarquer que si (toujours en supposant $R=1$) :

$$g[n-q-\tau] \simeq g[n-q]$$

alors :

$$X_1[q+\tau,k] \simeq X_1[q,k]$$

D'où :

$$C_{X_1,Y_1}[p,\tau,k] \simeq C_{X_1,Y_1}[p,0,k]$$

N.B. : Il est intéressant de noter que, puisque $C_{X_1,Y_1}[p,0,k] = C_{\tilde{X}_1,\tilde{Y}_1}[p,0,k]$, il n'est pas nécessaire de calculer les transformées de Fourier en convention passe-bas : on peut directement calculer les corrélations en convention passe-bande.

Approximation par décimation

Puisque les signaux issus de la transformation de Fourier à court-terme en convention passe-bas sont ramenés en bande de base avec une largeur de bande très inférieure à celle des signaux d'origine, et puisque, en vertu de l'approximation précédente, il n'est plus nécessaire de calculer l'intercorrélation pour des retards non nuls, on peut réintroduire la décimation, et écrire :

$$C_{X_1,Y_1}[p,0,k] \simeq \frac{1}{\sum_{q'=-\infty}^{+\infty} w[q'R]} . \sum_{q'=-\infty}^{+\infty} Y_R^*[q',k].X_R[q',k].w[p-q'R]$$

$...X_R[q',k]$ et $Y_R[q',k]$ désignant les transformées de Fourier à court-terme calculées avec le pas d'analyse R. Si l'on définit par w_R le résultat du sous-échantillonnage d'un facteur R de la fenêtre w et si l'on suppose que le facteur de normalisation $\sum_{q'=-\infty}^{+\infty} w[q'R]$ est égal à 1, l'équation précédente se réécrit sous la forme suivante :

$$C_{X_1,Y_1}[p,0,k] \simeq \sum_{q=-\infty}^{+\infty} Y_R^*[q,k].X_R[q,k].w_R[p-q]$$

La contrainte sur le pas d'analyse R est en revanche différente que celle exposée plus haut : il ne s'agit pas de préserver la reconstruction parfaite, mais de minimiser, lors du calcul de la corrélation, qui est en fait la convolution de $Y_1^*[p,k].X_1[p,k]$ par $w[p]$, la perte d'information par repliement spectral. Le taux de sous-échantillonnage maximum est donc déterminé par celui des deux termes de la convolution qui occupe la plus large bande. On rappelle que la largeur de bande de $X_1[p,k]$et $Y_1[p,k]$ est de l'ordre de F_e/N, F_e étant la fréquence d'échantillonnage, et N la taille de la fenêtre d'analyse de Fourier. Quant à la fenêtre $w[p]$, sa largeur de bande dépend forcément de sa forme. Pour une fenêtre exponentielle causale, la bande passante à -3 dB est égale pour des valeurs importantes de τ à :

$$BP\{w\}_{-3dB} \simeq \frac{1}{\pi\tau}$$

Dans ces conditions, le pas d'analyse maximum est défini par :

$$R \ll R_{max} = max(N, \pi\tau F_e)$$

Étant donné que l'on utilise des valeurs de τ typiquement comprises entre 50 et 100 millisecondes, c'est généralement la largeur de bande des transformées de Fourier qui limite le facteur de sous-échantillonnage. En pratique, des valeurs de R inférieures à $N/2$ pour une fenêtre de Hanning commencent à donner de bons résultats. **On choisit un pas d'analyse égal à $N/4$, soit un taux de recouvrement de 75%**.

Utilisation de ces approximations

On peut alors introduire, sur le modèle du périodogramme de Welch, un estimateur de l'interspectre de signaux non stationnaires, l'**interspectre à court-terme** :

$$S_{XY}[p,k] = C_{XY}[p,0,k] = \sum_{q=-\infty}^{+\infty} Y_R^*[q,k].X_R[q,k].w_R[p-q] \qquad \text{(IV.35)}$$

L'approximation conduit donc à :

$$C_{XY}[p,\tau,k] \simeq S_{XY}[p,k]$$

La même approximation est valable pour la puissance :

$$P_X\left[p+\frac{\tau}{R},k\right] \simeq S_{XX}[p,k]$$

Ces nouvelles définitions conduisent également à un estimateur de la cohérence de signaux non stationnaires, la **cohérence à court-terme** :

$$\Phi_{XY}[p,k] = \frac{S_{XY}[p,k]}{\sqrt{S_{XX}[p,k].S_{YY}[p,k]}} \tag{IV.36}$$

La fonction différence complexe peut donc être approchée par :

$$D_{\alpha,\tau}[p,k] \simeq S_{YY}[p,k] + \alpha^2.S_{XX}[p,k] - 2\alpha.\left|S_{XY}[p,k]\right|.cos\left(2\pi\frac{k}{N}\tau + \angle S_{XY}[p,k]\right) \tag{IV.37}$$

5.2 Minimisation

La recherche du couple (α,τ) optimal selon les deux méthodes de minimisation conduit aux résultats suivants :

– Minimisation de l'erreur absolue : l'estimation selon cette méthode conduit au système :

$$\begin{cases} \alpha_1[p,k] = (-1)^m \frac{|S_{XY}[p,k]|}{S_{XX}[p,k]} \\ \tau_1[p,k] = -\frac{N}{2\pi k}.\angle S_{XY}[p,k] + m.\frac{N}{2k} \end{cases} \tag{IV.38}$$

...m étant un entier relatif quelconque. L'erreur absolue minimale vaut :

$$D_{min}[p,k] = D_{\alpha_1,\tau_1}[p,k] = S_{XX}[p,k].\left(1-\left|\Phi_{XY}[p,\tau]\right|^2\right) \tag{IV.39}$$

– Minimisation de l'erreur normalisée

$$\begin{cases} \alpha_2[p,k] = (-1)^m \sqrt{\frac{S_{YY}[p,k]}{S_{XX}[p,k]}} \\ \tau_2[p,k] = -\frac{N}{2\pi k}.\angle S_{XY}[p,k] + m.\frac{N}{k} \end{cases} \tag{IV.40}$$

...m étant un entier relatif quelconque. L'erreur normalisée minimale vaut :

$$\varepsilon_{min}[p,k] = \varepsilon_{\alpha_2,\tau_2}[p,k] = \frac{1}{2}.\left(1-\left|\Phi_{XY}[p,\tau]\right|\right) \tag{IV.41}$$

De manière similaire au modèle gain-retard simple, on trouve que la validité du modèle en bande étroite dépend de la valeur absolue de la cohérence pour le canal considéré.

5.3 Lien avec la formule des interférences

On peut également noter que l'estimation obtenue peut être envisagée comme une écriture non stationnaire de deux cas particuliers de la formule des interférences. Pour s'en persuader, on peut se placer dans le cas où les signaux considérés sont stationnaires. On peut alors effectuer une analyse de corrélation à long-terme, c'est-à-dire considérer une fenêtre $w[p]$ constante sur toute la durée des signaux, de manière à stabiliser les estimations des statistiques d'ordre 2, tout en offrant la meilleure résolution fréquentielle possible[8]. On définit alors le gain complexe par bande :

$$\underline{H}[k] = \alpha[k].e^{j2\pi\frac{k}{N}\tau[k]}$$

[8]L'estimateur défini par l'équation IV.35 correspond alors rigoureusement à celui donné par la méthode du periodogramme de Welch.

Dans ce cas, les estimations IV.38 et IV.40 équivalent respectivement à :

$$\underline{H_1}[k] = \frac{S_{XY}[k]}{S_{XX}[k]}$$

et

$$\underline{H_2}[k] = \sqrt{\frac{S_{YY}[k]}{S_{XX}[k]}}.e^{j\angle S_{XY}[k]}$$

...qui sont deux écritures de la formule des interférences, dans le cas où le signal $y[n]$ résulte du filtrage du signal $x[n]$ par le filtre $h[n]$ de transformée de Fourier $H[k]$. La cohérence stationnaire est couramment utilisée pour vérifier la linéarité d'un canal (Max et Lacoume, 1996), et permet donc ici de juger de la validité du postulat d'existence du filtre $h[n]$.

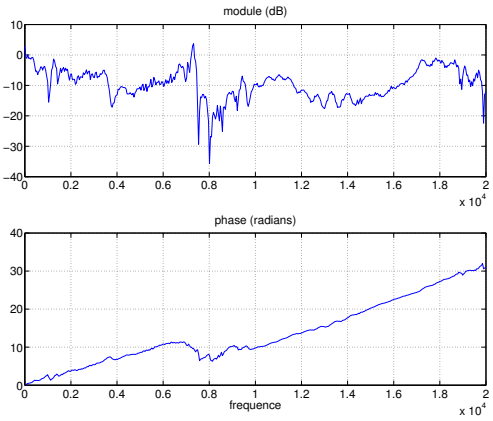
Le corollaire de cette remarque est que si l'on effectue une analyse temps-fréquence avec une fenêtre temporelle de durée au moins égale à celle des réponses impulsionnelles modélisant le trajet direct (c'est-à-dire de quelques millisecondes pour les HRIR, soit une fenêtre d'analyse d'au moins 200 échantillons à 44100 Hz), on obtient une estimation fidèle de la fonction de transfert intercanale en gain et en phase, avec pour chaque fréquence une indication de l'erreur.

Ainsi, le modèle de signaux sous-jacent au modèle gain-retard pour une représentation de Fourier à court-terme correspond à un filtrage linéaire, et la cohérence à court-terme définie par l'équation IV.36 en fournit à chaque instant une indication de la validité.

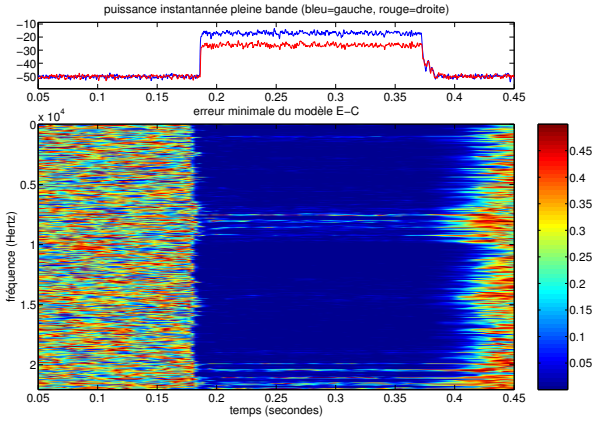
5.4 Exemple

Un exemple pratique d'estimation, dont les résultats sont représentés à la figure IV.8 permet d'illustrer ce propos : le signal analysé est un signal bicanal issu de la convolution d'une séquence de bruit blanc de 0,18 secondes par un couple de HRIR issues d'un des jeux de mesures de la base Listen (Vandernoot et Rio, 2003), pour un azimut de 30° dans le plan horizontal. Ces réponses impulsionnelles ont été calculées sur 512 points, soit 11,6 ms. A ce signal est superposé un bruit additionnel dont les deux voies sont indépendantes, le rapport signal sur bruit par voie étant de 30 dB. Le banc de filtres employé est la transformée de Fourier à court-terme, et l'on se placera dans l'hypothèse où la corrélation peut-être approchée par l'interspectre au sens de la définition IV.35. La transformée de Fourier à court-terme est calculée sur 1024 bandes de fréquences et pour une fenêtre de Hanning de 1024 points (23,2 ms), avec un taux de recouvrement de 7/8 (soit un pas d'analyse de 2,9 ms). La fenêtre d'analyse utilisée pour le calcul de la cohérence est une fenêtre exponentielle causale de constante de temps égale à 10 ms.

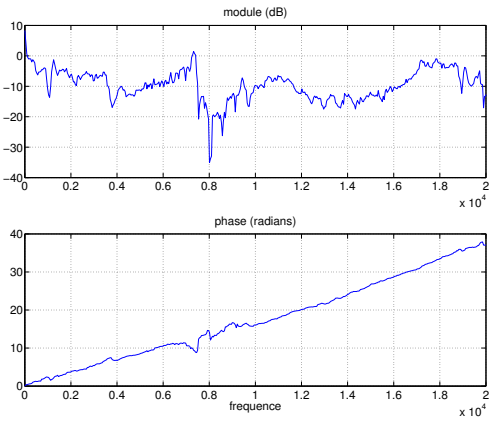
Le quotient des transformées de Fourier de chacune des deux voies de la réponse impulsionnelle, ou fonction de transfert interaurale (ITF), est représenté en haut et à gauche de la figure IV.8. En haut à droite est représentée, en plus de la puissance à court-terme des signaux pleine-bande (lissée par une fenêtre de Hanning de 1,5 ms), l'erreur $\varepsilon_{EC,1}[p,k]$ du modèle gain-retard optimale au sens de la l'erreur normalisée (équation IV.41). Pour la plupart des fréquences, le signal est détecté sans ambiguïté, l'erreur étant inférieure à 2%. Cependant, pour certaines zones de fréquences (en particulier entre 7 kHz et 9,5 kHz, et au dessus de 20 kHz) l'erreur est toujours importante, ce qui signifie que le modèle gain-retard n'y est pas une bonne approximation de la réalité, du moins pour ce banc de filtres. En se plaçant un court laps de temps après la détection (ici t=0,2 secondes, par exemple), on peut estimer le gain α_{min1} et la phase de l'interspectre (à partir de laquelle on estime le retard τ_{min}), et les comparer grâce à une représentation par diagramme de Bode au module et à la phase de l'ITF (en bas à gauche et en bas à droite de la figure IV.8). On peut observer que l'erreur d'estimation du gain et de la phase pour une fréquence donnée est étroitement liée à l'erreur du modèle : en particulier, les pics d'erreurs vers 8 kHz correspondent aux erreurs d'estimation les plus grossières. D'autre part, si la phase est correctement estimée en basses fréquences, l'estimation du retard de phase est instable pour des fréquences inférieures à 2 kHz, et ce alors que l'erreur reste faible à ces fréquences.



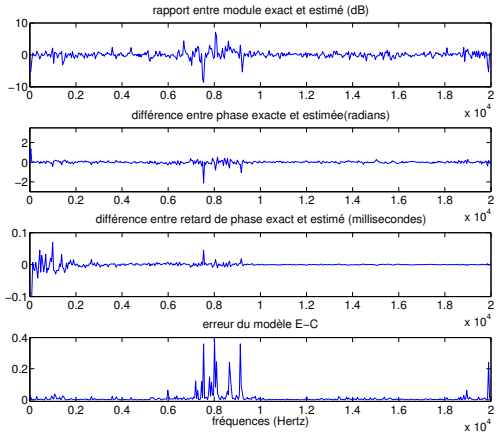
(a) Diagramme de bode de l'ITF exacte



(b) Puissance instantanée des signaux et indice normalisé de détection par bande



(c) diagramme de Bode de l'ITF estimée à l'instant t=0.2 ms



(d) différences entre les indices exacts et estimés

FIG. IV.8 – **Estimation de filtre par égalisation et annulation** par transformée de Fourier à court-terme, sur un signal (séquence de bruit blanc) spatialisé par synthèse binaurale, avec un bruit additionnel (rapport signal à bruit = 30 dB). L'indice normalisé de détection par bande (figure en haut à droite) permet de détecter le signal, ce qui permet alors d'estimer les gains et retards pour chaque fréquence (bas gauche) pour un instant donné (ici t = 0,2 s), estimations que l'on peut comparer à la fonction de transfert interaurale (ITF) exacte (haut gauche). On peut mettre en relation les erreurs d'estimations avec l'erreur d'égalisation et annulation (bas droite).

6 GÉNÉRALISATION DU PRINCIPE D'ÉGALISATION ET ANNULATION À PLUS DE DEUX CANAUX

Le modèle par égalisation et annulation a été initialement proposé comme modèle auditif de détection binaurale, ce qui implique naturellement que les signaux d'entrée soient au nombre de deux. Cependant, il n'y a aucune raison, en dehors du contexte de la modélisation de l'audition, de limiter son application à deux canaux uniquement. Bien que cette idée de généralisation ne soit pas développée dans le reste de l'étude puisque l'on s'est concentré sur l'étude d'enregistrements binauraux, il est utile d'en esquisser le principe, pour que les méthodes présentées dans ce travail puissent être appliquées à d'autres types de prise de son.

Une possibilité naturelle de généralisation consiste à effectuer l'égalisation et l'annulation sur tous les couples de canaux possibles, et de choisir comme erreur du modèle la somme de toutes les erreurs partielles. Ainsi, si l'on considère N signaux d'entrée notés $x_i[n]$, $n \in \{1...N\}$, et si l'on désigne par (α_{ij}, τ_{ij}) le couple de paramètres d'égalisation et annulation pour les signaux $x_i[n]$ et $x_j[n]$ (avec $i \neq j$, et en veillant à éviter de répéter deux fois la même combinaison permutée), on peut former la différence :

$$D_{\overline{\alpha},\overline{\tau}}[n] = \frac{1}{\sum_{i=1}^{N} \cdot \sum^{N}\, j = i+1 \quad (1)} \cdot \sum_{i=1}^{N} \cdot \sum_{j=i+1}^{N} \left[\sum_{p=-\infty}^{+\infty} \left(x_i[p] - \alpha_{ij}.x_j[p+\tau_{ij}]\right)^2 .w[n-p] \right]$$

$$D_{\overline{\alpha},\overline{\tau}}[n] = \frac{2}{N.(N-1)} \cdot \sum_{i=1}^{N} \cdot \sum_{j=i+1}^{N} D^{ij}_{\alpha_{ij},\tau_{ij}}[n]$$

...$\overline{\alpha}$ et $\overline{\tau}$ désignant les vecteurs de paramètres :

$$\overline{\alpha} = (\{\alpha_{ij},\ j > i\})$$

... et

$$\overline{\tau} = (\{\tau_{ij},\ j > i\})$$

Si aucune relation n'a été imposée entre les paramètres[9], les dérivées partielles de l'erreur $D_{\overline{\alpha},\overline{\tau}}[n]$ valent simplement :

$$\begin{cases} \frac{\partial D_{\overline{\alpha},\overline{\tau}}}{\partial \alpha_{ij}} = \frac{\partial D^{ij}_{\alpha_{ij},\tau_{ij}}}{\partial \alpha_{ij}} \\ \frac{\partial D_{\overline{\alpha},\overline{\tau}}}{\partial \tau_{ij}} = \frac{\partial D^{ij}_{\alpha_{ij},\tau_{ij}}}{\partial \tau_{ij}} \end{cases}$$

Donc la minimisation de l'erreur absolue $D_{\overline{\alpha},\overline{\tau}}[n]$ revient à minimiser chacune des erreurs partielles. On peut former de même l'erreur normalisée :

$$\varepsilon_{\overline{\alpha},\overline{\tau}}[n] = \frac{2}{N.(N-1)} \cdot \sum_{i=1}^{N} \cdot \sum_{j=i+1}^{N} \varepsilon^{ij}_{\alpha_{ij},\tau_{ij}}[n]$$

Si toutes les erreurs normalisées partielles sont comprises entre 0 et 1, alors c'est également le cas pour l'erreur normalisée totale. D'autre part, la minimisation de l'erreur normalisée totale est là aussi équivalente à la minimisation séparée de chacune des erreurs normalisées partielles. Cette formulation offre ainsi l'avantage de respecter le principe de superposition.

Cette généralisation suppose toujours l'existence d'une source unique à chaque instant, son apport principal résidant dans la multiplication du nombre de paramètres utiles, qui permettent ultérieurement de localiser la dite source avec une précision spatiale accrue. Il est cependant tout-à-fait possible d'envisager l'annulation simultanée de deux ou plusieurs

[9] Pour pouvoir imposer de telles relations, il est nécessaire de disposer de connaissances sur la position géométrique relative des microphones, ainsi que sur leurs caractéristiques de directivité.

sources, en utilisant des combinaisons linéaires plus complexes. On peut s'inspirer pour ce faire des travaux de Cheveigné (1993) sur l'estimation conjointe de plusieurs sources harmoniques mélangées dans un signal monophonique, en l'adaptant au cas de plusieurs canaux.

7 CONCLUSION

La méthode de détection développée ici, bien que se basant initialement sur un modèle très simple de relation entre les signaux d'entrée, permet, grâce à la possibilité qu'elle offre d'ajuster l'échelle temps-fréquence, de couvrir le plus grand nombre de situations possibles, de celle où les signaux ne diffèrent que d'un gain et d'un retard à celle où la relation s'apparente à un réel filtrage dont on cherche à estimer la fonction de transfert. Le choix du banc de filtres employé, de ses paramètres ainsi que de ceux de l'analyse par corrélations dépend donc étroitement du problème à traiter et de la précision souhaitée. D'autre part, le fait que cette méthode ne soit pas limitée dans le principe au cas de deux canaux permet d'envisager son application dans d'autres configurations de prise de son que les couples stéréophoniques ou les têtes artificielles.

D'autre part, il peut être intéressant de rappeler que l'outil de base, qui est la corrélation à court-terme, peut être lui aussi redéfini : ainsi, l'utilisation d'une définition symétrique de la corrélation permet de lier cette méthode temporelle aux méthodes d'analyse spectrale non stationnaire utilisant la transformée de Wigner-Ville. Ceci pose néanmoins deux problèmes : d'une part, celui de la causalité, car les événements sont dans ce cas détectés avec un peu d'avance (cela dit, celle-ci reste confinée dans des limites très raisonnables - inférieure à une milliseconde - dans le cas d'une prise de son binaurale ; d'autre part, celui du calcul des valeurs des signaux entre deux échantillons, qui nécessite soit une interpolation, soit un suréchantillonnage de rapport deux.

Ainsi, l'utilisation du principe d'égalisation et annulation en tant que cadre permet d'unifier de nombreuses méthodes d'estimation, tout en offrant plusieurs avantages supplémentaires : ainsi, l'erreur normalisée est un indice fort vis-à-vis de la qualité de l'estimation, puisqu'il permet de juger de la pertinence du modèle pour la détection **et** pour l'estimation des indices de localisation ; de plus, on dispose de méthodes de calcul rigoureuses des corrélations et puissances à court-terme, qui tiennent compte de la nature non stationnaire des signaux.

V

Application de la détection par égalisation et annulation aux milieux réverbérants

AYANT PRÉSENTÉ dans le chapitre précédent le principe général de la méthode de détection proposée dans cette étude, on étudie dans ce chapitre son application spécifique à des signaux réverbérés. On s'intéresse tout d'abord au comportement de cette méthode de détection en milieu réverbérant (section 1), en étudiant notamment l'effet des paramètres de l'analyse et de certains aspects spatiaux de la scène sonore, comme la distance à la source. Ensuite, on se penche sur un cas spécifique, qui est celui de signaux harmoniques à fréquence fondamentale constante par morceaux (section 2), en étudiant de quelle manière l'analyse peut bénéficier de cette information supplémentaire.

1 COMPORTEMENT DE LA DÉTECTION EN MILIEU RÉVERBÉRANT

1.1 Introduction

L'application du principe d'égalisation et annulation à la détection de source et de réverbération repose sur le principe présenté au chapitre III, et que l'on peut reformuler de la manière suivante : d'une part, la détection d'une source active correspond à repérer dans le signal les plages de temps et les bandes de fréquences pour lesquelles l'indice de détection est proche de zéro ; d'autre part, et *a contrario*, la détection de réverbération vise à repérer les plages de temps et les bandes de fréquences pour lesquelles aucune source n'est active (c'est-à-dire que l'indice de détection est élevé et proche de $\frac{1}{2}$), mais une ou plusieurs sources l'étaient dans un proche passé.

On dispose maintenant d'un cadre théorique pour effectuer cette détection, qui laisse une grande souplesse vis-à-vis des paramètres de l'analyse : notamment, on a la possibilité d'ajuster l'échelle en temps et en fréquence, grâce au choix du banc de filtres. De même, lorsque l'analyse est une transformation de Fourier à court-terme avec des bandes suffisamment étroites, il est également possible d'utiliser, en lieu et place de la corrélation à court-terme, la cohérence à court-terme, qui est bien moins coûteuse d'un point de vue computationnel. On justifie de fait l'usage de cette dernière représentation dans plusieurs recherches (Allen et al., 1977; Avendano et Jot, 2002) pour distinguer les zones temps-fréquences où l'influence des sources est prépondérante de celles où l'influence de la réverbération est prépondérante.

1.2 Évolution de l'indice de détection sur un signal réverbéré

Pour mettre en évidence le comportement de la méthode de détection proposée en milieu, on se place dans le cas d'un exemple synthétique simple, qui est celui d'un bruit large bande intermittent et réverbéré. La salle est celle choisie pour l'exemple de la section 3.3 du chapitre I. Il s'agit, on le rappelle, d'une salle vide d'un volume de $650\,m^3$, aux parois réfléchissantes, et au temps de réverbération de 3 secondes en basses fréquences et de 1 seconde à 10 kHz. Le signal source est donc constitué par une succession de séquences de bruit blanc de 3 secondes entrecoupées de silences de 1,5 secondes. Ces durées sont choisies de manière d'une part à ce que l'état stationnaire soit atteint au moment où le bruit s'interrompt, et d'autre part que la réverbération ne soit pas achevée au moment où la nouvelle séquence de bruit commence. Le signal observé est généré par convolution de ce signal source avec une réponse impulsionnelle binaurale mesurée dans la salle considérée à 1,8 m du haut-parleur, qui est frontal et dans le plan horizontal du point de vue du sujet). Un bruit blanc stationnaire a été ajouté dans chaque voie, de sorte que le rapport signal sur bruit moyen vaut 60 dB lorsque la source est active.

On peut observer sur la figure V.1 la valeur de l'indice de détection en sortie d'un banc de filtres en tiers d'octave, pour toutes les bandes d'une part, et pour la bande centrée sur 2 kHz d'autre part. On distingue trois étapes successives :

1. **Transitoire** : l'onde directe seule étant présente, la corrélation est maximale dans la plupart des bandes, et l'indice de détection est donc quasiment nul. C'est cette étape, très courte, qui permet la localisation, comme on le verra au chapitre VI.
2. **Entretien** : les réflexions successives font monter l'indice de détection jusqu'à un niveau d'équilibre, qui est déterminé par le niveau de corrélation en régime stationnaire. On sait déjà, en se référant à la section 3.2 du chapitre I, que ce dernier dépend entre autres de la fréquence et du rapport de l'énergie directe sur l'énergie réverbérée, et l'on retrouve cette dépendance sur la figure V.1 : l'indice de détection pendant la phase entretenue reste faible pour les dix premières bandes (soit pour des fréquences inférieures à 400 Hz), et pour les cinq bandes suivantes de 2 kHz à 5 kHz, mais est plus important (jusqu'à 0,2) pour les autres bandes. Cet aspect est approfondi en section 1.4 du présent chapitre.
3. **Décroissance** : lorsque la source fait silence, l'indice de détection monte à nouveau, car le son direct, puis les réflexions précoces, s'éteignent petit à petit, pour céder la place

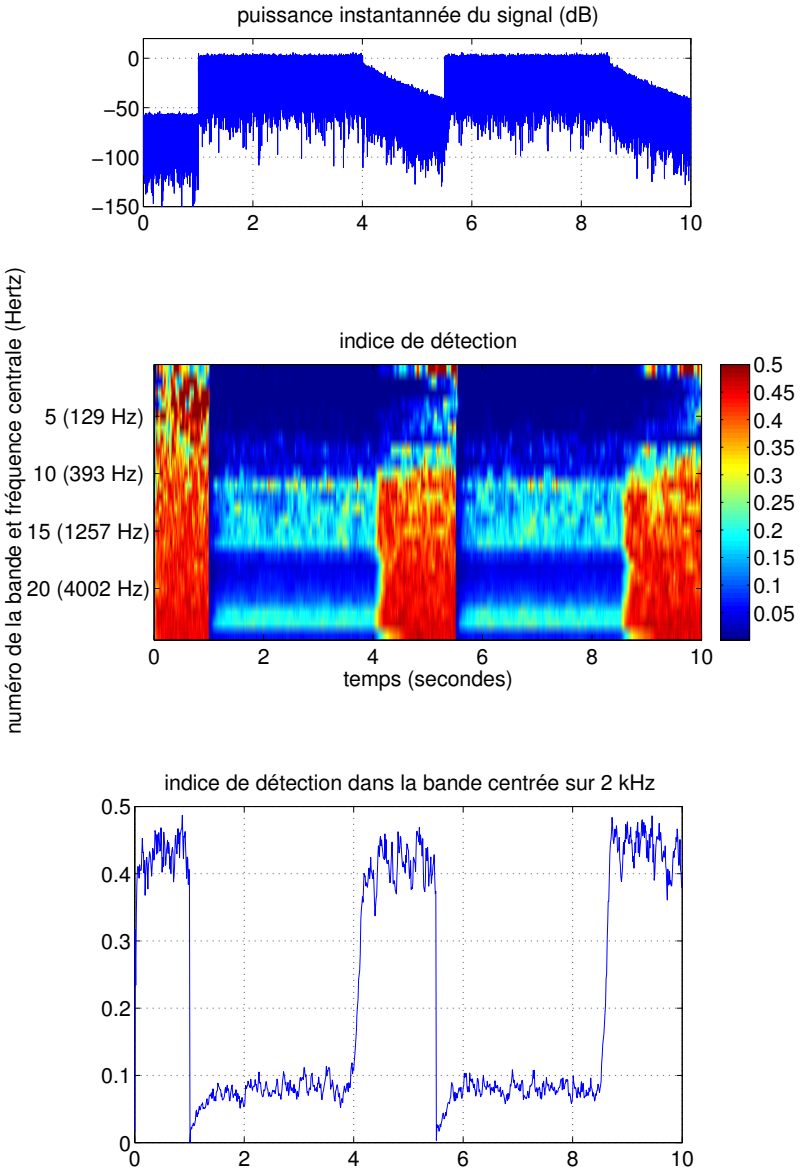


FIG. V.1 – **Principe de la détection en milieu réverbérant** : le signal analysé résulte de la convolution d'une succession intermittente de séquences de bruit blanc par une réponse binaurale de salle, avec addition de bruit. Sur la figure supérieure est représentée la puissance instantanée d'une des voies du signal observé. La figure intermédiaire représente l'indice de détection calculé en sortie d'un banc de filtres en tiers d'octave, et la figure inférieure représente l'indice de détection calculé dans la bande de fréquences centrée autour de 2 kHz. On note le comportement différent lors des transitoires, pendant la phase entretenue, et après l'extinction.

à la réverbération diffuse. En vertu du principe mentionné au chapitre I, la corrélation se stabilise autour du niveau de corrélation théorique en champ diffus, qui constitue le seuil inférieur de corrélation. Ainsi, l'indice de détection reste faible en basses fréquences (en dessous de quelques centaines de Hertz si l'on se fie à la théorie) même alors que la réverbération est fortement avancée. En revanche, plus la fréquence est élevée, moins les signaux sont corrélés.

Ce comportement, schématisé sur la figure V.2, permet de comprendre quel est la démarche adoptée pour la double détection :

- **détection de source** : en l'absence de signal cohérent, l'indice de détection reste proche de 0,5. Dès lors qu'il décroît jusqu'à franchir un seuil décidé à l'avance, on sait qu'une source active a été détectée. On peut alors s'intéresser aux gains et aux retards intercanaux, qui sont les indices sur lesquels repose la méthode de détection qui sera présentée au chapitre VI. Sans anticiper sur cet aspect du problème, on indique que puisque la précision de l'estimation de ces indices, et donc de la localisation, dépend étroitement de la qualité de la détection, qui est reflétée par la valeur de l'indice de détection, **la localisation sera basée sur les gains et retards estimés au moment où la détection est optimale**.
- **détection de réverbération** : la détection de réverbération nécessite d'estimer le plus précisément possible l'instant où la source fait silence. Éventuellement, cet instant peut-être estimé avec un certain retard ; cela revient à retarder également le début de la régression pour le calcul du temps de décroissance (la méthode d'estimation du temps de réverbération est présentée plus en détail au chapitre VII). En revanche, une estimation en avance sur le moment où la source fait effectivement silence serait préjudiciable, puisque l'on intègre alors une partie de l'énergie directe au calcul du temps de réverbération, si bien que celui-ci est alors mésestimé. Cette dernière remarque incite à choisir le seuil de détection avec précaution, ce qui n'est pas facilité par sa dépendance fréquentielle. L'idéal est de connaître la valeur du niveau de corrélation (ou de cohérence) stationnaire, de manière à déterminer le seuil en fonction.

On dégage ainsi dans le domaine temps-fréquences des zones pour lesquelles la source est jugée prépondérante, et celles pour lesquelles il est envisageable de décrire la réverbération, celle-ci n'étant pas masquée par une onde directe. Au bas des figures V.3 et V.4 sont représentées les résultats d'une telle détection sur le signal utilisé en exemple pour la figure V.1 , pour deux types de bancs de filtres : d'une part le même banc de filtres en tiers d'octave que précédemment, et d'autre part, pour comparaison, une transformation de Fourier à court-terme sur 128 bandes de fréquences. Le seuil de détection est calculé à partir de la connaissance de la corrélation (et donc de l'indice de détection) dans chaque bande de fréquences dans le cas stationnaire. Le seuil est choisi à la valeur de cet indice stationnaire majoré d'une marge de 0,1.

Les zones ainsi dégagées sont soumises à quatre types d'artefacts. Premièrement, puisque la corrélation/cohérence reste élevée en basses fréquences à n'importe quel stade de la réverbération, il y est quasiment impossible d'opérer une détection de bonne qualité, celle-ci nécessitant une dynamique importante de variation de l'indice de détection. Une remarque similaire peut être formulée lorsque l'indice de détection stationnaire est très élevé, comme c'est le cas avec la transformée de Fourier à court-terme, entre 6 kHz et 8 kHz : le seuil plafonne alors à 0,5, et aucune détection n'est possible. De plus, on remarque que l'indice de détection pendant la phase d'entretien est très fluctuant aux moyennes fréquences pour le banc de filtres en tiers d'octave. Cet effet est dû à la largeur de bande, très réduite à ces fréquences, de ces filtres, en comparaison avec la transformée de Fourier à court-terme. Il est responsable d'erreurs de détection, puisque même lors de la phase d'entretien, l'indice de détection peut passer au dessus du seuil. Le dernier artefact observable est la présence de traînées au dessus de 10 kHz sur la représentation de l'indice de détection calculé à partir des transformées de Fourier à court-terme. Cet aspect, qui est imputable à la durée de la fenêtre d'analyse, est développé ci-dessous.

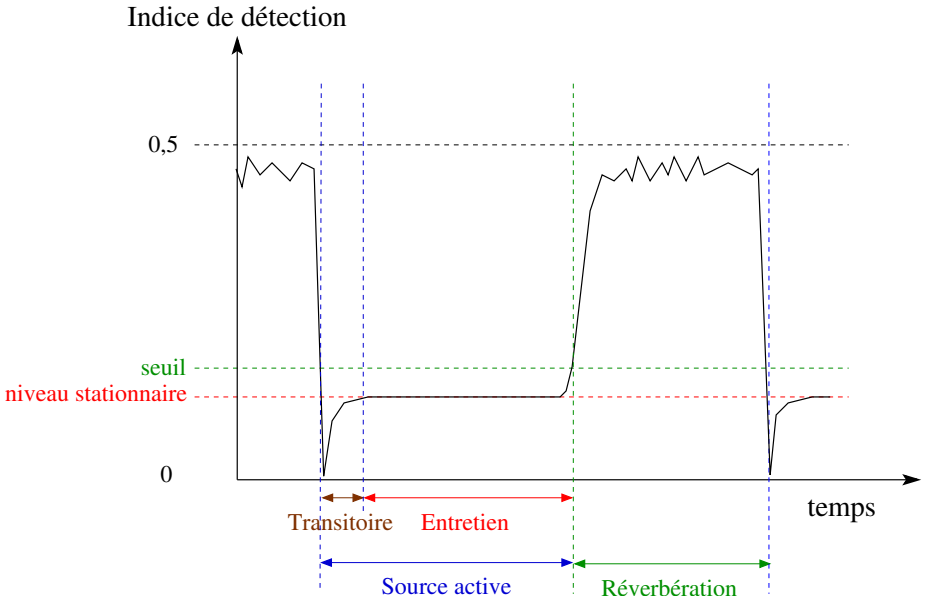


FIG. V.2 – **Principe de la détection de réverbération** : ci-dessus est représentée de manière schématique l'indice de détection dans une bande de fréquences donnée, lorsque le son émis dans la salle est un signal entretenu d'une durée suffisamment longue pour qu'un régime stationnaire se mette en place. On peut noter en particulier que le seuil permettant la détection de réverbération doit dépendre du niveau de l'indice de détection durant le régime stationnaire pour être efficace.

1.3 Influence de la durée de la fenêtre d'analyse

Comme indiqué au chapitre IV, le choix de la durée de la fenêtre d'analyse utilisée pour le calcul de la corrélation à court-terme est conditionné par deux contraintes, qui sont d'une part la stabilité du résultat par rapport à des fluctuations locales de la corrélation, et d'autre part la résolution temporelle de la détection. Dans le cas de la fenêtre exponentielle causale utilisée ici (définie par l'équation IV.15 du chapitre IV), la "durée"[1] est déterminée par la constante de temps ΔT. Puisque la fenêtre est discontinue en 0, la résolution temporelle de détection lors des attaques est très bonne quel que soit ΔT. En revanche, la détection des extinctions peut être mise à mal si ΔT est trop important. La figure V.5 permet de mettre en évidence cet aspect, à partir d'une analyse en bande limitée (bande en tiers d'octave centrée autour de 2 kHz).

Pour des valeurs faibles de ΔT, l'indice de détection est très fluctuant, et ce particulièrement en l'absence de source. Dans ce dernier cas, la valeur moyenne reste relativement élevée (de l'ordre de 0,35 lorsque $\Delta T = 10\ ms$), alors que les signaux sont en théorie complètement décorrélés. En revanche, la réactivité du détecteur tant à l'attaque qu'à l'extinction du signal source est très bonne. Si ΔT augmente, les fluctuations sont de plus en plus faibles, et la valeur moyenne tend vers 0,5 lorsque les signaux sont décorrélés, si bien que la discrimination entre les moments où la source est présente et ceux où elle est absente est bien plus aisée. Cependant, si la réponse du détecteur aux attaques est toujours très rapide (puisqu'il s'agit d'une fenêtre causale), la réponse à une extinction est d'autant plus lente que ΔT est important[2].

Ceci gène la détection de la réverbération, car si le silence avant l'arrivée d'un nouveau son est de trop courte durée, l'indice de détection n'a pas le temps de remonter jusqu'au niveau

[1] Il s'agit d'un abus de langage, car rigoureusement, la notion de durée n'a pas de sens ici, puisque la fenêtre est infinie.

[2] On peut noter que l'on retrouve la recommandation proposée en section 3.1 du chapitre IV sur la valeur minimale de ΔT à respecter pour ne pas avoir de fluctuations trop importantes : en effet, la bande considérée à une largeur à -3 dB de 460 Hz, si bien que la valeur minimale de la constante de temps est de l'ordre de $\Delta T_{min} = 10/460 = 21\ ms$.

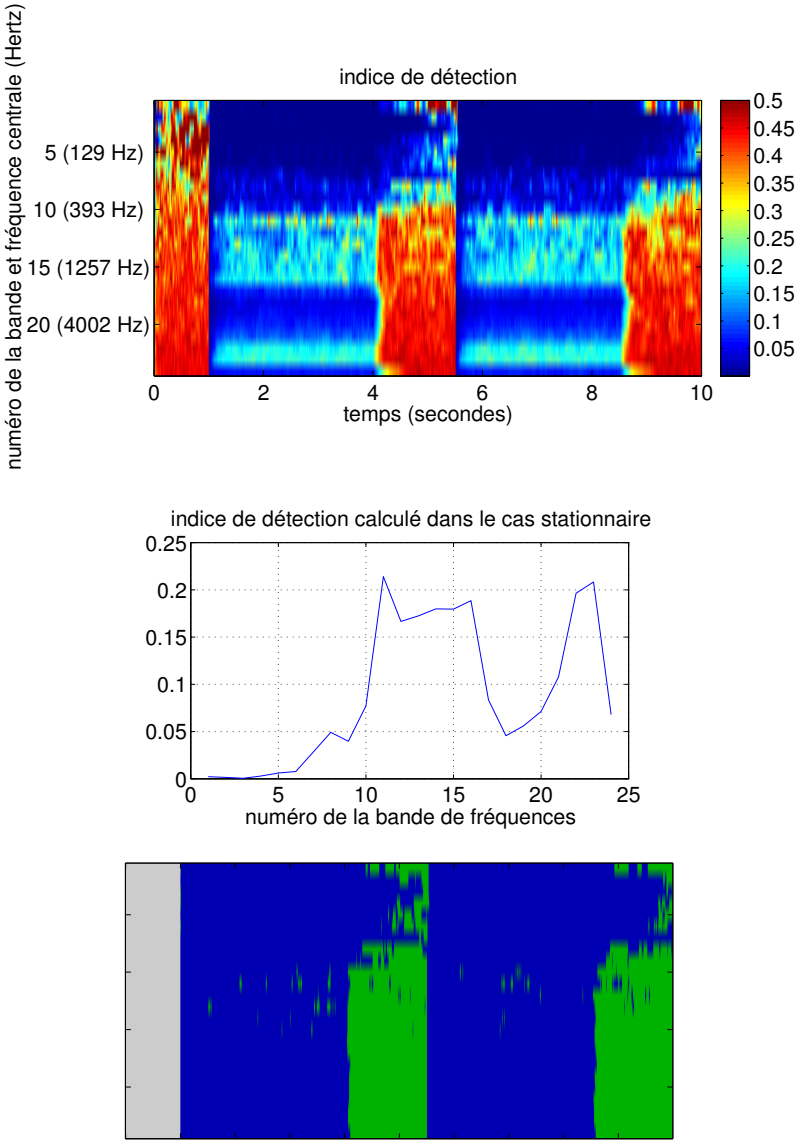


FIG. V.3 – **Exemple de détection de réverbération (banc de filtres en tiers d'octave)** : la figure supérieure représente à nouveau la corrélation à court-terme calculée après un banc de 24 filtres en tiers d'octave pour le même signal que celui dont l'étude est amorcée sur la figure V.1. La constante de temps de la fenêtre d'analyse de la cohérence vaut $\Delta T = 50\ ms$. La figure intermédiaire est la valeur de l'indice de détection pour une séquence de bruit stationnaire convoluée par la même réponse. La figure inférieure est le résultat de la détection, le seuil étant égal à la valeur stationnaire de l'indice majoré de 0,1. On y indique en vert (hachures verticales) quelles sont les zones où l'indice de détection est au dessus du seuil, et en bleu (hachures horizontales) celles où l'indice de détection est en dessous.

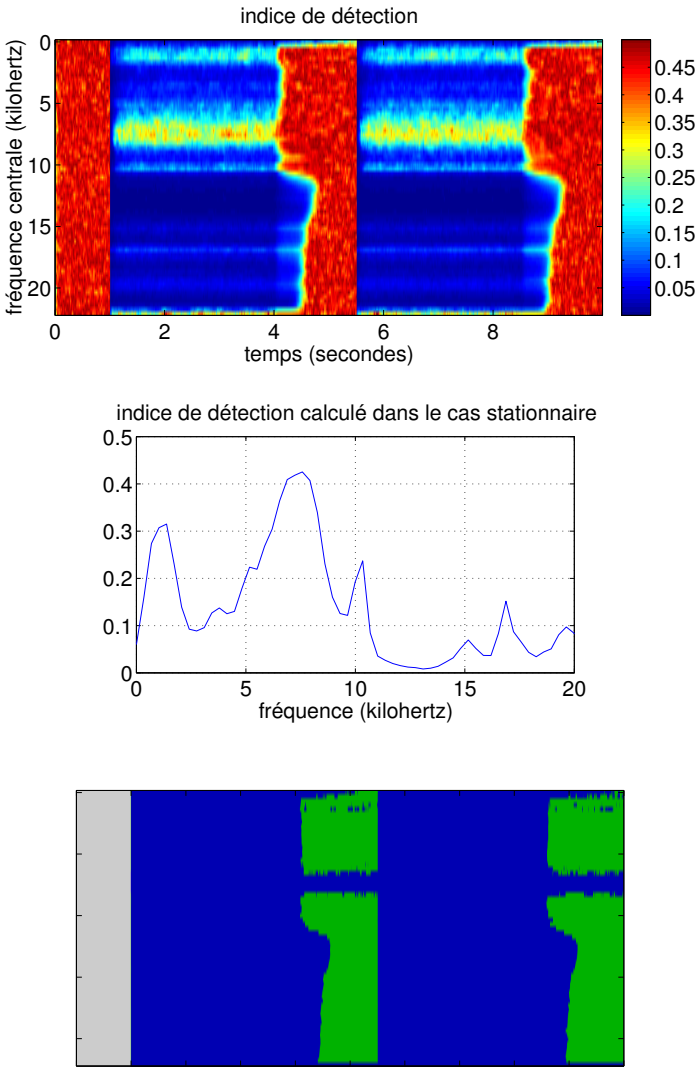
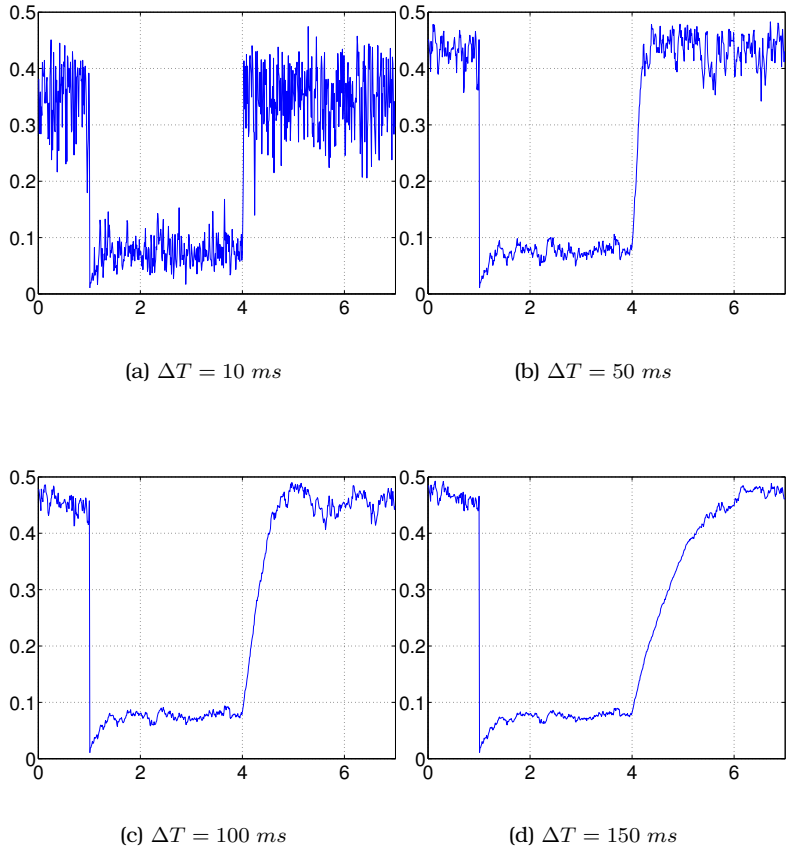


FIG. V.4 – **Exemple de détection de réverbération (transformation de Fourier à court-terme)** : la figure supérieure représente la cohérence à court-terme calculée à partir de la transformation de Fourier à court-terme du même signal que celui dont l'étude est amorcée sur la figure V.1. La fenêtre d'analyse temps-fréquence utilisée est une fenêtre de Hanning sur 128 points (2,9 ms). La constante de temps de la fenêtre d'analyse de la cohérence vaut $\Delta T = 50\ ms$. La figure intermédiaire est la valeur de l'indice de détection pour une séquence de bruit stationnaire convoluée par la même réponse. La figure inférieure est le résultat de la détection, le seuil étant égal à la valeur stationnaire de l'indice majoré de 0,1. On y indique en vert (hachures verticales) quelles sont les zones où l'indice de détection est au dessus du seuil, et en bleu (hachures horizontales) celles où l'indice de détection est en dessous.



(a) $\Delta T = 10\ ms$ (b) $\Delta T = 50\ ms$

(c) $\Delta T = 100\ ms$ (d) $\Delta T = 150\ ms$

FIG. V.5 – **Influence de la durée de la fenêtre d'analyse** : les courbes ci-dessus représentent l'indice de détection calculé en bande étroite sur une séquence de bruit interrompu convoluée à une réponse impulsionnelle de salle binaurale. La bande considérée est la bande en tiers d'octave centrée autour de 2 kHz. Le temps de réverbération y est de 2,75 secondes. Chaque courbe correspond à une valeur différente de la constante de temps ΔT de la fenêtre exponentielle utilisée pour le calcul de la corrélation à court-terme.

en champ diffus théorique. Ceci explique également les traînées mentionnées en section 1 : en effet, la réverbération est de faible énergie en hautes fréquences, et de plus décroît très vite ; son influence sur le calcul de la cohérence est donc négligeable par rapport à celle de l'onde directe, même plusieurs dixièmes de secondes après l'extinction de cette dernière. La fenêtre exponentielle agit comme un **masque** qui minimise l'effet d'évènements directement postérieurs à un évènement de forte énergie.

Le choix de la durée de la fenêtre résulte donc encore une fois d'un nécessaire compromis. En pratique, une valeur de ΔT comprise entre 50 ms et 100 ms s'avère adéquate pour décrire correctement la majorité des situations rencontrées.

Autres fenêtres

On peut se demander si un autre type de fenêtre ne convient pas mieux à la détection de réverbération : en particulier, la fenêtre rectangulaire, qui présente une double discontinuité, pourrait sembler idéale pour détecter les attaques **et** les extinctions avec une bonne résolution temporelle. Néanmoins il s'avère qu'elle a ses propres défauts : en effet, après plusieurs observations, il s'avère que si l'on souhaite que les fluctuations soient du même ordre avec une fenêtre rectangulaire qu'avec une fenêtre exponentielle, il faut que la fenêtre rectangulaire ait une longueur à peu près égale à $2,5 \times \Delta T$. Or on induit de fait un retard conséquent

sur la réaction aux extinctions, qui n'est pas souhaitable non plus.

On peut également envisager l'emploi de **fenêtres à durée variable** en fonction du niveau de corrélation (ou, de manière équivalente, de l'indice de détection) : en ajustant la durée de la fenêtre à une valeur relativement faible pour les zones pour lesquelles la corrélation est élevée (source active), et à une valeur importante pour les zones pour lesquelles la corrélation est faible (bruit et/ou réverbération seuls), on peut espérer à la fois minimiser les fluctuations en l'absence de source et offrir une réaction rapide aux extinctions. Une telle approche nécessite néanmoins un certain soin quant à l'implémentation, notamment pour s'assurer que la fenêtre est toujours à somme égale à 1 ; ainsi, cette idée n'a pas été mise en pratique ici, faute de temps.

1.4 Influence du niveau relatif d'énergie réverbérée

On a déjà mentionné que le niveau de corrélation en régime entretenu résulte d'un équilibre énergétique entre l'onde directe et la réverbération. Il suffit de considérer les cas extrêmes pour s'en persuader :

- En l'absence de réverbération, le coefficient de corrélation est par définition égal à la corrélation à long-terme mesuré sur les réponses impulsionnelles anéchoïques relatives à la configuration de microphones envisagée. Dans le cas d'une prise de son binaurale, il s'agit donc du coefficient de corrélation par bande de fréquences mesuré sur les HRTF pour la position de la source correspondante, qui excepté lorsque cette dernière est proche de l'axe interaural, vaut 1 à toutes les fréquences sauf au niveau des pôles et zéros (voir chapitre VI).
- Dans le cas limite où l'onde directe et les réflexions spéculaires sont noyées dans la réverbération, le niveau de corrélation correspond à la corrélation en champ diffus, qui reste très faible quelle que soit la fréquence centrale et la largeur de la bande considérée, excepté en très basses fréquences.

Dans une situation intermédiaire, le niveau de corrélation en régime entretenu sera donc fonction du niveau relatif d'énergie réverbérée. Les résultats de la figure V.6 confirment cette hypothèse, en indiquant l'évolution du module de la cohérence en configuration binaurale, en fonction de la distance du sujet à la source (la cohérence est calculée par la méthode du périodogramme de Welch sur la partie stationnaire du résultat de la convolution d'un bruit blanc par la réponse impulsionnelle binaurale). Il est visible que le niveau de cohérence est étroitement lié au rapport d'énergie directe sur l'énergie réverbérée. Du fait de la relation directe entre la cohérence et l'indice de détection en analyse de Fourier à court-terme, on retrouve le même phénomène pour ce dernier : **le niveau de l'indice de détection en régime stationnaire est d'autant plus élevé que le rapport d'énergie directe sur l'énergie réverbérée est faible**.

1.5 Choix du seuil

La dépendance entre l'indice de détection en régime stationnaire et le niveau relatif de la réverbération pose un réel problème en pratique : étant donné que l'on ne dispose *a priori* d'aucune information sur la scène sonore, on ne connaît pas la distance relative des sources par rapport à la distance critique, et on est donc incapable d'estimer à l'avance le rapport d'énergie directe sur l'énergie réverbérée pour la source concernée, et donc *a fortiori* la valeur de l'indice de détection stationnaire pour chaque fréquence. Cela rend bien plus difficile la détermination du seuil de détection.

On peut alors envisager deux types de recours : soit en fixant le seuil à un niveau indépendant de la fréquence, soit en estimant la valeur de l'indice de détection moyen au fur et à mesure de l'enregistrement. La deuxième approche, qui est la plus prometteuse, nécessite néanmoins une détection non plus en deux phases, mais en trois : il n'est plus question ici de repérer uniquement la présence d'une source (transitoire et entretien) ou son absence, mais de pouvoir distinguer le transitoire de la phase d'entretien et de la réverbération. On serait alors en mesure d'estimer au fur et à mesure le niveau de l'indice de détection dans le cas stationnaire, par exemple en calculant la moyenne de l'indice de détection pendant

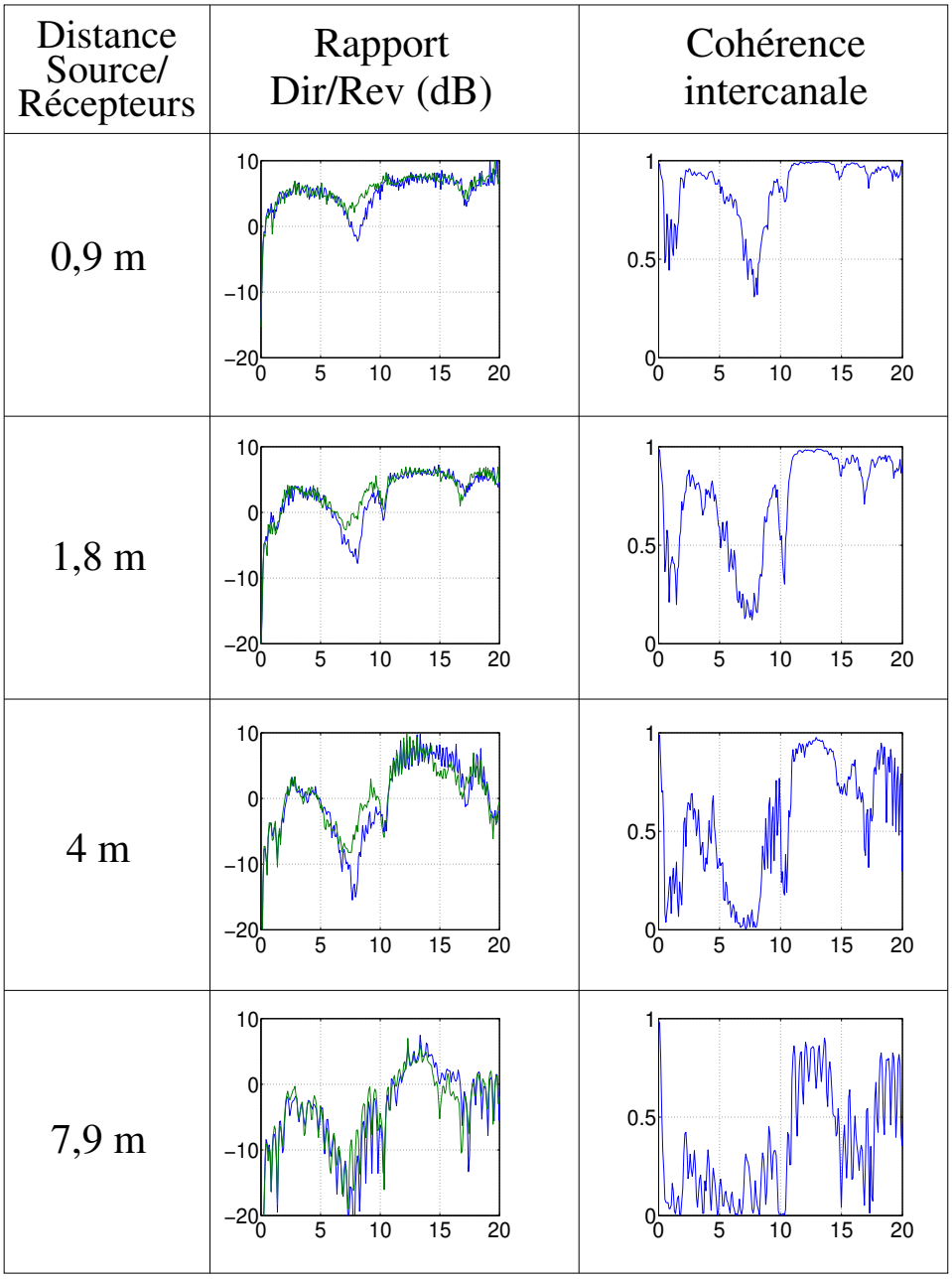


FIG. V.6 – **Influence du niveau relatif d'énergie réverbérée sur la cohérence stationnaire** : sont représentées, dans la salle étudiée depuis le début de ce chapitre, et pour quatre positions du sujet (qui fait toujours face à la source mais avec une distance plus ou moins importante), le rapport d'énergie du son direct sur l'énergie réverbérée totale en fonction de la fréquence en kilohertz (à gauche), et le module de la cohérence entre les deux microphones lorsque le signal source est un bruit blanc stationnaire, elle aussi fonction de la fréquence en kilohertz (à droite). Toutes les analyses sont basées sur des transformations de Fourier à court-terme avec une fenêtre de Hanning sur 512 points (11 ms) et un recouvrement de 75%.

les phases d'entretien. Cette piste n'a pas été suivie ici, faute de temps, et nécessite des recherches supplémentaires. Il s'agit néanmoins d'une question de grande importance, car en parvenant à estimer l'indice de détection stationnaire, et en menant à l'envers le raisonnement développé en section 1.4, on pourrait être en mesure d'estimer le rapport de l'énergie directe sur l'énergie réverbérée, qui est l'un des principaux indices permettant d'estimer la distance entre la source et le point d'écoute.

En l'absence d'une telle technique, il est donc nécessaire de choisir un seuil fixé à l'avance. Pour que ce seuil soit efficace en toutes circonstances, il doit se trouver au dessus de tous les niveaux de l'indice de détection en phase d'entretien dans les bandes de fréquences concernées, ce qui est difficile à envisager : il suffit de se référer aux résultats de la section précédente pour constater que puisque la corrélation stationnaire peut atteindre des valeurs proches de zéro lorsque la source est lointaine, il est impossible de définir un "seuil universel". Il est alors nécessaire d'émettre l'une ou l'autre des deux hypothèses suivantes : soit on suppose que la source reste proche du point d'écoute, de manière à ce que le niveau de l'indice de détection stationnaire ne soit pas trop élevé ; soit on suppose que les événements sonores émis par la source sont de très courte durée, si bien que dans ce cas, le régime stationnaire n'a pas le temps de s'installer, et il est alors possible de fixer le seuil à un niveau élevé. La seconde hypothèse facilite beaucoup la détection de réverbération, car au lieu de devoir fixer un seuil par rapport à un régime entretenu dont le niveau de l'indice de détection dépend énormément du contexte, on le fixe par rapport au régime transitoire, dont le niveau de l'indice de détection est plus stable par rapport à la proportion de réverbération. Pour la suite, et en admettant cette seconde hypothèse, le seuil est fixé à une valeur typiquement située entre 5 % et 10 %.

Il est utile de noter que cette question du choix du seuil ne concerne que la détection de la réverbération, puisque le niveau de l'indice de détection en régime transitoire reste proche de zéro (sauf pour des valeurs extrêmement petites du rapport d'énergie directe sur l'énergie réverbérée).

1.6 Exemple

La méthode de détection par seuil proposée ci-dessus est illustrée sur un exemple pratique (figure V.7). On considère comme signal source un extrait sec de voix parlée, le locuteur (masculin) prononçant la phrase "It does not enter his mind that he lives alone in the world". La réponse binaurale de salle employée, mesurée à une distance de 4 mètres de la source, fait partie du jeu de réponses servant depuis le début de ce chapitre d'exemple. L'analyse est effectuée par transformée de Fourier à court-terme sur 512 points, avec un recouvrement de 75 %. La constante de temps ΔT de la fenêtre exponentielle utilisée pour l'analyse par égalisation et annulation à court-terme vaut 100 millisecondes. Le seuil de détection est fixé à 10 %.

La comparaison des spectrogrammes du signal sec et d'une des deux voies du signal réverbéré permet de comprendre la limitation d'une analyse purement monophonique basée sur l'énergie dans un tel cas : concernant le spectrogramme du signal réverbéré, la représentation est moins précise dans le temps, le contraste bien plus faible au moment où survient un nouveau phonème, et seuls les événements les plus saillants (principalement pour les plosives et les fricatives) sont encore repérables.

Le calcul de l'indice de détection permet de lever, au moins en partie, cette incertitude, en fournissant des indices permettant de juger si la source a été active à un instant et dans une bande de fréquences donnée. Bien que les phonèmes soient peu visibles sur la représentation spectrographique, le fait qu'ils soient présents sur les deux voies au même moment fait chuter l'indice de détection. La comparaison par rapport au seuil fixé, soit 10 % dans ce cas, fournit un ensemble de zones correspondant aux instants et aux fréquences pour lesquels une activité de la source sonore a été détectée. Ces zones peuvent être mises en relation avec les maxima du spectrogramme du signal sec, et l'on observe qu'excepté dans les très basses fréquences, il y a un lien étroit entre ces deux représentations.

L'observation de l'analyse confirme une faiblesse prévue, qui est l'inefficacité de la détection en très basses fréquences : puisque la cohérence reste élevée à tous les stades de

la réverbération, l'indice de détection reste faible, et la détection ne fonctionne pas. Ceci explique la ligne bleue continue dans la représentation par zones de détection.

Il est également utile de noter que les zones en vert ne correspondent pas exactement aux zones pour lesquelles la réverbération a été détectée. En vertu des notions présentées au chapitre III, le problème de détection de réverbération ne peut être considéré rigoureusement comme le dual de celui de détection de source, car le cas où aucune source en activité n'est détectée peut aussi bien correspondre à la présence d'un champ réverbéré qu'à celui d'un silence ou d'une zone de bruit de fond. La mise en évidence des zones correspondant aux queues de réverbération nécessite donc une analyse supplémentaire, qui tienne compte à un instant donné de l'historique dans la bande considéré, conformément au schéma III.1 du chapitre III : y-a-t'il eu un événement sonore dans un proche passé ? L'énergie dans la bande considérée est-elle significative ? La section suivante présente un exemple d'une telle analyse dans le cas de signaux harmoniques.

1.7 Discussion

L'application d'un seuil sur l'indice de détection agit comme un masque sur la représentation temps-fréquence, en indiquant les zones pour lesquelles la source est jugée active. C'est d'ailleurs sur ce principe de masque que sont basées les techniques de déréverbération par cohérence à court-terme (Allen et al., 1977; Avendano et Jot, 2002). En revanche, on perd toute information sur l'aspect harmonique ou non des signaux considérés, et *a fortiori* sur la fréquence fondamentale. L'indice de détection, ou, de manière équivalente, la corrélation ou cohérence à court-terme, ne donne que très peu d'informations sur la structure d'un signal. Il ne permet que d'attester ou non de la présence simultanée sur les deux canaux d'un même signal.

2 APPLICATION AUX SIGNAUX HARMONIQUES

2.1 Introduction

Le chapitre IV et la section précédente ont permis de mettre en évidence les principes de la double détection de source et de réverbération principalement sur des séquences, interrompues ou non, de bruit blanc ou rose. Il s'agit de signaux simples à étudier, et ce à plus d'un titre : d'une part, ils sont à large bande passante, et contiennent donc des informations significatives à toutes les fréquences, alors que dans le cas de signaux réels, et en particulier des signaux harmoniques, certaines bandes ne contiennent que peu voire aucune énergie, et le rapport signal sur bruit y est de fait très défavorable ; d'autre part, les silences entre deux séquences successives dans les signaux étudiés en section 1 sont suffisamment longs (une seconde et demie) pour que la réverbération puisse se dérouler amplement, et pour que l'on puisse donc la décrire librement, sans interférences extérieures.

Pour des signaux réels, et en particulier des signaux musicaux, des silences d'une telle durée sont excessivement rares, et il faut faire face à un problème de **recouvrement temporel** des événements sonores par l'arrivée des événements sonores suivants. Ce recouvrement intervient même en supposant que les supports spectraux de l'événement étudié et de l'événement masquant sont disjoints pendant les phases entretenues, par exemple si les deux événements sont harmoniques et le rapport entre les fréquences fondamentales non rationnel. En effet, premièrement la résolution fréquentielle est limitée, et deuxièmement, il est très rare qu'un événement sonore soit parfaitement harmonique, et il est quasiment incontournable que la partie harmonique du signal soit accompagnée d'une composante non harmonique à large bande plus ou moins forte dans chacun des signaux. Dans le cas de signaux musicaux, cette composante non harmonique est typiquement due soit au geste lors de l'attaque (percussion, pincement), soit à celui nécessaire à l'entretien (souffle, frottement).

Ce recouvrement est une double source de gêne pour la détection. Pour la détection de source, la réverbération tardive d'un évènement antérieur se comporte vis-à-vis de la détection de l'évènement présent comme une source de bruit additionnelle, ce qui gène à la fois la détection proprement dite et l'estimation de la direction. La détection de réverbération, quant

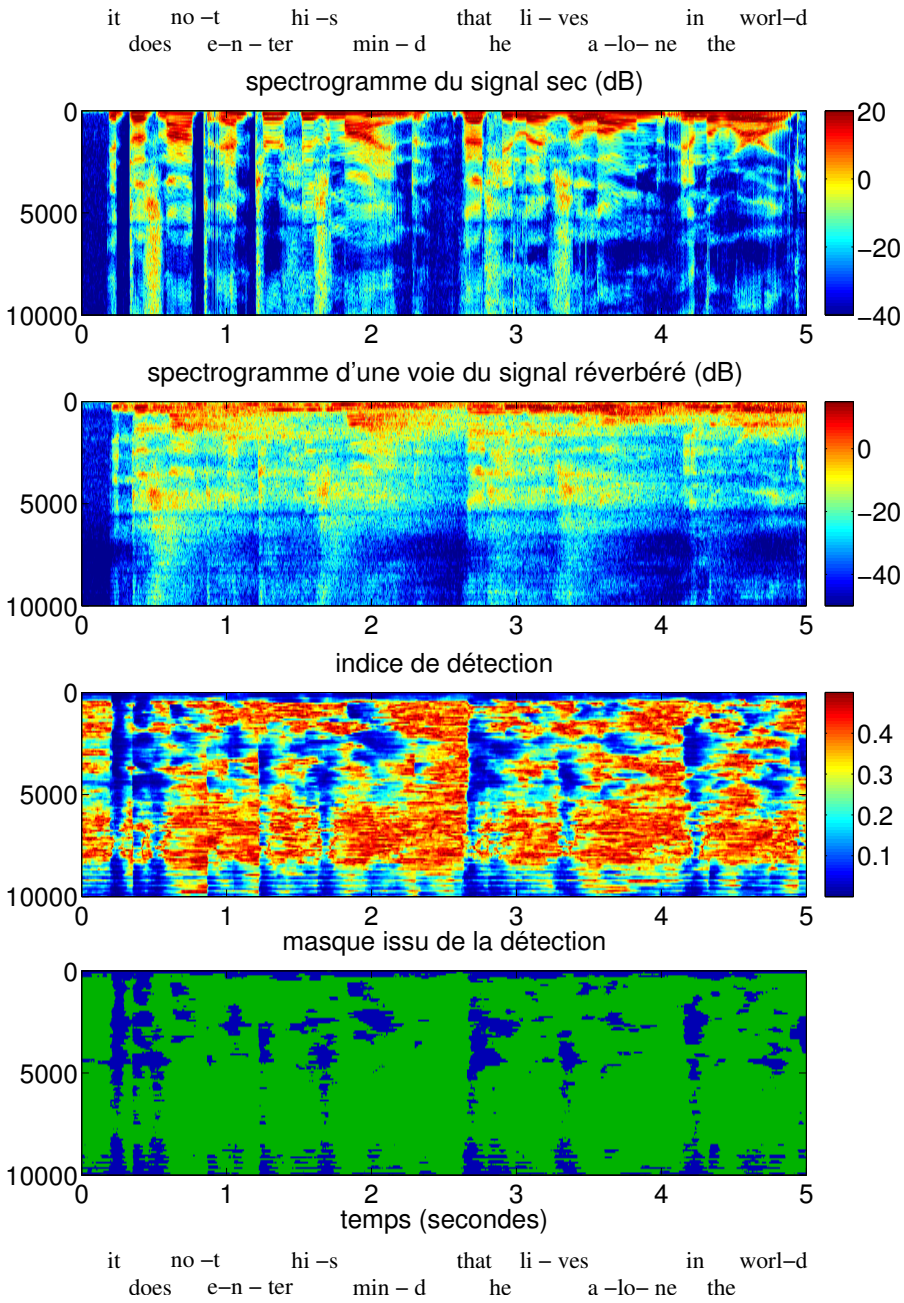


FIG. V.7 – **Exemple de détection** : le signal observé est un extrait de voix parlée, réverbéré par convolution avec une réponse binaurale de salle. Les transformées de Fourier à court-terme sont calculées avec une fenêtre de Hanning de 12ms (512 points), et un recouvrement de 75%. L'indice de détection est calculé à partir de la cohérence à court-terme, avec une fenêtre exponentielle de constante de temps $\Delta T = 100\ ms$. La figure inférieure représente en bleu les zones pour lesquelles l'indice de détection est en dessous de 0,1 (source détectée), et en vert celles pour lesquelles l'indice de détection est au dessus de 0,1 (réverbération ou bruit ou silence détecté).

à elle, est limitée dans le temps par la venue des événements sonores suivants, ce qui réduit la plage de temps sur laquelle l'estimation du temps de réverbération est possible.

Néanmoins, moyennant une ou plusieurs hypothèses supplémentaires sur la nature des signaux-sources, on peut parfois **rehausser** le signal cible par rapport au signal masquant, et ainsi améliorer le rapport signal sur bruit. Compte-tenu du fait que l'on ne connaît pas la position de la source (puisque c'est l'une des informations que l'on cherche), il n'est pas possible d'effectuer un rehaussement sur des critères spatiaux par pointage d'antenne. L'approche qui a été développée dans cette étude est monophonique, et se base sur l'hypothèse d'**harmonicité** des évènements sonores : on suppose que chaque évènement sonore est d'une part en grande partie harmonique, et d'autre part que la fréquence fondamentale instantanée est quasiment constante sur toute la durée du signal source. De fait, la réverbération tardive associée sera non plus harmonique (puisque la structure de la réverbération fait chuter l'harmonicité (Wu et Wang, 2003)), mais néanmoins à structure quasi-harmonique, c'est-à-dire constituée de bandes étroites dont les fréquences centrales suivent une loi harmonique. De fait, à supposer que l'on soit capable d'estimer la fréquence fondamentale de l'évènement sonore, on dispose d'une information très précieuse sur la nature de la réverbération tardive qui lui est associée.

L'estimation de la fréquence fondamentale proprement dite n'est pas l'objet de cette étude[3]. La technique utilisée ici est présentée dans une publication récente (Baskind et Cheveigné, 2003), qui peut être consultée en annexe E. Elle est le fruit d'une adaptation à des signaux monodiques réverbérés d'une technique d'estimation de fréquence fondamentale multiple récemment développée par Cheveigné (Cheveigné, 1993; Cheveigné et Baskind, 2003), sur le modèle d'un estimateur robuste de la fréquence fondamentale par égalisation et annulation développé précédemment (Cheveigné et Kawahara, 2002).

Ce rehaussement s'effectue nécessairement événement par événement, et nécessite donc une **segmentation temporelle** des signaux observés, de manière à appliquer sur chaque segment le traitement *ad hoc*. Cette segmentation nous procure un avantage supplémentaire, qui est de pouvoir détecter de manière plus fine les zones de réverbération par rapport au masque tel qu'il a été proposé dans la section précédente : en effet, si l'on suppose que chaque segment ne contient après rehaussement qu'un seul événement, la zone de réverbération correspond à la plage de temps, **juste après l'événement**, du moment où l'indice de détection passe à nouveau au dessus du seuil jusqu'à celui où il le franchit à nouveau (instant qui correspond à l'arrivée de l'événement suivant). Le schéma V.8 résume le principe complet de la détection pour de tels signaux.

2.2 Rehaussement de signaux harmoniques à fréquence fondamentale constante par morceaux

La première possibilité de rehaussement, en supposant que l'on connaît la fréquence fondamentale instantanée, consiste à appliquer un filtre sur l'événement masquant pour en supprimer l'essentiel de l'énergie. Les filtres les plus courants permettant de réaliser cette annulation sont des **filtres en peigne,** définis par une réponse impulsionnelle du type :

$$h_T[n] = \frac{1}{2}\left(\delta[n] - \delta[n-T]\right) \tag{V.1}$$

...où $\delta[n]$ est la fonction impulsion, et T est la période de l'événement que l'on cherche à annuler. Ceci suppose que l'événement est à fréquence fondamentale constante sur toute sa durée (ce qui exclut par exemple le cas de signaux de parole) ; en effet, bien qu'il soit possible d'appliquer un filtrage variant dans le temps pour suivre la fréquence fondamentale si celle-ci n'est pas constante, on ne maîtrise plus l'effet d'un tel traitement sur les harmoniques du signal à rehausser. Dans le cas d'un filtre invariant dans le temps, on peut assurer que l'atténuation de chaque harmonique est constante (ce qui préserve par exemple l'aspect exponentiel de la décroissance), et l'on est capable de quantifier cette atténuation.

[3] On pourra se reporter à la publication reproduite en annexe E pour une présentation d'une méthode d'estimation de la fréquence fondamentale pour des signaux monodiques réverbérés.

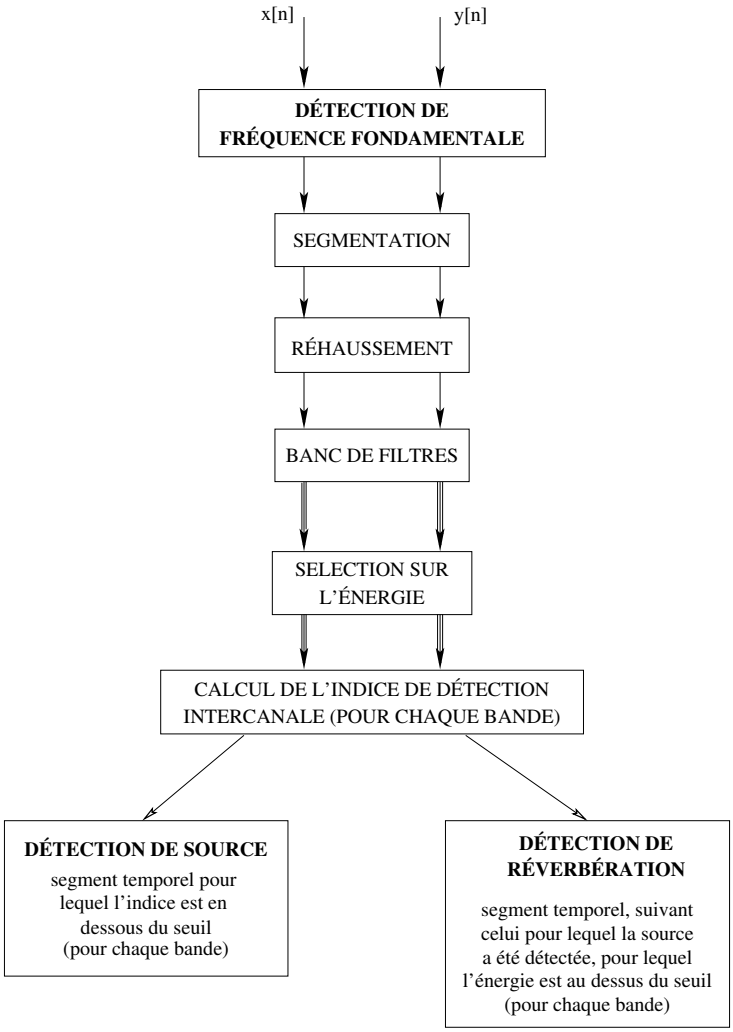


FIG. V.8 – **Schéma de principe de la détection sur des signaux harmoniques** (dans le cas où la fréquence fondamentale est constante par morceaux) : les deux voies $x[n]$ et $y[n]$ du signal observé sont tout d'abord soumises chacune à une détection de fréquence fondamentale. Ceci permet d'isoler chacune des notes, et de rehausser la note courante par rapport à la note précédente et à la note suivante. On sélectionne alors les bandes de plus forte énergie, avant d'appliquer la détection par égalisation et annulation qui permet d'isoler le moment où la source est active de celui où seule la réverbération est présente.

De plus, la formule V.1 suppose que la période fondamentale T est entière, c'est-à-dire que le retard est multiple de la période d'échantillonnage. Lorsque ce n'est pas le cas (et c'est rarement le cas), il est très utile d'avoir recours à des **retards fractionnaires**. Il existe de nombreuses manières de simuler un retard fractionnaire. Le choix qui a été fait ici consiste en des filtres FIR reposant sur l' **interpolation de Lagrange** (Välimäki et Laakso, 2000). Ce type d'implémentation est à complexité ajustable, en agissant sur l'ordre de l'interpolation. A l'ordre 1 on retrouve la classique interpolation triangulaire, c'est-à-dire :

$$h_T[n] = \frac{1}{2}\delta[n] - \frac{1}{2}\left((1 - T_{frac}) * \delta\left[n - T_e\right] + T_{frac} * \delta\left[n - T_e - 1\right]\right)$$

...où $T = T_e + T_{frac}$, T_e désignant la partie entière de T, et T_{frac} la partie décimale. Les filtres utilisés dans cette étude utilisent des interpolations d'ordre 2, ce qui est un compromis raisonnable entre performances et complexité.

Il faut noter que quelles que soient les performances du filtre, il ne peut que supprimer au

mieux la partie harmonique du signal masquant. La composante inharmonique est filtrée, mais comme le signal relatif à l'événement que l'on cherche à rehausser l'est également, le rapport d'énergie entre le signal utile et le bruit inharmonique est peu affecté par un tel traitement, qui sera donc d'efficacité réduite dans le cas d'instruments pour lesquels la composante inharmonique peut être très marquée, et en particulier sur les attaques, comme par exemple le piano.

La figure V.9 permet d'illustrer cette technique de rehaussement sur une simulation. Le signal source est composé de deux événements sonores identiques à une contraction temporelle près ; il s'agit d'une note de flûte traversière très courte (227 ms), choisie car la flûte est un instrument avec une composante harmonique très stable. La première note, non dilatée, est à la fréquence $F1 = 660\ Hz$ (mi4) ; la seconde est montée d'un ton, et a pour fréquence $F2 = 742\ Hz$. Le signal observé résulte de la convolution de ce signal source avec l'une des réponses binaurales de salles étudiées depuis le début de ce chapitre, qui est celle mesurée pour une distance de 4 mètres entre la source et le sujet. Sont représentés les puissances et les indices de détection pour les réponses avec ou sans rehaussement de l'une ou de l'autre note, et l'on compare ces données au cas où l'annulation est idéale, c'est-à-dire pour laquelle même la composante non harmonique est annulée.

Il est nécessaire de prévoir à l'avance l'effet d'un tel traitement sur le signal que l'on cherche à rehausser. Dans le meilleur des cas, le rapport entre les fréquences fondamentales de l'événement masquant et de l'événement à rehausser est non rationnel ; dans cette situation, les séries harmoniques relatives aux deux événements sont totalement disjointes, et les harmoniques de l'événement à rehausser sont atténuées mais aucune n'est annulée. En revanche, si le rapport entre les fréquences fondamentales est rationnel, une ou plusieurs harmoniques de l'événement à rehausser sont annulées.

2.3 Analyse pitch-synchrone

Ce type de rehaussement n'est envisageable que pour des bancs de filtres uniformes. Il vient en complément du premier, et vise à configurer l'analyse pour qu'elle concentre l'information dans quelques bandes, de manière à y maximiser le rapport signal sur bruit. Le principe consiste à faire en sorte que chacune des harmoniques du signal à étudier soient centrées dans la bande de fréquences qui les contient. Les fréquences centrales dans un banc de filtres uniforme obéissent à une loi du type :

$$f_k = \frac{k}{N}.F_e$$

...Nétant la longueur des filtres en échantillons, et F_e la fréquence d'échantillonnage.

Si f_0 désigne la fréquence fondamentale de l'événement à analyser, alors une analyse pitch-synchrone est obtenue si les fréquences f_k peuvent s'écrire :

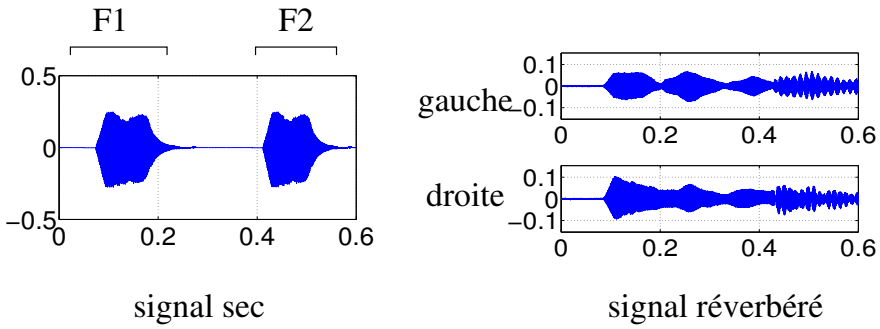
$$f_k = \frac{k}{M}.f_0$$

...M étant un facteur de suréchantillonnage fréquentiel. Ceci implique :

$$N = M.\frac{F_e}{f_0}$$

Si ce nombre n'est pas entier, on choisit l'entier le plus proche, l'approximation étant d'autant meilleure que M est grand. Cependant, lorsque M augmente, d'une part le coût de calcul augmente aussi, surtout si N n'est pas une puissance de deux, et d'autre part, la largeur de chaque bande diminue. Or on ne peut pas diminuer indéfiniment la largeur de bande, sous peine de perdre de l'information sur la réverbération. La largeur de bande doit en effet être supérieure à la largeur de bande de la réverbération d'un son pur interrompu, qui vaut, si l'on se place dans le cas d'une décroissance purement exponentielle :

$$\Delta F_{-3dB}^{reverb} = \frac{3.ln(10)}{\pi.Tr} \simeq \frac{2,2}{Tr}\ (Hz)$$



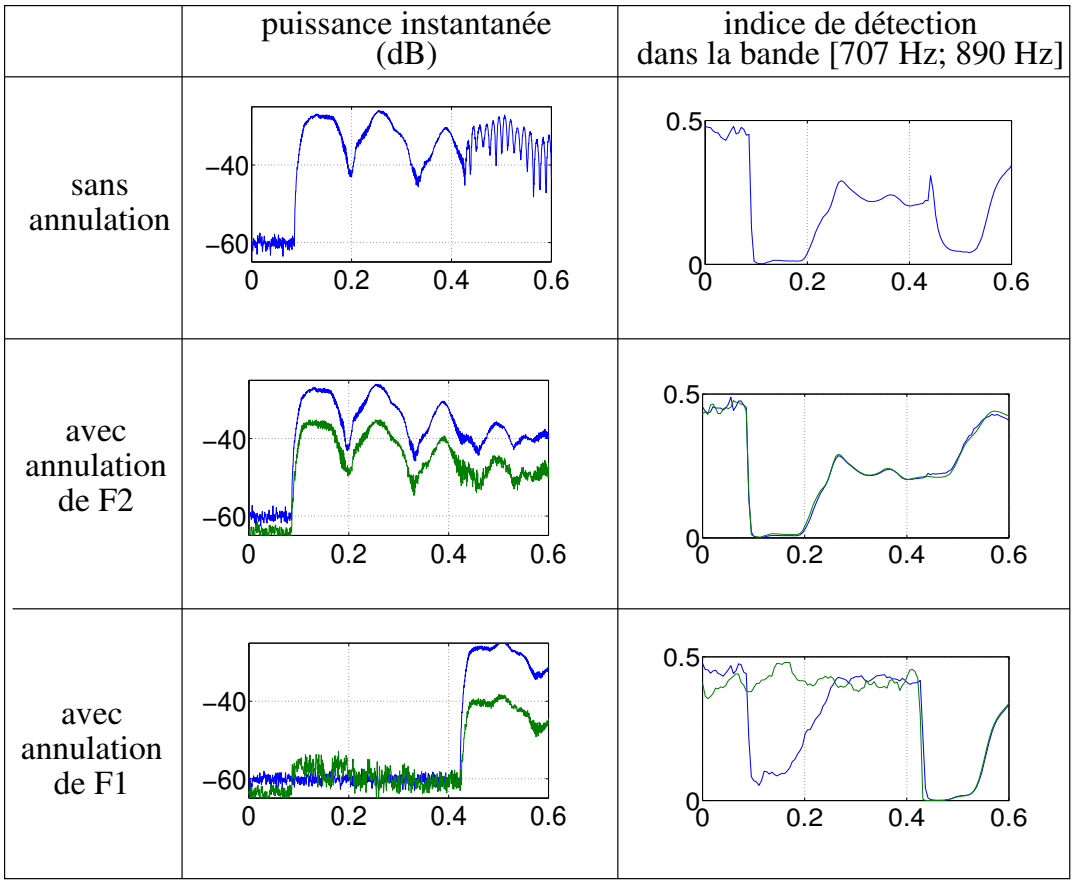


FIG. V.9 – **Exemple de rehaussement de signaux harmoniques** : le signal source (en haut à droite) est composé de deux événements sonores identiques à une contraction temporelle près. Le signal observé (en haut à droite) résulte de la convolution de ce signal source par une réponse binaurale de salle, avec addition de bruit (rapport S/B=40dB). Sur le tableau principal sont représentés la puissance instantanée pleine bande (calculée avec une fenêtre de Hanning de 1,5 ms) et l'indice de détection dans la bande en tiers d'octave centrée sur 793 Hz, pour trois configurations : 1-sans prétraitement ; 2-après rehaussement de la première note (annulation de F2) ; 3-près rehaussement de la seconde note (annulation de F1). Pour chacun, la courbe bleue représente le résultat après annulation idéale, et la courbe verte le résultat après annulation par filtrage en peigne.

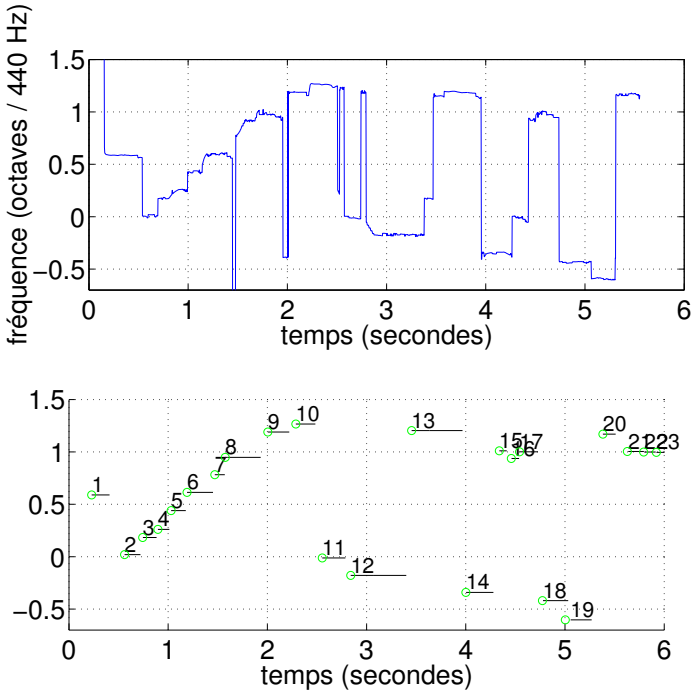


FIG. V.10 – **Exemple de détection sur un signal harmonique - estimation de la fréquence fondamentale** : la figure supérieure est le résultat de l'estimation sur le signal observé de la fréquence fondamentale. La figure inférieure est une effectuée à partir de cette information ainsi que de l'indice d'apériodicité fourni par l'estimateur, comme préalable à la détection.

D'autre part, grâce à une analyse pitch-synchrone, il n'est plus nécessaire d'effectuer la détection, puis les estimations de la position et du temps de réverbération, dans chacune des bandes. Puisque l'on sait que l'information utile est concentrée dans les bandes d'indice $k = pM+1$ (p étant un entier positif), il suffit d'effectuer l'analyse dans ces bandes-ci uniquement.

2.4 Exemple

L'exemple qui suit est un extrait de musique qui se prête bien à la méthode de rehaussement proposée ci-dessus. Il s'agit d'un enregistrement de la *Partita* de Bach pour flûte seule. Comme cela a déjà été indiqué, la flûte est un instrument avec une composante harmonique très stable en fréquence. La composante inharmonique, qui est due principalement au souffle dans ce cas, est ici d'importance raisonnable dans ce cas, comme le montre l'exemple de la figure V.9, pour lequel le segment à la base des deux événements sonores est en fait la première note de cet enregistrement. Il s'agit d'un enregistrement sec, qui a été réverbéré par convolution avec la même réponse impulsionnelle binaurale de salle que précédemment, qui est donc mesurée à 4 mètres de la source.

La première étape consiste à estimer la fréquence fondamentale, selon la méthode présentée en annexe E. La figure V.10 indique le résultat de cette estimation, ainsi que celui d'une analyse qui en découle, visant à découper le flux audio en notes, que l'on définit ici comme des segments de fréquence fondamentale stable. Il est alors possible d'appliquer les méthodes de rehaussement présentées ci-dessus. Pour chaque note, on applique sur les deux voies deux filtres en peignes accordés sur les fréquences fondamentales de la note précédente et de la note suivante, lorsque le rapport entre ces dernières et la fréquence courante n'est pas une puissance de 2.

Pour chacune de ces notes prétraitées, on applique la méthode de détection par égalisation et annulation proposée dans cette étude, avec une constante de temps $\Delta T = 100\ ms$. Le banc de filtres retenu est la transformation de Fourier à court-terme. Le plus adapté serait

d'employer une analyse pitch-synchrone, mais pour des raisons purement liées à la clarté de la présentation de cet exemple, on choisit une analyse à longueur de fenêtre constante. Ceci permet, à la figure V.11, de représenter le résultat de détection simultanément pour toutes les notes.

Par rapport à l'exemple de la section 1.6, on gagne énormément en clarté. Le fait d'effectuer l'analyse note par note, et non plus dans la globalité, permet pour chacune de faire une détection appropriée : avant tout, seules les harmoniques contenant suffisamment d'énergie sont étudiées ; de plus, l'utilisation des techniques de rehaussement présentées ci-dessus permet une analyse bien plus précise ; finalement, et contrairement à l'exemple précédent, on peut distinguer les zones de réverbération (en vert sur l'illustration graphique) des zones de bruit de fond (en gris). Alors que l'exemple précédent ne permettait pas de définir clairement les segments où effectuer l'estimation du temps de réverbération, on met en évidence, pour chaque note et pour chaque harmonique significative, un intervalle de temps suivant directement la plage de temps d'activité de la source sur lequel on peut décrire l'enveloppe de réverbération.

Bien que l'on recouvre de nombreuses informations utiles par rapport à une représentation spectrographique (il suffit de comparer, sur la figure V.11, les représentations du milieu et inférieure), il subsiste encore quelques zones d'ombre, notamment lors de successions trop rapides de notes de fréquence voisine. Ainsi, tous les événements ne sont pas détectés pendant la montée initiale (entre autres, le mi4 vers 1,4 seconde n'est pour ainsi dire pas détecté), ainsi que pendant le trille vers 6 secondes. Là encore, aucun événement n'a été détecté en basses fréquences, soit en dessous de 250 Hz.

3 CONCLUSION

Dans ce chapitre a été étudiée l'application du principe de détection par égalisation et annulation à des signaux réverbérés. Il y est montré que si la détection de source est soumise à peu d'ambiguïté dans la plupart des situations, la détection de réverbération est une tâche plus délicate dans le cas général : d'une part, le seuil de détection optimal dépend du contexte (notamment de la distance à la source), et d'autre part, il est nécessaire de distinguer les zones correspondant spécifiquement à des queues de réverbération de celles où le bruit de fond prédomine. La résolution de ces ambiguïtés nécessite une analyse plus fine, ainsi que des hypothèses supplémentaires sur le signal, comme par exemple celle consistant à admettre que celui-ci n'est constitué que d'événements sonores de relative courte durée, de manière à ce que l'état stationnaire n'ait pas le temps de s'installer.

Il est indispensable d'ajouter que l'efficacité de la méthode de détection dépend étroitement de la capacité à distinguer onde directe et réverbération selon des critères de cohérence spatiale. Ainsi, puisque celle-ci reste élevée en basses fréquences à tout instant, la détection y sera inopérante. La fréquence de coupure à partir de laquelle la détection fonctionne dépend de la différence de marche entre les microphones.

La méthode de détection proposée précise et généralise la méthode de séparation d'énergie directe et de réverbération proposée par Allen et al. (1977) et reprise par Avendano et Jot (2002). L'apport du paradigme proposé est double : d'une part, il permet de fixer une condition de durée sur la fenêtre d'analyse (celle-ci doit être de durée bien supérieure au retard maximal entre les deux microphones pour que la cohérence à court-terme puisse être considérée comme une approximation valide de la corrélation à court-terme) ; d'autre part, on place ainsi la méthode de détection de réverbération par cohérence à court-terme dans un cadre théorique, tout comme l'on a pu placer au chapitre précédent les méthodes d'estimation des différences intercanales de temps et d'intensité dans un cadre théorique plus général de détection et localisation de source.

On s'est également intéressés au cas des signaux harmoniques, et plus spécifiquement à celui où la fréquence fondamentale est constante par morceaux. Cette situation, qui correspond à beaucoup de signaux musicaux, permet, grâce à la connaissance de la fréquence fondamentale, de pousser l'analyse plus loin, d'améliorer notablement le rapport signal sur bruit, et d'affiner la détection de réverbération.

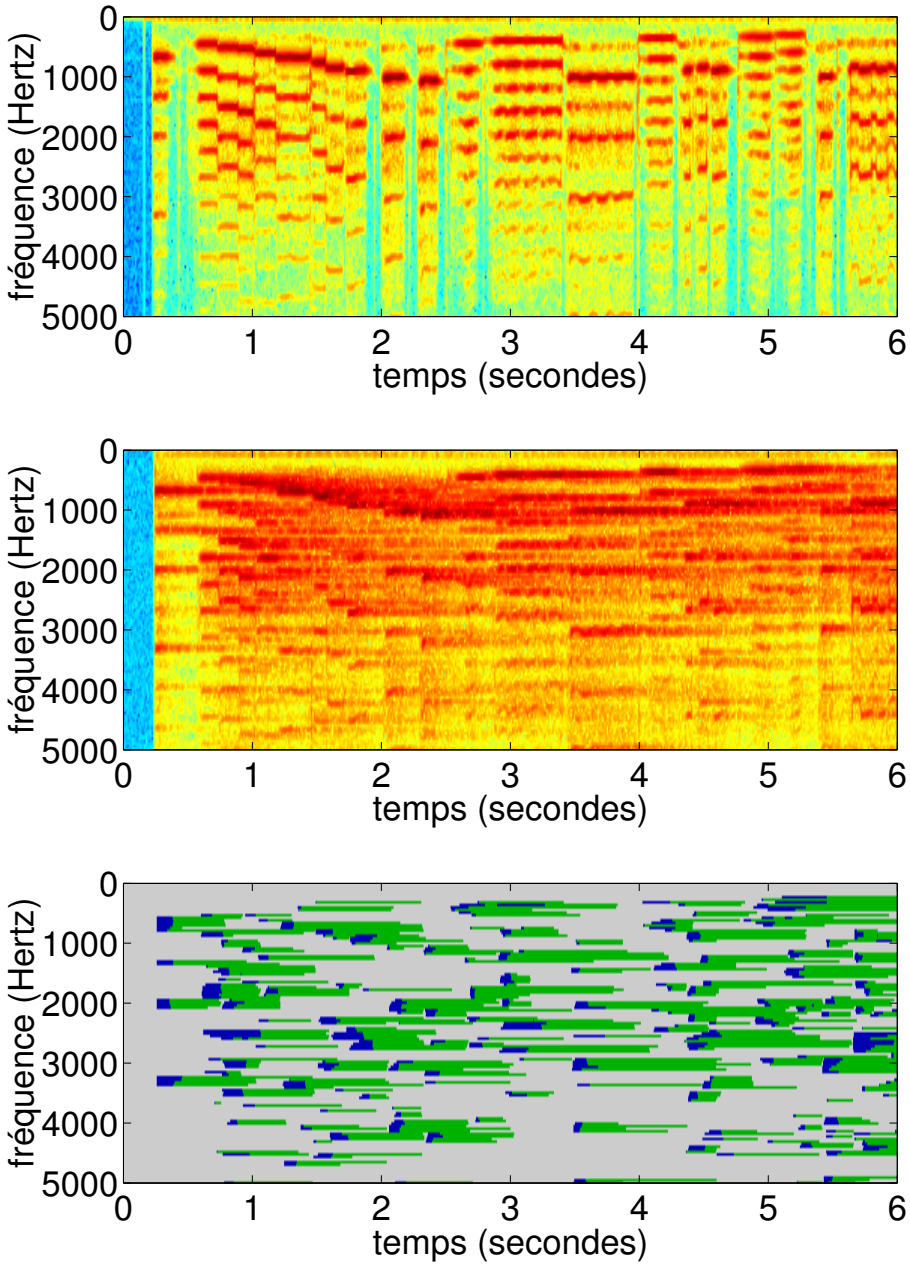


FIG. V.11 – **Exemple de détection sur un signal harmonique** : le signal observé résulte de la convolution d'un enregistrement très sec d'un extrait de la *Partita* de Bach pour flûte seule par une réponse binaurale de salle. Le signal obtenu est analysé par un détecteur utilisant à la fois la connaissance de la fréquence fondamentale instantanée et la cohérence pour révéler les zones où la source et la salle prédominent (pour des raisons de clarté de présentation, l'analyse effectuée n'est pas pitch-synchrone). Les figures supérieure et intermédiaire sont des représentations spectrographiques du signal source et d'une des voies du signal réverbéré, respectivement. La figure inférieure est le résultat de l'analyse, représentant en bleu les zones jugées dominées par l'onde directe, et en vert les zones jugées dominées par la réverbération.

On dispose donc à ce stade d'une estimation des zones temps-fréquences pour lesquelles la source ou la réverbération prédominent, ainsi que des valeurs optimales des paramètres d'égalisation et annulation (c'est-à-dire les retards et gains intercanaux) à chaque instant et dans chaque bande de fréquences. Il faut maintenant effectuer la description spatiale elle-même, c'est-à-dire d'une part tirer parti de la connaissance des retards et gains intercanaux aux instants auxquels la détection est la meilleure (c'est-à-dire les attaques) pour en déduire la position de la source physique, et d'autre part de se servir de la connaissance des plages de temps pour lesquelles la réverbération a été jugée prédominante pour décrire les caractéristiques temporelles de son enveloppe.

Troisième partie

Méthodes de description spatiale

VI

Estimation de la direction de la source

Le chapitre IV présente une méthode conjointe de détection bicanale d'évènements et d'estimation des paramètres physiques relatifs à leur position que sont les différences de temps et d'intensité par bande de fréquences entre les deux canaux. Il s'agit maintenant de pouvoir déduire de ces paramètres une véritable estimation géométrique de la position de la source, tâche qui se heurte à deux sources d'incertitudes : d'une part, les incertitudes liées à la configuration géométrique de la prise de son binaurale, et d'autre part, les incertitudes liées à la méthode d'estimation. En se basant sur les spécificités de la propagation et sur les méthodes d'estimation de la direction les plus connues dans le cas d'enregistrements binauraux (section 1), on propose une configuration du banc de filtres qui y soit adaptée (section 2) ; puis est présentée une méthode d'estimation de la direction qui, en étant étroitement liée à la méthode de détection et d'estimation des indices intercanaux, vise à dépasser ses principales sources d'ambiguité (section 3). Finalement, on propose en section 4 un exemple complet d'estimation de la direction à partir d'un court extrait d'un enregistrement binaural *in situ* de musique.

1 PARTICULARITÉS DE L'ESTIMATION DE DIRECTION À PARTIR D'ENREGISTREMENTS BINAURAUX

1.1 Spécificité de la prise de son binaurale

La plupart des techniques d'estimation de la direction de provenance d'une onde (sonore, électromagnétique ou autres) en l'absence de connaissances sur l'information qu'elle véhicule font appel à des antennes de capteurs (Nicol, 1996; Chen et al., 2002; Varma, 2002), dont la directivité est supposée maîtrisée dans une bande de fréquences la plus large possible. Ainsi, les traitements utilisant des antennes de microphones supposent souvent que ces derniers sont omnidirectionnels, approximation qui est souvent valide jusqu'à plusieurs kilohertz pour des capsules de bonne qualité, ou à directivité basée sur des harmoniques sphériques du premier ordre (bidirectionnels, cardioïdes et dérivés). Dans ce cadre, la relation entre la direction de provenance et les paramètres physiques liant les signaux de l'antenne peut être formulée simplement de manière analytique : lorsque les microphones sont omnidirectionnels et peu distants les uns des autres, il est possible de négliger la différence d'amplitude, et les signaux issus de la captation de l'onde sonore émise par une source ponctuelle ne diffèrent que par des retards. Inversement, si la prise de son est effectuée au moyen de microphones directionnels coïncidents (couples X-Y, microphones ambisoniques,...), les signaux issus de la captation de la même onde sont synchrones quelle que soit la direction de la source, mais n'ont pas la même intensité.

Cependant, la configuration de prise de son binaurale, qu'elle fasse appel à une tête artificielle ou un sujet réel, se distingue de ces configurations particulières de prise de son d'ambiance par plusieurs aspects.

D'une part, l'introduction de cet obstacle à géométrie complexe qu'est le buste (c'est-à-dire la tête et le torse) modifie grandement la directivité des microphones : le torse, l'effet d'ombre de la tête et le pavillon sont responsables de phénomènes d'interférences complexes qui modifient sensiblement la réponse en fréquence de chaque capteur. De fait, il devient très difficile d'établir un lien simple entre la direction de provenance de la source et les paramètres de gain et de retards dans chaque bande de fréquence. L'estimation de la direction de provenance devient alors une tâche à part entière.

D'autre part, une prise de son binaurale est intrinsèquement **individuelle** : alors que la majorité des systèmes de prise de son utilisent des microphones de série dont la directivité et la réponse en fréquence sont connues, il est difficile de connaître *a priori* les caractéristiques géométriques et le comportement mécanique et acoustique du buste, sauf s'il s'agit d'une tête artificielle de série (ou d'un obstacle apparenté, du type de la tête Charlin), dont les caractéristiques sont connues. Étant donnée l'étroite dépendance de la réponse de chaque capteur (et donc des indices interauraux[1]) à la géométrie de l'obstacle, l'estimation de la direction de provenance de la source nécessite une phase préliminaire d'**apprentissage**, qui vise à configurer la méthode de détection et/ou de localisation pour un buste donné. Cet apprentissage peut s'effectuer de deux manières différentes.

La première approche consiste à configurer un modèle paramétrique de localisation ou de latéralisation à partir d'expériences psychoacoustiques (Tollin, 1998; Lindemann, 1986a,b) : on configure le modèle de telle sorte à ce qu'il fournisse des résultats similaires aux jugements moyens de localisation. De fait, cette approche participe d'une modélisation de d'audition, visant à prédire la direction de provenance **perçue** d'un son, plutôt que la direction **objective** de la source physique.

La seconde approche consiste à configurer le modèle à partir de la connaissance la plus exhaustive possible des indices utiles, soit ici les indices interauraux. Ces indices peuvent être soit **calculés** par analyse des HRTF mesurées pour un grand nombre de directions de provenance (Gaik, 1993; Martin, 1995), soit **estimés** en fonction des dimensions géométriques de la tête, du pavillon et du torse. Cette estimation à partir de données géométriques nécessite un **modèle physique** de la propagation acoustique au voisinage du buste, qui soit suffisamment précis pour rendre compte des principaux aspects mentionnés en sec-

[1] Il suffit par exemple de songer à la relation entre les différences de temps et les dimensions de la tête pour s'en persuader.

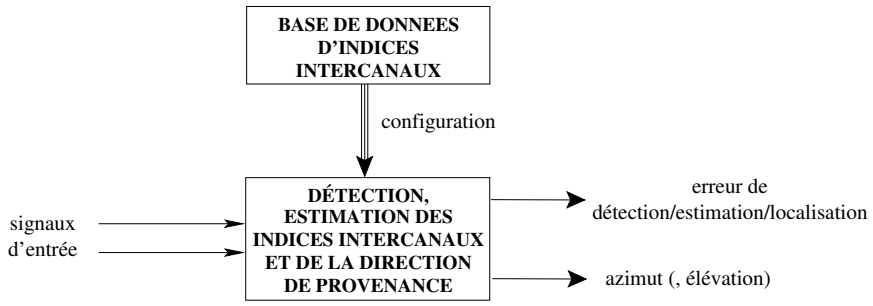


FIG. VI.1 – **Estimation de la direction de provenance par individualisation de la méthode de détection et d'estimation**

tion 2, ce qui peut être envisagé par l'essor récent de techniques de modélisation numérique (Katz, 1998). L'inconvénient d'utiliser des mesures de HRTF est que celles-ci ne sont envisageables que pour un nombre fini de positions de la source, mais on peut avoir recours à des techniques d'interpolation spatiale afin d'inférer la valeur des indices pour les directions inconnues[2]. Ces informations sont utilisées soit pour inférer une loi statistique (Gaik, 1993), soit pour établir une base de données qui permette de décider de la direction de provenance la plus probable de la source parmi toutes celles à disposition (Martin, 1995). C'est cette seconde approche qui sera employée ici, et l'on se basera sur la connaissance de HRTF mesurées sur chacun des sujets concernés.

La connaissance des HRTF en champ lointain pour un grand nombre de directions de source peut être exploitée à divers niveaux de l'analyse, soit directement au sein de la méthode de détection et d'estimation des indices interauraux, soit à la suite de l'estimation, dans un module de décision séparé.

1.2 Estimation de la direction de provenance à partir d'une représentation temps-espace

Le principe de ce type de méthode d'estimation de la direction de provenance est de modifier et de configurer la méthode de détection et d'estimation pour qu'elle fournisse une représentation qui puisse être interprétée dans chaque bande de fréquences, non plus comme une représentation temps-retard (modèles par corrélation) ou temps-retard-gain (modèles par égalisation et annulation), mais directement comme une représentation temps-espace (figure VI.1). Les différences de temps et d'intensité sont alors des paramètres pris implicitement en compte, mais non accessibles de manière explicite.

La majorité de ces méthodes découlent du modèle de coïncidence de Jeffress qui au cours des années a subi de nombreuses modifications, tant sous sa forme de réseau de neurones (Schauer et Paschke, 1999), que sous sa forme de réseau par corrélation. Par exemple, le modèle de Jeffress ne tient compte que des retards entre les deux voies, et ne tient pas compte des différences interaurales d'intensité, qui sont pourtant des indices tout aussi importants vis-à-vis de la localisation. Sayers et Cherry (1957) proposent, sans modifier le modèle lui-même, une transformation préalable qui vise à intégrer les différences d'intensité dans le processus de latéralisation. Lindemann (1986a; 1986b) propose une modification plus profonde du modèle, en incluant un principe d'inhibition contralatérale issu de théories sur les causes physiologiques de l'effet de précédence, qui permet à la fois de modéliser celui-ci et de tenir compte des différences d'intensité. Il semble également que ce mécanisme permette d'améliorer sensiblement la résolution du motif d'activation sur l'axe des retards. Gaik (1993) ajuste le modèle de Lindemann de manière à tenir compte de la relation entre les différences

[2]Ce problème de l'estimation des fonctions de transfert de tête pour des positions intermédiaires est bien connu en synthèse binaurale, et une méthode couramment employée, est l'interpolation entre deux ou trois positions connues (interpolation linéaire ou triangulaire), typiquement sur les paramètres d'un modèle ARMA+(retard pur) des HRTF (Larcher, 2001; Freeland et al., 2002).

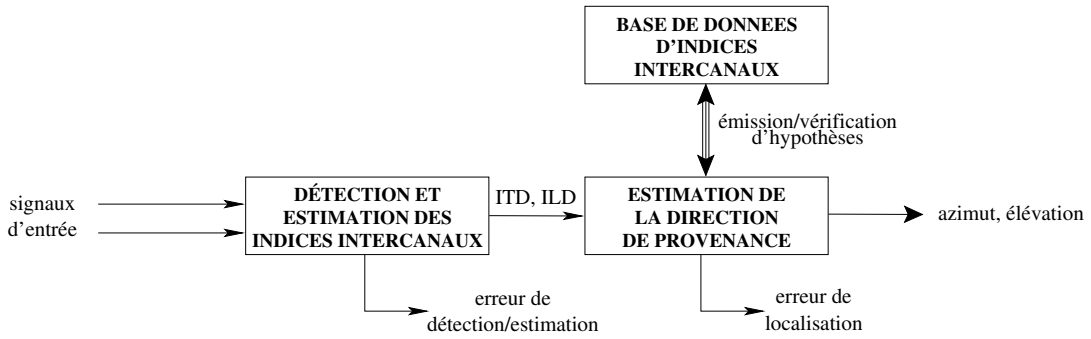


FIG. VI.2 – **Estimation de la direction de provenance distincte de la détection**

de temps et d'intensité. Ces différences de temps et d'intensité sont évaluées à partir de la connaissance des HRTF pour 122 directions possibles de la source, et servent comme données statistiques permettant de paramétrer une loi polynomiale de relation entre ITD et ILD par bande critique. Un ultime apport au modèle de Jeffress-Lindemann-Gaik est celui de Bodden (1993), qui consiste à transformer l'axe de latéralisation sur la représentation de Lindemann en un axe azimutal.

Ces modèles d'estimation de la direction sont des modèles de **latéralisation**, c'est-à-dire d'estimation de la position intracrânienne dans le cas d'une écoute dichotique, ou de l'angle du cône de confusion par rapport à l'axe interaural dans le cas d'une écoute binaurale. Ainsi, dans le modèle de Jeffress-Lindemann-Gaik, les fluctuations des différences interaurales au sein d'un même cône de confusion, qui certes sont de faible amplitude, sont gommées par l'analyse statistique. Pour obtenir une estimation complète de la direction de provenance de la source, il est nécessaire d'y adjoindre une estimation de l'angle dans le cône de confusion, par exemple basée sur des critères spectraux monauraux. La décision finale, qui se base sur les indices de latéralisation dans chaque bande critique et sur les indices monauraux, est dans ce cas l'objet d'une ultime étape de l'analyse (Chung et al., 2000; Searle et al., 1976).

1.3 Estimation de la direction de provenance par critère de distance

Cette méthode, proposée par Richard Duda et reprise par Martin (1995), sépare explicitement l'estimation des indices interauraux de celle de la direction de provenance en deux étapes distinctes (figure VI.2) : le premier module fournit au second une estimation la meilleure possible des indices interauraux pour les bandes critiques jugées significatives, et celui-ci en déduit la direction de provenance la plus probable possible, par comparaison avec les indices interauraux connus pour des positions données de la source. Il s'agit donc d'un problème de **décision** parmi un nombre fini de directions possibles.

L'algorithme de décision est la plupart du temps basé sur un modèle statistique liant les indices interauraux calculés pour une source donnée à la direction de provenance de la source : on modélise ainsi les retards de phase IPD_k, les retards d'enveloppe IED_k et les différences d'intensité ILD_k (k étant l'indice de la bande de fréquences concernée parmi les N bandes du banc de filtres) comme des variables aléatoires qui sont la somme de la valeur nominale (déterministe) de l'indice concerné pour la direction de la source, et d'un bruit d'estimation :

$$\begin{cases} IPD_k = ipd_k^0(\theta,\varphi) + B_{IPD_k}(\theta,\varphi) \\ IED_k = ied_k^0(\theta,\varphi) + B_{IED_k}(\theta,\varphi) \\ ILD_k = ild_k^0(\theta,\varphi) + B_{ILD_k}(\theta,\varphi) \end{cases} \forall k \in \{1...N\}$$

...θ et φ étant respectivement l'azimut et l'élévation de la source, $ipd_k^0(\theta,\varphi)$, $ied_k^0(\theta,\varphi)$ et $ild_k^0(\theta,\varphi)$ étant les valeurs nominales des indices, et $B_{IPD_k}(\theta,\varphi)$, $B_{IED_k}(\theta,\varphi)$ et $B_{ILD_k}(\theta,\varphi)$ étant les bruits d'estimation aléatoires. Si l'on suppose que les bruits d'estimation sont **gaussiens**, **centrés** et **indépendants de la direction de provenance**, une estimation au sens du

maximum de vraisemblance équivaut à chercher le couple (θ, φ) qui minimise l'erreur quadratique :

$$d^2(\theta,\varphi) = \sum_{k=1}^{N} \left[\frac{\left(IPD_k - ipd_k^0(\theta,\varphi)\right)^2}{\sigma_{IPD_k}^2} + \frac{\left(IED_k - ied_k^0(\theta,\varphi)\right)^2}{\sigma_{IED_k}^2} + \frac{\left(ILD_k - ild_k^0(\theta,\varphi)\right)^2}{\sigma_{ILD_k}^2} \right]$$

...σ_{IPD_k}, σ_{IED_k} et σ_{ILD_k} étant les écarts-type des bruits d'estimation de chacun des indices interauraux. Cette erreur peut être considérée comme une **mesure pondérée de distance** dans un espace à $3N$ dimensions entre un vecteur cible $(IPD_1...IPD_N, IED_1...IED_N, ILD_1...ILD_N)^T$ et des vecteurs candidats, chacun étant constitué par les indices interauraux pour une direction donnée. Cette distance peut également s'exprimer, grâce à une formulation du type du théorème de Pythagore, en fonction de distances par bande de fréquences :

$$d(\theta,\varphi) = \sqrt{\sum_{k=1}^{N} d_k^2(\theta,\varphi)} \tag{VI.1}$$

...avec

$$d_k(\theta,\varphi) = \sqrt{\frac{\left(IPD_k - ipd_k^0(\theta,\varphi)\right)^2}{\sigma_{IPD_k}^2} + \frac{\left(IED_k - ied_k^0(\theta,\varphi)\right)^2}{\sigma_{IED_k}^2} + \frac{\left(ILD_k - ild_k^0(\theta,\varphi)\right)^2}{\sigma_{ILD_k}^2}} \tag{VI.2}$$

Cette méthode offre l'avantage de permettre, si la précision l'autorise, une estimation de la direction de provenance en azimut et en élévation, sans forcément nécessiter des méthodes monaurales d'estimation de l'élévation. Elle permet également de pondérer le rôle respectif des retards de phase, d'enveloppe et des différences d'intensité, en pondérant cette mesure de distance séparément pour chaque terme. Cela dit, elle pose deux problèmes principaux, qui sont la détermination des écarts-type et l'absence de prise en compte de l'erreur d'estimation des indices interauraux.

Pour déterminer les écarts-type, Martin (1995) propose, puisqu'il vise par cette méthode à établir un modèle d'audition binaurale, d'employer les **seuils différentiels de perception** (*Just Noticeable Differences*, ou *JND*) des indices interauraux. Plus précisément, il utilise les JND en temps et en intensité mesurés par Hershkowitz et Durlach (1969) sur des sons purs à 500 Hz, en supposant qu'ils sont identiques pour toutes les positions de la source et toutes les fréquences. Or cette dernière hypothèse est peu réaliste, tant de point de vue de l'audition que de la méthode d'estimation utilisée : en effet, les différences de temps et d'intensité sont estimées respectivement par maximisation de la corrélation et par rapport des puissances, ce qui correspond au cadre présenté de l'estimation par égalisation et annulation. Or la remarque de la section 4.5 du chapitre IV, notamment, indique que la précision de l'estimation dépend étroitement de la fréquence centrale de la bande concernée, et ce quel que soit le contexte de prise de son. Dans le cas d'une prise de son binaurale, la section 2 montrera de plus que la précision de l'estimation, qui dépend de la position des pôles et des zéros de la fonction de transfert intercanale, est donc étroitement liée à la direction de provenance de la source.

La seconde critique que l'on peut adresser à cette méthode de localisation, qui est liée à la remarque précédente, est qu'elle ne tient pas compte de l'erreur d'estimation des indices interauraux pour la source considérée : les biais dûs à la présence de bruit additionnel, par exemple, ne sont pris en compte que dans le sens où ils tendent à augmenter la distance $d(\theta,\varphi)$, mais il est difficile de déterminer dans l'absolu si une distance minimale non négligeable est due à la présence de bruit additionnel, ou au fait que l'onde sonore provienne d'une direction non répertoriée dans la base de données, qui on le rappelle ne contient les informations sur les indices interauraux que pour un nombre fini de directions[3]. De même, il n'est pas tenu compte ici de l'ambiguïté sur l'estimation du retard de phase en bande étroite mentionnée à la section 4 du chapitre IV, ceci car le comportement du modèle a été étudié par Martin principalement pour des bruits à large bande.

[3]Typiquement, si la direction de la source se trouve entre deux directions connues, la distance $d(\theta,\varphi)$sera du même ordre et non nulle pour ces deux directions.

1.4 Discussion

Puisque l'on cherche à estimer la direction de provenance en azimut et en élévation, il est nécessaire de conserver les disparités entre les indices intercanaux au sein des cônes de confusion, et ceci est un argument d'importance en faveur de la méthode par mesure de distance. En revanche, il paraît important, et spécialement hors du contexte de modélisation de l'audition, de dépasser son défaut principal, qui est l'absence de prise en compte de la méthode d'estimation des indices intercanaux. En ce sens, une méthode plus intégrée sur le modèle de celles présentées en section 1.2 apporte un plus indéniable. La méthode que l'on proposera en section 3 résulte d'une volonté de lier les avantages de ces deux approches.

2 CONFIGURATION DE LA MÉTHODE DE DÉTECTION POUR DES ENREGISTREMENTS BINAURAUX

Comme indiqué au chapitre IV, les méthodes basées sur l'égalisation et l'annulation portent sur une caractérisation plus ou moins précise (en fonction de la résolution fréquentielle) de la fonction de transfert intercanale. Dans le cas d'enregistrements binauraux, cette fonction de transfert est la **fonction de transfert interaurale**, qui est égale au rapport des HRTF contralatérale et ipsilatérale pour la position donnée de la source. Le choix d'un banc de filtres pour l'analyse requiert avant tout une meilleure connaissance des particularités de ces fonctions de transfert, et donc une compréhension minimale des phénomènes de propagation.

2.1 Organisation structurelle des HRTF

La validité du modèle gain-retard dépendant de la régularité de la fonction de transfert intercanale dans chaque bande de fréquences, il est nécessaire avant toute chose de rappeler les principaux éléments constitutifs de la structure des HRTF. En première approximation, on peut considérer que les HRTF mesurées en conduit bouché[4] résultent de l'influence conjointe de trois éléments simples qui sont la tête (sans pavillons), le torse et les pavillons.

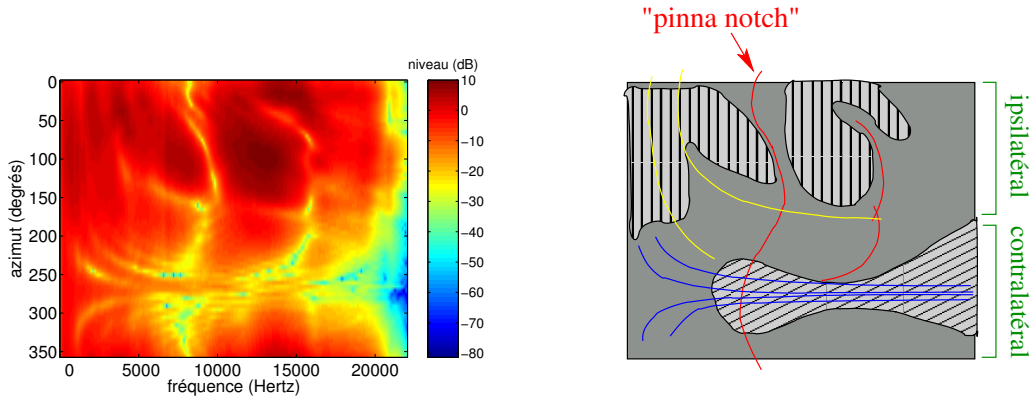
Influence de la tête

Si l'on écarte pour l'instant les pavillons, la principale influence de la tête est un effet de masque acoustique vis-à-vis de l'oreille contralatérale. Un modèle sphérique (Rayleigh, 1907; Kuhn, 1977), bien que très simple, permet de mettre en valeur les deux principaux aspects de cet effet d'ombre, que l'on peut observer par exemple à la figure VI.3, dans la zone où l'oreille considérée est contralatérale (c'est-à-dire pour des angles supérieurs à 180°) : en basses fréquences, les ondes acoustiques ayant contourné la tête par diffraction forment des franges d'interférences dont les fréquences caractéristiques et les amplitudes dépendent principalement de l'angle de la source par rapport à l'axe interaural ; en hautes fréquences, et en particulier lorsque la source se trouve près de l'axe interaural, la tête joue pleinement son rôle de masque, et le spectre est globalement atténué. On peut remarquer que ces deux aspects se retrouvent dans la fonction de transfert interaurale, sous forme de résonances et d'antirésonances dont l'amplitude varie entre 20 et 60 dB.

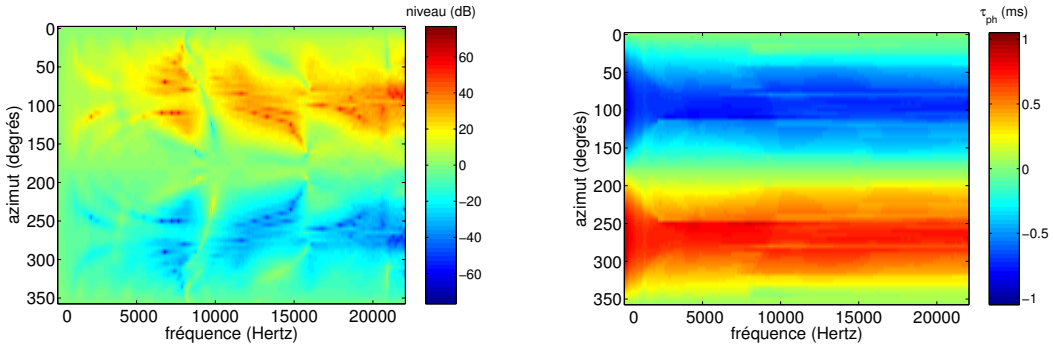
Influence du pavillon

Le pavillon joue un rôle primordial dans la localisation au sein d'un cône de confusion, et en particulier en hautes fréquences, pour lesquelles les ondes acoustiques, se réfléchissant sur les parois du pavillon et de la conque, créent des interférences dépendant de l'angle d'incidence. La manifestation la plus marquée est l'"antirésonance de pavillon" (*pinna notch*), intervenant typiquement entre 6 kHz et 8 kHz (Brown et Duda, 1998; Algazi et al., 2001a), correspondant aux interférences dans la conque pavillonnaire. L'observation des figures VI.3 et VI.4 montre que la position de cette antirésonance dépend sensiblement de la direction

[4] Si le conduit auditif est ouvert, il faut ajouter sa résonance qui intervient vers 2kHz.



(a) représentations réelle et schématique des HRTF gauches (plan horizontal)



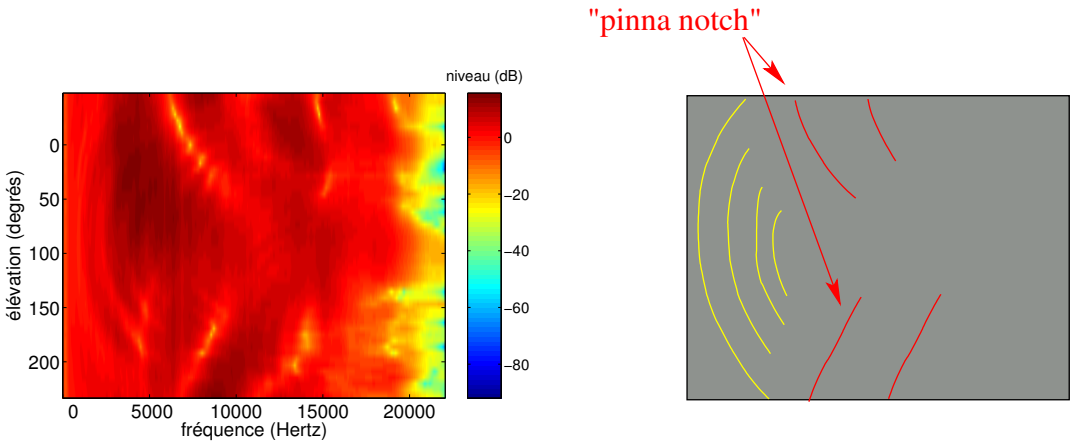
(b) amplitude et retard de phase des fonctions de transfert interaurales (plan horizontal)

FIG. VI.3 – **dépendance azimutale des HRTF dans le plan horizontal** : sur la figure en haut à gauche sont affichés les spectres d'amplitude des HRTF mesurées par Gardner et Martin (1994) sur l'oreille gauche d'une tête artificielle KEMAR (symétrique), et dans le plan horizontal. La figure en haut à droite est une représentation schématique qui en indique les aspects principaux : les zones hachurées verticalement correspondent aux maxima de la fonction de transfert ; les zones hachurées de manière oblique indiquent les fréquences atténuées par l'effet d'ombre de la tête lorsque la source est à droite ; les lignes bleues correspondent aux zéros dûs aux interférences relatives aux trajets multiples des ondes sonores vers l'oreille contralatérale, les lignes jaunes à ceux dûs aux interférences sur le torse, et les lignes rouges à ceux dûs aux interférences dans le pavillon. Les figures inférieures représentent l'amplitude et le retard de phase des fonctions de transfert interaurales.

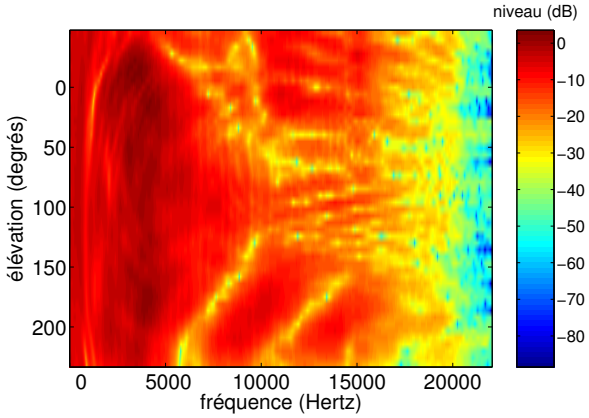
d'incidence, et en particulier de l'élévation. Ici encore, on retrouve cet effet dans la fonction de transfert interaurale.

Influence du torse

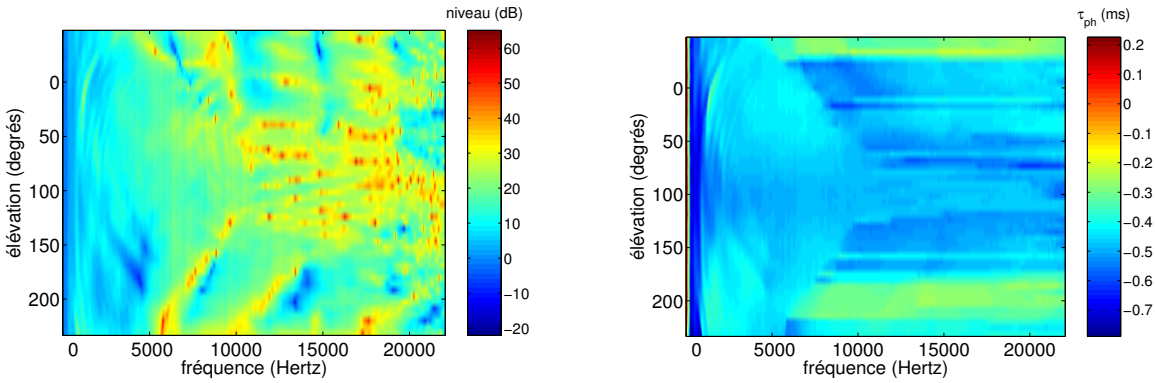
L'interaction des ondes acoustiques avec le torse est également une source d'informations utile, et ce particulièrement au sein d'un cône de confusion. Ainsi, on peut observer sur la figure VI.4 des franges d'interférences dépendant de l'angle d'élévation, et ce pour les trois types de fonctions de transfert (ipsi-, contralatérale et interaurale), mais elles sont particulièrement manifestes sur les fonctions de transfert ipsilatérales. En revanche, les interférences



représentation réelle et schématique des spectres d'amplitude des HRTF gauches



représentation réelle des spectres d'amplitude des HRTF droites



amplitude et retard de phase des fonctions de transfert interaurales

FIG. VI.4 – **dépendance à l'angle d'élévation δ des HRTF dans un cône d'axe interaural** : sur la figure en haut à gauche sont affichés les spectres d'amplitude ipsilatéraux des HRTF mesurées par Algazi et al. (2001b) sur un sujet réel, dans le cône faisant un angle de 45° avec l'axe interaural. La représentation schématique en haut à droite met à nouveau en valeur les principales antirésonances dues au pavillon (lignes rouges) et les franges d'interférences dues aux réflexions et à la diffraction sur le torse (lignes vertes). La figure intermédiaire représente les fonctions de transfert contralatérales, et les figures inférieures l'amplitude et le retard de phase des fonctions de transfert interaurales (pour une définition de l'angle δ et des autres angles utiles, on pourra se reporter à l'annexe A)

entre l'onde directe et celle réfléchie sur le torse sont d'importance relativement faible dans le plan horizontal (Algazi et al., 2002) : ainsi, la figure VI.3 indique que sur pour une tête artificielle KEMAR, l'amplitude des interférences, visibles surtout en dessous de 5 kHz et pour une source ipsilatérale, sont de l'ordre de 5 dB.

2.2 Variation des indices interauraux

Les remarques de la section précédente amènent à plusieurs constatations quant aux indices interauraux. Ainsi, on observe que les différences d'intensité, qui correspondent à l'amplitude de la fonction de transfert interaurale, sont comme on s'y attendait quasiment nulles en basses fréquences quel que soit l'angle d'incidence, mais présentent des variations significatives à partir de 1 kHz environ. Il est difficile de dégager une tendance globale quant aux niveaux relatifs ipsilatéraux et contralatéraux en fonction de la position, car les variations fréquentielles sont trop importantes. Une analyse à bande étroite trouve tout son sens ici, car elle permet de distinguer des directions au sein d'un même cône de confusion.

Quant au retard de phase, on peut rappeler que chaque résonance ou antirésonance provoque une discontinuité de phase dont l'amplitude varie en fonction du facteur de qualité et de l'ordre du pôle ou, respectivement, du zéro. On peut observer sur les figures VI.3 et VI.4, et surtout sur la figure VI.5, l'effet de l'antirésonance principale du pavillon sur la régularité du retard de phase. En moyennes fréquences, celui-ci est relativement régulier, mais il subit une nette discontinuité à la fréquence centrale du pôle ou du zéro (qui varie entre 5 kHz et 10 kHz), puis reste au delà relativement irrégulier, en raison du grand nombre de pôles et de zéros dûs principalement au filtrage pavillonnaire. En basses fréquences, le retard de phase est, conformément aux prévisions de Kuhn (1977), plus important d'un facteur 3/2 qu'en moyennes fréquences, la zone de transition se situant entre 500 Hz et 2 kHz.

Le retard de groupe est, de manière générale, beaucoup moins stable que le retard de phase, ce qui est compréhensible puisqu'il est issu d'une dérivation de la phase, opération qui a par nature tendance à amplifier les fluctuations hautes fréquences. En dehors des pôles et zéros (qui d'ailleurs sont repérables de façon évidente sur le retard de groupe), l'amplitude des fluctuations est de l'ordre de la milliseconde, soit la moitié de la marge possible de valeurs ! Puisqu'il est peu envisageable que l'analyse ait une précision fréquentielle suffisante pour suivre toutes ces fluctuations, on pourra s'attendre à une **mauvaise résolution sur l'estimation du retard de groupe**[5].

2.3 Choix du banc de filtres

Le choix du banc de filtres repose sur plusieurs notions : d'une part, il est nécessaire que l'approximation gain-retard par bande des fonctions de transfert interaurales soit la meilleure possible dans la majorité des bandes ; d'autre part, la densité de filtres doit être suffisamment élevée dans la zone de fréquences utile de manière à ne pas perdre d'information ; de plus, ce choix de la densité doit résulter d'un compromis vis-à-vis d'une part de la résolution temporelle, et d'autre part du coût de calcul. Cette "zone de fréquences utile" correspond simplement au domaine de fréquences dans lequel on est susceptible d'avoir les meilleurs rapports signal sur bruit dans la situation étudiée. Puisque l'on vise *in fine* à analyser des enregistrements musicaux ou de parole, il est naturel de privilégier le registre inférieur du spectre (jusqu'à quelques kilohertz), qui contient l'essentiel de l'information utile.

La détermination de la validité du modèle gain-retard en fonction du banc de filtres dépend de la largeur de bande, et donc du facteur de qualité de chaque filtre du banc. La figure VI.6 indique les facteurs de qualité (la bande passante étant calculée par le moment spectral

[5]Cet aspect est à rapprocher de la mauvaise résolution du retard de phase en basses fréquences : en effet, le retard de phase est défini comme la pente de la corde qui joint sur la courbe de la phase déroulée le point à la fréquence nulle au point à la fréquence concernée. Or, si l'on se rapproche des basses fréquences, ces deux points se rapprochent aussi, et la corde devient donc une approximation de plus en plus précise de la tangente, si bien que le retard de phase se rapproche alors de la définition du retard de groupe. On peut également expliquer ce problème de résolution simplement en rappelant que le retard de groupe est le retard d'enveloppe d'un signal à bande étroite, enveloppe dont le spectre est par définition à support dans les basses fréquences uniquement.

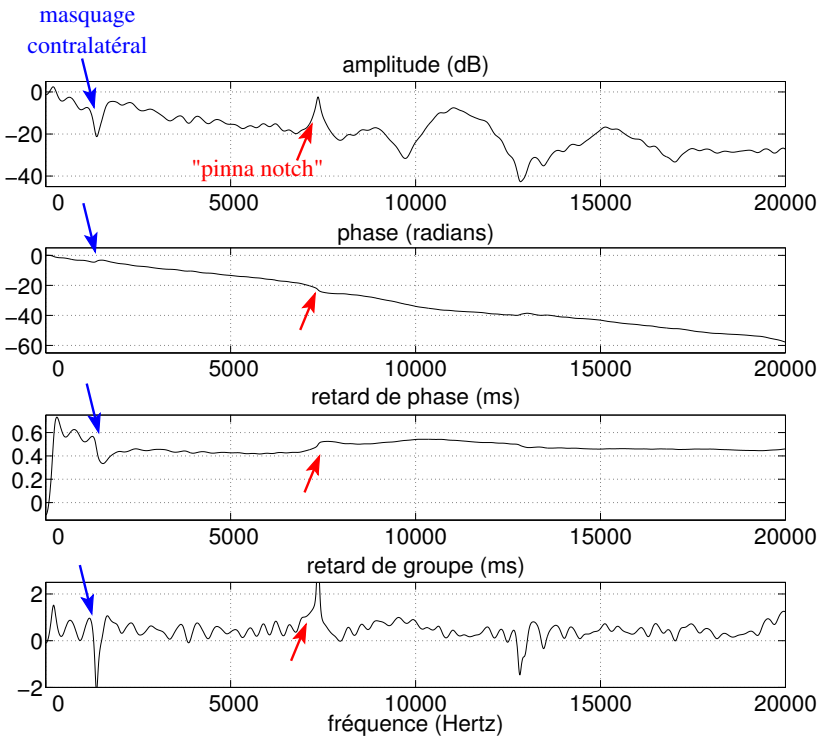


FIG. VI.5 – **Fonction de transfert interaurale** de la même tête qu'en figure VI.3, pour un azimut de -45° dans le plan horizontal. La fonction de transfert est calculée par le rapport de la fonction de transfert contralatérale sur la fonction de transfert ipsilatérale. Les flèches bleues indiquent le zéro dû à l'interférence destructive de l'effet d'ombre de la tête au niveau de l'oreille contralatérale ; les flèches rouges indiquent l'effet du *pinna notch*, qui est la principale influence du pavillon du côté ipsilatéral. On observe que le retard de phase est quasiment constant entre les deux, qu'il subit une augmentation d'un facteur 3/2 en basses fréquences, indépendamment des phénomènes d'interférences (le fait que le retard de phase tende vers zéro en très basses fréquences est un biais de mesure).

centré au deuxième ordre de la fenêtre, conformément à la définition du chapitre IV) pour les bancs de filtres étudiés.

La validité du modèle peut être étudiée plus attentivement en appliquant directement sur les réponses impulsionnelles (ce qui permet de s'affranchir de l'influence de la source) les méthodes à long-terme (c'est-à-dire avec une intégration sur tout l'axe temporel) d'égalisation et annulation proposées à la section 2 du chapitre IV. Les figures VI.7 et VI.8 présentent les résultats de l'estimation en fonction du banc de filtres employé[6] sur le couple de HRTF utilisé à la figure VI.5, c'est-à-dire celle correspondant à un azimut de -45° dans le plan horizontal, mesurée sur un tête réelle par Algazi et al. (2001b). Plusieurs aspects sont à dégager de l'observation de ces figures.

D'une part, aucun des bancs de filtres étudiés ici ne permet de rendre compte fidèlement des pôles et zéros de la fonction de transfert interaurale : ainsi, l'erreur est conséquente (au moins de l'ordre de 10 %) à 1,5 kHz et 7,5 kHz, et ce quel que soit le banc de filtres, alors que par exemple la transformée de Fourier à court-terme sur 256 points offre un bien meilleur facteur de qualité que les autres filtres à 7,5 kHz. Pour pouvoir tenir compte correctement de ces singularités, il faudrait des filtres bien plus étroits encore, ce qui pose le problème du coût de calcul (puisqu'il serait nécessaire dans ce cas de multiplier le nombre de filtres) et

[6]Les résultats pour le banc de filtres en tiers d'octave ne sont pas présentés ici, puisqu'ils sont assez similaires à ceux obtenus pour le banc de filtres gammatone, ce que l'on pouvait prévoir à partir de la figure VI.6, étant donné que les facteurs de qualité à une fréquence donnée sont assez proches.

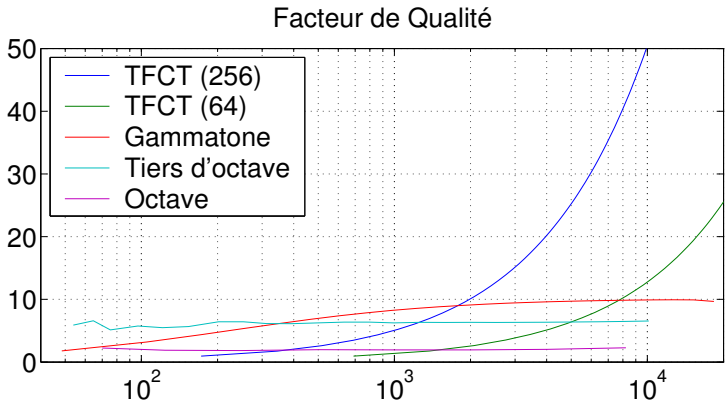


FIG. VI.6 – **facteurs de qualité** de quelques bancs de filtres : transformée de Fourier à court-terme sur 256 et 64 points, banc de 24 filtres gammatone sur une échelle ERB, banc de 24 filtres de Butterworth d'ordre 3 en tiers d'octave, et banc de 8 filtres de Butterworth d'ordre 3 en octave. La définition des facteurs de qualité employée est celle de la section 4 du chapitre IV.

de la résolution temporelle de détection.

D'autre part, comme on pouvait s'y attendre, l'erreur du modèle dépend du facteur de qualité du filtre dans la bande considérée : ainsi, en basses fréquences (c'est-à-dire en dessous du zéro à 1,5 kHz), elle est la plus faible pour le banc de filtres gammatone, qui a les bandes les plus étroites à ces fréquences. En hautes fréquences (au dessus du pôle à 7,5 kHz) c'est au contraire la transformée de Fourier à court-terme sur 256 points qui donne les meilleurs résultats, car les filtres gammatone y sont très larges et ne peuvent rendre compte de toutes les variations de module et de phase de la fonction de transfert interaurale.

On peut également remarquer qu'une détection d'enveloppe préalable (ici par transformation de Hilbert) donne toujours une erreur normalisée plus faible. Ce résultat peut paraître étonnant étant donné que le retard de groupe est estimé avec une moins bonne précision que le retard de phase, mais cela s'explique par le fait que la détection d'enveloppe ramène les signaux à bande étroite en bande de base, si bien que le modèle, comme cela est indiqué en section 4.5 du chapitre IV, est plus tolérant à une erreur d'estimation du retard.

Comme le prévoit la théorie, l'estimation du retard de phase se heurte à l'ambiguïté intrinsèque des signaux à bande étroite : ceci est particulièrement visible sur l'estimation après TFCT sur 256 points en hautes fréquences, car les filtres y sont particulièrement étroits : alors que l'erreur d'estimation de phase reste en valeur absolue inférieure à 0,1 radians dès 2 kHz excepté aux fréquences correspondant à des pôles ou zéros, l'estimation du retard de phase subit fréquemment des sauts d'un nombre entier de périodes, et ce même en dehors des fréquences de forte incertitude mentionnées ci-dessus.

Discussion

Si l'on cherche, conformément à une remarque précédente, à privilégier l'analyse dans un domaine de fréquences allant jusqu'à quelques kilohertz, l'alternative consiste soit à utiliser des bancs de filtres de type temps-échelle (comme des filtres en tiers d'octave) ou apparentés (comme les filtres gammatone basés sur l'échelle ERB ou Bark), soit des bancs uniformes avec un grand nombre de canaux. La seconde solution étant à la fois lourde en calcul et offrant une mauvaise résolution temporelle, on se dirige naturellement vers la première.

On a pu vérifier en pratique que ce type de banc de filtres convenait à une modélisation

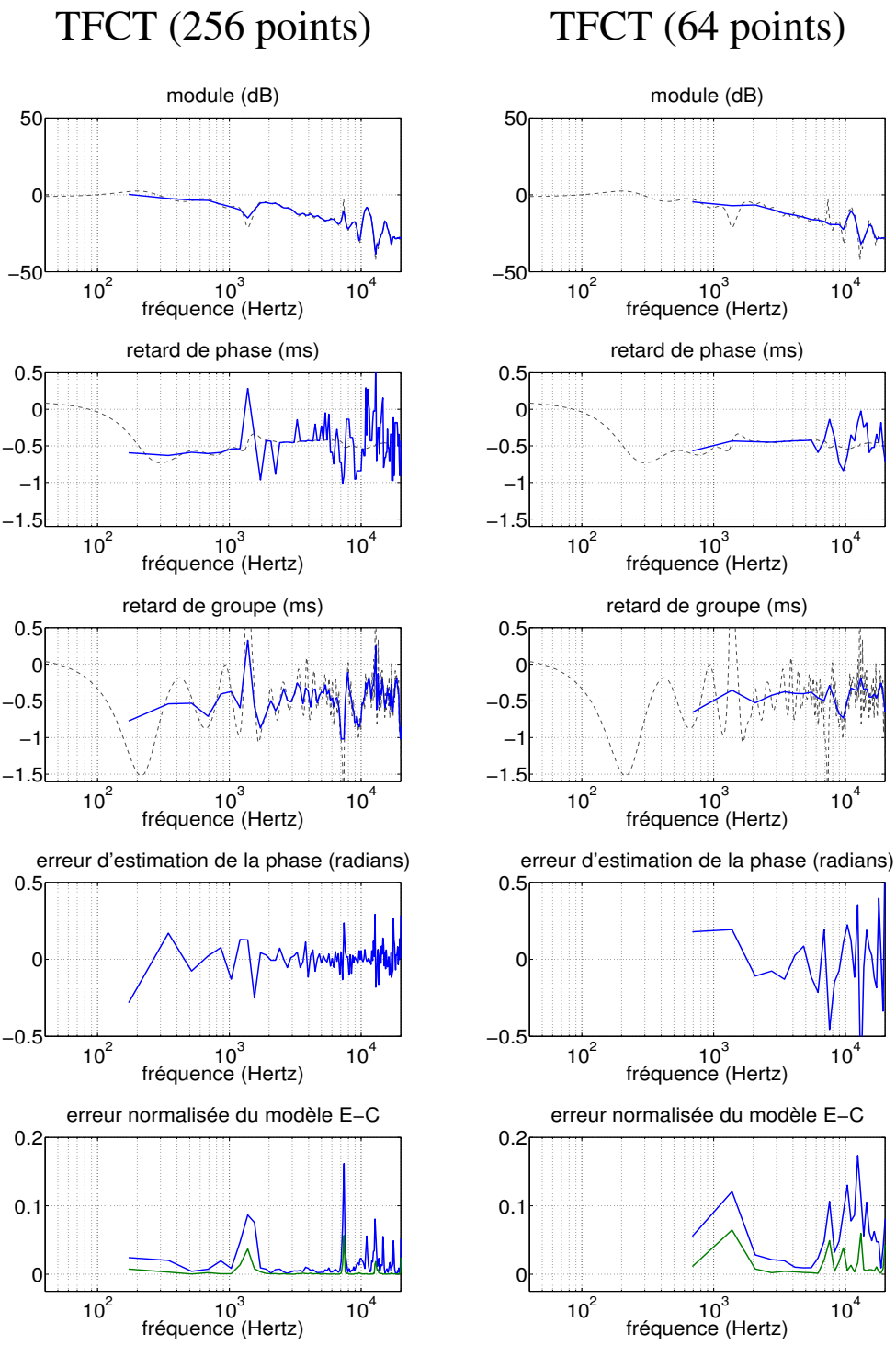


FIG. VI.7 – **estimation de fonction de transfert interaurale par égalisation et annulation** : l'estimation est effectuée directement sur les deux réponses impulsionnelles, à la sortie d'une transformée de Fourier à court terme sur 256 et 64 canaux. Sur les trois figures en haut de chaque colonne, le trait plein représente l'estimation du gain, du retard de phase et du retard de groupe dans chaque sous-bande, et le trait interrompu est la référence résultant d'une transformée de Fourier suréchantillonnée sur 4096 points. L'erreur d'estimation de phase (4e ligne) permet de distinguer les biais sur l'estimation de la phase des ambiguïtés du retard de phase à une période près. L'erreur du modèle E-C (5e ligne) est indiquée avec et sans détection d'enveloppe, ce dernier cas correspondant toujours à la courbe inférieure.

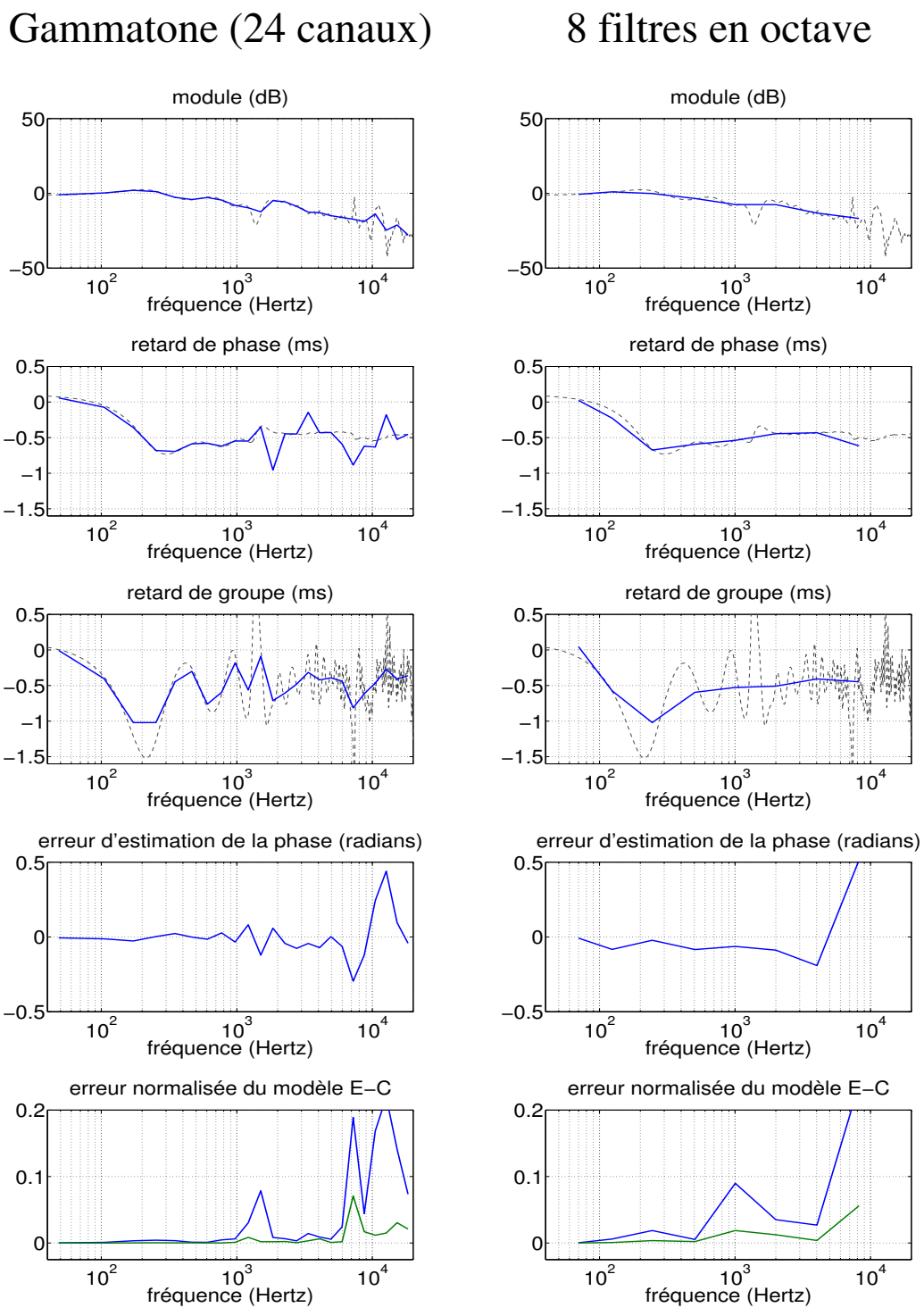


FIG. VI.8 – **estimation de fonction de transfert interaurale par égalisation et annulation (suite)** : l'estimation est effectuée directement sur les deux réponses impulsionnelles, à la sortie d'un banc de 24 filtres gammatone, et d'un banc de 8 filtres en octave (voir page 124 pour les commentaires additionnels).

par gains et retards de la fonction de transfert interaurale, si toutefois le facteur de qualité est suffisamment élevé (ce qui n'est pas le cas par exemple pour un filtrage en octaves). Le filtrage qui a été retenu pour la localisation est **le banc de filtres gammatone sur une échelle ERB**, qui présente un facteur de qualité légèrement supérieur à un filtrage en tiers d'octave dans la zone fréquentielle privilégiée, soit entre 100 Hz et 5 kHz : en deçà, la faible précision temporelle et l'absence de différences d'intensité rendent peu utile l'estimation, et au-delà de 5 kHz, l'influence du pavillon entraîne une multiplication du nombre de pôles.

2.4 A propos de la notion de cônes de confusion

Pour un modèle sphérique de tête, sans pavillon ni torse, et pour lequel les "oreilles" sont diamétralement opposées, les cônes de confusion sont définis comme les surfaces en champ lointain pour lesquelles les indices interauraux sont constants. Cette définition est sans ambiguïté dans ce cas, les cônes de confusion étant rigoureusement des demi-cônes d'axe interaural, et ayant pour sommet le centre de la sphère.

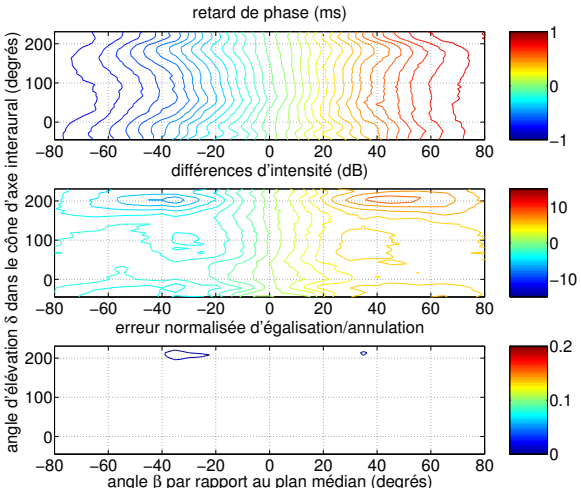
Dans un cas réel, la situation est plus complexe : en effet, la présence du torse, des pavillons, le fait que la tête ne soit pas à symétrie de révolution, et l'excentration de l'axe interaural ne permettent plus de définir de manière univoque des surfaces pour lesquels les différences de temps et d'intensité sont constantes à toutes les fréquences.

La figure VI.9 illustre cet aspect, grâce à la représentation pour quelques bandes de fréquences des courbes à retard de phase constants (que l'on appellera "courbes iso-ITD"), des courbes à différences d'intensité constantes (que l'on appellera "courbes iso-ILD"), et de l'erreur normalisée, sur un plan dont les axes sont l'angle β par rapport au plan médian et l'élévation δ dans le cône d'axe interaural[7]. Le jeu de mesures utilisé a été mesuré sur un sujet humain au centre Cipic de l'université de Californie-Davis par Algazi et al. (2001b). Si la symétrie de révolution était respectée, les courbes iso-ITD et iso-ILD seraient toutes des lignes verticales, quel que soit l'angle par rapport au plan médian.

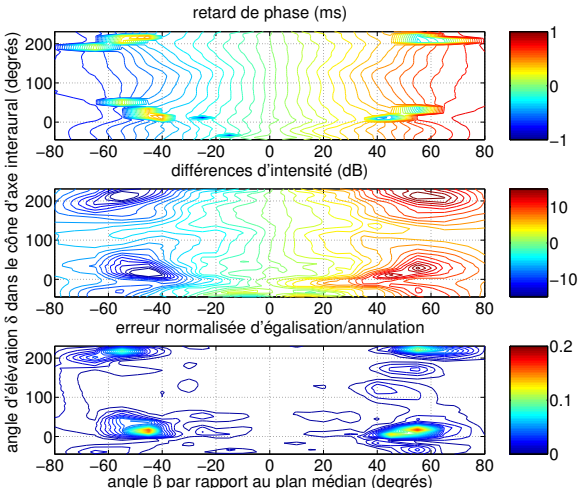
L'erreur normalisée permet de repérer les zones (correspondant à des pôles et zéros des fonctions de transfert interaurales dans les domaines fréquentiel et spatial) pour lesquels l'estimation du retard de phase et des différences d'intensité n'est pas fiable. Si l'on fait abstraction de ces zones, on peut remarquer que les courbes iso-ITD ne peuvent être considérées verticales qu'à proximité du plan médian ; au fur et à mesure que la source s'éloigne de ce dernier, les courbes iso-ITD se déforment, le retard de phase étant, pour un même cône de confusion, relativement moins élevé pour les faibles élévations et vers l'arrière (c'est-à-dire pour des angles d'élévation δ supérieurs à 90°). Ceci est dû au fait que les oreilles sont excentrées vers l'arrière et le bas par rapport au centre de la tête ; les courbes iso-ILD se déforment encore plus nettement lorsque l'on s'éloigne du plan médian, et n'ont rapidement plus rien à voir avec des lignes verticales ; bien qu'il soit toujours possible de dégager une tendance globale quant à la dépendance azimutale des différences d'intensité, l'allure des courbes iso-ILD est principalement conditionnée par la position des pôles et des zéros.

Un autre aspect nous concernant particulièrement est la dépendance fréquentielle de ces courbes : si toutes les courbes iso-ITD ont la même allure d'une bande de fréquences à l'autre (toujours en faisant abstraction des zones contenant pôles et zéros, dont le nombre augmente avec la fréquence), elles ne se superposent pas exactement. Si l'on suit une courbe iso-ITD à une fréquence, on peut observer des fluctuations de l'ITD dans une autre bande parfois supérieures à 100 ms. Ces variations ne sont pas négligeables, et l'on peut espérer qu'elles facilitent l'estimation de la direction au sein d'un cône de confusion. Les variations fréquentielles des courbes iso-ILD sont encore plus flagrantes, puisque les pôles et zéros ne se situent pas aux mêmes directions d'une fréquence à l'autre. Cela dit, le fait que les méthodes d'estimation par égalisation et annulation ne soient pas en mesure de modéliser correctement des pôles et zéros restreindra l'exploitation de ces variations, comme on le verra en section 3.7.

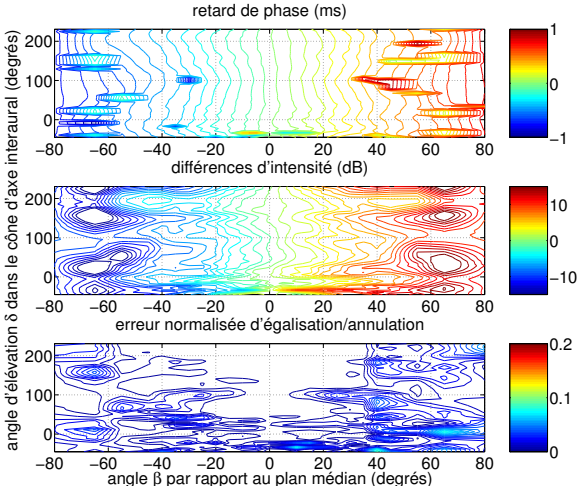
[7] Pour une définition des angles et systèmes de coordonnées utilisés, on pourra se reporter à l'annexe A.



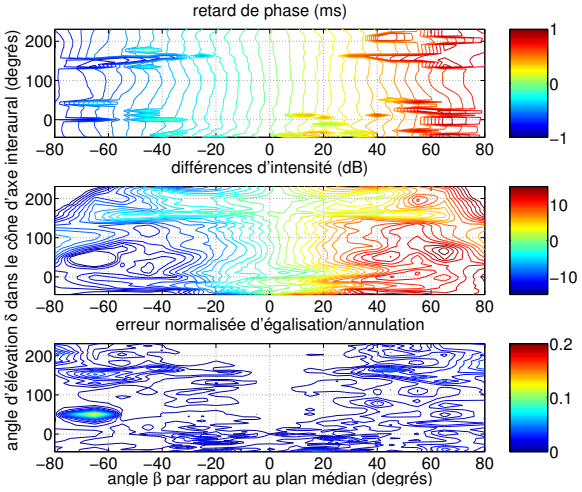
(a) Bande 18 (fréquence centrale 600 Hz)



(b) Bande 15 (fréquence centrale 1200 Hz)



(c) Bande 12 (fréquence centrale 2300 Hz)



(d) Bande 10 (fréquence centrale 3400 Hz)

FIG. VI.9 – **Courbes à retard de phase et différences d'intensité constants** : les indices interauraux sont estimés par méthode stationnaire d'égalisation et annulation sur un jeu de mesures de la base du centre Cipic/U.C.Davis, avec un banc de filtres gammatone à 24 canaux (puisqu'il s'agit de filtres initialement destinés à la modélisation de l'audition, la numérotation des bandes est effectuée dans l'ordre **décroissant** des fréquences, des hautes vers les basses). Les courbes iso-ITD sont affichées par pas de $50\ \mu s$; les courbes iso-ILD sont affichées par pas de 1 dB ; les courbes à erreur normalisée constante sont affichées par pas de 0,5 %.

3 ESTIMATION DE LA DIRECTION PAR DÉCISION SUR L'ERREUR D'ÉGALISATION ET ANNULATION

Après avoir rappelé les principales méthodes d'estimation de la direction dans le cas d'enregistrements binauraux, et précisé les particularités physiques de ce type de prise de son, nous sommes en mesure de proposer une méthode d'estimation de la direction visant à tirer le meilleur parti de l'information issue de la détection par égalisation et annulation.

3.1 Cadre et hypothèses de travail

La méthode d'estimation de la direction de provenance proposée ici est, comme on le verra, étroitement liée à la méthode de détection par égalisation et annulation développée au chapitre IV. Elle se concentre sur les instants pour lesquels l'estimation des retards et différences de niveau est la moins ambiguë, c'est-à-dire ceux pour lesquels la détection est optimale. D'un point de vue physique, ceci correspond aux moments pour lesquels seule l'onde directe a atteint les microphones à l'instant considéré. **L'analyse devra donc porter sur la plage de temps très courte qui sépare l'arrivée de l'onde directe de l'arrivée de la première réflexion**. Encore une fois, si la résolution temporelle est suffisante, cette plage de temps est assez bien repérable à partir de l'erreur d'égalisation et annulation, puisque l'arrivée de la première réflexion fait chuter le niveau de corrélation, et donc augmenter l'erreur E-C, qui est, on, le rappelle, l'indice de détection.

Sur cette plage de temps, les signaux observés $x_L[n]$ et $x_R[n]$ sont supposés pouvoir être modélisés par convolution d'un signal source par le couple de HRTF champ lointain correspondant à la direction de la source, avec un bruit additionnel participant à la fois du bruit de fond électronique et de la réverbération des évènements précédents. On suppose de plus que les réponses impulsionnelles peuvent être modélisées dans chaque bande de fréquences, avec ou sans détection d'enveloppe préalable, par un gain et un retard pur, et que l'on dispose de la connaissance des paramètres de ce modèle (soit le gain, le retard de phase et le retard d'enveloppe) ainsi que des erreurs de modélisation, le tout pour le plus grand nombre possible de directions de provenance. La méthode d'estimation de ces paramètres est celle proposée en section 2 du chapitre IV, c'est-à-dire l'estimation par égalisation et annulation à long-terme. Pour finir, il est supposé dans un premier temps que la source ayant émis l'onde sonore contenue dans les signaux $x_L[n]$ et $x_R[n]$ se trouve dans une direction qui coïncide avec l'une de ces directions connues.

Dans ces conditions, **et pour un niveau de bruit raisonnable**, les résultats de la section 1 de l'annexe B indiquent que les indices intercanaux (différences d'intensité, retards de phase et d'enveloppe) calculés par égalisation et annulation par bande de fréquences sont sensiblement égaux à ceux calculés sur les HRTF, moyennant les ambiguïtés et incertitudes propres à la méthode d'estimation employée.

Un des avantages de l'emploi conjoint de ce modèle de signaux et des méthodes par égalisation et annulation est qu'il est possible de juger de la pertinence de certaines de ces hypothèses : les erreurs d'égalisation et annulation par bande de fréquence calculées sur les HRTF indiquent si la modélisation par gains et retards est valide ; l'indice de détection, c'est-à-dire l'erreur calculée sur les signaux observés, nous renseignent sur le rapport signal sur bruit, et donc sur l'éventualité que l'estimation des différences d'intensité soit biaisée, par exemple par l'arrivée de la première réflexion.

Quant à l'hypothèse de décorrélation du bruit, celle-ci n'a de sens en espaces clos que si d'une part si l'on considère que l'information sonore ne provenant pas de la source à l'instant considérée soit intégralement due à un champ réverbéré qui puisse être considéré comme diffus, et d'autre part si l'on limite l'analyse à des fréquences supérieures à une fréquence de coupure de l'ordre de 300 Hz, en deçà de laquelle même un champ parfaitement diffus fournit des signaux corrélés pour cette distance entre les microphones[8].

[8]Cette limitation n'est valable qu'en présence de champ réverbéré. Si un temps suffisamment long sépare l'évènement considéré de l'évènement précédent pour que la réverbération soit négligeable, ou si l'espace considéré est anéchoïque, l'information en basses fréquences est utilisable.

3.2 Principe de l'estimation de la direction

Nous cherchons une méthode qui puisse rassembler les principaux avantages des deux approches mentionnées aux sections 1.2 et 1.3, c'est-à-dire :

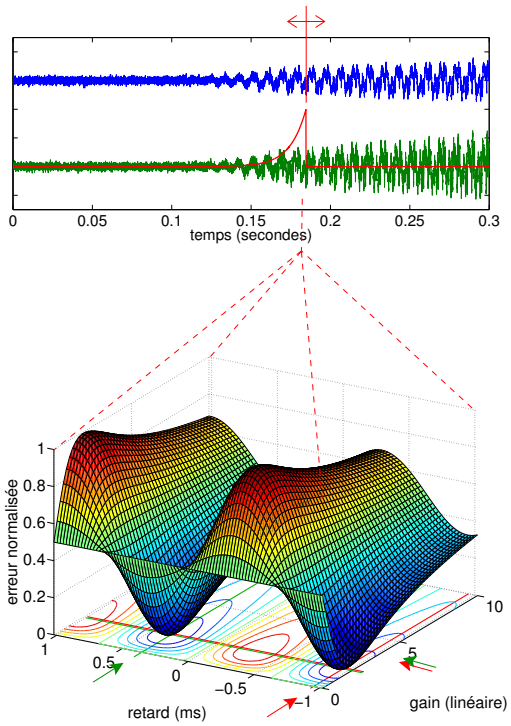
- qui tienne compte de la précision de la méthode d'estimation des indices interauraux et de ses ambiguïtés intrinsèques
- qui permette une estimation de la direction en azimut et en élévation, lorsque la précision est suffisante
- qui tienne compte de l'erreur de détection et d'estimation par bande, et pouvoir distinguer une mauvaise détection d'une incertitude sur la décision sur la direction de provenance

Des trois points mentionnés ci-dessus, le premier est peut-être à traiter avant tout, puisque le problème des ambiguïtés en basses fréquences et en bande étroite est commun aux deux méthodes. La faible résolution sur l'estimation du retard de phase en basses fréquences (et sur l'estimation du retard de groupe à toutes les fréquences) y limite la précision sur l'estimation de la direction ; il est nécessaire de tenir compte de cet aspect dans la décision en donnant un poids relativement faible à ces estimations. Le problème de l'ambiguïté de phase en bandes étroites est légèrement différent, puisque dans ce cas le retard modulo la période est souvent estimé avec une bonne précision, mais la présence d'éventuels sauts de période entraîne une ambiguïté sur l'estimation de la direction. De toutes les solutions qui ont été trouvées jusqu'ici pour lever cette ambiguïté, la plus efficace est sans doute celle consistant à effectuer une sommation sur toutes les bandes utiles du motif de corrélation. Néanmoins, il a été mentionné au chapitre IV que celle-ci nécessite dans le cas binaural quelques précautions, dûs à la variabilité du retard de phase. On se propose ici de contourner cette ambiguïté d'estimation du retard d'une autre manière, sans chercher explicitement à la lever.

Le principe réside dans le constat suivant : alors que l'écart entre le retard de phase estimé et le retard de phase nominal est important dès que l'estimation "se trompe" d'une ou de plusieurs périodes, **l'erreur d'égalisation et annulation entre ces deux minima locaux est quasiment identique** (c'est d'ailleurs la source de l'ambiguïté). Ainsi, si, comme c'est le cas de la méthode citée en section 1.3, l'indice d'erreur permettant de juger de la direction de provenance de la source dépend de l'écart absolu entre les indices interauraux estimés et les indices calculés sur les HRTF pour chaque direction, alors cette erreur est importante dès qu'un saut de période se produit, alors que l'erreur d'égalisation et annulation n'a presque pas varié.

La figure VI.10 illustre graphiquement cette remarque : le signal analysé est un segment de quelques dixièmes de secondes issu de synthèse binaurale. Le signal source est un enregistrement d'une note de clarinette basse dans le bas du registre (note A2, fondamental à 110 Hz), convolué par un couple de HRTF mesurées sur une tête KEMAR par Gardner et Martin (1994), et mélangé à une séquence de bruit blanc, avec un rapport signal sur bruit pleine bande moyen d'environ 20 dB pour les deux canaux. La représentation affichée est l'erreur normalisée d'égalisation et annulation calculée à 0,18 secondes du début du signal, avec une fenêtre exponentielle de constante de temps $\Delta T = 10\,ms$, et pour la bande 723Hz-822Hz (le banc de filtres est constitué de 24 filtres gammatone sur une échelle ERB). Cette bande contient quasiment exclusivement le 7e harmonique du signal, qui est donc à 770 Hz, ce qui correspond à une période de 1,3 ms. Alors que l'estimation du retard de phase se trompe justement de 1,3 ms (soit plus de la moitié de l'écart maximal admissible), l'erreur normalisée du modèle est de 0,43% pour le minimum, et de 0,48% pour les valeurs de retard de phase et de gain correspondant au couple de HRTF utilisé (l'erreur de modélisation pour la mesure du gain et du retard sur les HRTF dans cette bande est de 0,12 %).

Une remarque similaire peut-être formulée vis-à-vis des problèmes de résolution sur le retard en basses fréquences : sur la figure VI.11, le même signal est analysé au même instant, mais cette fois-ci pour la bande 35Hz - 62Hz : il s'agit de la toute première bande du banc, pour lequel le rapport signal sur bruit est mauvais (autour de 10 dB) puisque cette bande est en dessous du fondamental, et n'en contient qu'une petite partie, ainsi que le souffle. Ici, alors que l'estimation du retard de phase diffère du calcul direct sur le couple de HRTF



Bande 723Hz–822Hz

FIG. VI.10 – **Exemple d'ambiguïté de phase en bande étroite** : le signal est une note de clarinette basse (fondamental à 110Hz), spatialisé par synthèse binaurale au moyen d'un couple de HRTF du KEMAR à -30° d'azimut dans le plan horizontal, et mélangé à une séquence de bruit blanc (rapport signal sur bruit pleine bande moyen d'environ 20 dB pour les deux canaux). La représentation en dessous est l'erreur normalisée d'égalisation et annulation pour la bande 723Hz-822Hz calculée à 0,18 s, pour une fenêtre exponentielle avec une constante de temps de 10 ms, et sans détection d'enveloppe. Les flèches rouges indiquent le retard de phase et le gain estimés, et les flèches vertes indiquent le retard de phase et le gain mesurés sur le couple de HRTF .

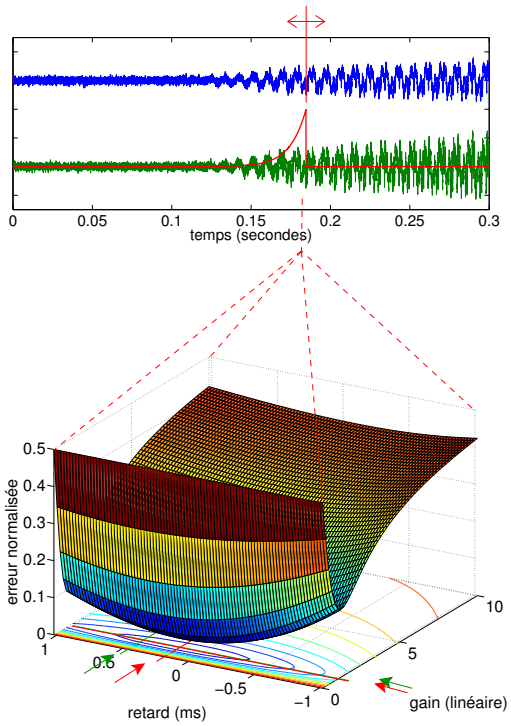
de 0,28 ms, l'erreur est de 2,5% pour le minimum, et de 3,3% pour les valeurs de retard de phase et de gain correspondant au couple de HRTF utilisé (l'erreur de modélisation pour la mesure du gain et du retard sur les HRTF dans cette bande est négligeable).

Ceci incite à utiliser comme mesure d'erreur au sein d'une bande de fréquence, en lieu et place du terme $(d_k(\theta,\varphi))^2$ défini par l'équation VI.2, une fonction directe de l'erreur normalisée d'égalisation et annulation [9], pour les valeurs des indices interauraux calculés à partir des HRTF pour la direction (θ,φ) dans la bande considérée. Puisque l'on désire distinguer l'erreur sur l'estimation de la direction de l'erreur de détection (c'est-à-dire le minimum de l'erreur normalisée), on choisit comme **indice de décision** par bande non pas l'erreur normalisée, mais sa variation par rapport à l'erreur normalisée minimale, qui permet de juger de la qualité du modèle gain-retard dans la bande k concernée et à l'instant n considéré, pour chacune des directions (θ,φ) répertoriées dans la base de données :

$$d_{EC,k}(\theta,\varphi)[n] = \varepsilon^k_{\alpha_k(\theta,\varphi),\tau_{ph,k}(\theta,\varphi)}[n] - \varepsilon^k_{min}[n] \tag{VI.3}$$

Dans cette expression, $\varepsilon^k_{\alpha,\tau}[n]$ est l'erreur normalisée d'égalisation et annulation dans la bande k, $\alpha_k(\theta,\varphi)$ et $\tau_{ph,k}(\theta,\varphi)$ sont respectivement le gain et le retard de phase calculés par

[9] Pour l'instant, il s'agit de l'erreur sans détection d'enveloppe préalable. Le problème de la prise en compte des retards d'enveloppe sera développé en section 3.4



Bande 35Hz–62Hz

FIG. VI.11 – **Exemple de mauvaise résolution sur l'estimation du retard en basses fréquences** : le signal est une note de clarinette basse (fondamental à 110Hz), spatialisé par synthèse binaurale au moyen d'un couple de HRTF du KEMAR à -30° d'azimut dans le plan horizontal, et mélangé à une séquence de bruit blanc (rapport signal sur bruit pleine bande moyen d'environ 20 dB pour les deux canaux). La représentation en dessous est l'erreur normalisée d'égalisation et annulation pour la bande 35Hz-62Hz calculée à 0,18 s, pour une fenêtre exponentielle avec une constante de temps de 10 ms, et sans détection d'enveloppe. Les flèches rouges indiquent le retard de phase et le gain estimés, et les flèches vertes indiquent le retard de phase et le gain mesurés sur le couple de HRTF

méthode E-C sur le couple de HRTF pour la direction (θ, φ), et $\varepsilon^k_{min}[n]$ est l'erreur normalisée minimale :

$$\varepsilon^k_{min}[n] = \min_{\alpha,\tau} \left\{\varepsilon^k_{\alpha,\tau}[n]\right\}$$

Tout comme la méthode rappelée en section VI.2, celle-ci peut être interprétée comme une décision au sens du **maximum de vraisemblance** : calculer l'indice $d_{EC,k}(\theta,\varphi)[n]$ pour une direction (θ, φ) donnée et à l'instant n revient à calculer l'erreur d'égalisation et annulation **dans l'hypothèse** où la source provient de cette direction. Dans cette hypothèse, cette erreur est liée à la puissance des bruits résiduels à droite et à gauche, qui représentent en fait les portions de signal non modélisables par égalisation et annulation avec les valeurs données de gains et retards pour la direction (θ, φ). La décision consiste donc à choisir la direction la plus probable au sens où elle **maximise le rapport signal sur bruit**.

Le principe de l'estimation de la direction ainsi proposé est résumé de manière graphique sur la figure VI.12. Pour saisir la différence avec les méthodes présentées en début de ce chapitre, il est utile de comparer ce schéma à ceux des figures VI.1 et VI.2. La différence principale avec les méthodes présentées en section 1.2 réside dans le fait que la représentation utilisée comme coeur de l'estimation, c'est-à-dire l'erreur normalisée d'égalisation et annulation, n'est pas une représentation temps-espace, mais temps-gain-retard. Par rapport aux méthodes présentées en section 1.3, on conserve l'idée d'une estimation de la direction

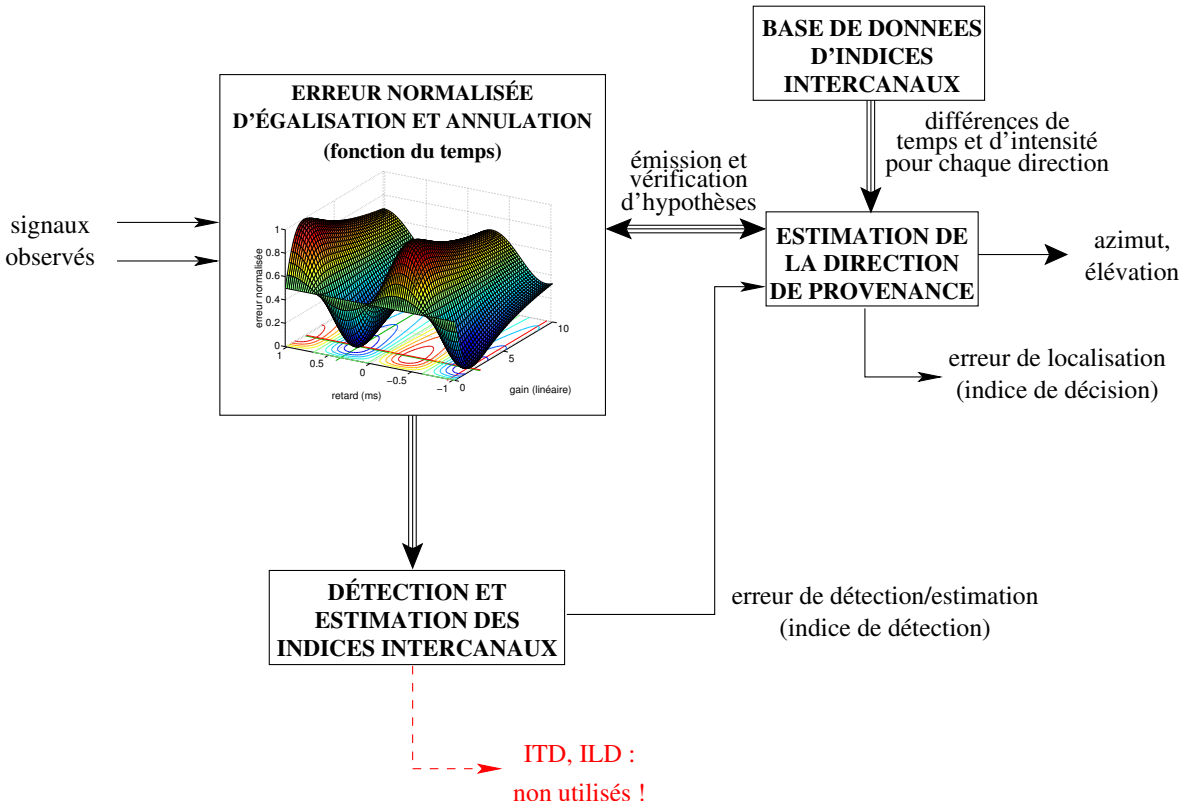


FIG. VI.12 – **Principe de l'estimation de la direction - unification avec la méthode de détection** : le coeur de l'analyse est constitué par le calcul de l'erreur normalisée d'égalisation et annulation, pour toute valeur du retard τ et du gain α. La détection se base sur la recherche du minimum de cette erreur, qui définit l'indice de détection. La direction de provenance n'est estimée que lorsque cet indice de détection est suffisament bas. Pour cet instant, il utilise la connaissance exhaustive de l'erreur normalisée pour tester chacune des directions connues. L'estimation du retard optimal (ITD) et du gain optimal (ILD) pour les signaux observés n'est pas utilisée pour l'estimation de la direction, et n'est effectuée que pour information.

au sens du maximum de vraisemblance, mais pour éviter les ambiguïtés d'estimation des différences de temps et d'intensité, la méthode repose directement sur la connaissance de l'erreur normalisée pour toute valeur du gain et du retard.

Détection et décision

En pratique, cet indice de décision n'est pas pas calculé à tous les instants n, d'une part car cela est inutile, et d'autre part une recherche systématique à tous les instants serait très coûteuse. En supposant que pendant toute la durée pour laquelle un évènement est détecté, la source physique est immobile, il est possible de ne calculer l'indice de décision qu'à un seul instant. Afin de fournir la meilleure estimation possible, on choisit comme instant de calcul celui qui **minimise l'erreur de détection**. Il est important de noter que cet instant (qu'on notera par la suite $n_0(k)$) dépend de la bande considérée, bien qu'il se situe (sauf en très basses fréquences, pour lesquelles la détection perd de sa précision) typiquement entre l'instant d'arrivée du son direct et celui de la première réflexion.

La figure VI.13 illustre ce principe sur un signal résultant de la convolution d'une séquence de bruit blanc par une réponse de salle élémentaire constituée d'un son direct à 30° et d'une réflexion à 60° (les HRTF sont issues des mesures sur une tête KEMAR par Gardner et Martin (1994), et les angles sont notés en convention horaire), atténuée de 10 dB

et retardée de 20ms. On peut observer notamment à la figure inférieure la valeur de l'indice de décision $d_{EC,k}(\theta,\varphi)[n_0(k)]$ calculé pour les 710 directions connues en azimut et en élévation. Celles-ci sont triées par rapport à l'angle qu'elles forment avec le plan médian, sachant que cet angle est positif lorsque la source est à droite, et négatif lorsque la source est à gauche. Les directions sont donc regroupées sur l'axe des abscisses par cône d'axe interaural, pour toutes les élévations disponibles, avec une tolérance $\pm 2.5^\circ$ sur l'angle par rapport au plan médian. Ainsi cette représentation sous forme de latéralisation est uniquement graphique et n'empêche pas de différencier deux directions formant le même angle avec le plan médian, mais d'élévations distinctes.

Il est manifeste que l'indice est minimal pour des angles de latéralisation avoisinant 30°[10], et ce quelle que soit la fréquence. De plus, on retrouve notamment le phénomène de perte de résolution sous forme spatiale, qui s'amplifie au fur et à mesure que la fréquence centrale de la bande diminue, ainsi que l'ambiguïté en bande étroite, qui se manifeste graphiquement par la présence de minima secondaires formant deux bandes de part et d'autre de la direction. Cela dit, celle-ci est minimisée par le fait que l'indice de décision tienne compte à la fois des retards et des différences de niveau (ces dernières n'étant pas soumises à cette incertitude). Un dernier aspect qui mérite d'être mentionné est la perte de pertinence de l'indice en hautes fréquences (ici pour les bandes 1 à 6, soit des fréquences supérieures à environ 7 kHz), qui n'est plus stable (si l'on utilise d'autres tirages de séquences de bruit blanc, on observe qu'il réagît de manière aléatoire). Ce dernier aspect est dû à la non validité du modèle gain-retard à ces fréquences (qui a été étudiée en section 2), qui empêche de pouvoir identifier les différences de temps et d'intensité estimées sur les signaux observés à celles calculées sur les HRTF. Cette dernière observation indique clairement qu'il sera nécessaire d'écarter pour toutes les directions connues les bandes de fréquences pour lesquelles le modèle gain-retard n'est pas valide.

La distinction entre qualité de la détection et qualité de l'estimation de la direction est permise par l'évaluation séparée de l'indice de détection $\varepsilon^k_{min}[n_0(k)]$ et de l'indice de décision sur la direction $d_{EC,k}(\theta,\varphi)[n_0(k)]$:

- Si l'indice de détection minimal $\varepsilon^k_{min}[n_0(k)]$ est faible, mais que l'indice de décision $d_{EC,k}(\theta,\varphi)[n_0(k)]$ reste important quel que soit la direction (θ,φ) envisagée, alors l'information dans cette bande de fréquences indique qu'une source est présente, mais qu'elle provient d'une direction non répertoriée dans la base de données (par exemple, une direction située entre deux direction connues)
- Si l'erreur minimale d'égalisation et annulation **et** l'indice de décision minimal sont faibles, alors l'information dans cette bande de fréquences indique qu'une source est présente et qu'elle provient vraisemblablement de la direction minimisant l'indice de décision
- Si l'erreur minimale d'égalisation et annulation est importante, l'information dans cette bande de fréquences ne permet pas de statuer de la présence d'une source à l'instant n considéré.

En pratique, ces différents cas seront traités par comparaison des erreurs avec deux seuils fixés à l'avance : le **seuil de détection** S_1, défini au chapitre V, est la valeur maximale de l'indice de détection $\varepsilon^k_{min}[n_0(k)]$ pour laquelle on considère qu'une source est présente dans la bande k ; le **seuil de pertinence de l'estimation de la direction** S_2 est la valeur maximale de l'indice de décision $d_{EC,k}(\theta,\varphi)[n_0(k)]$ pour laquelle on considère que la direction estimée correspond à la direction de la source. Ces deux seuils seront typiquement fixés à 10%.

3.3 Liens avec les méthodes usuelles d'estimation de la direction de provenance

La méthode d'estimation sur la direction de provenance proposée résulte, comme cela a été mentionné plus haut, de la volonté de lier les avantages des deux grandes classes de méthodes participant de modèles binauraux de localisation, présentées en sections 1.2 et 1.3. En fait, les points communs avec celles-ci vont au-delà de la simple analogie conceptuelle :

[10] Le problème de la sensibilité en élévation sera étudié en section 3.7

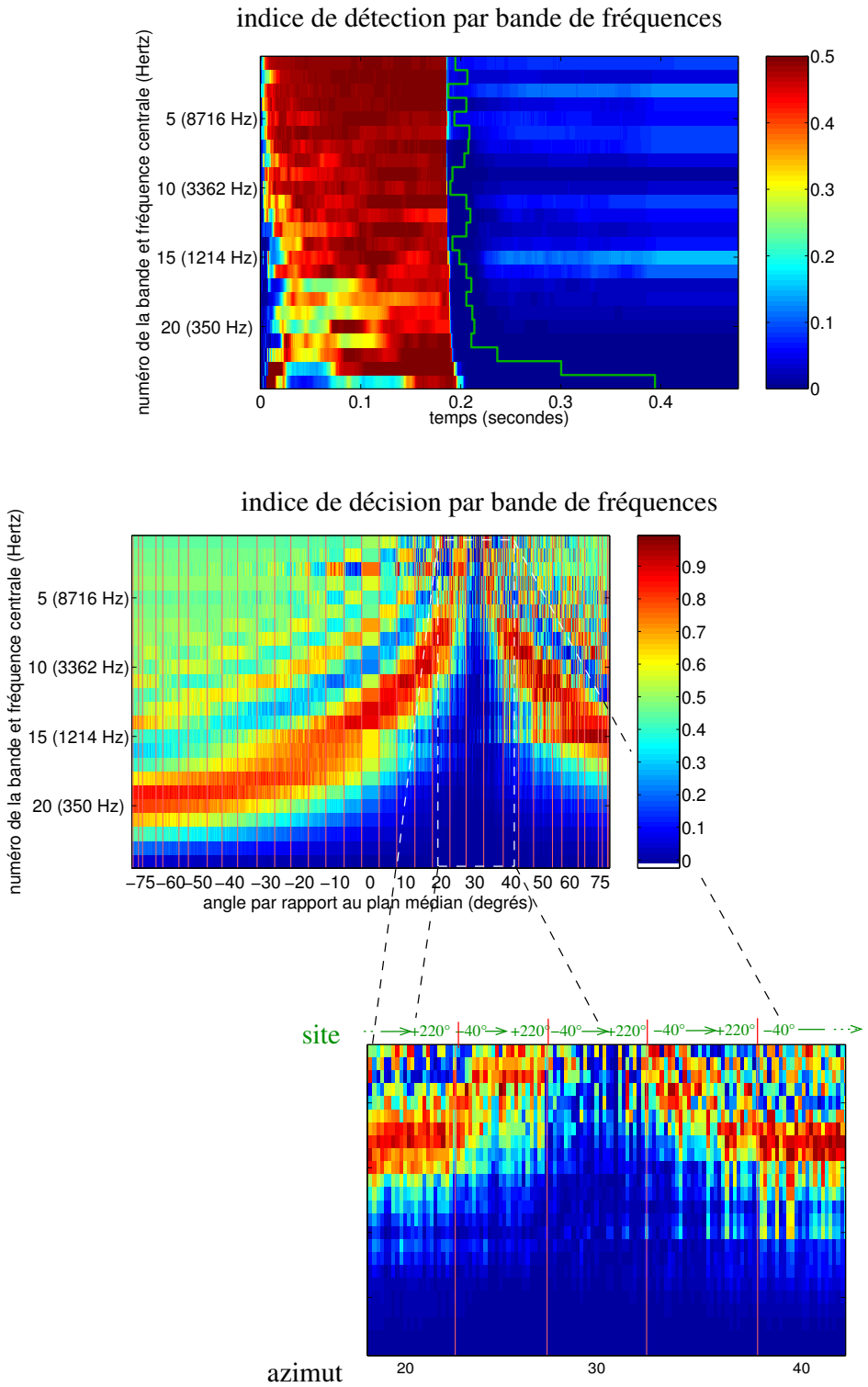


FIG. VI.13 – **Détection de source et décision sur la direction sans détection d'enveloppe** : le signal analysé est un bruit blanc débutant à 0,2s, convolué par une réponse impulsionnelle binaurale formée par un son direct à 30° d'azimut et une réflexion à 60° (tous deux sur le plan horizontal) atténuée de 10 dB et retardée de 20 ms. Les deux voies du signal sont mélangées à deux séquences de bruit blanc incohérentes (le rapport signal sur bruit moyen est de 32 dB à gauche et de 42 dB à droite). La figure supérieure indique l'indice de détection pour chacune des 24 bandes, et le trait vert l'instant $n_0(k)$ où cet indice atteint son minimum après la détection pour la bande k. En dessous on trouve l'indice de décision $d_{EC,k}\left(\theta,\varphi\right)\left[n_0\left(k\right)\right]$ calculé pour toutes les directions de la base de données (dans le plan horizontal ou non), triées par cône d'axe interaural (avec une tolérance de $\pm 2.5°$ sur l'angle par rapport au plan médian), puis par site. Finalement, est représenté un agrandissement de cette représentation autour du minimum.

D'une part, l'indice de décision $d_{EC,k}(\theta,\varphi)[n]$ est, pour une bande de fréquences donnée, une représentation temps-espace au même titre que celles rappelées en section 1.2. En particulier, les indices interauraux de temps et d'intensité ne sont plus des variables explicites, cédant la place aux variables de direction θ et φ. La différence principale avec les méthodes usuelles est que l'on ne se base pas sur une loi statistique de relation entre différences de temps et d'intensité, mais directement sur les indices individuels à chaque direction. Cette approche a le défaut de nécessiter une recherche systématique de similarité au sein de toutes les directions connues (ce qui pose entre autres un problème de temps de calcul), mais offre la possibilité d'une estimation en élévation basée sur des indices interauraux, alors que les estimations basées sur une représentation temps-espace, comme cela a déjà été mentionné, ne permettent généralement qu'une estimation de l'angle par rapport au plan médian.

D'autre part, le calcul de la section 2 de l'annexe B montre qu'au voisinage du minimum, l'erreur normalisée est une fonction quadratique de l'écart en gain et en retard par rapport à ce minimum : ainsi, l'application de l'équation B.3 au présent formalisme donne :

$$\begin{aligned} d_{EC,k}(\theta,\varphi)[n] \simeq \quad & \frac{1}{4}.\rho_{x_k,y_k}[n,\tau_{ph,k}(\theta,\varphi)].\left(\frac{ln(10)}{20}\right)^2.\left(\alpha_k^{dB}(\theta,\varphi)-\alpha_{min,k}^{dB}\right)^2 \\ & -\frac{1}{4}.\frac{\partial^2\rho_{x_k,y_k}}{\partial\tau^2}[n,\tau_{ph,k}(\theta,\varphi)].\left(\tau_{ph,k}(\theta,\varphi)-\tau_{min,k}\right)^2 \end{aligned} \tag{VI.4}$$

Dans cette expression, $\alpha_{min,k}^{dB}$ et $\tau_{min,k}$ sont les valeur du gain (en décibels) et du retard qui minimisent l'erreur normalisée dans la bande de fréquences considérée, et $\rho_{x_k,y_k}[n,\tau]$ est l'intercorrélation à court-terme des signaux observés filtrés par le filtre du canal fréquentiel considéré, dont on rappelle que la valeur pour le retard optimal $\tau_{ph,k}(\theta,\varphi)$ est liée à l'erreur $\varepsilon_{min}^k[n]$ par :

$$\varepsilon_{min}^k[n]=\frac{1}{2}.\{1-\rho_{x_k,y_k}[n,\tau_{ph,k}(\theta,\varphi)]\}$$

On met ainsi en évidence le fait que l'indice de décision $d_{EC,k}(\theta,\varphi)[n]$ équivaut à une distance pondérée dans le plan (α,τ), et donc que la méthode proposée ici est comparable, **au voisinage de la solution optimale**, à une mesure de distance sur le modèle de l'équation VI.2, pour laquelle les facteurs de pondération en gain et en retard de phase vaudraient :

$$\frac{1}{\sigma_{ILD_k}^2}=\frac{1}{4}\left(1-2.\varepsilon_{min}^k[n]\right).\left(\frac{ln(10)}{20}\right)^2$$

$$\frac{1}{\sigma_{IPD_k}^2}=\frac{1}{4}.\frac{\partial^2\rho_{x_k,y_k}}{(\partial\tau)^2}[n,0]$$

Une différence importante par rapport à la méthode de Martin (1995) est que le facteur de pondération en retard de phase dépend ici du signal, et donc entre autres de la bande utile du canal fréquentiel considéré : en particulier, il est intéressant de noter que plus les signaux observés sont basses fréquences, plus les variations de leur intercorrélation en fonction de τ seront lentes, et donc plus sa dérivée seconde sera faible. De fait, le facteur de pondération sur le retard de phase sera lui aussi faible, si bien que l'importance relative de celui-ci par rapport aux différences d'intensité sur la décision diminue... ce qui était l'un des aspects que l'on recherchait.

D'autre part, le facteur de pondération en gain est proportionnel à $\left(\frac{1}{2}-\varepsilon_{min}^k[n]\right)$, ce qui signifie qu'une importance maximale est accordée aux estimations pour lesquelles la détection est de bonne qualité $\left(\varepsilon_{min}^k[n]\to 0\right)$, alors que les estimations pour lesquelles la détection est mauvaise$\left(\varepsilon_{min}^k[n]\to\frac{1}{2}\right)$ ne sont pas ou peu prises en compte. La même remarque est valable pour le facteur de pondération en retard : la dérivée seconde de la corrélation au maximum est étroitement liée à l'amplitude de variation de cette fonction ; si les signaux sont décorrélés, la corrélation est par définition constante et égale à zéro pour tout retard, donc la dérivée seconde est nulle.

3.4 Prise en compte des retards d'enveloppe

L'indice de décision sur la direction défini par l'équation VI.3 peut être généralisé pour tenir compte des différences de temps et d'intensité entre les enveloppes des signaux :

$$D_{EC,k}(\theta,\varphi) = w_k.d_{EC,k}(\theta,\varphi) + w_k^{env}.d_{EC,k}^{env}(\theta,\varphi) \tag{VI.5}$$

avec

$$d_{EC,k}(\theta,\varphi) = \varepsilon^k_{\alpha_k(\theta,\varphi),\tau_{ph,k}(\theta,\varphi)}[n_0(k)] - \varepsilon^k_{min}[n_0(k)]$$

et

$$d_{EC,k}^{env}(\theta,\varphi) = \varepsilon^{k,env}_{\alpha_k(\theta,\varphi),\tau_{env,k}(\theta,\varphi)}[n_0^{env}(k)] - \varepsilon^{k,env}_{min}[n_0^{env}(k)]$$

Dans ces expressions, $\varepsilon^k_{\alpha,\tau}[n]$ désigne l'erreur normalisée d'égalisation et annulation dans la bande k sans détection d'enveloppe et $\varepsilon^k_{min}[n]$ son minimum, $\varepsilon^{k,env}_{\alpha,\tau}[n]$ l'erreur normalisée avec détection d'enveloppe et $\varepsilon^{k,env}_{min}[n]$ son minimum ; $\alpha_k(\theta,\varphi)$, $\tau_{ph,k}(\theta,\varphi)$ et $\tau_{env,k}(\theta,\varphi)$ sont respectivement le gain, le retard de phase et le retard d'enveloppe calculés par méthode E-C sur le couple de HRTF pour la direction ; $n_0(k)$ et $n_0^{env}(k)$ sont les instants pour lesquels l'erreur normalisée respectivement sans et avec détection d'enveloppe sont minimaux ; finalement, w_k et w_k^{env} sont des facteurs de pondération entre le calcul avec et sans détection d'enveloppe.

Il a déjà été mentionné que le comportement des méthodes par égalisation et annulation après détection d'enveloppe est à rapprocher celui qu'ont les mêmes méthodes sans détection d'enveloppe en basses fréquences, simplement car l'enveloppe d'un signal à bande étroite est ramenée en bande de base quelle que soit la fréquence de la porteuse. Ainsi, le problème de mauvaise résolution sur l'estimation du retard de phase en basses fréquences se retrouve à toutes les bandes pour le retard d'enveloppe, ce qui a été relevé en section 2, et l'erreur normalisée d'égalisation et annulation reste donc relativement faible quel que soit le retard. D'autre part, les différences d'intensité sont quasiment inchangées par détection d'enveloppe, dès lors que la bande passante est suffisamment étroite. La résultante de ces deux remarques est que la résolution spatiale est mauvaise et ne permet pas de décider avec précision de la direction de la source.

On trouvera une illustration de cet aspect à la figure VI.14 : le signal étudié est identique à celui de la figure VI.13, la seule différence étant qu'une détection d'enveloppe par transformation de Hilbert a été appliquée avant l'analyse des signaux issus du banc de filtres. On peut y constater d'une part que la dynamique de l'indice de détection (figure du haut) est bien plus réduite qu'en l'absence de détection d'enveloppe, et d'autre part que la résolution spatiale de la décision est bien plus mauvaise (figure du bas), bien que l'on s'affranchisse dans ce cas des phénomènes des franges secondaires dues aux ambiguïtés de phase.

Ceci indique de manière claire que les retards d'enveloppe ne permettent pas à eux seuls une estimation précise de la direction de provenance. Leur prise en compte a néanmoins un sens sur des signaux pour lesquels le rapport signal sur bruit est défavorable dans la majorité des bandes et les retards de phase y sont donc difficiles à estimer, puisqu'ils suppléent ce manque d'informations par l'apport d'indications (certes imprécises) sur la direction de la source. En pratique, pour une bande de fréquences k donnée, le poids w_k^{env}des indices d'enveloppe devra être de toute manière nécessairement plus faible que le poids w_k.

3.5 Intégration fréquentielle

L'indice $(d_{EC,k}(\theta,\varphi))^2[n]$ proposé ci-dessus ne permet de juger de la direction probable de la source que vis-à-vis d'une unique bande de fréquences. Il s'agit maintenant d'étendre ce principe, et de proposer une méthode d'estimation de la direction qui intègre l'information provenant de toutes les bandes utiles, c'est-à-dire celles pour lesquelles le rapport signal sur bruit est suffisamment favorable. Sur le même principe que celui exposé en section VI.2, on forme **l'indice de décision global**, simplement par sommation des erreurs quadratiques

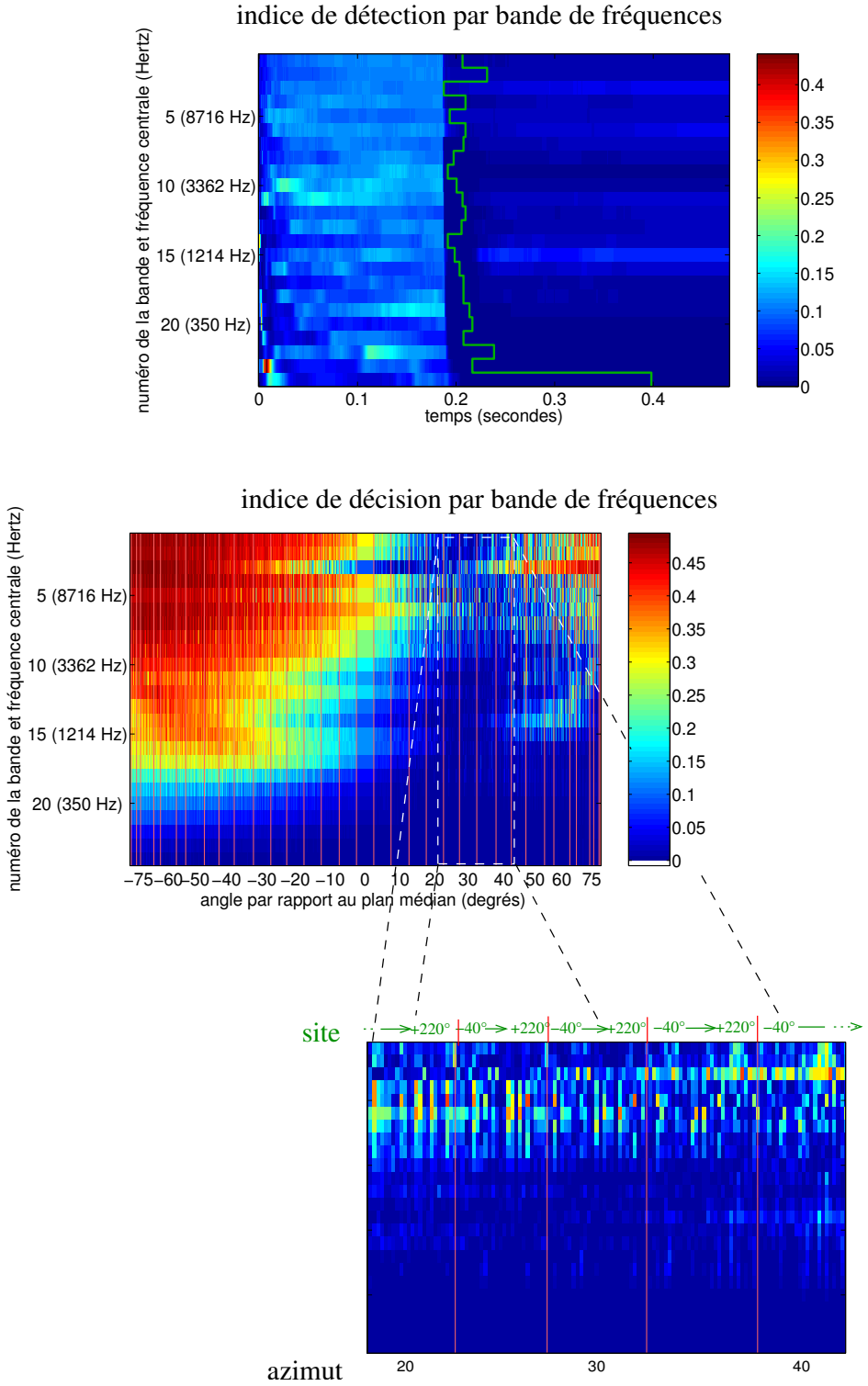


FIG. VI.14 – **Détection de source et décision sur la direction avec détection d'enveloppe** : se reporter à la figure VI.13 pour une description des signaux et figures.

élémentaires $D_{EC,k}(\theta,\varphi)$:

$$D_{EC}(\theta,\varphi) = \frac{\sum_{k=1}^{N}\left[w_k.d_{EC,k}(\theta,\varphi) + w_k^{env}.d_{EC,k}^{env}(\theta,\varphi)\right]}{\sum_{k=1}^{N}(w_k + w_k^{env})} \quad \text{(VI.6)}$$

La division par la somme des poids permet de normaliser l'indice de décision, qui sera ainsi compris entre 0 et 1 quelle que soit la valeur des poids w_k et w_k^{env} et le nombre de bandes.

Le schéma de principe global de la méthode de détection de source et d'estimation de sa direction est donné à la figure VI.15 : il fait apparaître deux seuils nommés S_1 et S_2, qui sont respectivement le seuil de détection, et le seuil de pertinence de l'estimation de la direction, tous deux définis en section 3.2.

Détermination des poids

Contrairement aux poids relatifs entre différences de temps et d'intensité au sein d'une même bande, qui sont fixés par la méthode (section 3.3), les poids d'une part entre les estimations avec et sans détection d'enveloppe, et d'autre part entre les bandes de fréquences, ne sont pas prédéterminés par le modèle. Ceci est dû au fait que le modèle gain-retard considère intrinsèquement que les paramètres sont indépendants de la fréquence. Le cas d'un filtrage complexe a pu être géré par l'introduction au chapitre IV d'un filtrage préalable en bandes étroites, mais les méthodes d'égalisation et annulation sont appliquées indépendamment sur chaque canal, puisqu'aucune relation n'est imposée entre les différents canaux. Il est donc nécessaire de déterminer une méthode permettant de choisir les poids. Ce choix est crucial vis-à-vis de la qualité de l'estimation, car il s'agit de donner le maximum d'importance aux informations les plus significatives. Ceci implique de minimiser l'influence :

- des bandes basses fréquences : d'une part, celles-ci n'apportent pas ou peu d'informations, puisque la notion même de retard de phase perd de sa pertinence, et que les différences d'intensité y sont quasiment nulles quelle que soit la direction de la source, si bien que l'erreur d'égalisation et annulation reste faible quelles que soient les valeurs du retard et du gain.
- des informations issues d'une analyse avec détection d'enveloppe, par rapport à celles issues d'une analyse sans détection d'enveloppe : en effet, les résultats de la section 3.4 nous indiquent que la résolution d'estimation des retards d'enveloppe est faible à toutes les fréquences
- des bandes pour lesquelles l'erreur de détection est relativement élevée

La solution qui a été choisie ici consiste à pondérer les indices élémentaires $d_{EC,k}(\theta,\varphi)$ et $d_{EC,k}^{env}(\theta,\varphi)$ par l'**amplitude de leur variation** sur l'ensemble des directions connues, c'est-à-dire :

$$w_k = \max_{\theta,\varphi}\{d_{EC,k}(\theta,\varphi)\} - \min_{\theta,\varphi}\{d_{EC,k}(\theta,\varphi)\}$$

$$w_k^{env} = \max_{\theta,\varphi}\left\{d_{EC,k}^{env}(\theta,\varphi)\right\} - \min_{\theta,\varphi}\left\{d_{EC,k}^{env}(\theta,\varphi)\right\}$$

En effet, cela permet d'une part de donner une faible importance aux bandes basses fréquences (puisque l'erreur y varie peu), aux indices d'enveloppe (puisque les variations de l'erreur d'égalisation et annulation y sont globalement deux fois plus faibles que sans détection d'enveloppe), et aux bandes pour lesquelles l'erreur de détection est élevée, puisque l'indice $\min_{\theta,\varphi}\left\{d_{EC,k}^{env}(\theta,\varphi)\right\}$ est de toute manière supérieur à l'erreur de détection[11].

La figure VI.16 donne un exemple d'intégration fréquentielle des indices de décision : on choisit cette fois-ci comme signal-source une séquence stationnaire de bruit blanc d'une demi-seconde, sans bruit additionnel. Celle-ci est convoluée par un couple de HRTF mesurées sur une tête artificielle KEMAR, pour un azimut de 60° dans le plan horizontal. On

[11] Cela dit, les bandes pour lesquelles l'erreur d'égalisation est vraiment élevée ont de toute façon déjà été écartées lors de la détection.

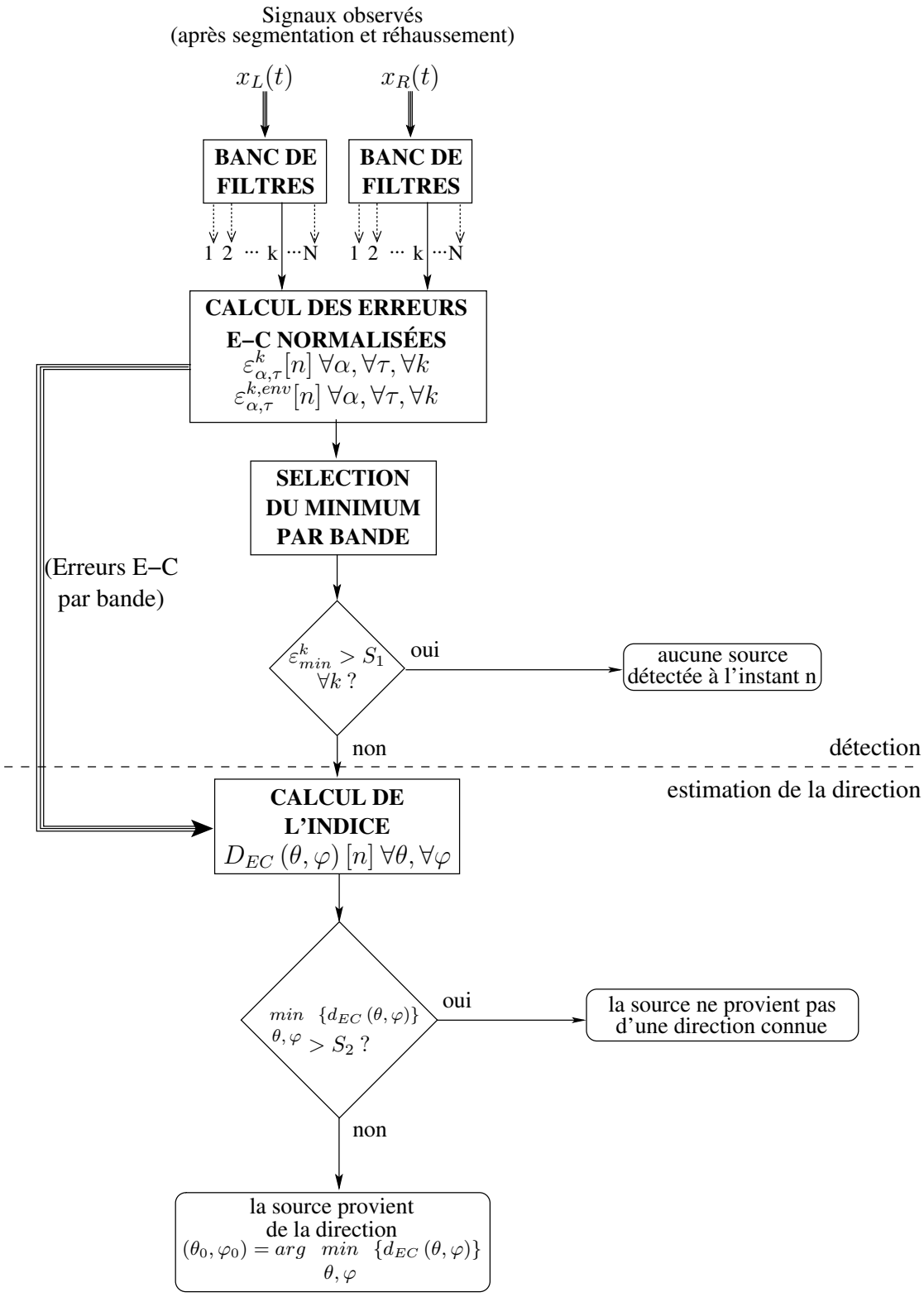


FIG. VI.15 – **Schéma de principe global de la méthode de détection et estimation de la direction par égalisation et annulation**

peut observer notamment que la pondération minimise comme on le souhaitait l'influence des canaux basses fréquences (c'est-à-dire de 20 à 24 pour le banc de filtres considéré[12]), et donne une plus grande importance aux indices de phase (c'est-à-dire aux indices calculés sans détection d'enveloppe) qu'aux indices d'enveloppe. La minimisation de l'indice de décision global obtenu permet de retrouver sans aucune ambiguïté la direction exacte de la source. On peut néanmoins ajouter que le fait que l'estimation de la direction est si précise est dû aux conditions expérimentales, qui sont particulièrement favorables : le signal est à large bande, stationnaire, sans bruit additionnel ni réflexions secondaires. Ce dernier aspect autorise l'estimation des puissances à être effectuée sans ambiguïté sur une large durée (la fenêtre est une fenêtre exponentielle avec une constante de temps de 100 millisecondes), contrairement à la situation d'un espace clos, pour laquelle (comme on a pu le voir sur l'exemple de la section 3.2) la durée utile d'estimation dépasse rarement la dizaine de millisecondes. Néanmoins, on verra sur un enregistrement réel (section 3.6) que l'estimation est toujours possible dans des situations bien plus délicates, comme le montrera l'exemple en section 4.

Rejet des bandes non valides

Tous les calculs de la section 1 de l'annexe B, sur lesquels on peut justifier l'identification faite entre les paramètres d'égalisation et annulation estimés sur les signaux observés et ceux calculés à partir des HRTF, supposent que la modélisation de ces dernières sous forme de gains et de retards par bande de fréquences soit valide. Si ce n'est pas le cas, il devient beaucoup plus difficile de prévoir le comportement des méthodes d'égalisation et annulation sur les signaux observés. Cette remarque amène à ne pas considérer, pour chaque direction éventuelle de la source, les bandes de fréquences pour lesquelles l'erreur d'égalisation et annulation sur les HRTF est importante, ce qui revient à annuler le poids w_k ou w_k^{env} correspondant pour la direction considérée uniquement.

Les bandes concernées sont celles qui contiennent un pôle ou un zéro de la fonction de transfert interaurale, et concernent donc principalement les hautes fréquences (au dessus de 5 kHz) pour le filtrage pavillonnaire, et la zone entre 1 kHz et 2 kHz pour les interférences dues aux réflexions sur le torse et au trajets multiples vers l'oreille contralatérale. La figure VI.17 illustre ce principe en indiquant les bandes rejetées pour un jeu de HRTF issu de la base du centre Cipic (Algazi et al., 2001b), avec un seuil sur l'erreur normalisée de 5%. On peut remarquer que la quasi-totalité des informations dans les deux bandes supérieures est jugée non valide. De plus, quel que soit l'angle par rapport au plan médian, les bandes à 5 kHz et plus sont rejetées pour des élévations faibles.

Comparaison avec les méthodes de latéralisation par corrélation

La formulation de l'équation VI.6 s'apparente à la méthode d'estimation du retard intercanal par sommation de l'autocorrélation sur tous les canaux fréquentiels, mentionnée notamment à la section 4.4 du chapitre IV. Cependant, la variable n'est ici pas le retard, mais la direction de provenance, qui est une variable bidimensionnelle. La différence est double : d'une part, on a ici la possibilité d'estimer la direction en latéralisation et en élévation, alors que la détermination "du" retard interaural ne permet au maximum que de juger de la latéralisation de la source ; d'autre part et surtout, on s'affranchit du problème dû aux fluctuations du retard interaural en fonction de la fréquence. Pour s'en persuader, on peut étudier la figure VI.18, et comparer l'intercorrélation par bande de la figure à l'indice de direction par égalisation et annulation, pour un même banc de filtres (Gammatone à 24 canaux sur une échelle ERB). Le retard qui maximise la fonction de corrélation dépend étroitement de la bande de fréquences, est soumis à de nombreux sauts de périodes (bandes 3 à 8 et bande 13) et est mal estimé pour la bande 2, qui contient un pôle. Le minimum de l'indice de direction $d_{EC,k}(\theta, \varphi)$ correspond quasiment à toutes les fréquences à la direction nominale, sans fluctuation d'une bande à l'autre, ni estimations mauvaises dûs à des sauts

[12] On rappelle que pour le banc de filtres gammatone, la numérotation des bandes s'effectue dans le sens décroissant des fréquences.

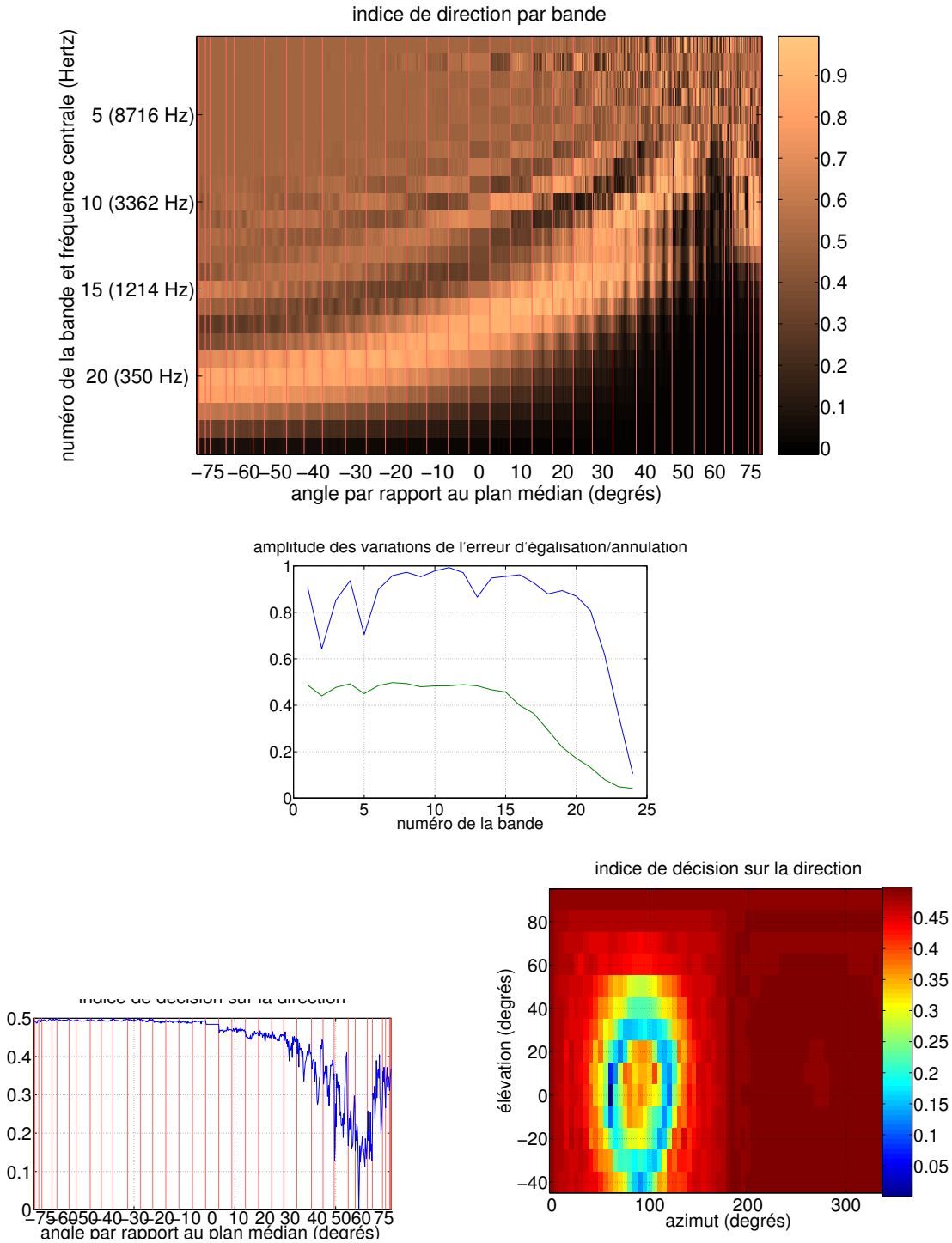


FIG. VI.16 – **Exemple de calcul de l'indice de décision par intégration fréquentielle** : le signal analysé est une séquence de bruit blanc convoluée à un couple de HRTF mesurées sur une tête artificielle KEMAR, pour un azimut de 60° dans le plan horizontal. Sur la figure supérieure est représentée l'indice de décision par bande sans détection d'enveloppe $d_{EC,k}(\theta,\varphi)$, les directions étant triées par cône d'axe interaural (avec une tolérance de $\pm 2.5°$ sur l'angle par rapport au plan médian), puis par site Sur la figure intermédiaire sont représentés les poids sur les indices de phase (en bleu) et d'enveloppe (en vert) calculés grâce à l'ambitus des variations de l'indice de décision. Sur les figures inférieures est représenté l'indice de décision global $D_{EC}(\theta,\varphi)$, d'une part sous forme de fonction monodimensionnelle de la direction triée par cône d'axe interaural, et d'autre part sous forme de fonction bidimensionnelle de l'azimut et de l'élévation.

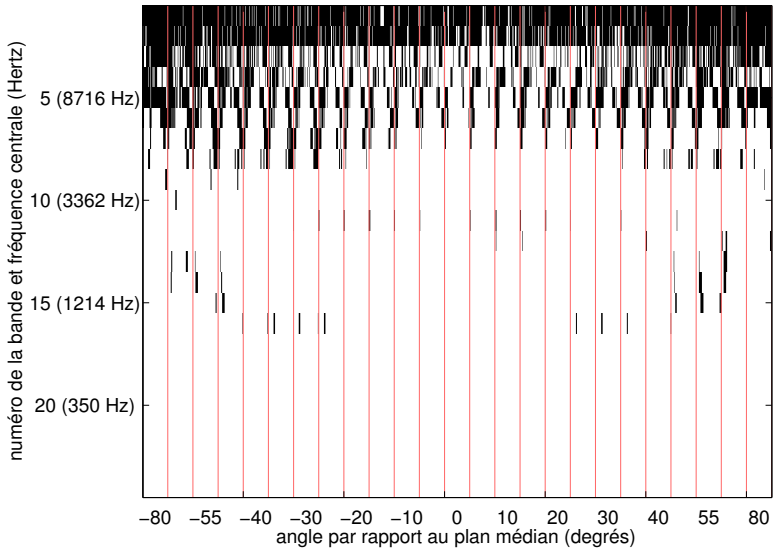


FIG. VI.17 – **Rejet des bandes de fréquences non valides** : pour chacune des directions (triées par cône d'axe interaural -avec une tolérance de $\pm 2.5°$ sur l'angle par rapport au plan médian-, puis par élévation), les zones noires indiquent que l'erreur normalisée d'égalisation et annulation sans détection d'enveloppe est supérieure à 5%. Les mesures de HRTF sont issues de la base de données constituée au centre Cipic par Algazi et al. (2001b).

de périodes. La direction est néanmoins mal estimée d'une part dans les bandes 2 et 3 (mais ceci est prévisible car l'erreur de détection y est importante, et donc évitable en baissant le seuil de détection), et d'autre part en basses fréquences (qui sont minimisées lors de la décision finale puisque l'amplitude des variations y est faible).

3.6 Intégration séquentielle

A ce stade, on est en mesure de détecter un évènement sonore, et d'estimer la direction de provenance à partir de cet unique évènement. La qualité de cette estimation dépend entre autres de la bande passante du signal, du rapport signal sur bruit, et de la durée utile d'estimation des paramètres d'égalisation et annulation. Si une telle estimation est satisfaisante pour des signaux à large bande comme les séquences de bruit utilisées jusqu'ici, elle perd de sa précision dans le cas de signaux à bande étroite et/ou harmoniques, simplement parce que le nombre de bandes de fréquences utiles dans le banc de filtres devient beaucoup plus réduit. Dans le cas où l'évènement sonore se superpose au champ réverbéré créé par l'évènement sonore précédent, le problème est encore un peu plus délicat, puisque d'une part le rapport signal sur bruit chute, et d'autre part le bruit (qui est principalement constitué par ce champ réverbéré), contrairement au modèle utilisé, n'est pas complètement décorrélé d'une voie à l'autre, en particulier dans les basses fréquences (voir chapitre I).

Il s'agit maintenant de proposer une extension de cette méthode qui puisse intégrer les informations provenant de plusieurs évènements sonores successifs, de manière à préciser l'information au cours du temps. Ceci suppose que chaque évènement sonore provienne de la même source physique, et que celle-ci ne se déplace pas.

Comme indiqué ci-dessus, une difficulté importante vis-à-vis de l'étude de signaux réels est que ceux-ci ne fournissent pas des indices interauraux non ambigus dans toutes les bandes de fréquences : concrètement, ceci se traduit par une absence de détection dans les bandes jugées non pertinentes, et donc par un indice de décision $d_{EC,k}(\theta, \varphi)$ non défini dans ces bandes. De fait, la qualité de l'estimation de la direction est limitée par ce manque d'informations.

L'hypothèse à la base de cette intégration séquentielle est l'**immobilité de la source**. De

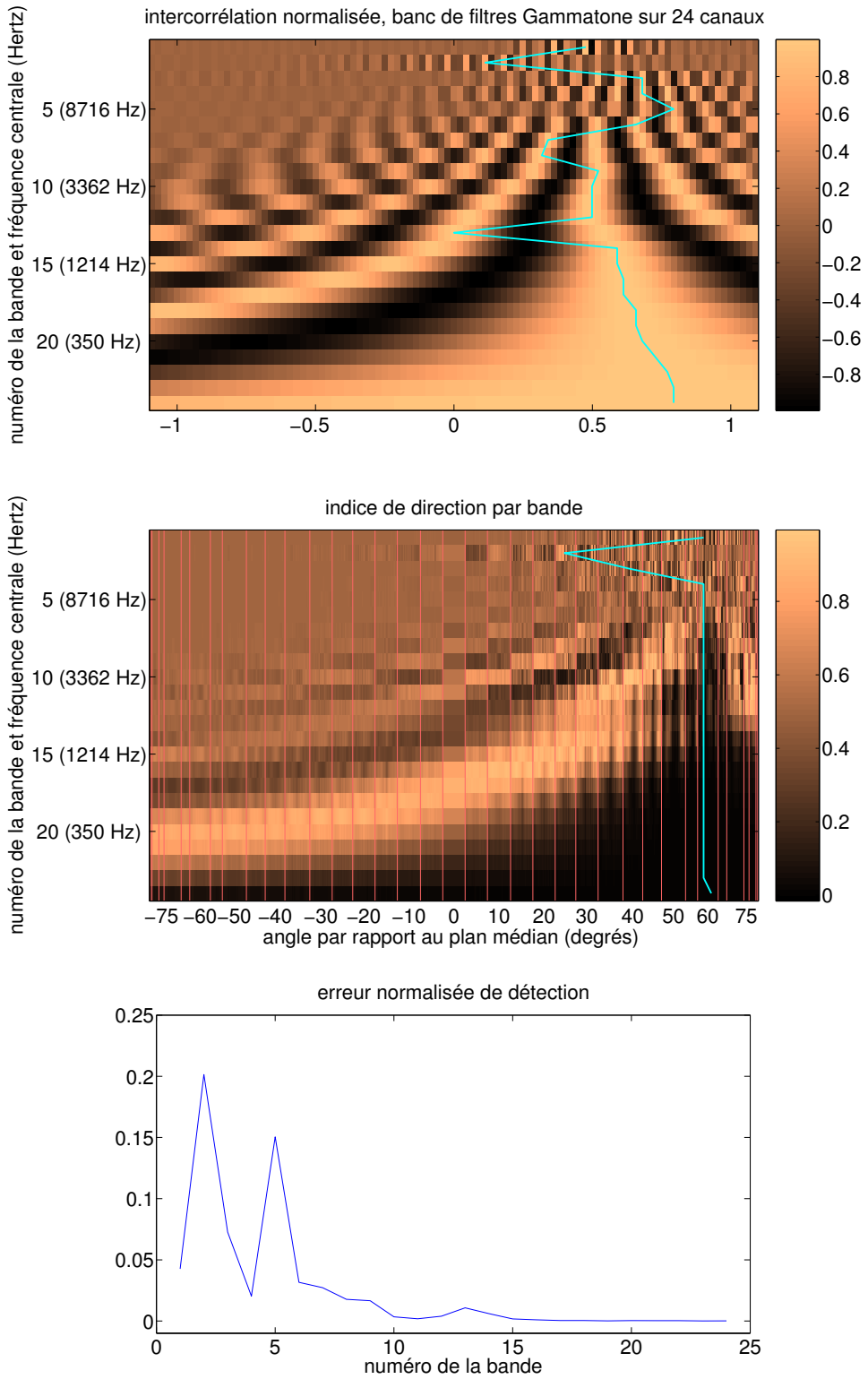


FIG. VI.18 – **Comparaison avec les méthodes de latéralisation par corrélation** : la figure supérieure représente l'intercorrélation stationnaire par bande du signal bicanal utilisé dans l'exemple de la figure VI.16 (sans détection d'enveloppe). La ligne continue suit le maximum par bande de cette intercorrélation, qui fournit l'estimation du retard . La figure intermédiaire représente l'indice de décision $d_{EC,k}(\theta, \varphi)$par bande. La ligne continue suit le minimum par bande, qui fournit l'estimation de la direction de la source en azimut et en élévation. La figure inférieure est l'erreur de détection par égalisation et annulation dans chaque bande.

fait, les indices de décision pour chaque événement sont plus ou moins valables en fonction des bandes, mais correspondent toujours à la même direction optimale. L'idée est alors d'utiliser les informations pertinentes parmi les estimations successives pour construire un indice de décision moyen par bande le plus complet possible, c'est-à-dire défini dans le plus de bandes possible, et avec une erreur minimale. Ceci est effectué, une nouvelle fois, par une simple opération de moyenne sur tous les motifs de décision individuels à disposition :

On suppose que l'on dispose de P évènements sonores successifs, pour lesquels la détection optimale et l'estimation ont été effectués individuellement. On note $d^{p}_{EC,k}(\theta,\varphi)$la valeur de l'indice de décision pour la direction (θ,φ)et la k-ième bande, relative au p-ième évènement sonore ($p \in \{1...P\}$). L'**indice de décision moyen par bande**, noté $d^{moyen}_{EC,k}(\theta,\varphi)$, est défini par :

$$d^{moyen}_{EC,k}(\theta,\varphi) = \frac{1}{N_k}.\sum_{p\in V_k} d^{p}_{EC,k}(\theta,\varphi) \tag{VI.7}$$

...V_k désignant l'ensemble des évènements sonores pour lesquels une information cohérente a été détectée dans la k-ième bande, et N_k désignant le nombre d'éléments de V_k. Une fois cet indice calculé pour chaque bande, on procède à l'intégration fréquentielle définie en section 3.5 pour décider de la direction la plus probable de la source.

3.7 Résolution au sein des cônes de confusion

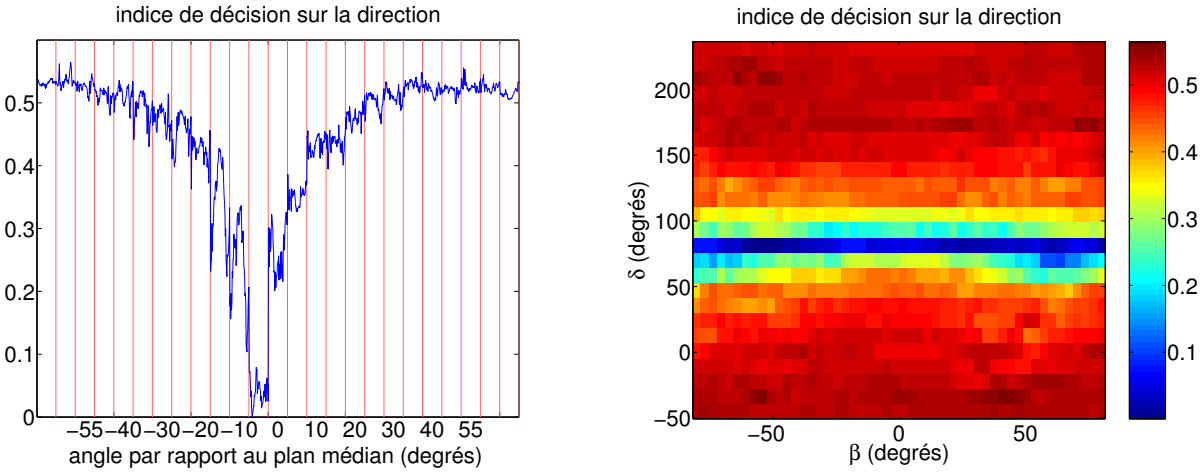
L'un des points du cahier des charges relatif à la mise en place de cette méthode était de pouvoir conserver la possibilité d'estimer l'angle d'élévation de la source, et pas seulement son angle par rapport au plan médian. Ceci présuppose que les variations des différences interaurales au sein d'un même "cône de confusion"[13], relevées en section 2.4, soient suffisamment importantes.

Distances inter-HRTF

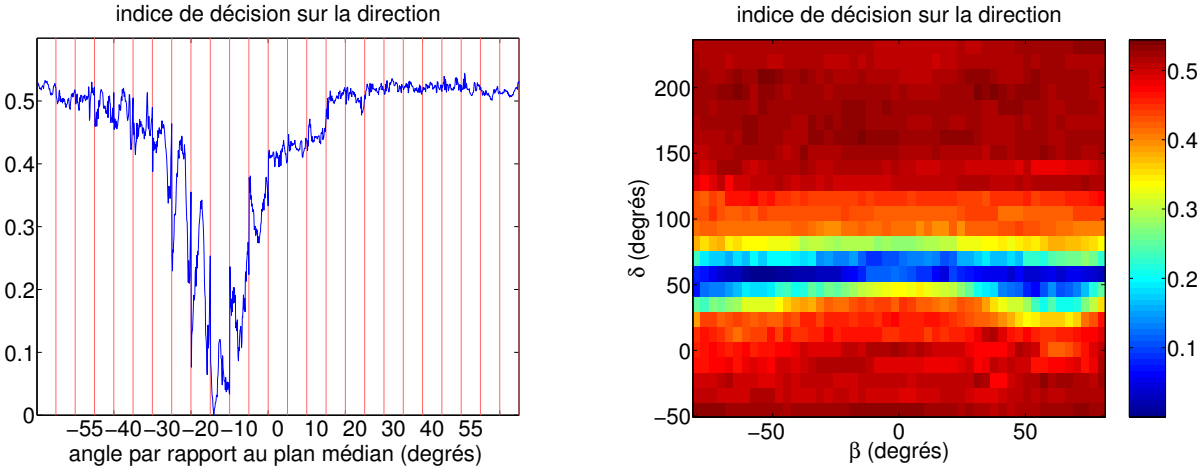
Il est possible d'étudier cet aspect en s'intéressant aux distances (au sens de l'indice de décision sur la direction) entre les HRTF : en effet, puisque les réponses impulsionnelles sont des signaux idéaux vis-à-vis de l'estimation de la direction, on fait ainsi abstraction de l'influence du signal source (notamment de sa bande passante) et du rapport signal sur bruit. En figure VI.19 sont ainsi représentées les valeurs des indices de décision sur la direction $D_{EC}(\theta,\varphi)$pour toutes les directions connues de la base de données, les signaux analysés étant des couples de HRTF pour trois directions dans le plan horizontal (d'azimuts respectifs -5°, -15° et -65°). Les indices sont affichés à chaque fois de deux manières : d'une part, en tant que fonction monodimensionnelle, les directions étant triées par cônes d'axe interaural ; d'autre part, en tant que fonction bidimensionnelle de l'angle β par rapport au plan médian et de l'angle δ par rapport au plan horizontal, ces angles étant considérés dans un repère sphérique dont l'axe polaire est l'axe interaural. Ce système de coordonnées est utilisé ici, d'une part car il permet de représenter les cônes de confusion de manière plus adéquate, et d'autre part car il s'agit du système de coordonnées naturel du jeu de mesures utilisé, qui est issu la base Cipic/U.C.Davis.

On observe aisément sur la représentation de droite les zones de confusion possibles : alors que la discrimination selon l'angle β est tout-à-fait correcte quelle que soit la direction de provenance de la source (l'indice variant de 0 à un peu plus de 0,5), il existe une zone couvrant toutes les valeurs de δ, correspondant plus ou moins à un cône d'axe interaural, pour laquelle l'estimation est intrinsèquement ambiguë, l'indice de décision y restant généralement inférieur à 0,05. Cette ambiguïté est d'autant plus forte que la source est proche du plan médian, où les différences interaurales sont quasiment nulles à toutes les fréquences. De plus, lorsque la source est proche du plan médian, la zone d'ambiguïté correspond effectivement à un cône d'axe interaural, mais plus la source s'éloigne du plan médian, plus cette zone se déforme, s'incurvant vers le plan médian pour des élévations proches de 90°.

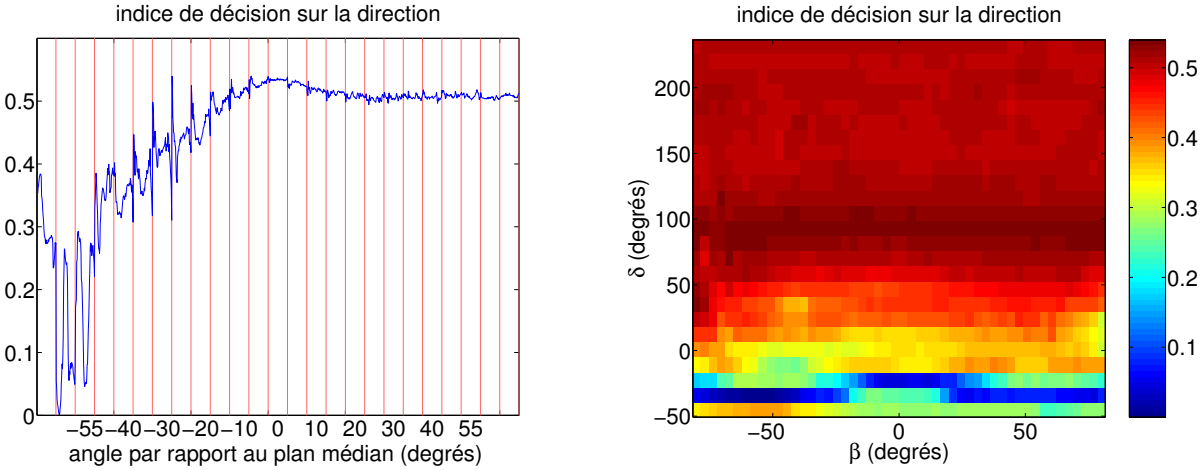
[13] La notion même de cône de confusion est peu précise en bande large, puisque la zone d'ambiguïté sur les indices interauraux dépend de la fréquence et du type d'indice (de temps ou d'intensité) considéré.



(a) Source à -5° d'azimut



(b) Source à -15° d'azimut



(c) Source à -65° d'azimut

FIG. VI.19 – **Indices de décision inter-HRTF** : l'indice de décision est calculé pour toutes les positions de la base de données, le signal observé étant un jeu de HRTF pour une direction donnée dans le plan horizontal. Les indices sont affichés triés par cône d'axe interaural (gauche), et dans le plan (β, δ) (droite). Les HRTF sont issues de la base Cipic/U.C.Davis.

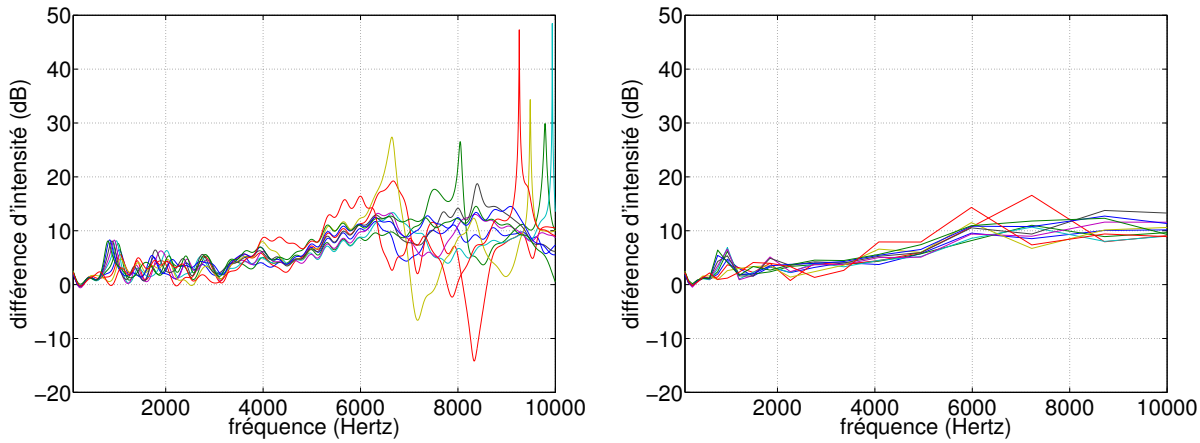


FIG. VI.20 – **Variation fréquentielle du spectre d'amplitude au sein d'un cône de confusion :** sur la figure de gauche sont affichés les modules des fonctions de transfert interaurales pour les dix directions les plus proches de la direction $(\theta = -15°, \varphi = 0°)$ pour le même jeu de HRTF qu'en figure VI.19. Sur la figure de droite sont indiquées les différences d'intensité calculées pour les mêmes directions par égalisation et annulation à la suite d'un banc de filtres gammatone (échelle ERB) à 24 canaux.

On retrouve ici la déformation des cônes de confusion due à l'excentration des oreilles, et mentionnée en section 2.4.

Cette faible résolution de l'indice de décision au sein du cône de confusion ne pose pas de problème tant que le signal n'est pas bruité et que l'onde sonore provient d'une direction connue, car dans ce cas l'estimation de la direction est toujours parfaite (c'est-à-dire que l'indice de décision minimal est nul). En revanche, si l'onde sonore provient d'une direction non répertoriée, ou si le rapport signal sur bruit est mauvais, l'estimation de la direction perd de sa précision, et notamment, la faculté de discrimination au sein du cône de confusion est dégradée.

La résolution en élévation fonction de la résolution fréquentielle

La faculté d'estimation de la direction en élévation est directement liée à la présence des pavillons et du torse, qui brisent la symétrie de révolution de la tête. Comme cela a été indiqué en section 2, leurs effets principaux sur les fonction de transfert monaurales (et donc interaurales) est l'introduction de pôles et de zéros dont les fréquences centrales dépendent étroitement de la direction de provenance de l'onde sonore. Or il est indiqué au chapitre IV que la qualité de la modélisation de pôles et zéros de la fonction de transfert intercanale par égalisation et annulation dépend étroitement de la résolution fréquentielle du banc de filtres employé en amont. On peut ainsi prévoir que la faculté de discrimination en élévation est donc fonction de la résolution fréquentielle.

La figure VI.20 permet d'illustrer ce problème sur un cas précis : après avoir déterminé quelles sont les dix directions qui, dans l'exemple précédent (figure VI.19), minimisent l'indice de décision $D_{EC}(\theta, \varphi)$pour un source située à un azimut de -15° dans le plan horizontal, on affiche pour chacune le module de la fonction de transfert interaurale, que l'on compare aux différences d'intensité calculées par égalisation et annulation dans un banc de filtres gammatone à 24 canaux. On peut observer d'une part que les petites fluctuations de quelques décibels jusqu'à 6 kHz sont pas ou peu modélisées, et surtout, que les pôles et zéros en hautes fréquences sont purement et simplement gommés par l'analyse, qui d'ailleurs dans ces bandes de fréquences ne sera pas jugée fiable, car l'erreur du modèle y est trop importante.

Ce phénomène, qui peut être également constaté lorsque la fonction de transfert interaurale contient des résonances ou antirésonances en basses fréquences, résulte directement du

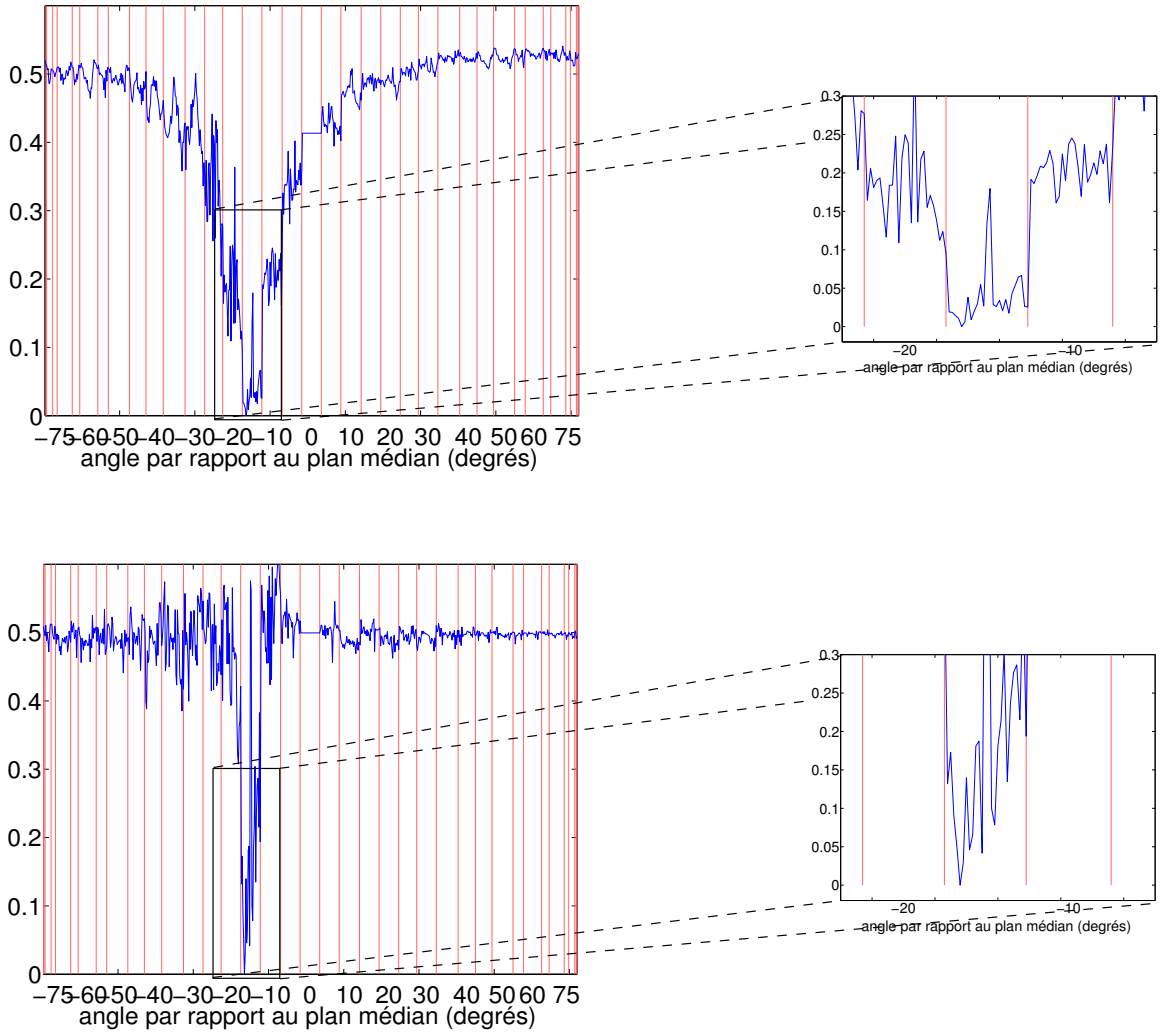


FIG. VI.21 – **Dépendance de la résolution en élévation à la résolution fréquentielle** : les deux figures représentent l'indice de décision en fonction de la direction, pour une source impulsionnelle à incidence dans le plan horizontal, pour un azimut de -15°. Les mesures de HRTF sont issues de la base établie par Gardner et Martin (1994) sur une tête KEMAR. Les bancs de filtres utilisés sont un banc de filtres gammatone à 24 canaux (haut), et un banc de filtres STFT à 1024 canaux (bas).

choix du banc de filtres (voir section 2.3), qui résulte d'un nécessaire compromis entre précision et coût de calcul, avec la contrainte d'un bonne précision dans les fréquences moyennes. Pour s'en assurer, on peut s'intéresser à un cas précis, illustré à la figure VI.21, montrant l'influence du banc de filtres sur la précision de l'estimation de la direction : la source est une impulsion, convoluée à un couple de HRTF à -15° d'azimut et pour une élévation nulle, pour une tête artificielle KEMAR. Deux bancs sont comparés : le banc de filtres gammatone à 24 canaux, utilisé jusqu'ici, et un banc de filtres STFT à 1024 canaux, qui offre une résolution fréquentielle comparable aux filtres gammatone dans les basses fréquences, mais bien sûr bien meilleure en moyennes et hautes fréquences. La comparaison est équivoque : non seulement la résolution sur l'angle par rapport au plan médian est bien meilleure, mais on gagne également beaucoup en termes de discrimination dans le cône de confusion.

4 EXEMPLE

L'exemple qui suit présente de manière synthétique l'analyse d'un enregistrement *in situ*, utilisant l'ensemble des méthodes développées au sein de ce chapitre et de la seconde partie de ce document.

4.1 Contexte

L'enregistrement analysé est un enregistrement binaural du premier mouvement de la fantaisie n° 2 de Telemann en la mineur pour flûte seule, effectué dans l'Espace de Projection de l'IRCAM. La flûte a été choisie pour plusieurs raisons : d'une part, il s'agit d'un instrument monodique, et à fréquence fondamentale assez stable, ce qui permet d'appliquer les techniques de rehaussement présentées en section 2.2 du chapitre V ; d'autre part, même si les sources sonores sont multiples, les dimensions de l'instrument sont suffisamment faibles pour que celui-ci puisse être considéré comme ponctuel à quelques mètres ; pour finir, bien que la flûte soit un instrument au spectre relativement pauvre, les attaques sont suffisamment marquées pour que l'estimation de la direction soit envisageable. L'enregistrement à une durée totale de 2 minutes, mais seules les 25 premières secondes sont analysées.

L'Espace de Projection est une salle rectangulaire dont on peut faire varier d'une part le volume, en agissant sur la hauteur du plafond, et d'autre part sur les propriétés acoustiques des parois, qui sont tapissées de prismes rotatifs dont les trois faces sont des surfaces offrant des propriétés acoustiques distinctes (absorbantes, diffusantes ou réfléchissantes). Dans cette situation, l'Espace de Projection était configuré avec un plafond en position basse, à environ 3,5 mètres de hauteur. Le temps de réverbération vaut 2 secondes à 1kHz. Toutes les parois (plafond compris) excepté le mur sud sont configurées pour être au maximum diffuses. Le mur sud est au contraire choisi réfléchissant sur toute sa surface, si bien que les deux réflexions principales viennent du mur sud et du sol (qui est très réfléchissant).

Les positions respectives et dimensions sont indiquées sur la figure VI.22. La flûtiste se trouve au sud-est de l'Espace de Projection, le sujet effectuant l'enregistrement en est éloigné de 3,46 mètres dans la direction nord-ouest/ouest, est orienté vers l'est, de telle manière à ce que l'azimut relatif de la flûtiste soit de -30° ; de plus, la flûte se trouve à une hauteur inférieure d'une dizaine de centimètres à celle des oreilles du sujet (qui se trouvent à 1,80 m du sol), longueur très faible par rapport à la distance, si bien que le site peut être considérée nul. De ces considérations géométriques, on peut déduire que la réflexion du sol arrivera à l'auditeur 3.9 ms après le son direct, avec une incidence de 30° en azimut et de -43° en site, et que la réflexion sur le mur sud interviendra 23 ms après le son direct, avec une incidence de -74° en azimut dans le plan horizontal.

Les microphones utilisés sont ceux employés pour les mesures de HRTF en chambre anéchoïque : il s'agit d'une paire de Knowles FG3329, qui sont des microphones de mesure omnidirectionnels à électret. Ceux-ci ont été insérés dans des supports plastiques moulés à la forme des conduits auditifs, de manière à ce que ceux-ci soient bouchés. Les enregistrements ont été effectués sur support informatique direct-to-disk, avec une fréquence d'échantillonnage de 44,1 kHz et une quantification sur 16 bits. Étant donné qu'il s'agit de microphones de mesure, le bruit de fond électronique est assez élevé : selon les données techniques du constructeur, le bruit équivalent après pondération A est de 30 dB SPL.

La base de données utilisée pour le calcul des HRTF est issue de la campagne relative au projet européen Listen (Vandernoot et Rio, 2003). Les HRTF ont été mesurées pour une cinquantaine de sujets, sur 187 directions, couvrant 10 angles en site (de -45° à +90° par pas de 15°), et 24 angles en azimut (de 0° à 360° par pas de 15°) pour des sites faibles, le nombre de pas en azimut décroissant à partir d'un site de 60°. Cette base de données a été analysée au préalable, pour le sujet ayant participé à l'enregistrement dans l'Espace de Projection, par la méthode d'estimation des indices intercanaux par égalisation et annulation proposée dans la section 2 du chapitre IV, en utilisant un banc de filtres gammatone à 24 canaux sur une échelle ERB, et une détection d'enveloppe par transformée de Hilbert.

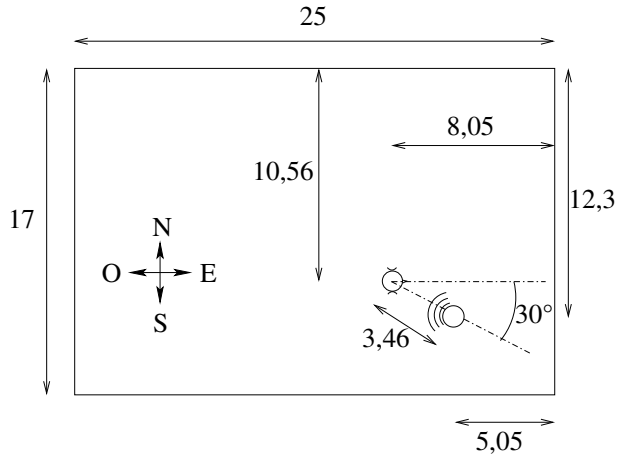


FIG. VI.22 – **Configuration de l'enregistrement** (les dimensions sont indiquées en mètres)

4.2 Estimation de fréquence fondamentale

La première étape consiste à détecter des événements harmoniques et estimer, si cela a un sens, la fréquence fondamentale instantanée, ceci de manière à repérer chaque note. Cette première analyse a un but double : d'une part, elle permet de trouver approximativement les attaques successives du signal, et donc d'éviter d'avoir à effectuer la détection bicanale (qui est relativement coûteuse en calcul) sur la totalité de l'enregistrement ; d'autre part, la connaissance de la fréquence fondamentale moyenne au sein de chaque note permet d'effectuer un rehaussement de chaque attaque selon le principe formulé en section 2 du chapitre V.

La figure VI.23 (à gauche) présente le résultat de la détection de fréquence fondamentale. Les fréquences sont représentées en octaves relativement à 440 Hz. Ces données sont ensuite analysées de manière à détecter les notes, qui correspondent à des plages de temps pour lesquelles la fréquence fondamentale est stable. Le résultat de cette modélisation est indiqué en figure de droite : sur la durée de l'enregistrement (23,5 secondes), 31 notes ont été détectées ; deux notes (15 et 27) sont des erreurs de l'algorithme dues aux superpositions de deux signaux harmoniques lorsque les notes sont trop rapides ; de plus, l'ornement (sol-fa-mi-fa) vers 17,5 secondes n'a pas été détecté car trop rapide. La première note présente de nombreuses octaviations, mais elle a été néanmoins détectée sur toute sa longueur, puisque l'algorithme de modélisation gère ce type de situations.

4.3 Estimation de la direction

Une fois la fréquence fondamentale détectée ainsi que le début, la fin et la fréquence fondamentale moyenne de chaque note, on est en mesure de procéder à la détection par égalisation et annulation. Celle-ci est opérée sur chacune des notes indépendamment, la détection proprement dite étant effectuée sur des plages de temps de 400 ms autour du début de chaque note. Pour chaque bande de fréquences pour laquelle une information cohérente sur les deux canaux a été détectée (le seuil de détection est fixé à 10 % pour les signaux bruts, et à 2 % pour les enveloppes), l'indice de décision fréquentiel $d_{EC,k}(\theta, \varphi)$ est calculé pour les 187 directions de la base de données. Ceci permet de calculer un indice de décision "pleine bande" pour chaque évènement sonore.

Ensuite, un indice de décision global est calculé sur la totalité des notes, sur le principe de la moyenne par bande des indices individuels présenté en section 3.6. Celui-ci sera l'indice final permettant de juger de la direction probable de la source.

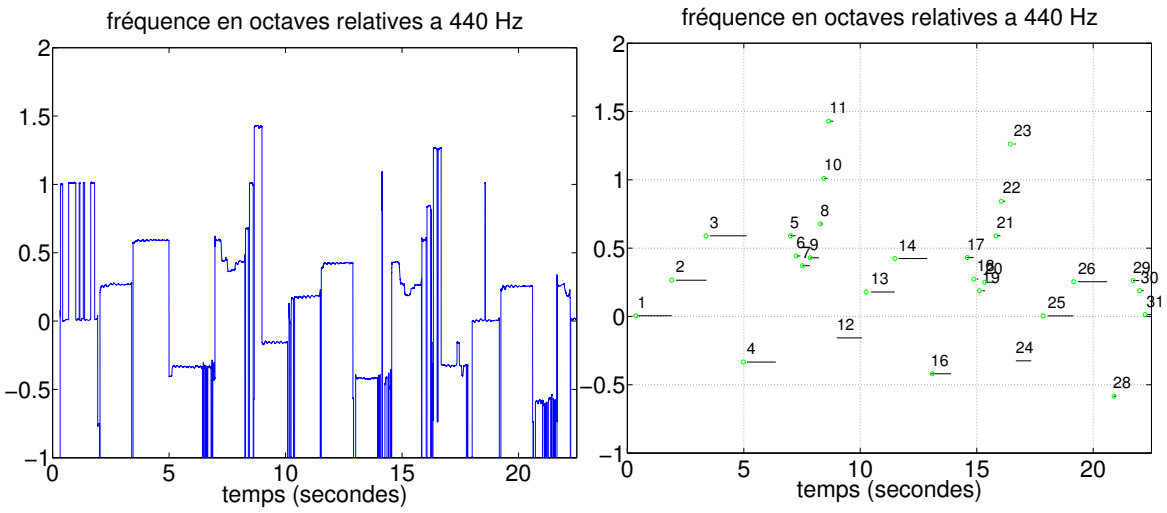


FIG. VI.23 – **Détection de fréquence fondamentale** : à gauche est représenté le résultat brut issu de la détection de fréquence fondamentale. A droite est représentée une modélisation sous forme d'évènements sonores à fréquence fondamentale constante.

4.4 Résultats

La figure VI.24 présente le résultat de l'estimation de la direction pour cet exemple. Sur la partie supérieure sont représentés les indices de décision (par bande et pleine bande) pour quelques évènements sonores. Les échelles, qui n'ont pas été indiquées par souci de lisibilité, sont identiques aux échelles des figures de la partie inférieure. On observe sur chacun de ces exemples, d'une part que l'indice de décision reste élevé (supérieur à 0,1) quelque soit la direction, et d'autre part que la direction est estimée de manière très floue : il n'y a pas un minimum mais une zone de minima, correspondant aux directions formant un angle par rapport au plan médian allant de -10° à -30°, et il est impossible d'estimer le site. Le fait que l'indice de décision reste élevé est dû au faible rapport signal sur bruit sur la totalité de l'enregistrement : bien que les queues de réverbération aient été pour la plupart largement atténuées (parfois de plus de 20 dB), la présence d'un important bruit de fond électronique constitue une importante gène.

La partie inférieure de la figure VI.24 indique le résultat de l'intégration séquentielle : la technique proposée permet de construire un indice de décision fréquentiel beaucoup plus complet que les indices individuels : sur les 24 bandes, seules 5 n'ont jamais été détectées sur les signaux sans détection d'enveloppe, alors qu'au maximum 10 bandes ont pu être détectées sur chaque évènement sonore. L'information obtenue est de fait plus riche, et permet de résoudre une grande partie des incertitudes observées sur les analyses individuelles à chaque évènement sonore : bien que l'indice de décision minimal reste relativement important (toujours à cause de la quantité de bruit), l'estimation de la direction est bien moins ambiguë, car non seulement les directions formant un angle de -10° à -20° avec le plan médian ne sont plus considérées comme candidates possibles, mais l'indice de décision obtenu permet une discrimination au sein du cône de confusion, et un rejet des directions arrières (partie droite de la bande à -30° sur la figure en bas à gauche. La source est finalement localisée en direction à -30° en azimut, et -15° en site.

Il est certain que la réflexion au sol, très présente, joue sur la précision de la localisation, puisqu'une observation des instants pour lesquels la détection a été jugée optimale indique que celle-ci est choisie fréquemment après l'arrivée de la première réflexion. La rapidité des attaques joue beaucoup sur cet aspect, puisque dans le cas de transitions très rapides, la détection peut avoir lieu avant la première réflexion, ce qui permet une estimation de la direction moins ambiguë. Cependant, si la première réflexion biaise relativement beaucoup l'estimation des différences d'intensité, elle n'a que peu d'influence sur les différences de

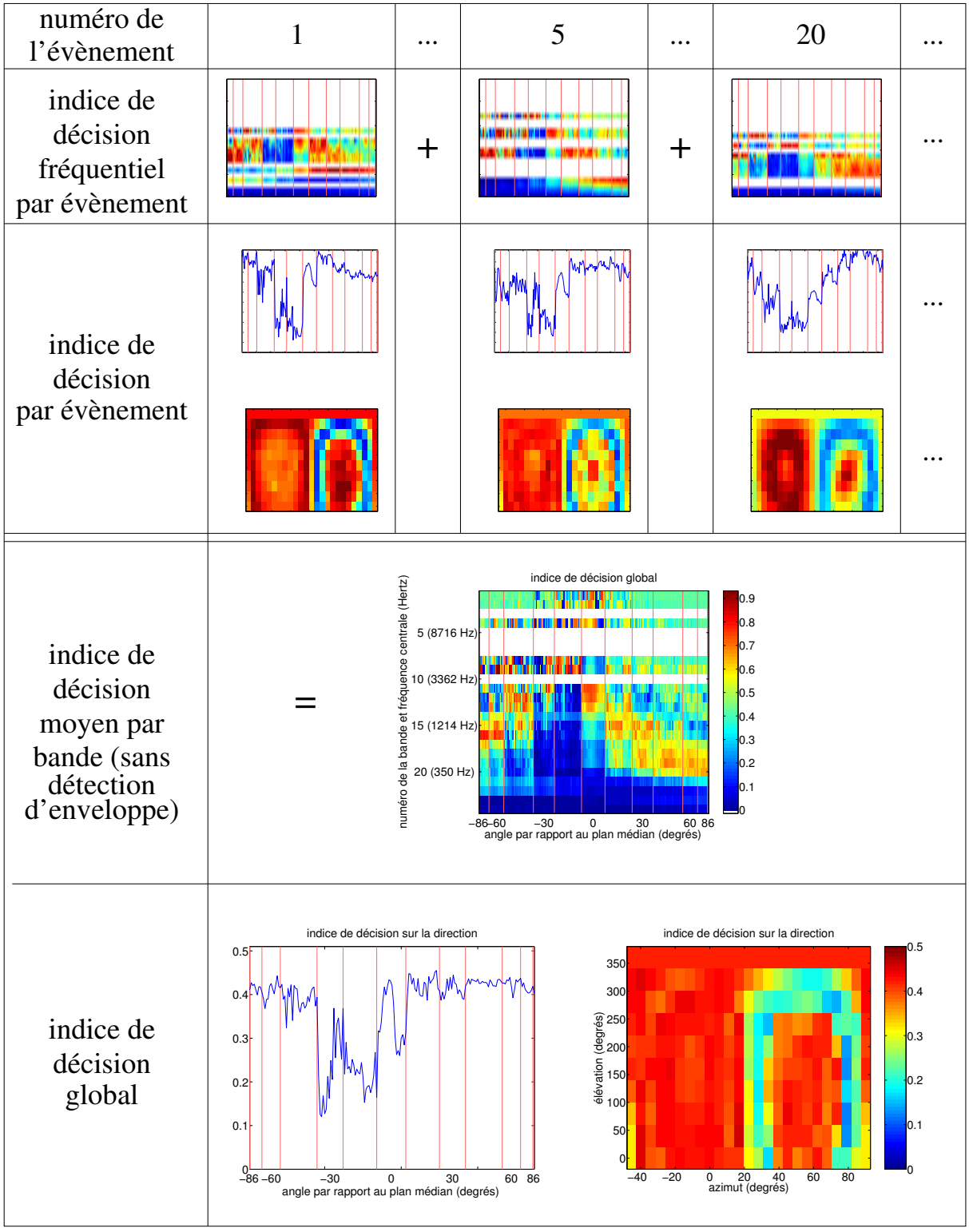


FIG. VI.24 – **Exemple complet d'estimation de la direction de la source à partir d'un enregistrement *in situ*** : seules les 20 premières secondes de l'enregistrement, contenant 25 notes, sont analysées ; pour chacune des notes, le calcul de l'indice de décision fréquentiel $d^{p}_{EC,k}(\theta,\varphi)$ est effectué aux instants pour lesquels la détection est optimale. Ces indices permettent de calculer un indice de décision moyen par bande, puis l'indice de décision global, qui est à la base de l'estimation de la direction. Un indice de décision pour chaque évènement est également calculé, et la comparaison avec l'indice de décision global permet de mettre en évidence le gain de précision apporté par une analyse sur plusieurs évènements.

temps, puisque comme il s'agit typiquement d'une réflexion sur le sol ou le plafond, elle provient du même azimut que l'onde directe, et donc pratiquement du même angle par rapport au plan médian.

5 CONCLUSION

La méthode d'estimation de la direction proposée ici résulte d'une volonté de combiner les avantages des principalement méthodes connues à ce jour participant ou non d'une modélisation de l'audition, tout en tâchant de s'affranchir au mieux des ambiguïtés intrinsèques aux outils d'estimation utilisés. Cette approche a conduit à une méthode d'estimation de la direction étroitement imbriquée à la méthode de détection élaborée au chapitre IV, notamment car elle n'utilise pas l'estimation explicite des différences de temps et d'intensité, mais repose sur la représentation qui permet de calculer ces derniers, c'est-à-dire l'erreur d'égalisation et annulation.

Elle repose sur la connaissance des différences de temps et d'intensité à partir des HRTF pour le plus grand nombre possible de directions de provenance, et sur l'évaluation systématique de l'erreur du modèle pour chacune de ces directions. Cette approche à l'avantage de laisser la possiblité d'estimer la direction hors du plan horizontal. Même s'il a pu être montré que la discrimination est bien meilleure en azimut qu'au sein d'un même cône de confusion, une estimation en élévation est tout de même envisageable lorsque le rapport signal sur bruit n'est pas trop défavorable.

En revanche, le fait de calculer l'erreur pour toutes les directions de provenance possibles est lent, et ne permet pas de juger efficacement de la direction lorsque celle-ci se trouve entre deux directions connues. On peut proposer deux améliorations pour corriger l'un ou l'autre de ces défauts. La première consisterait en une recherche non systématique sur toutes les directions connues, mais tâchant de converger rapidement vers la meilleure direction possible. On peut par exemple envisager de ne s'attacher en premier lieu, comme le propose Richard Duda, qu'à la localisation en azimut, puis à résoudre l'ambiguité au sein du cône de confusion. Ceci sous-entendrait une première estimation grossière basée sur une valeur moyenne des différences de temps et d'intensité, à laquelle succède une estimation avec la précision maximale, mais au sein d'un nombre limité de possibilités.

Une seconde amélioration possible, qui y fait directement suite, consisterait à interpoler les indices intercanaux entre deux positions connues, de manière à pourvoir juger de la direction même lorsque celle-ci n'est pas répertoriée dans la base de données. Ce type d'interpolation suppose néanmoins que l'échantillonnage spatial soit suffisament fin. On peut également envisager une autre manière de concevoir la base de données : actuellement, celle-ci est basée sur un modèle **gain-retard**, contenant les différences de temps et d'intensité pour le plus grand nombre possible de directions. Or on a vu que le modèle gain-retard perdait de sa pertinence au voisinage d'un pôle ou d'un zéro, alors que ceux-ci sont inévitables en présence d'obstacles comme le torse et la tête. Il est possible d'envisager que le modèle sous-jacent à la base de données soit au contraire un modèle **pôles-zéros**, qui viserait à répertorier tous les pôles relatifs à la propagation, ainsi que leurs trajectoires spatiales et fréquentielles. Dans cette hypothèse, il serait possible de **reconstruire** la fonction de transfert intercanale pour n'importe quelle direction, et donc d'estimer les différences de temps et d'intensité pour n'importe quel banc de filtres, de manière à effectuer l'estimation de la direction selon la méthode présentée dans ce chapitre.

VII

Description de la réverbération

Ce chapitre a pour objet de proposer des méthodes de description de l'enveloppe de la réverbération à partir de signaux musicaux. Après avoir développé en section 1 la question de la distinction entre aspects spatiaux et temporels de la réverbération, on s'applique en section 2 à rappeler quelques unes des principales méthodes d'estimation de la durée de réverbération à partir de réponses impulsionnelles ou de signaux musicaux ; puis, la section 3 propose un couplage de ces méthodes d'estimation avec les techniques de détection de réverbération par égalisation et annulation développées aux chapitres IV et V. Finalement, la section 4 présente deux exemples d'analyse, l'un sur une scène sonore virtuelle, l'autre sur un enregistrement *in situ*.

1 LA RÉVERBÉRATION : ASPECTS SPATIAUX ET TEMPORELS

Comme le rappellent les chapitres I et II, la réverbération est un phénomène à la fois spatial et temporel. L'aspect spatial correspond à la formation progressive d'un champ diffus : d'une onde sphérique ou plane provenant d'une direction précise, on bascule petit à petit vers un champ pour lequel l'énergie provient de partout. Ce champ, perçu comme enveloppant par un auditeur humain, ne permet plus de localiser la source sonore. La dimension temporelle est entre autres caractérisée par le temps nécessaire à l'énergie confinée dans la salle pour être absorbée par les parois. Cette notion de durée porte donc sur l'enveloppe de la réverbération, et non pas sur le détail de l'effet de salle, c'est-à-dire sur l'instant d'arrivée de chacune des réflexions. En fait, bien que ce dernier aspect ait un effet perceptible pour les réflexions précoces, notamment sur la sensation de distance à la source ou de taille de la salle, la durée de l'enveloppe est quasiment la seule caractéristique (en l'absence d'échos flottants) de la réverbération qui soit consciemment perçue comme étant de nature temporelle.

Ces deux aspects ne se manifestent pas de la même manière dans les signaux observés, et il est donc nécessaire de les considérer distinctement. Comme on a pu le voir au chapitre III, les principaux indices sur la distribution spatiale de l'effet de salle sont à rechercher dans la similarité, ou l'absence de similarité, entre les signaux de pression enregistrés en deux points distincts. Il ne s'agit cependant plus ici d'un problème de détection de source, mais du problème quasiment dual, qui est la **détection de réverbération** : on recherche dans l'enregistrement les plages de temps pour lesquelles la réverbération n'est pas masquée par l'onde directe. Si la diffusion et le niveau du champ réverbéré sont suffisants, ces plages de temps correspondent aux moments pour lesquels la superposition d'ondes provenant de multiples directions ôte toute cohérence entre les signaux observés. La méthode de détection par égalisation et annulation a donc pour vocation de **distinguer le flux d'arrière-plan du flux de premier plan**, en se référant au vocabulaire employé notamment par Griesinger (1997) (et introduit au chapitre II).

En revanche, la notion de durée de réverbération ne dépend pas des relations intercanales. Pour s'en persuader, on peut se reporter aux résultats de la section 3.3 du chapitre I : dès lors que l'on parvient à un champ raisonnablement diffus, la cohérence à court-terme varie peu, alors que l'énergie continue de décroître de manière exponentielle. D'ailleurs, la durée de réverbération (qu'il s'agisse de la durée de réverbération tardive ou précoce) est habituellement caractérisée à partir de réponses impulsionnelles monophoniques. Cette deuxième tâche, qui constitue l'**estimation de la durée de réverbération** proprement dite, nécessite donc d'être menée par une autre méthode que par celles présentées jusqu'ici.

Ainsi, contrairement au problème de l'estimation de la direction, il n'est pas possible ici d'effectuer la détection et l'estimation en une seule opération, puisque les paramètres du modèle utilisé pour la détection (qui est également le modèle utilisé pour la détection de source) n'ont pas de rapport direct avec la durée de la réverbération. La figure VII.1 résume ce principe de distinction entre les deux tâches.

2 MÉTHODES D'ESTIMATION DES DURÉES DE RÉVERBÉRATION

On rappelle avant tout dans cette section les principales méthodes permettant de caractériser l'enveloppe de la réverbération à partir de mesures, ceci car ceci constitue le fondement de la méthode proposée ultérieurement, qui n'en est qu'une extension à des signaux musicaux. Puis, sont présentées plusieurs recherches sur l'estimation de la durée de réverbération à partir de signaux musicaux, avec ou sans la connaissance du signal source.

2.1 Estimation des durées de réverbération à partir de mesures

Méthode du bruit interrompu

Le premier type de durée de réverbération à laquelle les acousticiens se sont intéressés est le **temps de réverbération**, qui, on le rappelle, caractérise l'enveloppe de décroissance tardive, qui est supposée exponentielle (ce qui exclut le cas de doubles décroissances). La

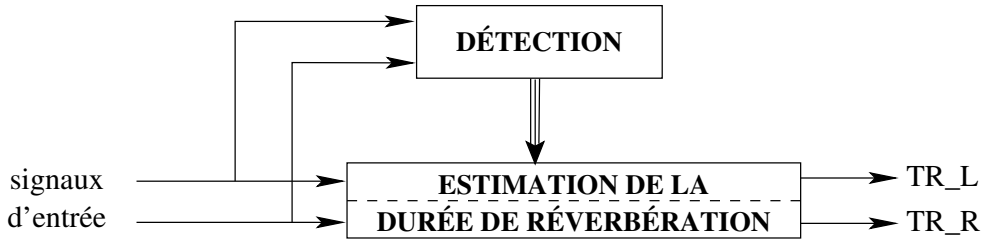


FIG. VII.1 – **Schéma de principe de la description de la réverbération** : la détection est effectuée à partir de la connaissance simultanée des deux voies du signal observé (après filtrage par le banc de filtres). L'estimation de la durée de réverbération est effectuée de manière distincte pour chacune des deux voies, et fournit donc deux estimations du temps de réverbération, TR_L et TR_R.

méthode de mesurage du temps de réverbération initialement proposée, qui a fait autorité pendant longtemps et constitue toujours le principe à la base de la grande majorité des méthodes de mesurage actuelles, et celle du **bruit interrompu** (Kuttruff, 2000; AFNOR, 2000) : après avoir excité la salle au moyen d'une séquence de bruit stationnaire diffusée par un haut-parleur de puissance suffisante, on interrompt celle-ci brutalement, tout en enregistrant la puissance instantanée de la pression acoustique au point d'écoute considéré. Le temps de réverbération est estimé à partir de l'étude de la décroissance sur cet enregistrement. Cette mesure est répétée de nombreuses fois pour le même point d'écoute, ce qui permet soit d'effectuer une moyenne des temps de réverbération estimés, soit de la courbe de décroissance, de manière à obtenir une estimation du temps de réverbération offrant la variance le plus faible possible. Puisque le temps de réverbération dépend de la fréquence, la séquence de bruit utilisée comme signal source est très souvent à bande limitée, typiquement en octave ou en tiers d'octave, si bien que le nombre de mesures s'en trouve décuplé par le nombre de bandes sur lesquelles on désire faire l'estimation.

Méthode de la décroissance intégrée

Schroeder (1965) a montré qu'il était possible d'effectuer la même estimation en une seule mesure. Cependant, celle-ci est plus complexe, puisqu'elle consiste à estimer la réponse impulsionnelle caractérisant la propagation de la source au point d'écoute considéré, ce qui nécessite soit d'employer des signaux très courts (sons purs, tirs au pistolet, éclatement de ballon), ce qui pose le problème de la limitation en puissance, soit d'effectuer une déconvolution par un signal source *ad hoc* (par exemple un sinus glissant ou une séquence à longueur maximale (Schroeder, 1979)). On appelle **courbe de décroissance en énergie** (*energy decay curve - EDC*) ou **décroissance intégrée** l'intégration rétrograde du réflectogramme, c'est-à-dire la puissance instantanée de la réponse impulsionnelle :

$$EDC_h(t) = \int_{\tau=t}^{+\infty} h^2(\tau).d\tau$$

Schroeder montre que la décroissance intégrée dans une bande de fréquences donnée est proportionnel à la décroissance idéale obtenue par la méthode du bruit interrompu, c'est-à-dire la moyenne de la puissance instantanée après extinction de la source lorsque le signal émis par celle-ci est une séquence stationnaire de bruit dans la bande considérée :

Soit $n(t)$ une séquence de bruit, de variance σ_n^2 , stationnaire pour $t < 0$, et interrompue en $t = 0$. Soit $y(t)$ le signal observé, qui, si l'hypothèse d'une propagation linéaire et à temps invariant est valide, est égal au produit de convolution de $n(t)$ par la réponse impulsionnelle $h(t)$. On montre alors que sa puissance instantanée moyenne vaut :

$$\left\langle y^2(t) \right\rangle = \left\langle (h * n)^2 (t) \right\rangle = \sigma_n^2 . EDC_h(t)$$

Le principe de l'estimation du temps de réverbération par la méthode du bruit interrompu ou, de manière équivalente, par décroissance intégrée, est à rapprocher de l'hypothèse de locale stationnarité de la réverbération tardive mentionnée au chapitre I. En effet, on peut montrer (voir la section 2 de l'annexe C) que dans ce cadre, l'intégration rétrograde de la puissance instantanée de la réponse impulsionnelle est proportionnelle à celle de l'enveloppe, et est donc elle aussi exponentielle.

Cependant, l'utilisation de la décroissance intégrée ne se limite pas à la réverbération tardive ; en effet, on l'utilise depuis déjà longtemps pour caractériser également la partie précoce de la réverbération, même alors que l'hypothèse de stationnarité locale n'est plus valable. Cette démarche a été encouragée par deux facteurs : d'une part, on sait que la sensation perceptive de réverbérance est mieux corrélée avec la durée de réverbération initiale qu'avec le temps de réverbération (voir chapitre II) ; d'autre part, le gain de précision apporté par la méthode de la décroissance intégrée par rapport à la méthode du bruit interrompu permet une bien meilleure estimation de la première partie de la décroissance, puisque l'on s'affranchit des fluctuations dues à la nature aléatoire de l'excitation (Schroeder, 1965).

Estimation en bandes étroites

La résolution fréquentielle de l'estimation du temps de réverbération selon la méthode exposée ci-dessus est intrinsèquement limitée par celle du banc de filtres employé. Ce constat à incité certains chercheurs à généraliser la démarche de Schroeder à une véritable représentation temps-fréquence de l'enveloppe de réverbération. Ainsi, Jot (1992) a défini le **relief de décroissance** comme suit :

$$EDR_h(t,f) = \left| \int_{\tau=t}^{+\infty} h(\tau).e^{-j2\pi f\tau}.d\tau \right|^2$$

Cette définition du relief de décroissance, appelée **spectre courant futur**, peut être vue comme l'intégration rétrograde d'une distribution appelée distribution de Page-Levin :

$$EDR_h(t,f) = \int_{\tau=t}^{+\infty} \rho_h(\tau,f).d\tau$$

avec

$$\rho_h(t,f) = -\frac{\partial}{\partial t} \left| \int_{\tau=t}^{+\infty} h(\tau).e^{-j2\pi f\tau}.d\tau \right|^2$$

Jot a pu montrer que la distribution de Page-Levin satisfait à des propriétés très intéressantes vis-à-vis de la réverbération : notamment, sa marginale en fréquence et sa marginale en temps sont respectivement égales au spectre et à la puissance instantanée de la réponse impulsionnelle, si bien que l'intégration rétrograde de cette distribution s'avère être un déploiement temps-fréquence de la courbe de décroissance de Schroeder en bande large. En généralisant, il montre que si l'on considère le filtrage de la réponse impulsionnelle par un filtre passe-bande à phase linéaire et à support temporel limité, il est possible d'en retrouver la courbe de décroissance comme la marginale en temps pondérée du relief de décroissance : soit $w(t)$ le filtre considéré, de transformée de Fourier $W(f)$ et supposé à retard de groupe τ_w constant ; alors la courbe de décroissance de la réponse impulsionnelle $h(t)$ filtré vaut :

$$EDC_{h*w}(t) = \int_{-\infty}^{+\infty} |W(f)|^2 .EDR_h\left(t-\tau_w, f\right).df$$

Ce résultat permet en théorie de retrouver la courbe de décroissance dans n'importe quelle bande de fréquences à partir du relief de décroissance.

En pratique, Jot substitue à la définition du spectre courant futur celle du **spectrogramme cumulé** :

$$CS_h[p,k] = \sum_{i=p}^{+\infty} STFS_h[i,k]$$

Dans cette expression, $STFS_h$ désigne le spectre de Fourier à court-terme (numérique) de la réponse impulsionnelle considérée, et p et k respectivement les indices temporel et fréquentiel considérés. Cette définition ne satisfait pas exactement les propriétés mentionnées ci-dessus, mais elle offre plusieurs avantages : d'une part, le cumul en énergie permet de réduire les fluctuations de la représentation, qui sont très importantes pour le spectre courant futur (on peut montrer à partir d'un modèle stochastique de la réverbération tardive que l'écart-type du spectre courant futur est de l'ordre de sa moyenne (Baskind, 1999) !) ; d'autre part, l'utilisation du spectrogramme comme représentation temps-fréquence initiale permet d'effectuer des transformations très utiles comme du débruitage (Jot et al., 1997), ce qui améliore sensiblement la qualité de l'estimation pendant l'analyse, et permet de synthétiser une réponse impulsionnelle artificielle, similaire à la réponse initiale, mais exempte de bruit de fond.

En fait, si l'on se souvient que la transformée de Fourier à court-terme (en convention passe-bande) peut être considérée comme la résultante d'un banc de filtres, on en conclut que le spectrogramme cumulé ne diffère pas dans le principe du calcul de décroissances intégrées bande de fréquences par bande de fréquences, alors que le spectre courant futur résulte d'un point de vue distinct, issu d'une représentation temps-fréquence continue, qui est la distribution de Page-Levin. Si cette dernière présente de nombreux atouts théoriques, elle est peu adaptée au cadre de cette étude, qui vise à rester la plus générale possible par rapport au choix de l'analyse temps-fréquence. En revanche, dans le cas où la détection a été effectuée par transformation de Fourier à court-terme, la définition du relief de décroissance par spectrogramme cumulé paraît toute adaptée : d'une part, il n'est pas nécessaire de calculer la représentation temps-fréquence initiale puisque cela a déjà été fait pour la détection, et d'autre part, on minimise l'amplitude des fluctuations, ce qui sera d'une grande importance étant données les conditions particulièrement délicates de l'estimation.

2.2 Estimation des durées de réverbération à partir de signaux musicaux

Le problème de la caractérisation de la réverbération à partir de signaux musicaux dépasse le simple intérêt intellectuel : il se justifie avant tout par la nécessité de décrire la réverbération d'une salle pleine. En effet, il est difficilement envisageable d'effectuer des mesures acoustiques, quelles qu'elles soient, lorsque la salle est pleine, c'est-à-dire la plupart du temps en situation de performance ou de conférence publique. Or puisque la grande majorité des indices objectifs permettant de qualifier la salle (et donc en particulier les durées de réverbération) sont très sensibles au taux d'occupation de la salle, la connaissance de ces indices uniquement pour un espace vide est d'une utilité assez limitée. L'une des approches envisagées consiste à estimer le temps de réverbération en salle pleine à partir des données en salle vide (Barron, 1993; Beranek, 1996). Ceci pose néanmoins le problème de la précision du modèle envisagé, et ne permet de caractériser que la réverbération tardive. Une autre solution, proposée par Schroeder (1979), vise à effectuer un mesurage de réponses impulsionnelles au cours du concert, en diffusant à un faible niveau des séquences de longueur maximale : celles-ci, étant perçues comme stationnaires par un auditeur, sont au mieux masquées, ou du moins atténuées par le son principal ; le problème du rapport signal sur bruit, forcément extrêmement défavorable, est compensé par le fait qu'il est possible de répéter un grand nombre de fois ces mesures pendant la durée du concert, à supposer bien sûr que la propagation soit invariante dans le temps, avec toutes les réserves que cette hypothèse implique, soulevées par Schroeder et mentionnées au chapitre I. Exceptées ces deux solutions, la seule manière de caractériser la réverbération en salle pleine est de l'estimer à partir d'enregistrements de musique (si il s'agit d'un concert) effectués lors de l'évènement.

Méthode de l'accord interrompu

Historiquement, la première approche a visé à généraliser la méthode du bruit interrompu à des signaux de musique. Cette technique, dite de l'**accord interrompu** (*stop chord*) (Barron, 1993; Beranek, 1996), consiste à repérer les instants pour lesquels la musique s'interrompt

brusquement (par exemple, juste après un accord *tutti*), et à analyser la décroissance qui s'en suit. Cette technique nécessite donc une supervision humaine pour détecter les instants utiles. Puisque l'on ne fait plus appel à un signal source aléatoire, on ne peut pas effectuer de moyenne statistique. Cependant l'utilisation de décroissances intégrées peut se justifier même sur des décroissances ne suivant pas une séquence de bruit stationnaire, encore une fois grâce à l'hypothèse d'ergodicité locale de la réponse impulsionnelle. En effet, les calculs de la section 3 de l'annexe C indiquent que l'intégration rétrograde de la décroissance suivant un signal quelconque s'interrompant brusquement est de nature exponentielle.

Approches par déconvolution

Le principe global de ces approches est de retrouver dans chaque bande de fréquences la courbe de décroissance associée à "la" réponse impulsionnelle caractérisant la propagation, de manière à effectuer par la suite l'analyse de la même manière que dans une situation de mesure. Il s'agit de problèmes de déconvolution, puisque l'on cherche à estimer à partir de signaux supposés résultant d'une convolution le filtre ou du moins l'information utile dans celui-ci, c'est-à-dire l'enveloppe de décroissance pour le problème présent.

L'approche systématique consiste à estimer la totalité de la réponse impulsionnelle relative à la propagation, ce qui est le problème dual de celui de la déréverbération. On suppose en général que le signal source est inconnu, si bien qu'on se trouve face à un problème de déconvolution aveugle. Une technique courante pour tenter de le résoudre est la déconvolution homomorphique (Baskind et Warusfel, 2002; Bees et al., 1991; Petropulu et Subramaniam, 1994; Liu et al., 1995), qui revient à effectuer une moyenne dans le domaine cepstral. Le cas où le signal source est connu peut être traité de la même façon, avec bien entendu plus de facilité. Conformément aux remarques émises à la section 1.3 du chapitre I, on note que cette approche trouve encore une fois ses limites dans l'hypothèse d'un canal acoustique invariant dans le temps.

Une manière de contourner ce problème de variation temporelle du canal est de ne chercher à estimer non pas la totalité de la réponse impulsionnelle, mais uniquement son enveloppe par bande de fréquences, ce qui suffit à estimer le temps de réverbération, par exemple après intégration rétrograde. Ainsi, Polack (Polack, 1982; Polack et al., 1984) montre qu'il est possible de considérer l'enveloppe (ou "modulation") d'un signal réverbéré comme résultant d'un filtrage (bruité) de l'enveloppe du signal source par l'enveloppe de la réverbération, celle-ci étant définie dans le domaine fréquentiel comme la **fonction de transfert de la modulation** (Schroeder, 1981). Cette technique nécessite néanmoins la connaissance du signal source, ou du moins son enveloppe.

Une limitation importante aux approches par déconvolution est que celles-ci supposent qu'une seule source soit active dans l'espace considéré, car elles ne permettent pas de gérer des mélanges. Pour les mêmes raisons, il est peu envisageable de traiter de la sorte (surtout par déconvolution systématique) des enregistrements fortement bruités.

Autres approches

Toutes les techniques d'estimation proposées jusqu'ici se basent sur l'analyse de bruits interrompus ou de réponses impulsionnelles. Ceci dit il est tout-à-fait envisageable d'estimer le temps de réverbération à partir de représentations moins usuelles.

Ainsi, Hansen (1995) a montré que l'autocorrélation d'une réponse impulsionnelle de salle présente en théorie une décroissance exponentielle (par bande de fréquences) tout-à-fait similaire à celle de la réponse impulsionnelle elle-même. Ceci l'a amené à proposer une méthode d'estimation aveugle[1] du temps de réverbération à partir de l'autocorrélation de sections successives (recouvrantes) d'enregistrements réverbérés. Il s'agit d'une méthode systématique, pour laquelle une durée de décroissance est estimée quasiment pour chaque segment, que celui-ci corresponde ou non à une réverbération libre. En supposant que l'énergie ne peut décroître plus vite que la réverbération, il propose de choisir comme estimation du temps de réverbération la plus faible de toutes ces durées de décroissances dans chaque

[1] c'est-à-dire ne nécessitant pas la connaissance du signal source

bande. La méthode de Hansen repose donc elle aussi sur l'hypothèse convolutive, et nécessite donc qu'une seule source soit présente à la fois.

Wu et Wang (2003) ont récemment proposé une approche originale, basée non pas sur l'analyse temporelle de décroissances, mais sur l'étude de l'effet de la réverbération sur l'harmonicité des signaux : en effet, la réverbération provoque une fluctuation aléatoire des phases de chaque partiel d'un signal (Blesser, 2001), ce qui entraîne une chute de l'harmonicité et fait fluctuer les estimations de la fréquence fondamentale instantanée autour de la valeur nominale. L'hypothèse de Wu et Wang est que l'étalement de la distribution de ces fluctuations est fonction du temps de réverbération. Ainsi, il serait possible d'estimer ce dernier uniquement grâce à un algorithme de suivi de hauteur. Cette idée est à rapprocher des méthodes de détection de signal basées sur l'harmonicité mentionnées en section 2.2 du chapitre III.

Discussion

Les méthodes présentées ici diffèrent dans le principe, mais la plupart se ramènent *in fine* à l'estimation du temps de réverbération à partir d'une décroissance intégrée. Ce choix consensuel n'est pas dû au hasard, et s'explique par le fait qu'il s'agit d'un mode de représentation particulièrement adapté à l'étude de la réverbération, ou plus généralement, de décroissances exponentielles, imputables ou non à des modes de résonance. En effet, il suffit de rappeler que l'intégration rétrograde d'une exponentielle décroissante est égale à un facteur près à la même exponentielle décroissante pour comprendre que l'on peut par ce biais atténuer largement les fluctuations sans modifier la nature de l'enveloppe. Cette technique est donc utilisée dans cette étude.

Pour que la méthode proposée reste la plus générale possible, on choisit d'écarter les techniques de déconvolution, qui d'une part nécessitent la connaissance de tout ou partie de l'information véhiculée par la source pour offrir une dynamique suffisante, et d'autre part ne permettent pas de gérer la présence simultanée de plusieurs sources. En ce sens, les techniques utilisant une analyse temps-fréquence sont attrayantes, car elles n'excluent pas la possibilité d'analyser la réverbération relative à une source même lorsqu'une nouvelle source rayonne, du moment que les deux sources ne partagent pas les mêmes bandes de fréquences. La technique proposée par Hansen effectue ce type d'opération implicitement, puisqu'il s'agit d'une méthode systématique (c'est-à-dire effectuant une analyse pour toutes les sections successives du signal, sans choix préalable de celles jugées les plus significatives), et il est tout-à-fait possible d'envisager l'extension dans cet esprit du champ d'application de la méthode *stop chord.*

En revanche, aucune de ces deux méthodes ne propose une détection "intelligente" des queues de réverbération : en effet, l'analyse d'accords interrompus nécessite une supervision manuelle pour décider du début et de la fin des segments à analyser, et la méthode de Hansen contourne ce problème, encore une fois puisqu'elle est systématique. Or l'application au chapitre V du principe de détection par égalisation et annulation au cas de signaux réverbérés nous permet d'y apporter une solution, en tirant parti des deux voies de l'enregistrement.

En section suivante est étudiée l'application conjointe de la méthode de détection par égalisation et annulation et de la méthode d'estimation basée sur le principe d'intégration rétrograde.

3 ESTIMATION DU TEMPS DE RÉVERBÉRATION BASÉE SUR LA DÉTECTION PAR ÉGALISATION ET ANNULATION

On se propose dans cette section d'adapter le principe de l'intégration rétrograde à des signaux quelconques, en se basant étroitement sur la détection de réverbération telle qu'elle a été formulée au chapitre V. Après avoir énoncé le principe, on indique quelles sont les principales difficultés auxquelles on doit s'attendre lorsque l'on analyse des signaux autres que des réponses impulsionnelles.

3.1 Principe

Lors d'une analyse de réponses impulsionnelles par la méthode de l'intégration rétrograde énoncée par Schroeder, la borne supérieure de l'intégration devrait être en théorie $+\infty$. Ceci est en pratique impossible, et ce pour deux raisons : la première, évidente, est qu'une réponse impulsionnelle n'est jamais mesurée pour une durée infinie, si bien que le domaine d'intégration est limité à la durée de l'enregistrement de la réponse ; la seconde raison est que la présence de bruit de fond limite la zone utile sur laquelle il est possible de caractériser la réverbération. En pratique, la borne supérieure de l'intégration est choisie de préférence après l'émergence du bruit de fond, pour éviter les effets de troncature (voir la section 3.2 de ce chapitre, ainsi que la section 4 de l'annexe C).

Dans le cas présent, on cherche à appliquer le principe de l'intégration rétrograde non plus à des mesures de réponses, mais à des enregistrements *in situ*. Or ce cas est très différent de celui de réponses impulsionnelles, la distinction majeure étant que dans la majorité des situations, ce n'est plus le bruit de fond qui limite la dynamique utile, mais l'arrivée d'une nouvelle onde directe. **Il est alors complètement exclu de fixer la borne supérieure de l'intégration rétrograde après cet instant** : en effet, alors qu'on maîtrise bien l'effet d'un bruit de fond stationnaire additionnel sur l'allure de la décroissance intégrée, on ne peut pas prévoir l'influence d'un signal source inconnu. La solution optimale pour maximiser la dynamique de la décroissance, tout en évitant ce problème, est de placer la borne supérieure de l'intégration **juste avant l'émergence de la nouvelle onde directe**.

C'est à ce stade qu'intervient la détection de réverbération : son rôle étant justement de préciser les zones sans source active susceptibles de correspondre à des queues de réverbération, on se fie complètement à elle pour décider des limites inférieures et supérieures de l'intégration rétrograde. Les figures VII.2 et VII.3 illustrent cet aspect : le signal observé, synthétique, est issu de la convolution par une réponse binaurale de salle d'une succession de deux séquences de bruit blanc séparées par un silence relativement long (1,5 secondes). Les séquences de bruit blanc sont volontairement choisies très courtes, de manière à adhérer aux hypothèses de travail élaborées au chapitre V. La réponse de salle choisie est issue du corpus étudié à ce même chapitre, lorsque le sujet se trouve à 1,8 mètres de la source. Le seuil de détection est absolu et fixé à 10 % (0,1) quelle que soit la fréquence. L'intégration rétrograde est effectuée à partir de la fin de la zone pour laquelle la réverbération a été détectée (en vert sur la figure), juste avant l'émergence de la nouvelle séquence de bruit, repérée en bleu sur la figure. L'estimation du temps de réverbération est effectuée par régression linéaire sur la décroissance intégrée, les limites temporelles étant définies par la détection de réverbération. On observe que ces limites semblent bien correspondre à la réverbération tardive dans la majorité des cas. Cependant on perd toute possibilité d'estimer le temps de réverbération en basses fréquences (en dessous de 200 Hz), puisque la détection de réverbération ne fonctionne pas sur cette plage, pour les raisons évoquées au chapitre V.

3.2 Effet de la troncature

La décroissance observée sur la figure VII.3 permet de rappeler qu'il n'est pas possible de supposer le résultat de l'intégration rétrograde comme linéaire (en représentation logarithmique) sur toute sa longueur, même lorsque le segment temporel considéré correspond à la réverbération tardive. En effet, le fait que la limite supérieure de l'intégration ne soit pas repoussée à l'infini biaise la courbe par rapport à une exponentielle parfaite, ce biais étant d'autant plus important que l'on s'approche de la fin du segment temporel considéré.

Intuitivement, on peut imaginer deux manières de tenir compte de ce biais. Une première solution consisterait à effectuer une **régression non linéaire** sur la totalité du segment : on tient alors explicitement compte de la troncature dans l'expression du modèle théorique de décroissance, qui est alors de la forme :

$$(d_{th}[n])_{dB} = A.n + B + 10 * log_{10}\left(1 - e^{-\frac{N_{tronc}-n+1}{F_e.T_r}}\right)$$

Dans cette expression, N_{tronc} est l'indice où est effectuée la troncature, F_e est la fréquence d'échantillonnage, et T_r est le temps de réverbération. Ce modèle théorique de décroissance

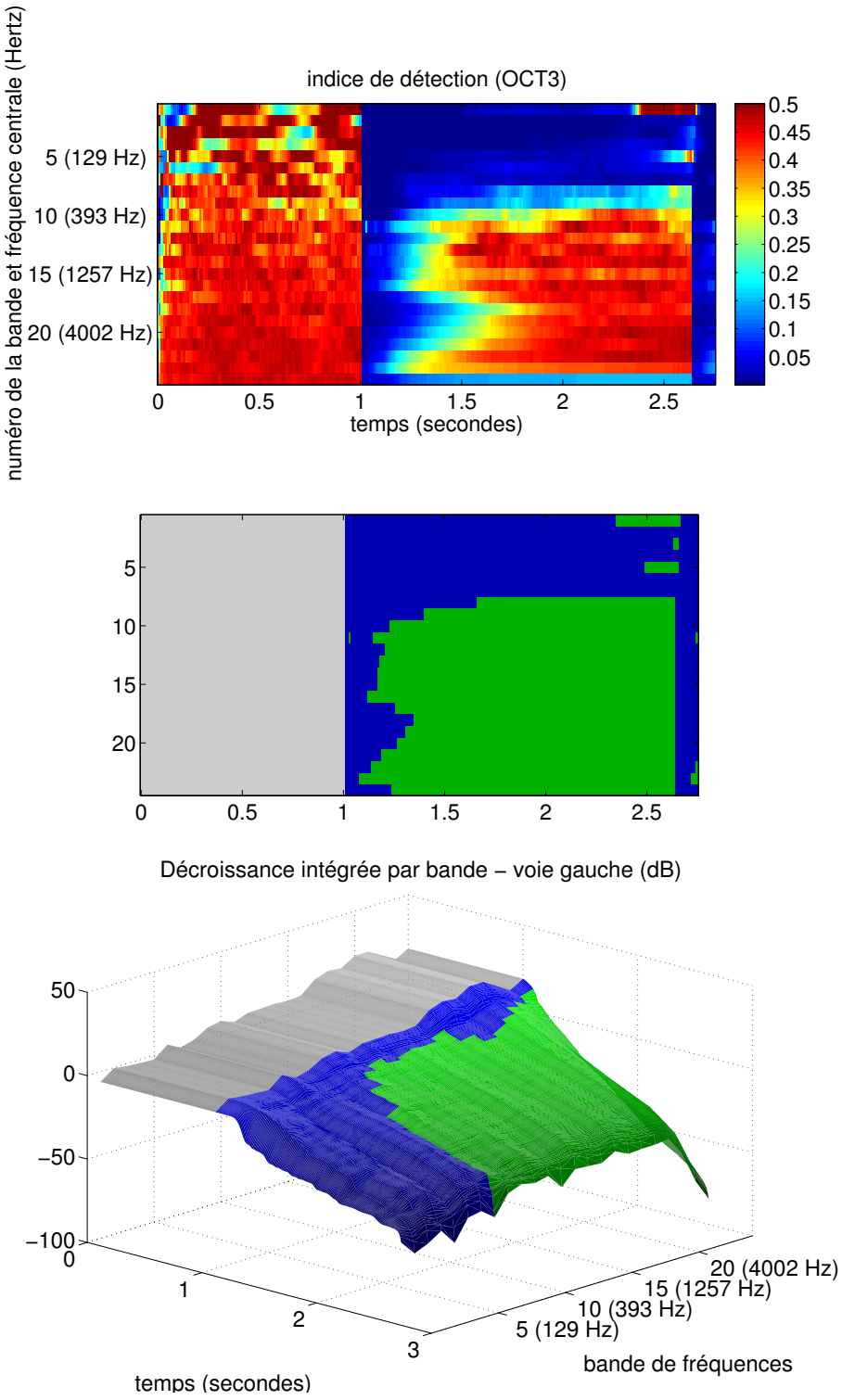


FIG. VII.2 – **Principe de la description de réverbération basée sur la détection (1)** : le signal observé résulte de la convolution par une réponse binaurale de salle d'une séquence courte (125 ms) de bruit blanc suivie d'un silence d'une seconde et demie, puis d'une nouvelle séquence de bruit. La figure supérieure est l'indice de détection calculée sur 24 bandes de fréquences en tiers d'octave. La figure intermédiaire en est l'interprétation par le détecteur (seuil de détection à 0,1). La figure inférieure est la décroissance intégrée calculée à partir de l'instant où la seconde séquence de bruit s'active. Les zones de détection y sont plaquées, de manière entre autres à indiquer que de l'information pertinente n'est pas utilisée en basses fréquences.

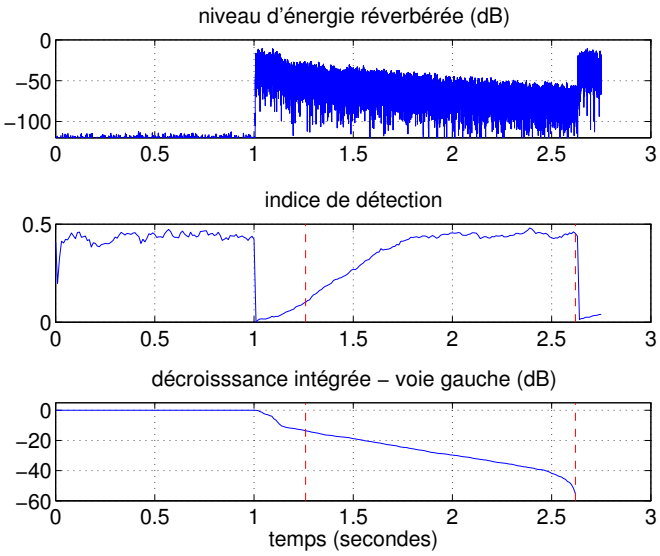


FIG. VII.3 – **Principe de la description de réverbération basée sur la détection (2)** : le signal observé est identique à celui de la figure précédente. Le résultat de l'analyse est cette fois représenté pour une seule bande, qui est la bande en tiers d'octave centrée sur 2 kHz. Les lignes verticales interrompues représentent le début et la fin de la zone de réverbération dégagée par le détecteur.

diffère du modèle linéaire par le troisième terme, qui est d'autant plus important que n est proche de N_{tronc}. Ce principe a été testé en utilisant la méthode de descente de gradient de type Gauss-Newton, mais n'a pas donné de résultats satisfaisants, car l'approximation inhérente à la méthode (les dérivées secondes sont négligées) ne convient pas à la fonction considérée.

Une autre solution, qui est celle qu'on choisit pour la suite, revient à effectuer une régression linéaire, mais pas sur la totalité du segment : en prenant une **marge de retrait** par rapport à la limite temporelle supérieure, on tâche de se ramener au cas où l'on peut négliger l'effet de la troncature. La figure VII.4 illustre l'effet de la marge de retrait sur la régression linéaire : plus la marge est importante, plus la régression linéaire est fidèle à la décroissance réelle.

La figure VII.5 illustre plus précisément cette influence de la marge de retrait sur l'estimation du temps de réverbération : pour des valeurs faibles de la marge de retrait, le biais de l'estimation diminue progressivement de 250 millisecondes pour une marge nulle jusqu'à 30 millisecondes pour une marge de 0,3 secondes. Puis le biais stagne autour de cette valeur palier, pour des valeurs de la marge allant jusqu'à 1,1 seconde. Au-delà, la durée du segment temporel sur lequel est effectué la régression n'est plus suffisante pour que l'estimation soit de bonne qualité, et cette dernière diverge.

La section 4 de l'annexe C permet de fixer un ordre de grandeur de la marge de retrait minimale pour négliger les effets de la troncature : si celle-ci est supérieure au dixième du temps de réverbération, alors l'effet de la troncature sur la décroissance logarithmique est inférieur à 1 dB. Puisque l'on ne connaît bien sûr pas le temps de réverbération par avance, il est nécessaire d'effectuer l'estimation par récurrence sur plusieurs itérations, selon l'algorithme suivant :

1. On fixe une valeur initiale arbitraire TR_0 pour le temps de réverbération (exemple : 3 secondes)
2. Régression linéaire avec une marge de retrait $\Delta t_1 = 0,1 * TR_0$
3. Calcul de la première estimation TR_1 du temps de réverbération à partir de la pente de la régression
4. Retour à l'étape 2, avec la marge de retrait $\Delta t_2 = 0,1 * TR_1$
5. ...

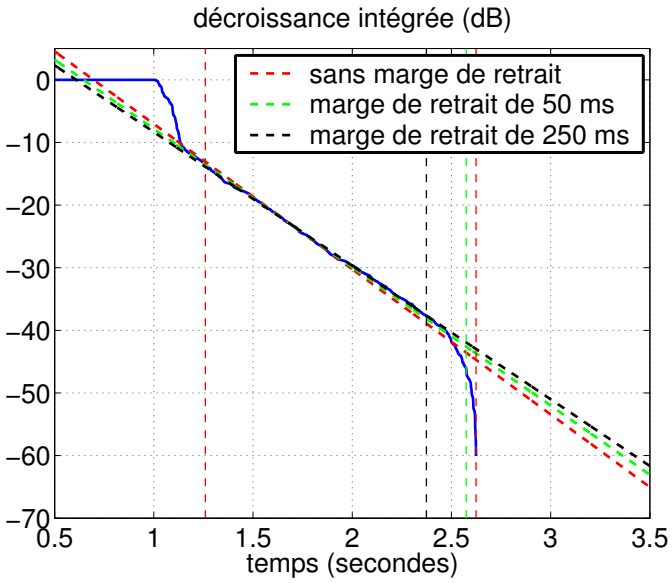


FIG. VII.4 – **Influence de la marge de retrait sur la régression** : la courbe en bleu est la décroissance intégrée, identique à celle de la figure VII.3, calculée pour la bande en tiers d'octave centrée sur 2 kHz. Trois régressions linéaires, représentées en traits interrompus rouges, verts et noirs, sont effectuées pour trois valeurs différentes de la marge de retrait, de 0 à 250 ms.

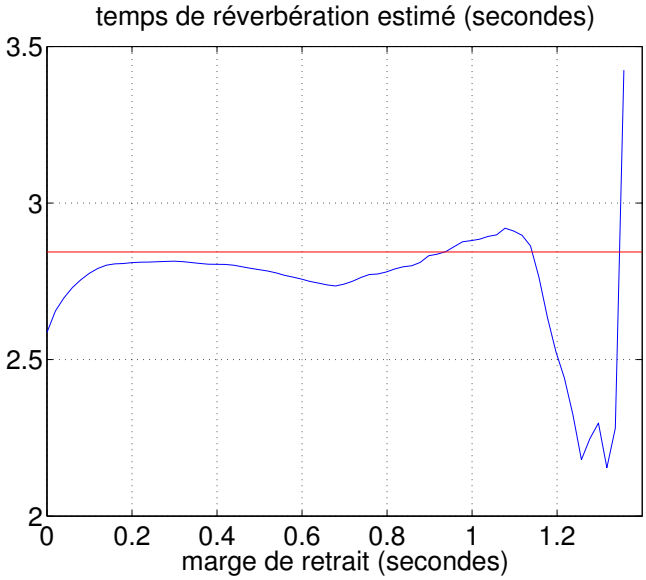


FIG. VII.5 – **Influence de la marge de retrait sur la qualité de l'estimation** : la courbe en bleu représente le temps de réverbération estimé par régression linéaire sur la décroissance de la figure VII.4, fonction de la marge de retrait temporelle par rapport à la fin du segment. La ligne rouge est le temps de réverbération estimé dans la même bande (en tiers d'octave centrée sur 2 kHz) à partir de la réponse impulsionnelle.

En pratique, deux itérations suffisent dans la majorité des cas pour stabiliser l'estimation du temps de réverbération.

3.3 De la difficulté d'estimer la durée de décroissance initiale

La méthode exposée ci-dessus vise à estimer le temps de réverbération, c'est-à-dire à caractériser la réverbération **tardive**. On peut se demander s'il est par extension possible de caractériser la réverbération **précoce**, au travers par exemple de la durée de décroissance initiale EDT_{10} ou EDT_{15}.

La figure VII.6 apporte une ébauche de réponse à cette question. Le signal observé est identique à celui étudié depuis le début de cette section, c'est-à-dire une séquence très courte de bruit blanc réverbérée. Si l'on applique la définition stricte de l'EDT_{10} à la décroissance calculée à partir du bruit, on surestime largement ce dernier : alors que la mesure à partir de la réponse impulsionnelle donne un résultat théorique de 0,72 secondes, l'estimation, qui vaut 2,4 secondes, est très proche du temps de réverbération dans cette bande (2,84 secondes)[2].

La raison de cette divergence réside dans la différence de nature entre les deux décroissances : en effet, on peut rappeler que l'intégration rétrograde du réflectogramme est égale à la moyenne statistique de la puissance instantanée (sans intégration rétrograde) enregistrée après extinction d'une séquence de bruit stationnaire. Ainsi, effectuer une intégration rétrograde de cette puissance instantanée correspond à effectuer une **double intégration rétrograde** sur le réflectogramme ! En pratique, si l'on n'altère pas la nature exponentielle de l'enveloppe de la réverbération tardive, on modifie largement la distribution de l'énergie précoce, notamment en adoucissant les discontinuités dues aux réflexions précoces.

Il est nécessaire de noter que ce problème de mauvaise estimation de l'EDT, qui est déjà présent lorsque l'extinction est immédiate comme c'est le cas ici, est accru lorsque la source fait silence progressivement, comme c'est le cas pour toutes les sources physiques comprenant un résonateur (comme la majorité des instruments de musique). En effet, dans ce cas, l'énergie décroissant plus lentement, l'EDT peut être sous- ou surestimé, selon l'instant auquel est fixé le début de l'intégration, et ce de manière difficilement prévisible.

Ainsi, même si l'on était capable de détecter avec précision l'instant exact auquel la source fait silence, il est difficile d'estimer correctement la durée de décroissance initiale à partir d'une intégration rétrograde effectuée sur des signaux non impulsionnels. Tout au plus, peut-on caractériser sa **tendance** par rapport au temps de réverbération : l'EDT est-il inférieur, du même ordre, ou supérieur au TR ?

Le temps de réverbération semble la durée caractéristique la plus stable, et nous portons désormais la plupart de nos efforts sur son estimation. Cela implique entre autres de choisir un seuil de détection relativement élevé, de manière à s'assurer que le début de la régression soit bien situé dans la réverbération tardive. Néanmoins, on verra sur l'étude des exemples en section 4 qu'il n'est pas toujours possible, lorsque le régime diffus tarde à s'installer, de fixer un seuil permettant de s'affranchir de l'influence des premières réflexions.

3.4 Influence de la bande passante

La caractérisation de la réverbération repose traditionnellement sur l'analyse de signaux à large bande passante, de manière à ce que toute l'information recherchée y soit contenue. L'analyse proprement dite est effectuée en bandes limitées (typiquement en octaves ou tiers d'octaves), car les indices recherchés dépendent de la fréquence, mais ces bandes ont une largeur suffisante pour qu'un grand nombre de modes y soient superposés, si bien que les méthodes statistiques restent valables.

Cependant, l'application des mêmes méthodes d'analyse à des signaux réels pose des problèmes dès lors que ces derniers sont à structure harmonique, ou plus généralement à spectre discret : il n'est en effet pas rare que dans cette situation, la bande de fréquences concernée ne contienne qu'un seul partiel, ou un nombre réduit de partiels proches les uns

[2] La même tendance, mais moins flagrante, peut être constatée pour l'indice EDT_{15}.

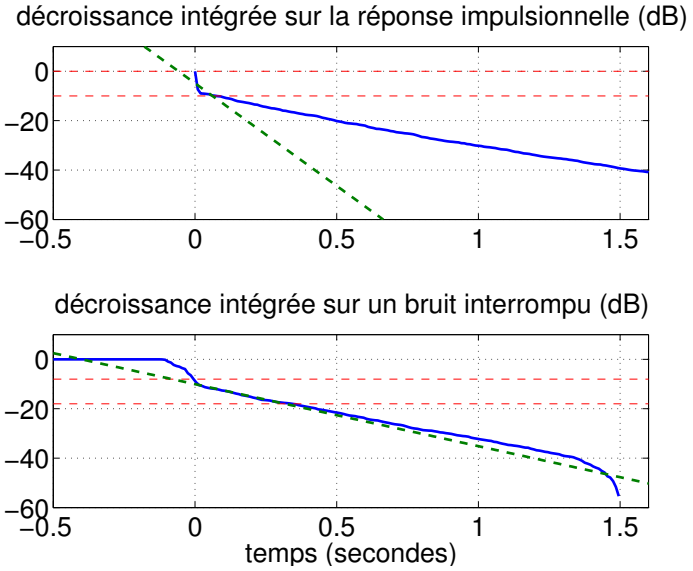


FIG. VII.6 – **De la difficulté d'estimer la durée de décroissance initiale** : sont représentées les décroissances intégrées calculées dans la bande en tiers d'octave centrée sur 2 kHz à partir de la voie gauche de la réponse impulsionnelle (haut), et de la voie gauche de la réponse à un bruit interrompu (bas). Pour ce dernier, l'origine temporelle est ramenée au moment où le signal source fait silence. Pour chaque graphe, les deux lignes horizontales déterminent la plage de calcul de l'EDT_{10}, c'est-à-dire la valeur de la décroissance à l'extinction, et après soustraction de 10 dB. La ligne oblique est le résultat de la régression linéaire sur cette plage.

des autres. Or, dans ce cas, il se peut que le nombre de modes excités par un tel signal ne soit pas suffisant pour que la réverbération puisse être considérée comme localement stationnaire, si bien que les calculs proposés en annexe C perdent de leur validité, et que la décroissance intégrée n'est plus proportionnelle à l'enveloppe exponentielle de décroissance, mais soumise à des fluctuations supplémentaires, dont l'amplitude dépend de la bande passante.

Cet aspect est illustré sur la figure VII.7 : on y observe aisément que les fluctuations de la décroissance intégrée par rapport à une exponentielle pure (c'est-à-dire une droite en représentation logarithmique) sont directement fonction du facteur de qualité de la bande passante des signaux considérés. On se heurte alors à un problème d'importance, qui est celui de la fiabilité de l'estimation du temps de réverbération sur de telles décroissances. La précision de l'estimation est donc liée d'une part à la bande passante des signaux observés, et d'autre part à la durée sur laquelle la régression linéaire est effectuée. Or les signaux musicaux sont très souvent à structure harmonique, et les silences y sont de durée relativement courte. On se place donc dans un cas d'emblée très défavorable. Une solution consiste à multiplier le nombre d'analyses au cours du temps et à en déduire une estimation moyenne pour chaque bande, puisque l'on peut raisonnablement supposer que l'estimation du temps de réverbération est en moyenne égale au temps de réverbération exact, si l'on se place à chaque fois dans la zone de réverbération tardive. Une manière supplémentaire de réduire ces fluctuations est d'avoir recours à des moyennes spatiales du temps de réverbération ou des décroissances intégrées.

3.5 Effet d'un bruit de fond additionnel

Comme dans une situation de mesure, il est nécessaire de tenir compte de l'inévitable présence de bruit de fond. Cependant, la situation est ici plus délicate, et ce pour deux raisons.

Premièrement, le bruit de fond est en proportion bien plus importante dans un enregistre-

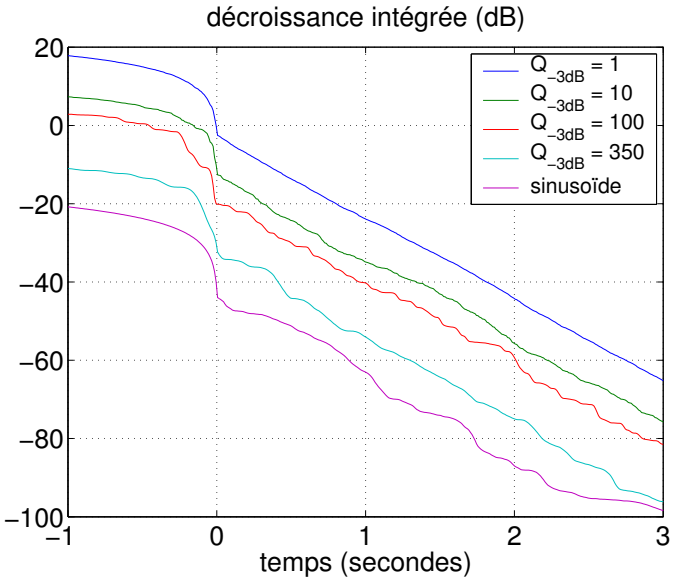


FIG. VII.7 – **Influence de la bande passante sur la linéarité de la décroissance** : est représentée l'allure de la décroissance intégrée (dans une zone où les effets de troncature sont négligeables) après un bruit interrompu brusquement à $t = 0$, et dont la bande passante est de largeur variable (largeur quantifiée par le facteur de qualité à -3 dB Q_{-3dB}), et centrée sur 2 kHz. On représente également le cas limite, qui est la décroissance suivant une sinusoïde pure brusquement interrompue.

ment brut que dans une réponse impulsionnelle mesurée, pour un rapport signal sur bruit acoustique équivalent, car la connaissance parfaite du signal source dans une situation de mesure permet, puisque le bruit de fond acoustique et électronique est en théorie décorrélé du signal, d'améliorer grandement le rapport signal sur bruit, et ce d'autant mieux que la durée du signal source est importante.

De plus, **il est d'autant plus difficile de distinguer le bruit de fond de la réverbération dans la décroissance intégrée que la durée d'intégration est faible** : cet aspect est mis en évidence sur la figure VII.8 : pour une durée d'intégration de 3 secondes et pour un rapport signal sur bruit supérieur à 25 dB, on peut distinguer dans la courbe deux portions correspondant aux zones de prédominance respectives de la réverbération et du bruit de fond, la première étant linéaire, et la seconde logarithmique. La transition entre les deux zones correspond à un brusque changement de pente de la courbe. En revanche, pour une durée d'intégration de 1,1 seconde, ce changement de pente n'est plus aussi marqué, et il devient très difficile de séparer les deux zones : **la courbe paraît toujours linéaire, alors que sa pente est modifiée par la présence de bruit de fond**.

Ce dernier aspect est très gênant, car il nous indique qu'il est alors impossible de juger du niveau du bruit de fond à partir de l'allure de la décroissance intégrée, alors qu'il a une influence notable sur l'estimation, comme l'atteste la figure VII.9. Or la méthode de détection proposée dans cette étude ne permet pas non plus de distinguer les zones pour lesquelles la réverbération prédomine sur le bruit de fond, car, comme cela a été mentionné au chapitre V, les deux présentent une faible corrélation intercanale. Il est donc nécessaire d'estimer la puissance du bruit de fond à un moment où l'on est assuré qu'il n'y a pas du tout de réverbération, comme par exemple au tout début de l'enregistrement, ou après un silence de plusieurs secondes sans événement sonore. Cela suppose bien sûr que le bruit de fond est stationnaire et conserve la même puissance tout au long de l'enregistrement. Une fois le bruit de fond estimé, on peut alors sélectionner les zones de réverbération valides, en éliminant celles pour lesquelles la dynamique entre le début de la réverbération et l'émergence du bruit de fond n'est pas assez importante.

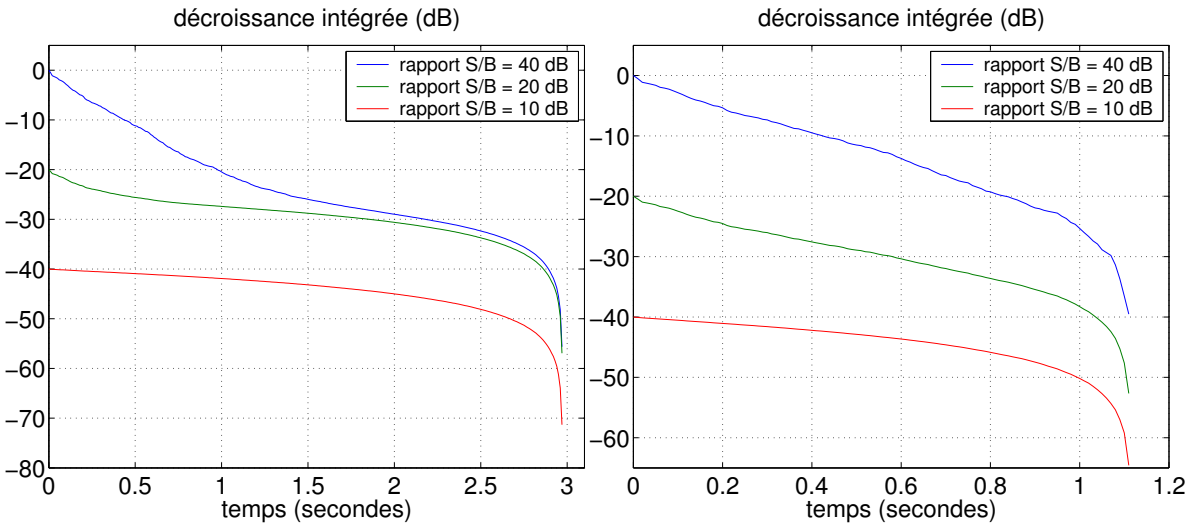


FIG. VII.8 – **Influence conjointe de la durée du segment et du bruit de fond sur l'allure de l'estimation** : les deux figures représentent les décroissances intégrées dans la bande en tiers d'octave centrée sur 2 kHz après un bruit interrompu, en présence d'un bruit de fond stationnaire additionnel en proportion variable. La figure de gauche correspond à une durée d'intégration de 3 secondes, et celle de droite à une durée de 1,1 seconde. Le rapport signal sur bruit est un rapport de puissances, la puissance du signal étant mesurée dans la portion stationnaire (avant l'interruption). Le rapport d'énergie directe/réverbérée est de 0 dB dans la bande considérée.

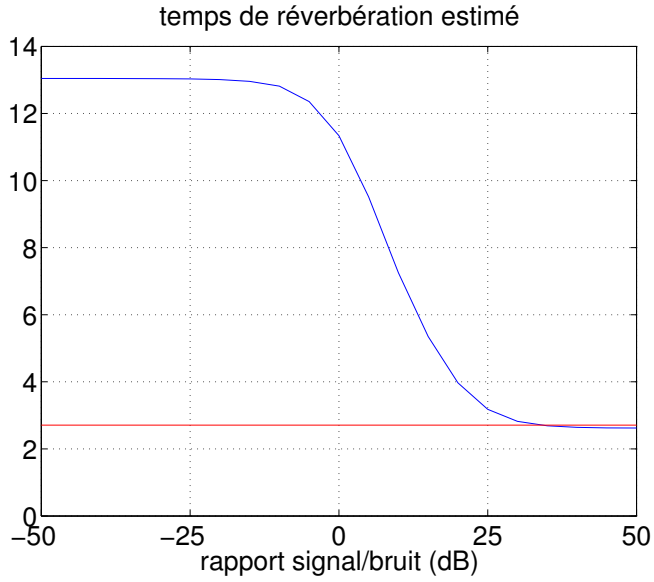


FIG. VII.9 – **Influence de la quantité de bruit de fond sur l'estimation du temps de réverbération** : est représentée ci-dessus l'estimation du temps de réverbération par régression linéaire en deux passes, en fonction du rapport signal sur bruit en puissance. Le signal source est encore une fois une séquence interrompue de bruit stationnaire en bande limitée. La ligne rouge est le temps de réverbération TR_{20} mesuré sur la réponse impulsionnelle.

3.6 Utilisation de moyennes spatiales

En situation de mesure, il est fréquent d'avoir recours à des moyennes spatiales des décroissances intégrées pour minimiser leurs fluctuations, ou bien directement des estimations du temps de réverbération, pour stabiliser ces dernières. Ceci suppose un niveau de diffusion suffisant pour que le temps de réverbération soit indépendant de la position, ce qui exclut en particulier le cas de décroissances multiples. De même, l'utilisation de moyennes spatiales n'est justifié que pour la réverbération **tardive**, et n'a aucun sens pour la description de la décroissance précoce.

On peut appliquer la même idée au cas présent, en s'intéressant à la moyenne des décroissances intégrées calculées sur les deux voies. L'hypothèse d'un niveau de diffusion suffisant n'est plus aussi stricte ici, puisque dans le cas d'enregistrements binauraux, les microphones sont suffisamment rapprochés pour que l'enveloppe de réverbération tardive soit considérée comme identique sur les deux voies.

Le schéma de principe complet de la méthode d'estimation du temps de réverbération pour un segment donné est représenté sur la figure VII.10. On y distingue la phase de détection qui permet de définir les bornes temporelles d'intégration, une phase de sélection destinée à éliminer les segments pour lesquels la dynamique de la décroissance est trop faible (sélection sur l'énergie) et ceux restant à un niveau de corrélation trop élevé (sélection sur l'indice de détection), puis finalement la phase d'estimation proprement dite par régression linéaire basée sur la décroissance intégrée sur chacune des deux voies, et sur la décroissance moyenne.

4 EXEMPLES

On présente ici deux exemples d'analyse, l'un d'une scène sonore virtuelle (c'est-à-dire issue de la spatialisation d'un signal source sec monophonique) avec aucun bruit de fond, et l'autre d'un enregistrement réel. Il s'agit dans les deux cas d'un enregistrement de flûte, cet instrument ayant été choisi préférentiellement dans le corpus à notre disposition, car ses caractéristiques permettent un rehaussement efficace du segment considéré, comme cela a été indiqué au chapitre V. Dans les deux cas, l'analyse est basée sur une transformation de Fourier à court-terme pitch-synchrone[3] avec un facteur de suréchantillonnage fréquentiel $M = 10$. Ce facteur est fixé à une valeur volontairement élevée pour deux raisons : premièrement, on adhère de fait aux conditions nous permettant d'utiliser la cohérence à court-terme en lieu et place de la corrélation à court-terme, bien plus coûteuse ; de plus, le fait de réduire la largeur des bandes d'analyse permet d'isoler d'autant mieux les notes les unes des autres.

4.1 Exemple d'analyse sur une scène sonore virtuelle

Le premier exemple analysé est celui de la section 2.4 du chapitre V : on rappelle qu'il s'agit du résultat de la convolution par une réponse binaurale de salle d'un extrait d'une interprétation de *la Partita* pour flûte seule de Bach. La réponse impulsionnelle fait partie du jeu de réponses étudiées tout au long du chapitre V. Elle a été mesurée dans une salle rectangulaire très réfléchissante d'un volume d'environ $650\ m^3$, sur un sujet humain se trouvant à 4 mètres de la source et en face d'elle (cette dernière se situant dans un coin de la pièce), ce qui implique, si l'on se réfère à la figure V.6, un rapport d'énergie directe/réverbérée de -15 dB à +10 dB en fonction de la fréquence et de la voie concernée.

Sur la figure VII.11 sont représentés les réflectogrammes ainsi que la cohérence à court-terme, cette dernière étant calculée par FFT sur 64 points, avec une constante de temps de 10 ms. On y montre notamment que la transition vers le régime diffus est très tardive, car on trouve encore des réflexions spéculaires à 190 ms (le temps de mélange est de l'ordre de 25 ms). Les valeurs du temps de réverbération RT_{30} et de la durée de décroissance initiale EDT_{15} mesurées en bandes de tiers d'octaves sur les réponses impulsionnelles sont données à la figure VII.12.

[3]Voir la section 2.3 du chapitre V pour plus de détails sur l'analyse pitch-synchrone.

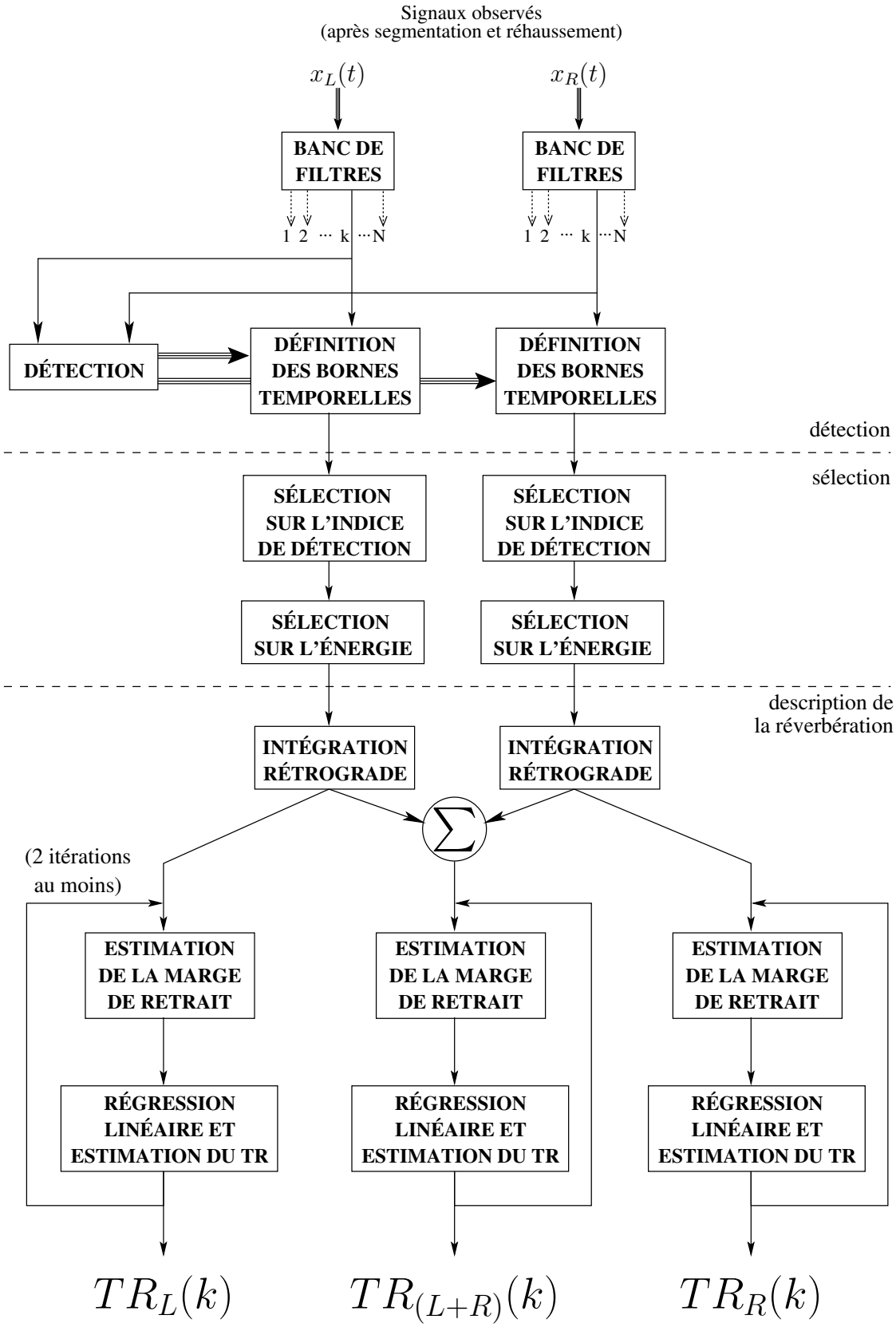


FIG. VII.10 – **Schéma de principe d'estimation du temps de réverbération** : l'algorithme fournit, pour chaque événement considéré et pour chaque bande de fréquences $k, k \in \{1...N\}$, une estimation du temps de réverbération à partir de la décroissance intégrée sur la voie gauche ($TR_L(k)$), de celle sur la voie droite ($TR_R(k)$), et de la somme des deux ($TR_{L+R}(k)$).

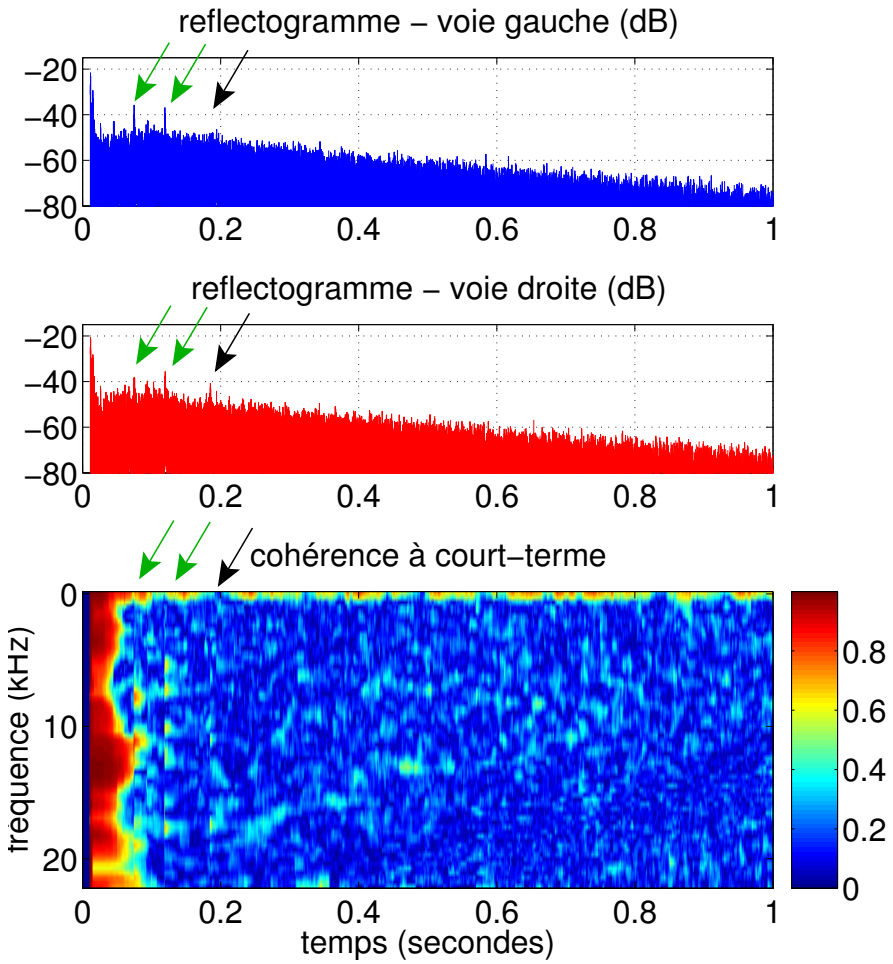


FIG. VII.11 – **Exemple d'analyse d'une scène sonore virtuelle - réflectogrammes et cohérence à court-terme** : les réflectogrammes permettent de visualiser les réflexions spéculaires les plus saillantes : en plus des toutes premières réflexions (provenant du sol et des murs les plus proches), on note deux impulsions plus tardives à 70 ms et 120 ms (flèches vertes), qui correspondent aux trajets aller et retour après réflexion sur les murs opposés. La cohérence à court-terme et l'observation du réflectogramme de la voie droite révèle une autre réflexion à 190 ms (flèche noire). Au-delà, l'absence d'informations cohérentes permet de considérer que la réverbération est diffuse.

La cohérence intercanale stationnaire est inférieure à 0,65 pour toutes les fréquences en dessous de 10 kHz, mais l'hypothèse de notes courtes[4] permet de fixer le seuil de détection à un niveau relativement faible, ici 10% dans un premier temps. La fenêtre d'analyse pour la détection est une fenêtre exponentielle avec une constante de temps de 100 ms.

La détection de fréquence fondamentale permet d'isoler, sur les 62 secondes que dure l'extrait, 329 événements, qui correspondent relativement bien aux notes de la partition, sauf lorsque le rythme est trop rapide et/ou les notes trop rapprochées en fréquences. Certaines notes et la plupart des ornements ne sont pas détectés, mais le seul effet néfaste est une diminution de la précision de la méthode de rehaussement spectral.

Pour chacun de ces événements, on applique un rehaussement par filtrage en peigne, puis la détection de réverbération par égalisation et annulation avec une fenêtre exponentielle de constante de temps $\Delta T = 100\ ms$, pour calculer les décroissances intégrées pour chacune des harmoniques jugées significatives, et l'on estime finalement trois valeurs du temps de

[4]Voir chapitre V.

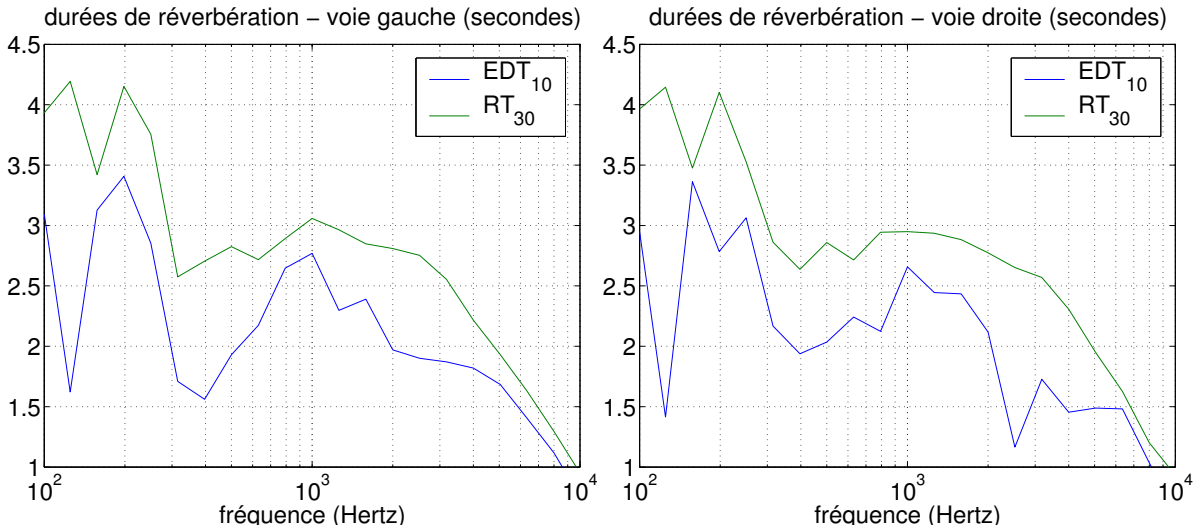


FIG. VII.12 – **Exemple d'analyse d'une scène sonore virtuelle - valeurs mesurées des durées de réverbération** : l'analyse est effectuée en bandes de tiers d'octave, pour chacune des voies de la réponse impulsionnelle employée.

réverbération pour la fréquence donnée : à partir de la voie gauche uniquement, de la voie droite uniquement et à partir de la moyenne des décroissances intégrées gauche et droite.

Les estimations sont alors regroupées, en fonction de la fréquence auxquelles elles ont été effectuées, en bandes de tiers d'octave, de manière d'une part à effectuer des moyennes statistiques, et d'autre part à comparer ces moyennes avec les valeurs mesurées des durées de réverbération. On peut observer sur la figure VII.13 le résultat de cette analyse : on y compare dans chaque bande l'estimation moyenne au RT_{30} et à l'EDT_{10} mesurés sur les réponses impulsionnelles, en indiquant à chaque fois les premiers et troisièmes quartiles, ainsi que les valeurs minimales et maximales des estimation

La première observation que l'on peut formuler est que dans la majorité des cas, l'estimation moyenne se situe entre l'EDT_{10} et le RT_{30}. Si la plage de variation des estimations au sein d'une même bande est parfois importante (1,3 seconde au maximum), celles-ci sont assez bien regroupées autour de la moyenne, comme l'atteste le faible écart entre le premier et le troisième quartile dans la plupart des cas. La précision de l'estimation, et le biais par rapport au RT_{30}, diminue généralement avec le nombre d'estimations individuelles.

On peut ensuite remarquer que compte tenu du nombre élevé de notes analysées, le nombre d'estimations (72 au total pour la voie gauche, et 60 pour la voie droite) est relativement réduit. En fait, la plupart des événements ont été jugés non viables pour l'estimation du temps de réverbération. La cause principale est que la plupart des segments restent trop cohérents d'une voie à l'autre, et l'indice de détection ne remonte pas à une valeur proche de son maximum, 0,5. Cette cohérence est due au fait que les silences entre deux notes sont souvent de trop courte durée pour que l'indice de détection ait le temps de remonter jusqu'à sa valeur en champ diffus ; la technique de rehaussement permet en théorie d'effacer une ou plusieurs des notes suivantes, mais en pratique, la proportion de la partie inharmonique des signaux sources (qui est principalement due au souffle) est assez importante, si bien que l'annulation par filtrage en peigne n'est pas toujours de bonne qualité.

Ainsi, alors que l'on cherche à caractériser la réverbération tardive uniquement, on obtient des estimations qui se basent en partie sur la décroissance précoce. Ceci s'explique par le fait qu'à cause de la présence de réflexions secondaires assez marquées, la transition vers le régime diffus est très tardive, comme cela a été mentionné ci-dessus, et comme le montre la figure VII.11. L'adéquation au temps de réverbération dépend alors étroitement de la valeur du seuil de détection : si celui-ci est faible, la régression linéaire commence très tôt, et l'intégration de réflexions précoces dans le calcul fausse l'estimation du temps

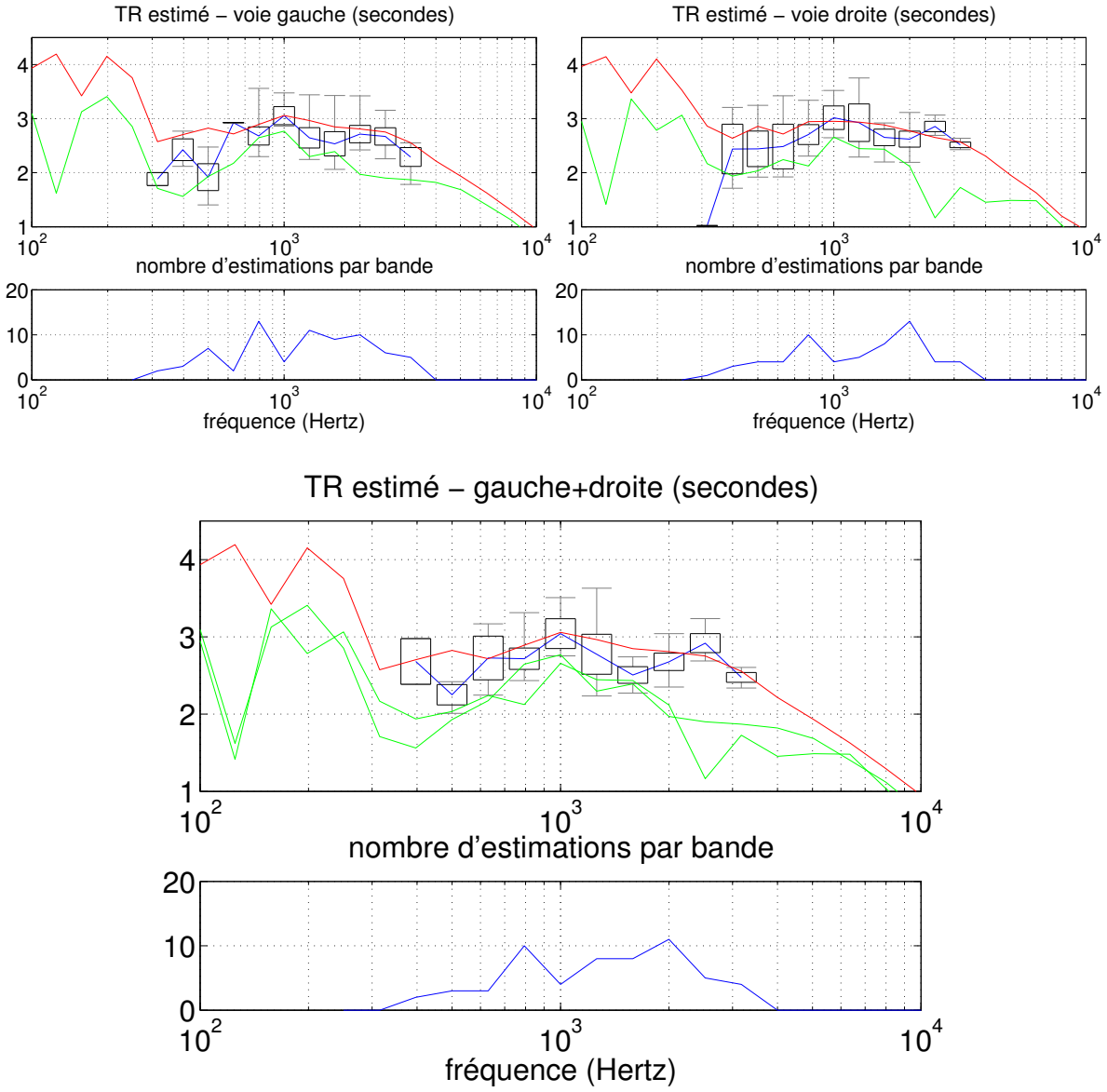


FIG. VII.13 – **Exemple d'analyse d'une scène sonore virtuelle - estimation du temps de réverbération** : l'estimation est effectuée pour la voie gauche seule, la voie droite seule, et à partir des moyennes sur les deux voies des décroissances intégrées. Pour chaque figure, sont représentés sur la partie supérieure la moyenne par bande des estimations (en bleu), les premier et troisième quartiles (indiqués par les rectangles), les valeurs minimales et maximales des estimations (traits gris), et les valeurs mesurées de l'EDT_{10} et du RT_{30}(en vert et rouge, respectivement). Sur la partie inférieure est indiqué le nombre total d'estimations par bandes. Le seuil de détection est dans ce cas à 10 %.

de réverbération (en le sous-estimant dans le cas présent) ; si le seuil est plus élevé, on s'affranchit de l'influence des premières réflexions, mais en contrepartie, on diminue la durée de la régression, et donc sa précision.

Pour information, on représente sur la figure VII.14 le résultat de la même analyse, mais avec un seuil beaucoup plus faible, soit 2 %. Si cela permet d'augmenter sensiblement le nombre d'estimations jugées valides (192 pour la voie gauche, et 183 pour la voie droite), l'estimation moyenne fournie se rapproche plus de la durée de décroissance initiale que du temps de réverbération. L'adéquation avec l'EDT_{10} est d'ailleurs assez remarquable pour plusieurs bandes (notamment entre 600 Hz et 1 kHz), mais si l'on se souvient des réserves émises en section 3.3, on comprend qu'il vaut mieux rester prudent quant à une conclusion plus générale.

4.2 Exemple d'analyse sur un enregistrement *in situ*

Dans ce deuxième exemple, on analyse du point de vue de la réverbération l'enregistrement déjà étudié en section 4 du chapitre VI où l'on à chercher à estimer la direction de la source. Il s'agit donc d'un extrait de la seconde fantaisie de Telemann en la mineur pour flûte seule, enregistré dans l'Espace de Projection de l'IRCAM, dont les murs ont été mis en configuration partiellement diffuse, sauf une partie du mur sud qui lui était parfaitement réfléchissant. Un extrait un peu plus long de l'enregistrement a été analysé dans ce cas (38 secondes au lieu de 20), de manière à disposer d'un nombre suffisant d'estimations. Par ailleurs, des réponses impulsionnelles ont été mesurées, pour la même position du sujet, avec une enceinte Tannoy 600 placée à la même position que l'était l'embouchure de la flûte pendant l'enregistrement. Le signal source utilisé pour la mesure est un sinus glissant, car il permet de minimiser, par rapport à des séquences de Golay, l'effet de petites variations de la propagation dues aux mouvements de tête. Les réflectogrammes correspondants et la cohérence à court-terme sont représentés en figure VII.15.

La réponse de salle à une allure assez différente que dans le cas précédent : bien que l'on peut encore noter la présence de deux réflexions spéculaires tardives (à 110 ms et 150 ms), celles-ci sont de faible énergie, les parois étant en grande partie diffuses. Leur effet sur la cohérence et sur la décroissance intégrée est assez faible. Les deux réflexions ayant une réelle incidence sur la distribution en énergie et sur la cohérence sont la réflexion sur le sol intervenant 4 ms après l'onde directe, et celle sur le mur sud (visible principalement sur la voie droite), arrivant 23 ms après l'onde directe.

Une manière supplémentaire de juger du caractère diffus de cet espace est de comparer les durées de réverbération précoce et tardive pour chacune des deux voies de l'enregistrement, ce qui est fait en figure VII.16. On note que si l'EDT_{15} et le RT_{20} sont très semblables sur la voie gauche, ils diffèrent de plusieurs dixièmes de millisecondes, en fonction de la fréquence, sur la voie droite[5]. Ceci confirme le rôle majeur joué par la réflexion sur le mur sud : en effet, celle-ci est principalement visible sur la voie droite, car elle est contralatérale vis-à-vis de l'oreille gauche, et est donc largement atténuée.

La détection est effectuée avec une constante de temps de la fenêtre est de 100 ms, et le seuil de détection est fixé à 5%. Compte tenu du fait que certaines des notes sont de longue durée, on s'écarte des hypothèses de travail élaborées au chapitre V, et il sera donc difficile d'effectuer la détection sur ces quelques événements. Cependant, il y a suffisamment de notes de courte durée dans l'enregistrement pour contrebalancer ce problème. Le nombre total d'estimations valides reste néanmoins très faible (25 pour la voie gauche, 30 pour la voie droite).

Sur la figure VII.17 est indiqué le résultat de l'analyse. On observe avant tout que l'estimation moyenne est en meilleure adéquation avec le temps de réverbération que dans le cas précédent, et ce quelle que soit la voie considérée (gauche, droite, ou somme des deux). Ceci s'explique en partie par l'usage du RT_{20} comme indice "cible" au lieu du RT_{30}, mais cette raison est insuffisante, car elle ne permet pas de rendre compte de la bonne stabilité de l'estimation d'une voie à l'autre : puisque la durée de décroissance initiale est très inférieure

[5]Le RT_{20} a dû être employé dans cette analyse à la place du RT_{30}, car le bruit de fond est trop important dans les réponses impulsionnelles pour assurer la dynamique nécessaire de 35 dB sur la décroissance.

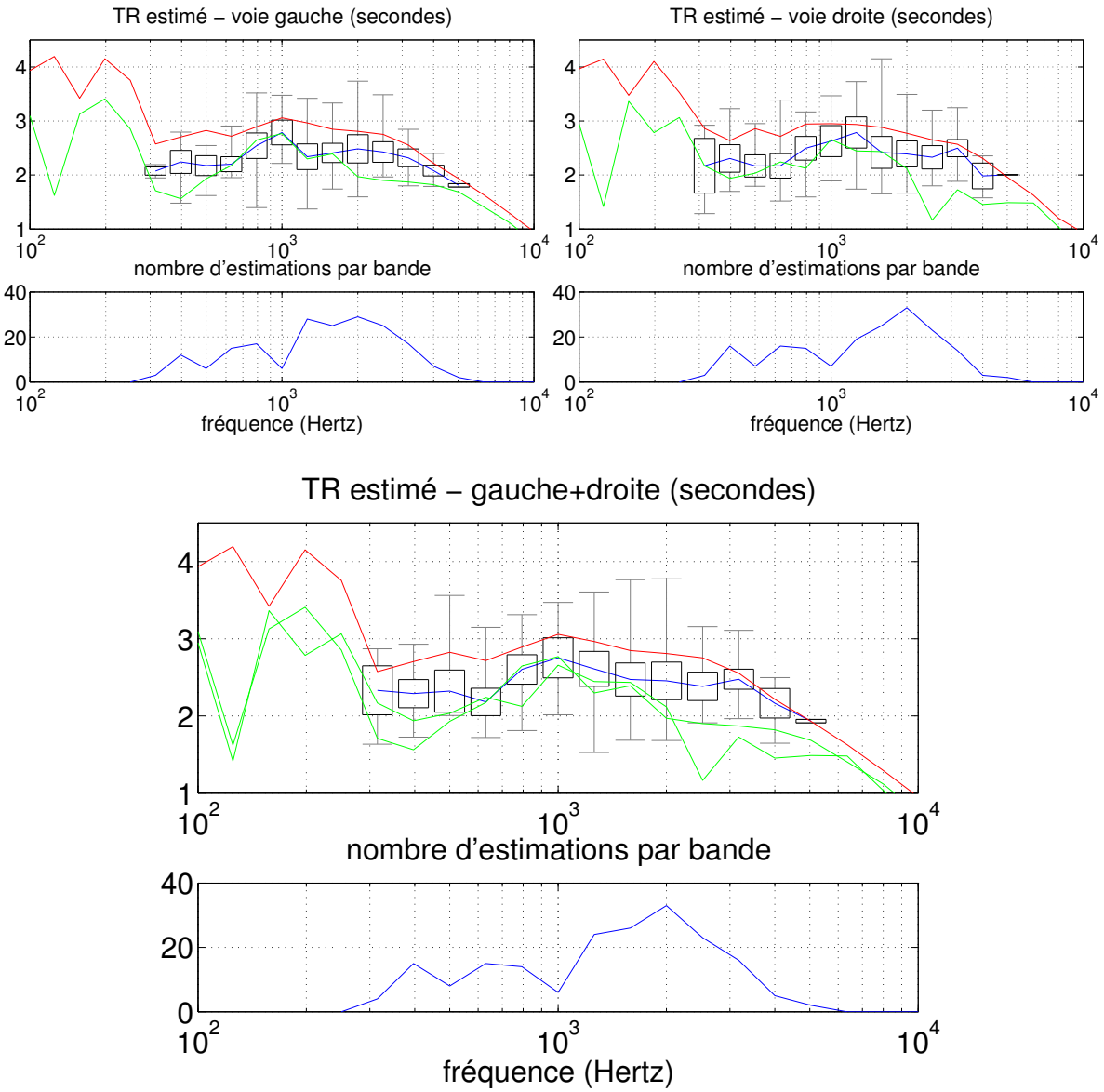


FIG. VII.14 – **Exemple d'analyse d'une scène sonore virtuelle - estimation du temps de réverbération** avec seuil de détection à 2% : se reporter à la figure VII.13 pour la légende.

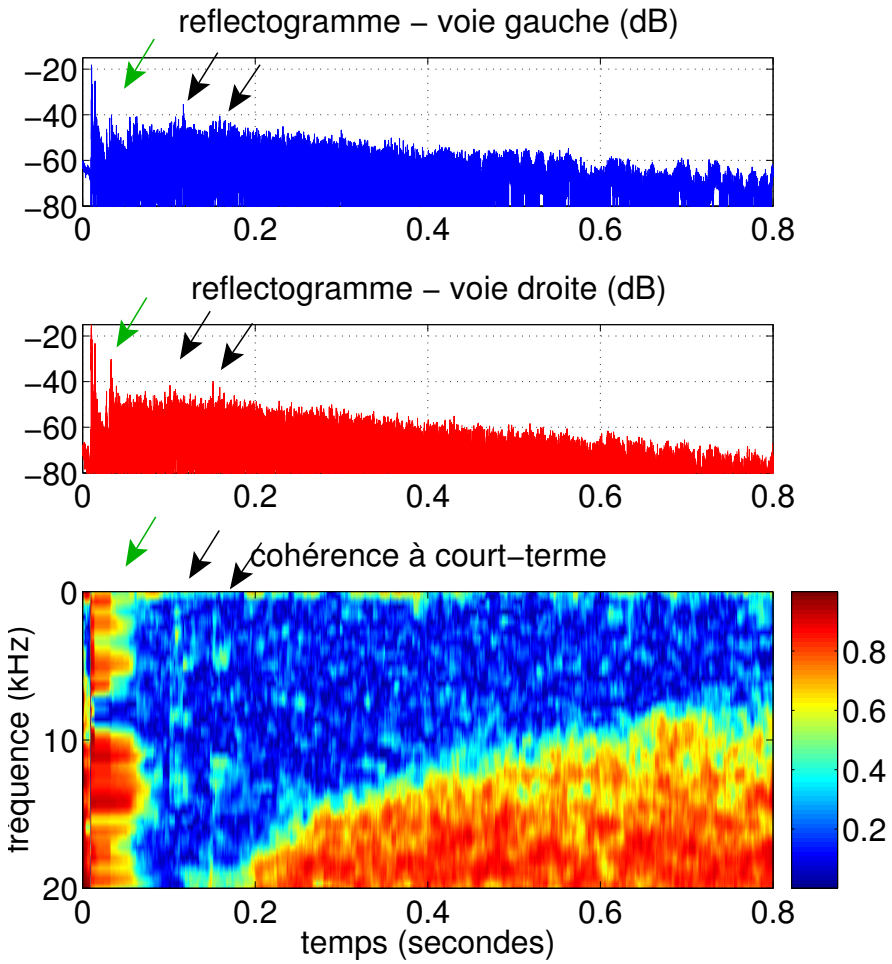


FIG. VII.15 – **Exemple d'analyse d'une scène sonore *in situ* - réflectogrammes et cohérence à court-terme de la réponse impulsionnelle** : en plus de la toute première réflexion provenant du sol, on note à 23 ms une forte réflexion provenant du mur sud (flèches vertes), ainsi que deux réflexions de plus faible énergie à 100 ms et 140 ms (flèches noires). On rappelle que la zone de forte cohérence en hautes fréquences débutant à t=200 ms correspond à l'émergence du bruit de fond.

à droite à gauche, on devrait s'attendre à ce que l'estimation du temps de réverbération soit sous-estimée à droite, ce qui n'est pas le cas. Cette stabilité est attribuable au fait que l'indice de détection est suffisamment élevé dans la situation donnée pour que la régression linéaire soit affranchie de l'influence des premières réflexions. Elle justifie *a posteriori* l'utilisation de moyennes sur les deux voies, qui n'auraient aucun sens si les estimations étaient sensibles à la distribution précoce de l'énergie, puisque celle-ci n'est dans ce cas pas identique d'une voie sur l'autre.

On peut néanmoins noter plusieurs problèmes de l'estimation, dues principalement au nombre limité d'estimations : lorsque le nombre d'estimations par bande est vraiment trop faible (3 ou moins), la notion de moyenne perd de son sens, et il est à craindre que l'estimation soit faussée : c'est le cas par exemple à gauche et à droite pour les bandes centrées sur 314 Hz et 2,52 kHz, ou encore à gauche pour la bande centrée sur 1 kHz. Si le nombre d'estimations augmente, la moyenne se stabilise, mais les estimations ne sont toujours pas assez nombreuses pour que les statistiques soient significatives :ainsi, pour la bande centrée sur 1,26 kHz, la différence entre le premier et le troisième quartile reste importante (de l'ordre de 0,5 secondes) quel que soit la voie considérée.

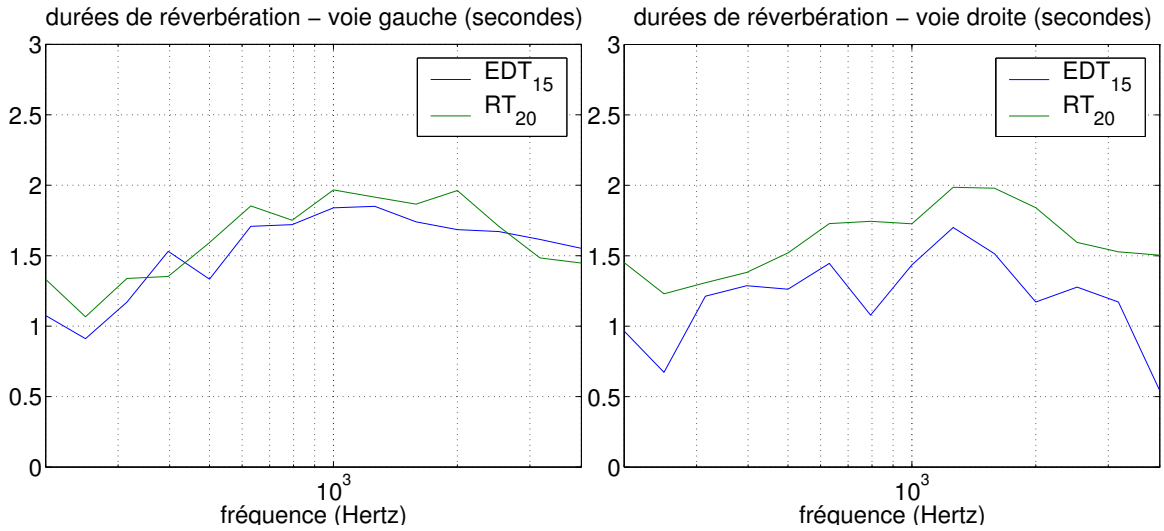


FIG. VII.16 – **Exemple d'analyse d'une scène sonore *in situ* - valeurs mesurées des durées de réverbération** : l'analyse est effectuée en bandes de tiers d'octave, pour chacune des voies de la réponse impulsionnelle employée.

5 CONCLUSION

Dans ce chapitre a été étudiée une méthode complète de description temporelle de l'enveloppe de réverbération à partir de signaux musicaux. Elle repose sur deux principes : premièrement, la description elle-même est basée sur une technique éprouvée d'estimation des durées de réverbération, qui est la régression linéaire sur la décroissance intégrée, en l'appliquant à des signaux autres que des réponses impulsionnelles. Cette généralisation nécessite des précautions supplémentaires, notamment quant à la validité de la définition des durées de décroissance initiale.

Le second principe à la base de la méthode de description proposée est le couplage avec la technique de détection de réverbération mise en place aux chapitres IV et V, de manière à déterminer à quels moments effectuer les estimations. On est ainsi en mesure de pouvoir juger dans chaque bande si seule la queue de réverbération est présente, ou si d'autres événements cohérents s'y superposent, et on peut donc assurer une estimation sur des segments où seule la réverbération est présente.

Par rapport à une analyse sur des réponses impulsionnelles, les causes possibles de biais (c'est-à-dire la présence de bruit de fond, ou de fluctuations si l'analyse est effectuée en bandes étroites) sont décuplées, étant donnés les signaux sources considérés (souvent harmoniques) et les conditions d'enregistrement. D'autre part, l'analyse sur des signaux musicaux entraîne une cause supplémentaire de biais spécifique, qui est la durée limitée de l'intégration et de la régression linéaire. Pour toutes ces raisons , il est nécessaire de disposer d'un grand nombre d'estimations pour stabiliser les statistiques, ce qui nécessite un analyse sur une longue durée. L'adéquation de l'estimation avec la durée de réverbération précoce ou tardive dépend de nombreux facteurs, certains étant intrinsèques à la situation étudiée (vitesse de transition vers le champ diffus, rapport d'énergie directe/réverbérée), et d'autres liés aux paramètres d'analyse, comme la valeur du seuil de détection.

Pour compléter la description, il faudrait maintenant s'intéresser plus étroitement aux aspects **spatiaux** de la réverbération, c'est-à-dire le niveau de diffusion, qui peut être mis en relation avec la sensation perceptive d'enveloppement par la salle. Ceci signifie étudier plus précisément l'évolution de la corrélation/cohérence intercanale au cours de la réverbération, au delà de la simple notion de détection par seuillage. D'autre part, il serait utile d'effectuer une estimation du temps de réverbération en basses fréquences, alors que dans l'état actuel, la détection, qui ne fonctionne pas dans ce domaine, l'interdit. Une solution pourrait

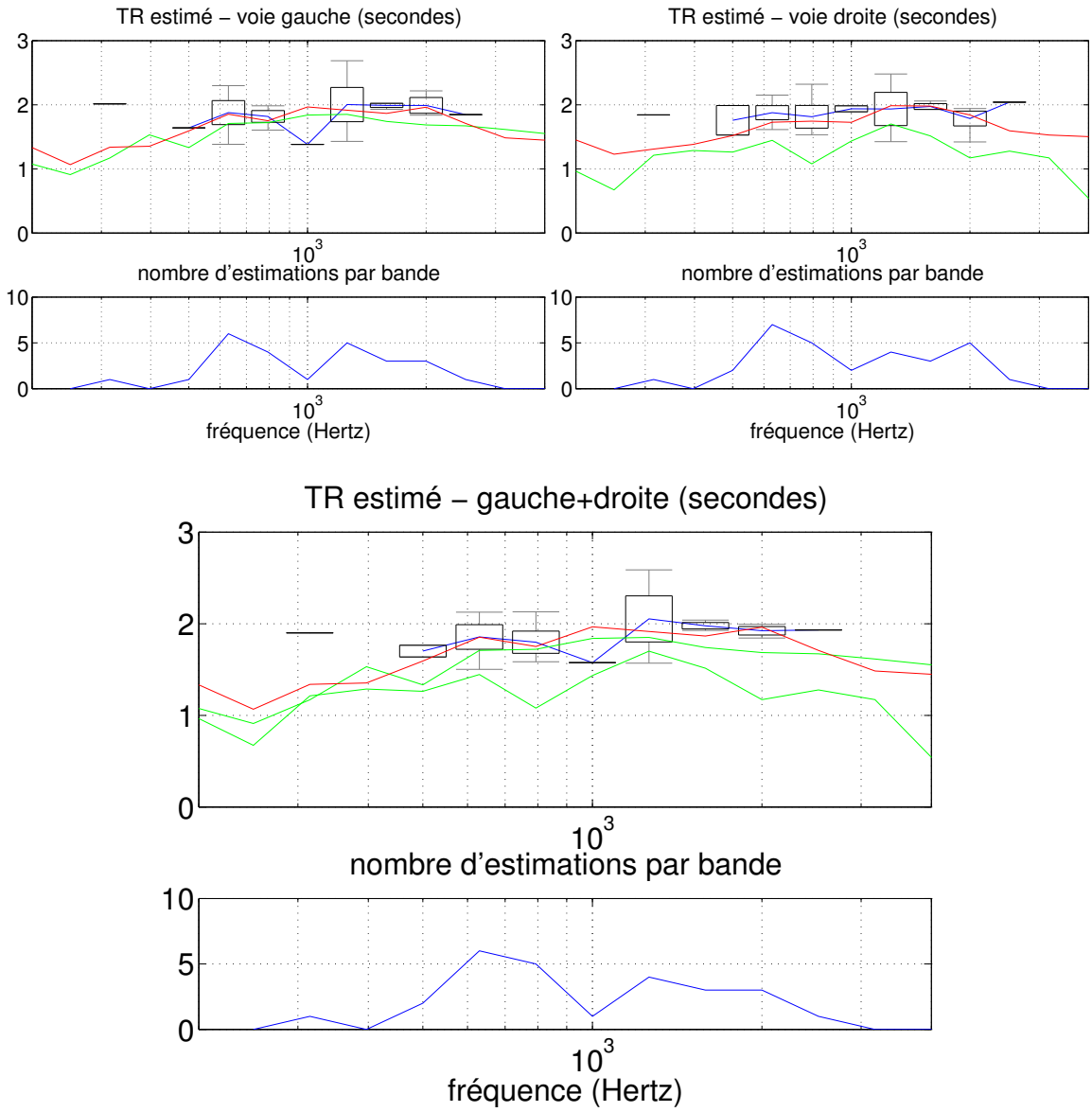


FIG. VII.17 – **Exemple d'analyse d'un enregistrement *in situ* - estimation du temps de réverbération** avec seuil de détection à 5% : se reporter à la figure VII.13 pour la légende.

consister à extrapoler en basses fréquences la valeur des bornes temporelles calculées en moyennes fréquences, puis effectuer l'estimation de manière usuelle.

ANNEXE : ESTIMATION À PARTIR DE LA PUISSANCE À COURT-TERME

Au cours de ce chapitre, il a pu être souligné à plusieurs reprises l'intérêt de caractériser la décroissance non pas à partir de la puissance instantanée, mais de son intégration rétrograde : celle-ci atténue largement les fluctuations, tout en conservant l'allure exponentielle de l'enveloppe de réverbération tardive. L'anticausalité, induite par le fait que l'intégration soit effectuée dans le sens rétrograde, permet de s'assurer que la décroissance dans la partie tardive ne dépend pas de la distribution des premières réflexions, si bien qu'elle conserve en théorie son caractère exponentiel en toutes circonstances. En pratique, ce dernier aspect est parfois mis en défaut par l'inévitable présence de bruit, et par l'effet de la troncature si celle-ci intervient trop tôt, ce qui est systématiquement le cas ici.

On peut s'interroger à titre prospectif sur l'utilisation d'autres représentations pour estimer le temps de réverbération. Puisque le principe de l'intégration de la puissance instantanée semble parfaitement adéquat, on peut proposer, en lieu et place de l'intégration rétrograde, la **puissance à court-terme** définie au chapitre IV. Celle-ci présente plusieurs atouts qui permettent d'envisager son emploi : premièrement, celle-ci est déjà calculée, puisqu'elle fait partie des indices nécessaires à la détection par égalisation et annulation ; de plus, on peut espérer que son lien étroit avec l'erreur de détection permette de définir les seuils de détection de manière moins arbitraire qu'avec l'intégration rétrograde ; finalement, il s'agit d'une représentation causale, ce qui permet d'envisager son emploi dans le cadre d'une application à temps réel.

L'aspect causal de cette représentation est pour l'estimation proprement dite à la fois un atout et un défaut. Son avantage principal est que l'**on s'affranchit du biais dû à la troncature** en fin de segment : la décroissance exponentielle calculée par puissance à court-terme est tronquée au même moment que le signal lui-même, sans présenter cette divergence caractéristique des décroissances calculées par intégration rétrograde. En revanche, et conformément à la remarque énoncée ci-dessus, on intègre, au moins au début de la décroissance, l'énergie relative aux réflexions spéculaires, ce qui écarte la décroissance de l'allure exponentielle, non seulement pendant la période où la réverbération n'est pas encore diffuse (comme c'est déjà le cas pour l'intégration rétrograde), mais aussi après, le temps que les réflexions ne soient plus prises en compte par l'intégration à court-terme.

La figure VII.18 permet de se rendre compte visuellement des différences entre les deux représentations. Dans cet exemple, les durées sur lesquelles sont finalement effectuées les régressions sont équivalentes (environ 1,2 seconde dans les deux cas), mais il est visible que la puissance à court-terme est plus étroitement liée à l'indice de détection que ne l'est l'intégration rétrograde : notamment, le segment temporel, défini par les lignes verticales rouges, où doit être effectué la régression, correspond à la période pendant laquelle l'indice de détection est à son palier supérieur : le début de ce palier correspond en effet au moment où les réflexions passées ne sont plus prises en compte dans l'intégration à court-terme, et la fin au moment où la source est à nouveau active.

On peut envisager de réduire la durée de la fenêtre d'intégration pour rallonger la durée de régression, mais on se retrouve alors face au dilemme mentionné en section 1.3 du chapitre V : si l'on diminue la durée de la fenêtre, le détecteur réagit plus rapidement, mais les fluctuations sont plus importantes. Il avait alors été suggéré d'employer une **fenêtre à durée variable**, celle-ci étant courte lorsque l'indice de détection est faible (et que l'on désire donc une forte réactivité aux extinctions), et longue lorsque l'indice de détection est élevé (de manière à minimiser les fluctuations).

Rien ne s'oppose dans le principe à cette idée, si ce n'est le respect d'une contrainte essentielle sur la durée de la fenêtre : en effet, on montre en section 5 de l'annexe C qu'**il est nécessaire que la durée de la fenêtre soit largement inférieure au quinzième du temps de réverbération** pour s'assurer que la décroissance conserve son caractère exponentiel. Cette condition est contraignante en particulier dans les salles petites et moyennes faiblement réverbérantes, et pour les hautes fréquences, pour lesquelles le temps de réverbération est généralement faible.

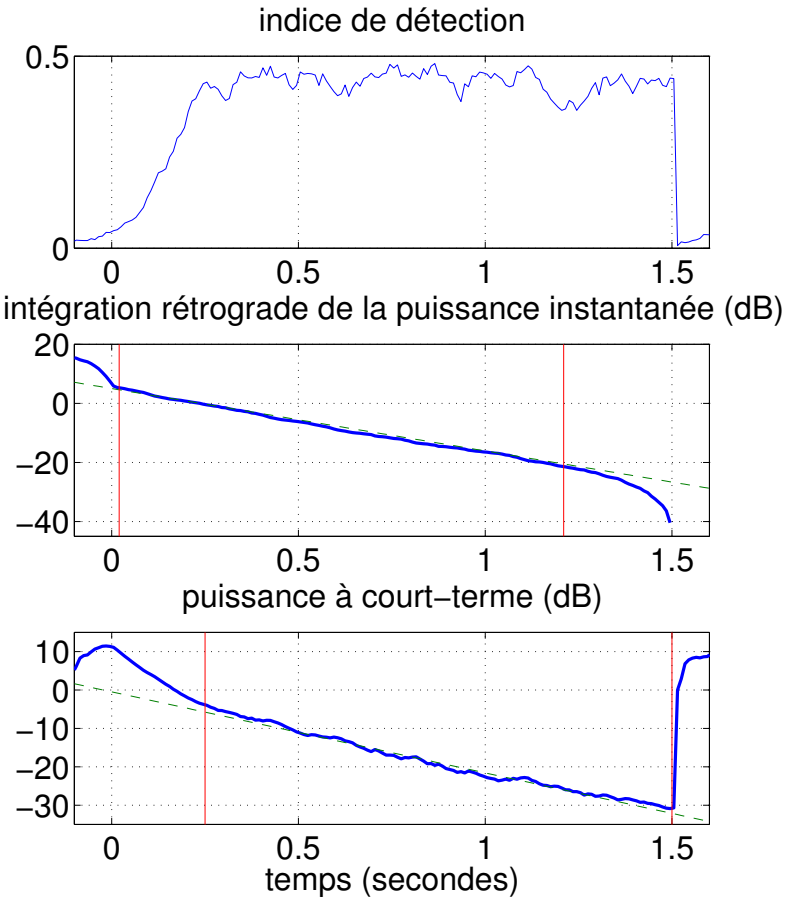


FIG. VII.18 – **Estimation du temps de réverbération à partir de la puissance à court-terme** : on reprend l'exemple de la succession de deux courtes bouffées de bruit de la section 3.1, en représentant dans la bande en tiers d'octave centrée sur 2 kHz l'indice de détection, l'intégration rétrograde pour la voie gauche, et la puissance à court-terme, toujours pour la voie gauche. La ligne verte en trait interrompu correspond à la décroissance idéale, c'est-à-dire que sa pente est définie par le temps de réverbération mesuré sur la réponse impulsionnelle. Les barres verticales représentent les limites temporelles de la régression linéaire, qui diffèrent pour l'intégration rétrograde et pour la puissance instantanée. Les puissances et corrélations à court-terme sont calculées avec une fenêtre exponentielle de constante de temps $\Delta T = 50\ ms$

Conclusion

Conclusion

Au cours de cet exposé, ont été proposées un certain nombre de méthodes de description spatiale de scènes sonores simples, reposant principalement sur un nombre restreint d'outils de bas-niveau basés sur des statistiques à court-terme du deuxième ordre. Plusieurs études ont visé à proposer des méthodes de description spatiale d'une scène sonore, mais la plupart d'entre elles se concentrent sur la localisation horizontale en situation anéchoïque. La recherche décrite ici tente d'aller un peu plus loin, en étudiant le cas des espaces réverbérants, et en étendant la description à un modèle plus complet de localisation, couplé à une méthode d'estimation de la durée de réverbération à partir de signaux musicaux.

Ces méthodes sont justifiées par un cadre théorique cohérent, qui permet entre autres la définition d'un modèle simple de relations entre les voies des signaux observés en présence d'une source active. L'un des intérêts majeurs de l'adéquation entre ce modèle et les méthodes proposées est de pouvoir juger à tout moment de l'erreur de modélisation, erreur qui est à la base du principe de détection de source.

Un autre des apports de cette étude est d'intégrer de manière formelle à ces méthodes la possibilité d'ajuster la résolution en temps et en fréquence de l'analyse, et d'en étudier l'effet sur la précision de la détection et de l'estimation. Cette capacité d'ajustement permet également de rappeler les liens étroits qui unissent les statistiques du deuxième ordre dans le domaine temporel (corrélation) et spectral (cohérence), et ainsi d'unifier plusieurs méthodes connues de détection en champ réverbéré. Cette unification est d'un grand intérêt d'un point de vue pratique : en effet, les calculs à base de corrélation à court-terme, bien qu'optimisés dans le cas de fenêtres rectangulaires ou exponentielles, restent très lourds, et sont décuplés par le nombre de bandes sur lesquelles porte l'analyse (sans compter le coût de calcul du banc de filtres lui-même). Or, la cohérence à court-terme peut être programmée de manière très efficace, notamment au moyen de transformées de Fourier rapides, et son calcul peut même être envisagé en temps réel.

Une des difficultés du problème est que si l'analyse des relations intercanales instantanées permet de juger de la présence d'une source active localisable, elle n'est pas directement adaptée au problème de détection de champ réverbéré ; tout au plus peut-on attester de l'absence d'une source active suffisamment proche pour être localisable, mais il est impossible sur cette seule information d'en déduire si la scène sonore est silencieuse, si elle ne contient que du bruit de fond, un champ réverbéré, ou encore une source trop lointaine pour être localisable. Il a paru nécessaire pour pallier ce manquement de compléter le cadre théorique, en incluant notamment la notion de stationnarité du bruit de fond et de causalité entre activité d'une source et réverbération tardive.

Nous nous sommes concentrés au cours de ce travail sur des enregistrements binauraux, mais en gardant à l'esprit l'application éventuelle des méthodes proposées à d'autres configurations de prise de son. Il serait maintenant utile de vérifier en pratique si cette généralisation est valide, en étudiant par exemple des enregistrements effectués au moyen de couples stéréophoniques, coïncidents ou non, ou encore des systèmes de prise de son à plus de deux canaux. Quelle que soit la configuration de prise de son utilisée, on doit garder à l'esprit qu'il est nécessaire d'en avoir une connaissance la plus complète possible pour pouvoir effectuer la description, en particulier pour pouvoir estimer précisément la direction

de provenance d'une onde sonore donnée. En effet, la précision spatiale est conditionnée par l'exhaustivité de cette connaissance. Dans le cas le plus défavorable pour lequel seules quelques informations sur la nature et la distance entre les microphones sont disponibles, la détection, et donc l'estimation du temps de réverbération, sont toujours efficaces, mais il faut se contenter d'une estimation très vague de la direction de la source.

Quelle que soit la technique de prise de son considérée, on se heurte aux mêmes problèmes en termes d'echelle de précision spatiale que pour des enregistrements binauraux : plus la configuration du système de prise de son est connue de manière exhaustive (directivité des microphones, position respective, géométrie d'un éventuel obstacle), plus la localisation est précise. Ainsi, il paraît difficile de lever l'ambiguïté de localisation au sein d'un cône de confusion sur des enregistrements binauraux, si l'on ne connaît que trop grossièrement la tête et le torse considérés. Quoi qu'il en soit, il est nécessaire de disposer d'une information minimale sur le type de directivité des microphones et leur position respective, sans quoi toute localisation est impossible. En revanche, la qualité de la détection n'est elle a priori pas affectée par le niveau de connaissance du système de prise de son : il est toujours possible de juger de la validité du modèle gain-retard, et donc de la présence d'une source unique (ou de plusieurs dans le cas de configurations multicanales), mais pas forcément de localiser cette dernière.

Le fait de disposer de plusieurs canaux donne plus de latitude, et autorise la description de scènes sonores plus complexes, ou une description plus approfondie de scènes simples. En effet, puisque le nombre de sources détectables augmente avec le nombre de canaux, on peut choisir soit d'améliorer la précision spatiale de l'estimation de la direction d'une source unique, soit de décrire avec la même précision spatiale le cas de plusieurs sources mélangées, soit enfin décrire de manière plus complète le cas d'une source unique : chaque réflexion pouvant être associée à une source image, il est envisageable d'estimer la direction de provenance non seulement de l'onde directe, mais également d'une ou plusieurs ondes réfléchies.

Annexes

A

Conventions de notations de coordonnées en prise de son binaurale

On DÉFINIT ici les axes, plans et angles utilisés dans le cadre d'une prise de son binaurale, qui correspondent aux conventions usuelles. Notamment, on rappelle les deux systèmes sphériques les plus utilisés, l'un prenant pour axe polaire l'axe vertical, et l'autre l'axe interaural.

1 DÉFINITION DES PLANS STANDARD

La figure A.1 indique les plans caractéristiques utilisés dans l'étude de l'écoute binaurale, et par extension dans les recherches sur les techniques de synthèse binaurale :

- Le **plan médian** est le plan vertical médiateur du segment joignant les deux oreilles.
- Le **plan horizontal** est le plan horizontal qui contient le segment joignant les deux oreilles.
- Le **plan frontal** est le plan vertical qui contient le segment joignant les deux oreilles

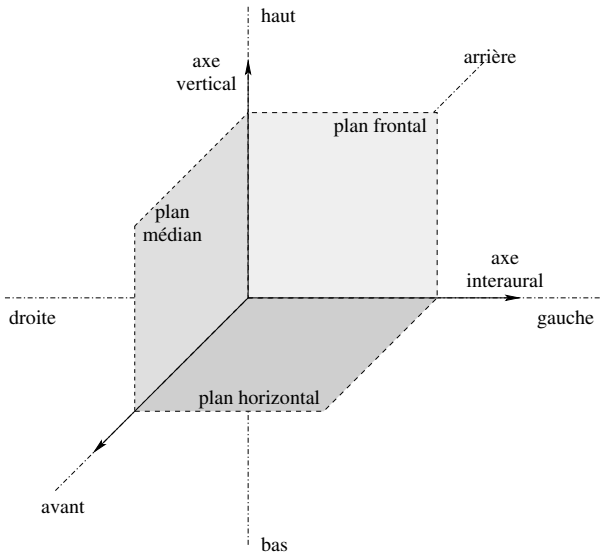


FIG. A.1 – **définition des plans standard**

2 CONVENTIONS SUR LES ANGLES

Il existe deux repères sphériques standard relatifs à l'écoute binaurale permettant de préciser la position d'une source dans l'espace. Ces deux repères ont pour origine le milieu du segment qui joint les deux oreilles. Ils diffèrent par l'axe polaire choisi :

1. Le **système de coordonnées d'axe polaire vertical** (figure A.2), choisi notamment par Gardner et Martin (1994) pour les mesures de HRTF sur la tête KEMAR, repère la direction d'une source par les angles suivants :
   - la *latéralisation* est quantifiée par l'azimut θ, qui est la latitude (repérée par rapport au plan médian)
   - l'*élévation* est quantifiée par le site φ, qui est la longitude (par rapport au plan horizontal).

2. Le **système de coordonnées d'axe polaire interaural** (figure A.3) est utilisé par exemple par le centre Cipic de l'Université de Californie/Davis (Algazi et al., 2001b). La direction d'une source y est repérée par les angles suivants :
   - la *latéralisation* est quantifiée par l'angle β, qui est la longitude (par rapport au plan médian). Il s'agit du complémentaire de l'angle par rapport à l'axe interaural.
   - l'*élévation* est quantifiée par l'angle δ, qui est la latitude (repérée par rapport au plan horizontal).

   En première approximation, on peut considérer que les cônes de confusion sont les demi-cônes d'axe interaural, c'est-à-dire les surfaces d'angle de latéralisation β constant.

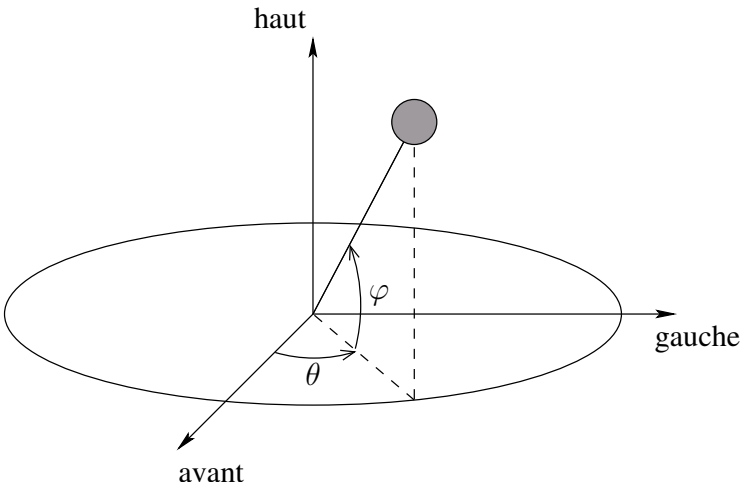


FIG. A.2 – **Système de coordonnées d'axe polaire vertical** : la latéralisation de la source est quantifiée par l'angle θ, appelé *azimut*, et l'élévation par le site φ .

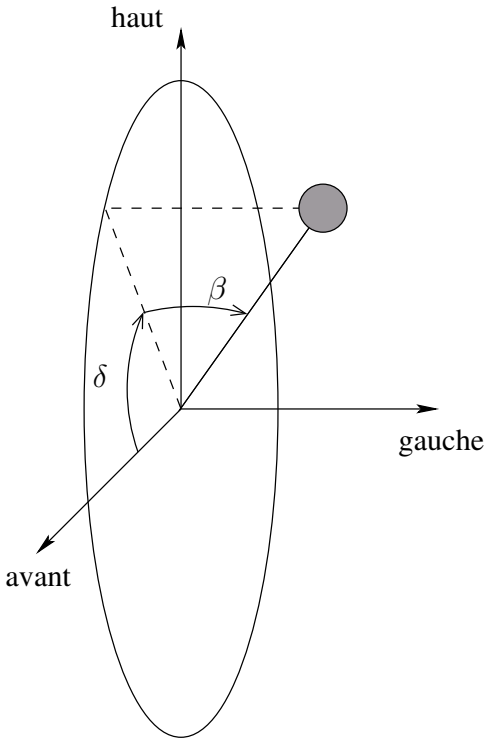


FIG. A.3 – **Système de coordonnées d'axe polaire interaural** : la latéralisation de la source est quantifiée par la longitude β, et l'élévation par la latitude δ.

B

Calculs relatifs à la méthode de détection et estimation par égalisation et annulation

DANS LES PAGES QUI SUIVENT, sont développés des calculs dont les résultats sont discutés au sein des chapitres IV et VI. La section 1 traite de la robustesse de la détection et de l'estimation en présence d'un bruit stationnaire, et la section 2 justifie le parallèle fait au chapitre VI entre la méthode d'estimation de la direction présentée dans cette étude et les méthodes basées sur une distance quadratique.

1 ÉGALISATION ET ANNULATION EN PRÉSENCE DE BRUIT

Les calculs proposés ici visent à prédire les valeurs de l'indice de détection et des paramètres optimaux d'égalisation et annulation, en fonction du niveau de bruit. On met notamment en évidence qu'en fonction des caractéristiques de ce dernier, l'estimation du gain et du retard sont plus ou moins biaisées.

1.1 Formalisation du problème

On suppose que les signaux observés obéissent au système IV.2, rappelé ici :

$$\begin{cases} x\left[n\right] = s\left[n\right] + b_x\left[n\right] \\ y\left[n\right] = A.s\left[n-\Delta\right] + b_y\left[n\right] \end{cases}$$

Dans cette écriture, $s[n]$ désigne le signal émis par la source physique, et $b_x\left[n\right]$ et $b_y\left[n\right]$ correspondent au bruit de fond acoustique et électrique. On suppose pour tous les calculs à venir que ces bruits sont stationnaires, d'énergies respectives E_{b_x} et E_{b_y}, et tous deux décorrélés du signal source $s[n]$.

On ne considère que l'égalisation et annulation à **long-terme**, et donc que les définitions à long-terme des puissances et corrélations. Cependant, le principe reste identique lorsque les calculs sont à court-terme, et également lorsque l'analyse est menée en bande étroite.

On cherche les deux couples qui minimisent respectivement l'erreur absolue et normalisée en présence de bruit. Ce calcul est mené dans deux cas : dans un premier temps, on suppose que les deux voies du bruit $b_x\left[n\right]$ et $b_y\left[n\right]$ sont décorrélées entre elles, et dans un second temps on s'intéresse à l'effet d'une éventuelle corrélation sur l'estimation.

1.2 Détection et estimation lorsque les deux voies du bruit sont décorrélées

Détection par minimisation de l'erreur absolue

Le système IV.7 indique que le couple (α_1, τ_1) qui minimise l'erreur absolue est défini par :
- le retard τ_1 qui maximise l'intercorrélation $C_{xy}\left[\tau\right]$ des deux voies du signal observé
- le gain $\alpha_1 = \frac{C_{xy}[\tau_1]}{E_x}$, E_x étant l'énergie du signal $x[n]$.

Si les bruits sont décorrélés, l'intercorrélation vaut simplement :

$$C_{xy}\left[\tau\right] = A.C_{ss}\left[\tau - \Delta\right]$$

Donc son maximum est obtenu pour $\tau_1 = \Delta$, et vaut $C_{xy}\left[\tau_1\right] = A.E_s$. D'autre part, les énergies valent :

$$E_x = E_s + E_{b_x}$$

et

$$E_y = A^2.E_s + E_{b_y}$$

Donc le couple (α_1, τ_1) vaut :

$$\begin{cases} \tau_1 = \Delta \\ \alpha_1 = A.\frac{E_s}{E_s + E_{b_x}} \end{cases}$$

Si l'on définit les rapports signal sur bruit en énergie sur les deux voies par :

$$\begin{cases} SNR_x = \frac{E_s}{E_{b_x}} \\ SNR_y = \frac{A^2.E_s}{E_{b_y}} \end{cases} \tag{B.1}$$

... on trouve que le couple (α_1, τ_1) vaut :

$$\begin{cases} \tau_1 = \Delta \\ \alpha_1 = A.\frac{1}{1+1/SNR_x} \end{cases}$$

Détection par minimisation de l'erreur normalisée

Le système IV.10 indique que si le couple (α_2, τ_2) minimisant l'erreur normalisée est défini par :

– le retard τ_2 qui maximise l'intercorrélation $C_{xy}[\tau]$ des deux voies du signal observé

– le gain $\alpha_2 = \sqrt{\frac{E_y}{E_x}}$, E_x étant l'énergie du signal $x[n]$.

Il est inutile de refaire le calcul pour τ_2, qui par définition est égal à τ_1. En revanche, le gain optimal vaut maintenant :

$$\alpha_2 = \sqrt{\frac{A^2.E_s + E_{b_y}}{E_s + E_{b_x}}}$$

En utilisant les rapports signal sur bruit définis ci-dessus, on trouve :

$$\alpha_2 = A.\sqrt{\frac{1 + 1/SNR_y}{1 + 1/SNR_x}}$$

Valeur de l'indice de détection

L'indice de détection est défini comme le minimum de l'erreur normalisée. Il vaut donc, en vertu de l'équation IV.11 :

$$\varepsilon_{min} = \varepsilon_{\alpha_2,\tau_2} = \frac{1}{2}\left(1 - \frac{1}{\sqrt{(1 + 1/SNR_x)\,.\,(1 + 1/SNR_y)}}\right)$$

On met donc en évidence le fait qu'en présence de bruit, même décorrélée d'une voie à l'autre, l'indice de détection dépend du rapport signal sur bruit. D'autre part, quelle que soit la méthode de minimisation, l'estimation du gain est biaisée.

En revanche, d'une part l'estimation du retard n'est pas affectée, et d'autre part l'indice de détection en l'absence de signal vaut $\frac{1}{2}$, quelle que soit la puissance du bruit (il suffit de faire tendre SNR_x et SNR_y vers zéro dans la formule précédente pour s'en convaincre).

1.3 Détection et estimation lorsque les deux voies du bruit sont corrélées

Deux causes physiques lors de l'enregistrement sont susceptibles de faire monter le niveau de corrélation du bruit :

– puisque le bruit est partiellement dû à la réverbération des sources antérieures, il n'est forcément plus décorrélé dans les très basses fréquences. En effet, il a été rappellé au chapitre I que même un champ idéalement diffus possède toujours une forte cohérence spatiale en basses fréquences, la fréquence de coupure étant fonction de l'écartement des microphones

– comme cela a été mentionné au chapitre I, le bruit d'origine électromagnétique est quasiment identique d'une voie à l'autre puisque les liaisons électriques des deux voies sont la plupart du temps collées, ou du moins très proches. Si l'on ne prend pas de précautions pour s'en affranchir (par exemple, au moyen de liaisons symétriques), le bruit est forcément corrélé d'une voie à l'autre

Le principe reste le même, mais il faut exprimer à nouveau l'intercorrélation, qui vaut dans ce cas :

$$C_{xy}[\tau] = A.C_{ss}[\tau - \Delta] + C_{b_x b_y}[\tau]$$

... $C_{b_x b_y}[\tau]$ étant l'intercorrélation des deux voies du bruit. Si l'on fait intervenir les intercorrélations normalisées et les rapports signal sur bruit données en B.1, on trouve :

$$C_{xy}[\tau] = A.E_s.\left(\rho_{ss}[\tau - \Delta] + \frac{1}{\sqrt{SNR_x.SNR_y}}.\rho_{b_x b_y}[\tau]\right)$$

Les expressions des énergies ne sont elles pas affectées par une corrélation du bruit, et restent donc identiques à celles données en section précédente.

Pour poursuivre plus en avant le calcul, il serait nécessaire de connaître plus précisément la nature des signaux considérés et du bruit. On peut néanmoins indiquer que **le retard τ_0 qui maximise l'intercorrélation n'est, dans le cas général, plus égal à Δ !** L'ampleur du biais dépend étroitement de deux causes :

- le rapport signal sur bruit sur chacune des voies
- le support de l'intercorrélation du bruit $\rho_{b_x b_y}[\tau]$ sur l'axe des retards. Si celui-ci contient le retard $\tau = \Delta$, il y a toutes les chances pour que le maximum de la corrélation soit déplacé.

De plus, l'indice de détection en l'absence de signal est lui aussi affecté. En effet, on trouve dans ce cas que l'erreur normalisée vaut :

$$\varepsilon_{min}^{bruit\ seul} = \frac{1}{2}\left(1 - \left|\rho_{b_x b_y}\right|\right)$$

En fait, si les deux voies du bruit sont corrélées, on peut quasiment considérer celui-ci comme un deuxième signal source, puisque du point de vue du signal observé, **on est incapable de faire la différence entre le signal issu d'une source physique réelle et celui issu d'une source de type électromagnétique ou, en très basses fréquences, du champ diffus** : tous ces types de signaux sont corrélés d'une voie à l'autre, et font donc baisser l'indice de détection.

2 DÉVIATION DE L'ERREUR NORMALISÉE AU VOISINAGE DU MINIMUM

On s'intéresse ici au comportement de l'erreur normalisée au voisinage du minimum, et notamment, à sa sensiblité aux variables de retard et de gain. Les calculs sont encore un fois menés dans un formalisme à long-terme, mais sont transposables à court-terme.

On s'intéresse donc à l'erreur normalisée d'égalisation et annulation appliquée à deux signaux $x[n]$ et $y[n]$. Celle-ci s'écrit, de manière générale (équation IV.9) :

$$\varepsilon_{\alpha,\tau} = \frac{1}{2} - \frac{\alpha C_{xy}[\tau]}{E_y + \alpha^2 E_x}$$

On suppose pour ce calcul que l'intercorrélation $C_{xy}[\tau]$ est définie comme fonction **continue** de τ. Ceci revient, lorsque les signaux sont numériques comme c'est le cas ici, à effectuer une conversion numérique-analogique à partir de l'intercorrélation calculée pour des valeurs discrètes du retard. Le but ici n'est cependant pas de proposer une méthode pratique. On cherche uniquement à mettre en évidence de manière théorique le comportement de l'erreur normalisée au voisinage de son minimum.

Soit le minimum (α_2, τ_2) de l'erreur normalisée. Il est défini, on le rappelle (équation IV.10), par :

$$\left\{\begin{array}{l} \tau_2 = arg\left\{\begin{array}{c} max \\ \tau \end{array}\ \left(C_{xy}(\tau)\right)\right\} \\ \alpha_2 = sgn\left(C_{xy}(\tau_2)\right).\sqrt{\frac{E_y}{E_x}} \end{array}\right. \tag{B.2}$$

On effectue un développement limité d'ordre 2 autour de ce minimum :

$$\begin{aligned} \varepsilon_{\alpha,\tau} = \varepsilon_{\alpha_2,\tau_2} &+ (\alpha-\alpha_2)\left.\frac{\partial \varepsilon_{\alpha,\tau}}{\partial \alpha}\right|_{\alpha_2,\tau_2} + (\tau-\tau_2)\left.\frac{\partial \varepsilon_{\alpha,\tau}}{\partial \tau}\right|_{\alpha_2,\tau_2} \\ &+\frac{1}{2}(\alpha-\alpha_2)^2\left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \alpha^2}\right|_{\alpha_2,\tau_2} + \frac{1}{2}(\tau-\tau_2)^2\left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \tau^2}\right|_{\alpha_2,\tau_2} \\ &+o\left((\alpha-\alpha_2)^3, (\tau-\tau_2)^3\right) \end{aligned}$$

Les dérivées partielles d'ordre 1 s'écrivent :

$$\left\{\begin{array}{l} \left.\frac{\partial \varepsilon_{\alpha,\tau}}{\partial \alpha}\right|_{\alpha,\tau} = \frac{\left(\alpha^2.E_x - E_y\right).C_{xy}[\tau]}{\left(E_y+\alpha^2 E_x\right)^2} \\ \left.\frac{\partial \varepsilon_{\alpha,\tau}}{\partial \tau}\right|_{\alpha,\tau} = -\frac{\alpha}{E_y+\alpha^2 E_x}.\frac{dC_{xy}}{d\tau}[\tau] \end{array}\right.$$

Si l'on applique les formules données en IV.10, on trouve qu'elles sont toutes les deux nulles en (α_2, τ_2). Les dérivées partielles d'ordre 2 s'écrivent :

$$\begin{cases} \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \alpha^2}\right|_{\alpha,\tau} = \frac{2.\alpha.E_x.C_{xy}[\tau].\left(3.E_y - \alpha^2 E_x\right)}{\left(E_y + \alpha^2 E_x\right)^3} \\ \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \tau^2}\right|_{\alpha,\tau} = -\frac{\alpha}{E_y + \alpha^2 E_x} . \frac{d^2 C_{xy}}{d\tau^2}[\tau] \end{cases}$$

Elles valent donc en (α_2, τ_2) :

$$\begin{cases} \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \alpha^2}\right|_{\alpha_2,\tau_2} = \frac{\sqrt{E_x.E_y}.C_{xy}[\tau_2]}{2(E_y)^2} \\ \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \tau^2}\right|_{\alpha_2,\tau_2} = -\frac{1}{2.\sqrt{E_x.E_y}} . \frac{d^2 C_{xy}}{d\tau^2}[\tau_2] \end{cases}$$

En faisant intervenir l'intercorrélation normalisée, on trouve finalement :

$$\begin{cases} \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \alpha^2}\right|_{\alpha_2,\tau_2} = \frac{E_x}{2.E_y}.\rho_{xy}[\tau_2] \\ \left.\frac{\partial^2 \varepsilon_{\alpha,\tau}}{\partial \tau^2}\right|_{\alpha_2,\tau_2} = -\frac{1}{2}\frac{d^2 \rho_{xy}}{d\tau^2}[\tau_2] \end{cases}$$

Donc le développement limité s'écrit :

$$\varepsilon_{\alpha,\tau} = \varepsilon_{min} + \frac{1}{4}.\rho_{xy}[\tau_2].\frac{(\alpha-\alpha_2)^2}{(\alpha_2)^2} - \frac{1}{4}.\frac{d^2\rho_{xy}}{d\tau^2}[\tau_2].(\tau-\tau_2)^2 + o\left((\alpha-\alpha_2)^3, (\tau-\tau_2)^3\right)$$

Au voisinage de la solution optimale, le déviation de l'erreur normalisée est donc une fonction quadratique des déviations en gain $(\alpha - \alpha_{min2})$ et en retard $(\tau - \Delta)$. Il s'agit des gains linéaires, mais une formulation équivalente peut-être trouvée pour les gains en décibels :

On désigne par α^{dB} et α_2^{dB} les valeurs en décibels des gains α et α_{min2} :

$$\alpha^{dB} = 20.log_{10}(\alpha)$$

...et

$$\alpha_2^{dB} = 20.log_{10}(\alpha_2)$$

Dans ce cas :

$$\begin{aligned} \frac{(\alpha-\alpha_2)^2}{(\alpha_2)^2} &= \left(\frac{\alpha}{\alpha_2} - 1\right)^2 \\ &= \left(10^{\frac{1}{20}\left(\alpha^{dB} - \alpha_{min2}^{dB}\right)} - 1\right)^2 \\ &= \left(e^{\frac{ln(10)}{20}\left(\alpha^{dB} - \alpha_{min2}^{dB}\right)} - 1\right)^2 \end{aligned}$$

En effectuant un développement limité de cette expression autour de $\alpha = \alpha_2$, on trouve :

$$\frac{(\alpha-\alpha_2)^2}{(\alpha_2)^2} \simeq \left(\frac{ln(10)}{20}\right)^2 \left(\alpha^{dB} - \alpha_{min2}^{dB}\right)^2$$

Donc la déviation de l'erreur s'écrit, au voisinage du minimum :

$$\varepsilon'_{\alpha^{dB},\tau} - \varepsilon'_{\alpha_2^{dB},\tau_{min}} \simeq \frac{1}{4}.\rho_{xy}[\tau_2].\left(\frac{ln(10)}{20}\right)^2.\left(\alpha^{dB} - \alpha_{min2}^{dB}\right)^2 - \frac{1}{4}.\frac{d^2\rho_{xy}}{d\tau^2}[\tau_2].(\tau-\tau_2)^2 \qquad \text{(B.3)}$$

Dans cette expression, $\varepsilon'_{\alpha^{dB},\tau}$ désigne l'erreur normalisée calculée pour un gain en décibels :

$$\varepsilon_{\alpha,\tau} = \varepsilon'_{20.log_{10}(\alpha),\tau}$$

C

Calculs de décroissances sur un modèle localement stationnaire de réverbération tardive

LES CALCULS PROPOSÉS ICI se réfèrent au chapitre VII. À travers un modèle très simple de réverbération tardive (section 1), on rappelle le bien fondé de l'usage du principe d'intégration rétrograde sur des réponses impulsionnelles 2 et on le justifie pour des signaux quelconques 3. L'effet de la troncature est étudié de manière quantitative en section 4. Finalement, nous nous intéressons en section 5 à la décroissance calculée non plus par intégration rétrograde, mais grâce à la puissance à court-terme.

1 MODÉLISATION À BANDE ÉTROITE DE LA RÉVERBÉRATION TARDIVE

Le modèle considéré ici découle de celui proposé par Polack (1988), et rappelé en section 2.2 du chapitre I. Il consiste à considérer la réverbération tardive comme une superposition de processus à bande limitée, aléatoires, localement stationnaires et localement ergodiques[1]. On considère une bande suffisamment limitée pour que le temps de réverbération puisse être supposé indépendant de la fréquence. Dans le cadre de ces hypothèses, la réponse impulsionnelle correspondant à la réverbération tardive s'écrit à temps discret sous la forme :

$$h_{rev}[n] = b[n].env[n] \tag{C.1}$$

, $b[n]$ désignant un processus aléatoire, stationnaire et ergodique d'autocorrélation $R_b[m]$, et $env[n]$ l'enveloppe (déterministe) à variations lentes, définie par :

$$env[n] = e^{-\alpha n}$$

...où $\alpha = \frac{3.ln(10)}{F_e.T_r}$ est la constante de décroissance (à temps discret), F_e étant la fréquence d'échantillonage, et T_r le temps de réverbération dans la bande considérée. Il s'agit d'une modélisation évidemment très simplifiée, et qui n'est pas valable pour la partie précoce de l'effet de salle. Pour qu'il soit néanmoins applicable à des situations réelles, on choisit de découper artificiellement la réponse impulsionnelle en deux segments jointifs. Si $h[n]$ désigne la réponse impulsionnelle complète, et si il existe un indice N_{mix} (noté ainsi par analogie avec le temps de mélange T_{mix}) à partir duquel les hypothèses ci-dessus sont valables, alors la réponse impulsionnelle peut s'écrire :

$$h[n] = \begin{cases} 0 \; si \; n < 0 \\ h_{spec}[n] \; si \; 0 \leq n \leq N_{mix} \\ h_{rev}[n] \; si \; n > N_{mix} \end{cases} \tag{C.2}$$

...$h_{spec}[n]$ désignant la première partie de la réponse, dominée par les réflexions spéculaires, et $h_{rev}[n]$ désignant un processus de réverbération tardive du type défini précédemment.

Les calculs qui suivent reposent largement sur l'hypothèse de locale ergodicité de la réponse diffuse, qui permet de séparer le processus stochastique de l'enveloppe dans le calcul des statistiques temporelles à court-terme du deuxième ordre. En pratique, ces hypothèses de travail pourront être parfois mises en défaut, et en particulier en bande étroite ainsi que pour de très faibles temps de réverbération, car dans ces deux cas (non disjoints), l'hypothèse d'ergodicité locale est infirmée : la durée minimale d'intégration pour que les statistiques du deuxième ordre soient stables est alors trop importante vis-à-vis des fluctuations de l'enveloppe.

Il est également utile de mentionner que dans la totalité des calculs qui suivent, on considère que le bruit de fond est négligeable.

2 DÉCROISSANCE INTÉGRÉE APRÈS UNE IMPULSION

Ce calcul est un rappel justifiant l'usage de l'intégration rétrograde sur la puissance instantanée de la réponse impulsionnelle.

On rappelle la définition de la courbe de décroissance en énergie :

$$EDC_h\,[n_0] = \sum_{n=n_0}^{+\infty} h^2[n]$$

[1]Les notions de stationnarité et d'ergodicité locales sont ici à prendre au sens **temporel**, qui sont à distinguer de la stationnarité et de l'ergodicité **spatiales**. Un champ diffus est par définition stationnaire (au sens large) dans le domaine spatial, mais pas nécessairement stationnaire dans le domaine temporel. Un champ réverbérant peut ou non, en fonction de la géométrie de la salle, être considéré comme stationnaire et ergodique dans le domaine spatial (Polack, 1992), mais ne peut de toute façon être considéré au mieux comme **localement** stationnaire et ergodique dans le domaine temporel.

En appliquant le modèle défini à l'équation C.2, et en se plaçant aux instants pour lesquels le modèle localement stationnaire défini à l'équation C.1 est valide, c'est-à-dire pour $n_0 > N_{mix}$, on obtient :

$$EDC_h\left[n_0\right] = \sum_{n=n_0}^{+\infty} b^2[n].e^{-2\alpha n}$$

Si l'on considère que les variations de l'enveloppe exponentielle sont lentes par rapport à celles du processus stationnaire $b[n]$, il est possible de séparer ces deux facteurs en réécrivant cette équation sous la forme approchée :

$$EDC_h\left[n_0\right] \simeq R_b[0].\sum_{n=n_0}^{+\infty} e^{-2\alpha n}$$

Ainsi, l'intégration rétrograde de la puissance instantanée de la réponse impulsionnelle est proportionnelle à celle de l'enveloppe. Puisque cette dernière est de nature exponentiellement décroissante, alors l'intégration rétrograde le sera également : en effet,

$$\sum_{n=n_0}^{+\infty} e^{-2\alpha n} = \frac{1}{1-e^{-2\alpha}}.e^{-2\alpha n_0}$$

D'où finalement :

$$EDC_h\left[n_0\right] \simeq \frac{R_b[0]}{1-e^{-2\alpha}}.e^{-2\alpha n_0} \ \forall n_0 > N_{mix} \tag{C.3}$$

Ce résultat justifie le mesurage du temps de réverbération à partir de la décroissance intégrée.

3 DÉCROISSANCE INTÉGRÉE APRÈS UN SIGNAL QUELCONQUE INTERROMPU

On se propose ici de généraliser le résultat précédent, et de montrer que dans le cadre des hypothèses de travail, la décroissance intégrée est exponentielle (toujours en laissant un laps de temps nécessaire pour le mélange) après n'importe quel signal s'interrompant brusquement.

Soit $x[n]$ un signal quelconque, nul pour $n \geq 0$. Soit $y[n]$ la résultante de la convolution de ce signal par une réponse impulsionnelle obéissant au modèle de l'équation C.2. La décroissance intégrée s'écrit :

$$EDC_y\left[n_0\right] = \sum_{n=n_0}^{+\infty} y^2[n] = \sum_{n=n_0}^{+\infty} \sum_{p_1=-\infty}^{0} \sum_{p_2=-\infty}^{0} x\left[p_1\right].x\left[p_2\right].h\left[n-p_1\right].h\left[n-p_2\right]$$

Si l'on ne considère que le cas où $n_0 \geq N_{mix}$, alors la partie précoce de la réponse, nommée $h_{spec}[n]$, ne contribue pas à la courbe de décroissance. Dans ce cas celle-ci vaut :

$$\begin{aligned} EDC_y\left[n_0\right] &= \sum_{n=n_0}^{+\infty} \sum_{p_1=\infty}^{0} \sum_{p_2=-\infty}^{0} x\left[p_1\right].x\left[p_2\right].b\left[n-p_1\right].b\left[n-p_2\right].e^{-2\alpha n}.e^{+\alpha.(p_1+p_2)} \\ &= \sum_{p_1=\infty}^{0} \sum_{p_2=-\infty}^{0} x\left[p_1\right].x\left[p_2\right].e^{+\alpha.(p_1+p_2)}.\sum_{n=n_0}^{+\infty} b\left[n-p_1\right].b\left[n-p_2\right].e^{-2\alpha n} \end{aligned}$$

Sur le modèle de la section 2, il est possible de fournir une écriture approchée du dernier facteur :

$$\sum_{n=n_0}^{+\infty} b\left[n-p_1\right].b\left[n-p_2\right].e^{-2\alpha n} \simeq R_b\left[p_1-p_2\right].\sum_{n=n_0}^{+\infty} e^{-2\alpha n}$$

D'où :

$$EDC_y\left[n_0\right] \simeq \left\{\sum_{p_1=\infty}^{0} \sum_{p_2=-\infty}^{0} x\left[p_1\right].x\left[p_2\right].R_b\left[p_1-p_2\right].e^{+\alpha.(p_1+p_2)}\right\}.\left\{\sum_{n=n_0}^{+\infty} e^{-2\alpha n}\right\}$$

Le premier facteur étant indépendant de n_0, la décroissance intégrée s'écrit finalement sous la forme :

$$EDC_y\left[n_0\right] \simeq \frac{C\left(x, R_b, \alpha\right)}{1-e^{-2\alpha}}.e^{-2\alpha n_0} \ \forall n_0 > N_{mix} \tag{C.4}$$

Là encore, on trouve que la décroissance intégrée suivant un signal quelconque interrompu est de nature exponentielle. En guise de vérification, on peut expliciter le cas particulier où le signal source est une impulsion :

si

$$x[n] = \delta[n]$$

alors

$$C\left(x, R_b, \alpha\right) = R_b[0]$$

On retrouve ainsi l'expression C.3.

4 INTÉGRATION SUR UNE DÉCROISSANCE TRONQUÉE

Il est nécessaire, si l'on souhaite appliquer le principe de l'intégration rétrograde à des silences de courte durée, d'étudier l'effet d'une troncature sur la courbe de décroissance. Si la troncature n'est pas trop importante (c'est-à-dire, si elle intervient après que le mélange soit effectif), alors les calculs précédents sont inchangés, à ceci près que la borne supérieure de l'intégration rétrograde n'est plus $+\infty$ mais l'indice de troncature, qu'on appellera N_{tronc}. L'intégration rétrograde de l'enveloppe s'écrit alors :

$$\sum_{n=n_0}^{N_{tronc}} e^{-2\alpha n} = \frac{1}{1-e^{-2\alpha}}.\left(e^{-2\alpha n_0} - e^{-2\alpha(N_{tronc}+1)}\right)$$

Donc l'équation C.4 devient :

$$EDC_y^{N_{tronc}}\left[n_0\right] \simeq \frac{C\left(x, R_b, \alpha\right)}{1-e^{-2\alpha}}.\left(e^{-2\alpha n_0} - e^{-2\alpha(N_{tronc}+1)}\right) \forall n_0 > N_{mix}$$

L'intégration rétrograde d'une décroissance intégrée n'est donc pas exponentielle, et ne sera donc pas linéaire en représentation logarithmique. Il est possible de définir un critère d'erreur par rapport au cas non tronqué, fonction de l'indice $n_0$considéré. Par exemple, on cherche les indices n_0 pour lesquels le rapport en décibels de la courbe de décroissance avec troncature sur la courbe de décroissance sans troncature est inférieure en valeur absolue à un certain seuil S_{dB} en décibels :

$$10.log_{10}\left(\frac{EDC_y^{N_{tronc}}\left[n_0\right]}{EDC_y\left[n_0\right]}\right) \leq -S_{dB} \tag{C.5}$$

En développant, on trouve :

$$\begin{aligned} \text{C.5} &\iff 1-e^{-2\alpha(N_{tronc}-n_0+1)} \leq 10^{-\frac{S_{dB}}{10}} \\ &\iff 2\alpha\left(N_{tronc}-n_0+1\right) \geq -ln\left(1-10^{-\frac{S_{dB}}{10}}\right) \end{aligned}$$

On définit la marge de retrait $\Delta T_{min}|_{S_{dB}}$ comme étant la durée (en secondes) minimale à laisser entre le dernier indice considéré et la fin de la décroissance pour que l'approximation linéaire soit valable à S_{dB} décibels près. Cette marge est définie simplement par :

$$\Delta T_{min}|_{S_{dB}} = \frac{1}{F_e}\left(N_{tr} - n_{0,max} + 1\right)$$

Puisque

$$\alpha = \frac{3.ln\left(10\right)}{F_e.T_r}$$

alors :

$$\text{C.5} \Longleftrightarrow \Delta T_{min}|_{S_{dB}} = -\frac{1}{6}.log_{10}\left(1 - 10^{-\frac{S_{dB}}{10}}\right).T_r$$

Pour fixer un ordre d'idées, on calcule la marge de retrait relative à 1 décibel près :

$$\frac{\Delta T_{min}|_{1_{dB}}}{T_r} \simeq 0,11$$

Il est donc nécessaire de laisser jusqu'à la fin de l'intégration rétrograde une marge au moins égale à une durée de l'ordre du dixième du temps de réverbération pour que la courbe de décroissance puisse être considérée comme linéaire à 1 décibel près. Bien entendu, on ne mentionne ici que le biais dû à la troncature. Si les hypothèses de travail, et notamment l'hypothèse d'ergodicité locale, ne sont pas respectées (et c'est généralement le cas en bande étroite), la courbe de décroissance est soumise à des fluctuations résiduelles supplémentaires.

5 UTILISATION DE LA PUISSANCE À COURT-TERME

On s'intéresse ici à l'évolution de la puissance à court-terme sur le modèle proposé de réverbération. En admettant que l'hypothèse d'ergodicité locale conduit selon le même principe qu'en section 3 à déduire que la puissance à court-terme d'un signal est proportionnelle, pendant la réverbération tardive, à la puissance à court-terme de son enveloppe (exponentielle), on développe le calcul de cette dernière.

Si l'on choisit comme fenêtre d'intégration une fenêtre exponentielle de constante de temps ΔT (en échantillons), on obtient comme valeur de la puissance à court-terme de l'enveloppe de réverbération :

$$P_{rev}\left[n_0\right] = \sum_{n=-\infty}^{n_0} e^{-2\alpha n}.e^{\frac{n-n_0}{\Delta T}}$$

...soit après calculs :

$$P_{rev}\left[n_0\right] = \frac{1}{1 - e^{2\alpha - \frac{1}{\Delta T}}}.e^{-2\alpha.n_0}$$

Donc, **si** ΔT **est constant**, la puissance à court-terme pendant la queue de réverbération est encore une fois proportionnelle à l'enveloppe exponentielle. En revanche, si ΔT est variable, comme cela a été suggéré par exemple en fin du chapitre VII, le dénominateur $\left(1 - e^{2\alpha - \frac{1}{\Delta T}}\right)$est lui aussi variable, et comme α est inconnu, on ne peut prévoir son évolution. Une manière de minimiser les effets de ces variations est de s'assurer que l'exponentielle reste négligeable devant 1, soit :

$$\Delta T << \frac{1}{2.\alpha} = \frac{F_e.Tr}{6.ln(10)}$$

Le même type de condition peut être trouvé pour une fenêtre rectangulaire : il faut que sa durée (en secondes) reste inférieure à $\frac{Tr}{6.ln(10)} \simeq \frac{Tr}{13,8}$ quelles que soient ses variations pour que la puissance reste d'allure exponentielle.

D

Publication 1

METHODS FOR BLIND COMPUTATIONAL ESTIMATION OF PERCEPTUAL ATTRIBUTES OF ROOM ACOUSTICS

ALEXIS BASKIND AND OLIVIER WARUSFEL

IRCAM, Room Acoustics Team
1 place Igor-Stravinsky, 75004 Paris, France
Alexis.Baskind@ircam.fr
Olivier.Warusfel@ircam.fr

Here is presented a set of signal processing methods, which aim at providing a blind estimation of a low-level description of the room effect related to a recorded audio scene, in order to derive a perceptual characterization of its spatial features. In this case, "blind" means that only the recording itself is provided, and that no other information (i.e. a geometrical description of the room, or a set of room impulse responses) is available. Two single channel tools are proposed, in order to take into account the different natures of the early and late parts of the room effect, that estimate the relevant information in each of them. Then the estimations of the early parts of the responses are simultaneously processed by a coincidence detector, in order to estimate the binaural cues of the sound scene.

INTRODUCTION

Although the room effect plays a major role in the perception of our environment, it is rarely the main topic in computer-based audio scene analysis problems. As a matter-of-fact, most of the research in this field (for instance speech recognition, sound categorization, or sound separation) is focused on sources and, from this point of view, the room effect is considered as a disturbance. Standard room acoustics analysis techniques, which rely on the knowledge of one or several room impulse responses [5, 13, 3], are not suited when considering “real-life” source signals such as music or speech, since they are not able to isolate the information on the room from the information on the sound source. Some work has been achieved on the measurement of reverberation time with musical signals [19], but it relies on the knowledge of the source signal.
Our goal is to find ways to automatically estimate all the spatial attributes of a sound scene, that is, not only those related to the source itself (azimuth and elevation), but also, and mainly, those related to the room (reverberance, apparent source width, listener envelopment, perceptual distance...). We assume that this sound scene is represented by a dummy-head, a stereophonic or a multichannel recording only, and that no extra information is given, neither of the room (i.e. geometrical description, impulse responses), nor of the source, so that the estimation is completely “blind”.
We decided to focus on a the derivation of a perceptual description of the room rather than a geometrical or signal description. As a matter-of-fact, deriving a geometrical description (i.e. materials and shape of the walls) is not currently feasable given a limited number of channels. Also a perceptual representation provides a concise description of the room effect, keeping only the most relevant information needed for the task at hand (since the final purpose is not dereverberation, a complete physical or signal description is not needed). For instance, the minute details of the late reverberation that can be found in a signal description are not relevant for audition, which is only sensitive to the mean energy decay.
Such a computational room acoustics description device would provide a powerful tool for automatic labelling and database indexation, and could be applied to many technical and artistic issues :

- production : automatic mixing of complex sound scenes,
- post-production : sound or language dubbing in cinema or music with respect to the original room and spatial effects,
- augmented reality : adding artificial sounds to a natural sound scene.

In all those cases, one or several dry sounds must be mixed to a pre-recorded audio scene. In order to obtain a result that sounds homogeneous, they must be processed by a room effect similar to the original one.
In the first part, main features and natural limitations of this problem will be addressed and assumptions chosen as a starting point will be presented. Then, the analysis will be presented step by step: estimation of the early reflections for each channel, binaural coincidence processing, and estimation of the reverberation envelope.

1. PROBLEM OVERVIEW AND ARCHITECTURE PROPOSAL

In order to propose an architecture that is flexible enough to adapt to most cases, the problem has first to be well defined. In particular, some major obstacles must be taken into account in order to propose a realistic working hypotheses and architecture.

1.1. Inherent obstacles

First, the physical complexity of the propagation of sound in a room must first be addressed: source and receiver may move; directivity patterns can drastically vary as a function of frequency, and thus as a function of the emitted signal [22]; many particular cases, like flutter echoes or coupled rooms, modify the shape of the room impulse response.
Moreover, within the room effect, discrete reflections and the reverberant diffuse field have very different natures : the former are closely linked to the position of the sources and the receiver, whereas the latter has a stochastic behaviour. However, a closer look at samples of the late part of the impulse response shows that their values are highly sensitive to any modification in the propagation channel. This implies that a room effect can hardly be considered as a time-invariant filter. Furthermore, the multiplicity of sources in a given sound scene, as well as mixing and editing processes in the case of an artistic recording, make the final recording not purely convolutive.

1.2. The perceptual approach

Following the preceding remarks (and according to the points mentionned in the introduction), a perceptually-based mode of description has been choosen. In such an approach, the sound scene is described by objective indices (such as ITD, IACC or reverberation time) which are closely linked to the perceptual characteristics of the corresponding auditory event (respectively localization, spaciousness or reverberance).
The perceptual description naturally takes into account the different natures of the early and late parts of the room effect : for instance, the observation scale depends on the criterion. Moreover, it is flexible enough to allow little modifications of the scene (such as source movements), whereas a signal processing description would require as many sets of impulse responses as positions of sources and receivers.
Such an approach also allows us to rely on performances and behaviour of audition, in order to design analysis tools that are as close as possible to it. In particular, the quantity of information that is available to the listener (and thus to our automatic analyzer) closely depends on the nature of the emitted sound : for example, a stationary sound won't provide enough information to the audition to evaluate any spatial cue (distance, size of the room...). In the same way, late reverberation can only be perceived when the source becomes suddenly silent.
Audition behaviour also depends on the context. A well known example is the case of the vibrato, that enhances the spatial impression in comparison with the strict stationary case. Griesinger [10] proposed that this effect is due to fluctuations of interaural time differences (ITD). Also, as mentioned in the introduction, audition is not sensitive to all the details in the reverberant decay, but only to its time-frequency envelope.

1.3. Working hypotheses

Concerning the analysis, our approach consists of taking into account as many of the features mentioned in 1.1 as possible in the analysis, and making simplifying assumptions concerning the others.
First, the signal has to be "as convolutive as possible", i.e. the sound scene will consist of a unique source and room, and this source as well as the microphones must remain fixed or slowly moving, in order to make the propagation channel nearly constant over a time period necessary to estimate the room effect.
Second, the recording is considered to be a "binaural" recording, made with a dummy head or synthesized artificially using head-related transfer functions. If not (i.e. stereophonic or multichannel recording), the recording should first be transcoded to a binaural format (thanks to virtual loudspeakers, for example) or, if we have extra information on the sound recording process (such as microphone positions), it could be taken into account in the analysis.

1.4. Architecture proposal

Following the remarks of section 1.1, we propose that the system matches the following points:

- monaural and binaural cue estimations will be carried out separately
- early reflections and late reverberation, because of their different natures, will be analyzed using different tools
- the analyses should focus on time segments of the recorded signals with the most relevant information, depending on the actual centre of interest, that is onsets for the first reflections, and sudden silences for late reverberation.

We propose the following architecture: first, the early part of the room effect is analyzed by homomorphic deconvolution; then, binaural coincidence processing is applied on the result of the former. Reverberation time is estimated separately, using energy decay plots.

2. ESTIMATION OF THE EARLY REFLECTIONS

The goal of this step is to estimate as accurately as possible the early part of the room effect, roughly the first 100 ms. Such an estimation is a prerequisite for deriving cues related to early reflections and direct sound, for example *localization*, *perceptual distance*, *auditory source width*, or *listener envelopment*.

Assuming that the recording, at least for the early part of the room effect, is the result of a time-invariant filter over the analysis duration, the estimation of this filter could be considered as a blind deconvolution problem. There are several ways to perform blind deconvolution. We have opted for homomorphic deconvolution, mainly because it provides a good time resolution, whereas parametrical techniques such as bayesian estimation (as proposed in [11]) needs very high orders to provide enough accuracy in the time domain, resulting in high computation costs. Moreover, homomorphic deconvolution naturally isolates redundant information (the room) from fluctuant information (the sound emitted by the source), contrary to linear techniques like wavelet analysis.

2.1. Principle of homomorphic deconvolution

2.1.1. Cepstral processing, homomorphic deconvolution

Homomorphic deconvolution, or cepstral analysis, consists in working with the cepstrum of the signal, which is defined as the inverse Fourier transform of the complex logarithm of the signal Fourier transform [16]:

$$C_x[k] = IFT\left\{Log\left[FT\left\{x[n]\right\}\right]\right\}$$

One of the major properties of the cepstrum is that it handles convolution as an addition : if $y_n = h * x_n$ (where h is the filter, and x_n one of the source signals), then in the cepstral domain, $C_{y_n} = C_h + C_{x_n}$. Thus, deriving the average cepstrum of several segments y_n of the signal will enhance redundant information (the filter) in comparison with fluctuant information (the source signal) [17]. However, cepstral processing, and especially cepstral averaging, has several drawbacks.

First, homomorphic deconvolution needs in theory the processed signals to be purely convolutive, which is not the case here, because of segmentation. As a matter-of-fact, each segment is mixed to the reverberant tail of the preceding segment, which thus can be considered as an additive disrupting source; in the same way, the reverberant tail of the present segment is put away in the next segment. However, practical experiments show that the technique allows a small additive disturbance : the S/N ratio must be greater than approximatively 40-60 dB when the noise is stationary white; however, the constraint is not so critical with exponentially decaying noise, and experiments showed that the minimum delay between two segments should be at least one third of RT_{60}. The main effects of an additive noise are an artificial enhancement of the first sample of the estimated impulse response and a global decrease of the signal-to-noise ratio.

Moreover, when using cepstral averaging, only the amplitude spectrum, and thus the minimum-phase part of the signal can be estimated, because such a technique is not able to estimate the wrapped phase with enough accuracy. This process entails working with short windows (an extra reason to focus only on the early part of the response) in order to minimize the energy of the all-pass component. In the same way, cepstral processing does not handle the pure delay component of the signals: therefore, the information on global time delay between both channels (and thus all interaural time delays between sound events) is lost, and must be estimated in another way.

2.1.2. Mathematical formulation

An ideal segmentation on both channels of the binaural recording would produce N segments for each channel which are the perfect convolution of the corresponding segments of the dry source by each impulse response:

$$\left\{\begin{array}{l} y_{L_n} = h_L * x_n \\ y_{R_n} = h_R * x_n \end{array}\right. \forall n \in \{1...N\}, \forall k$$

so that in the cepstral domain :

$$\begin{pmatrix} C_{y_{L_1}}(k) \\ \vdots \\ C_{y_{L_N}}(k) \\ C_{y_{R_1}}(k) \\ \vdots \\ C_{y_{R_N}}(k) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \\ 1 & 0 & & & 1 \\ 0 & 1 & 1 & & \\ \vdots & \vdots & & \ddots & \\ 0 & 1 & & & 1 \end{pmatrix} \cdot \begin{pmatrix} C_{h_L}(k) \\ C_{h_R}(k) \\ C_{x_1}(k) \\ \vdots \\ C_{x_N}(k) \end{pmatrix} \forall k$$

that is

$$\underline{Y}(k) = \underline{\underline{M}}.\underline{X}(k) \forall k \quad (1)$$

Since $\underline{\underline{M}}$ is not invertible, equation 1 cannot be directly resolved. However, one can consider the associated least-squares problem, which consists in minimizing the error

$$E\left(\underline{\hat{X}}(k), k\right) = \left(\underline{Y}(k) - \underline{\underline{M}}.\underline{\hat{X}}(k)\right)^T . \left(\underline{Y}(k) - \underline{\underline{M}}.\underline{\hat{X}}(k)\right)$$

for each quefrency sample k ("quefrency" is the cepstral domain axis) . The minimum error is reached if and only if

$$\underline{\underline{M}}^T.\underline{Y}(k) = \underline{\underline{M}}^T.\underline{\underline{M}}.\underline{\hat{X}}(k) \quad (2)$$

An exact solution $\underline{\hat{X}}^0(k)$ to eq. 2 can be proposed, which generalizes traditional cepstral averaging :

$$\left\{\begin{array}{l} \hat{C}^0_{h_L}(k) = \frac{1}{N}\sum_{n=1}^{N} C_{y_{L_n}}(k) \\ \hat{C}^0_{h_R}(k) = \frac{1}{N}\sum_{n=1}^{N} C_{y_{R_n}}(k) \\ \hat{C}^0_{x_p}(k) = \frac{1}{2}\left(C_{y_{L_p}}(k) + C_{y_{R_p}}(k)\right) \\ \qquad\qquad -\frac{1}{2N}\sum_{n=1}^{N}\left(C_{y_{L_n}}(k) + C_{y_{R_n}}(k)\right) \end{array}\right.$$

$\forall k, \forall p \in \{1...N\}$. Since an exact solution could have been found, there is no need to use pseudo inversion algorithms, such as truncated singular value decomposition (TSVD), which are not accurate enough for our purpose. All the solutions to eq. 2 can be obtained by adding to the particular solution $\underline{\hat{X}}^0(k)$ any vector of the kernel of $\underline{\underline{M}}^T.\underline{\underline{M}}$, that is :

$$\underline{\hat{X}(k)} = \underline{\hat{X}}^0(k) + \alpha(k).\begin{pmatrix} 1 & 1 & -1 & \cdots & -1 \end{pmatrix}^T \forall k \tag{3}$$

where $\alpha(k)$ is any real cepstrum.
In time or frequency domain, this corresponds to filtering the estimates of the impulse responses $\hat{h}_L^0(t)$ and $\hat{h}_R^0(t)$ by a common invertible filter $A(f)$, and simultaneously filtering the estimates of the sources $\hat{x}_1^0(t)$ to $\hat{x}_N^0(t)$ by the common inverse filter $\frac{1}{A(f)}$. The exact and unique solution to eq. 1, if it exists, is obtained for

$$\alpha(k) = -\frac{1}{N}\sum_{n=1}^{N} C_{x_n}(k) \tag{4}$$

, which is the mean cepstrum of the dry sounds over all the recording. Of course, the problem is that this filter cannot be computed, since the sources are not known. Without any extra algorithm to estimate it, a rough assumption has to be formulated. Ours consists in assuming that the mean source cepstrum $\frac{1}{N}\sum_{n=1}^{N} C_{x_n}(k)$ contains mainly spectral envelope information (i.e. low quefrencies), and that the fast variations are cancelled by averaging. This assumption relies on the fact that whereas a room impulse response is traditionaly broadband, most everyday sounds have a 1/f shaped spectrum [8]. Then the postfilter that was chosen consists in whitening the spectrum of the estimations. Of course, the spectral envelope of the source is no longer present in the estimation of the room impulse responses, but also all the information on the spectral shape of the room effect is lost, and we won't be able to derive spectral cues such as source spectral color. However, interaural spectral differences can be computed, since the estimation bias is identical for both channels, and would allow us to estimate the *elevation* of the source.

2.2. Practical issues

2.2.1. Performing the segmentation

The segmentation must be performed carefully, since it will drive the quality of the estimation. First, it must be able to enhance the S/N ratio, where the noise here is due to truncation errors, and reverberant tail of the previous part of the signal. Second, it must concentrate most of the energy in the early part of each segment, otherwise much of the information will be lost with the all-pass part of each segment. For both those reasons, segmentation must be performed on the onsets. It is worth noticing that this result is akin to human audition.

2.2.2. Exponential windowing

An exponential window can be applied to each segment in order to minimize the effects of truncation and to pick-up information over the unity circle [4]. It does not disturb the convolution relation : as a matter-of-fact,

$$e^{-\frac{t}{\tau}}.(x(t) * h(t)) = e^{-\frac{t}{\tau}}.x(t) * e^{-\frac{t}{\tau}}.h(t)$$

After deconvolution, we obtain an estimate of the response multiplied by the window, thus an inverse window has to be applied. The time constant τ, as well as the size of the window, has to be carefully choosen, since when applying this inverse window, a high damping will entail a very poor precision at the end of the estimation. The choice that has been made is to set τ to a duration corresponding to the expected duration of the estimation, and the length of the window to 5τ. The end of the estimation will diverge after inverse windowing, but only the first part is kept.

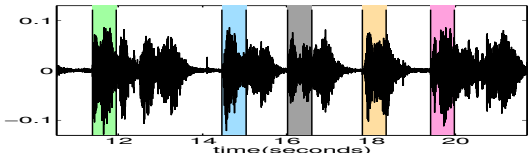
2.2.3. Requirements on the source signal and the impulse response

A good segmentation implies that the source signal contains many silences, so that the reverberant tail of the former segment does not disturb the current segment too much. This is often the case for speech, but rarely music. It also implies that the onsets are well marked, making the energy of the direct sound high compared to the remaining energy of the impulse response. If not, the onsets are hard to detect, since they do not correspond to a fast transient. Such a requisite can be easily related to the minimum-phase or mixed-phase nature of the impulse response : as a matter-of-fact, the minimum-phase part of a signal corresponds to the maximum concentration of the energy in the first samples for a given amplitude spectrum.
For this reason, and because homomorphic deconvolution is not able to estimate the all-pass part of the filter, we have to assume that the effect of the all-pass part of the impulse response is weak, compared to the minimum-phase part. From a physical point of view, it corresponds to a sufficiently high Dir/Rev level or, equivalently, to a small distance between source and receiver [21].

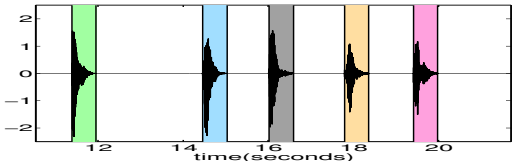
2.2.4. Example

In order to check the efficiency of this technique, we applied it to synthetic test signals. In the example provided here, the source signal is a long recording of anechoic speech, and the room effect is a synthetic monophonic impulse response that contains early coherent reflections (from 0 to 100ms) as well as a reverberation tail

($RT_{60} = 1\,s$, $Dir/Rev = 10\,dB$). The signal is first segmented and windowed (fig 1), then the cepstral averaging is performed. Finally, a spectral whitening is applyied on the estimation. Since we try to estimate the first 100 ms of the impulse response, the damping coefficient of the exponential window was fixed to 100 ms, and the window duration to 500 ms.
The results of this estimation are shown in fig. 2 : the original impulse response is compared to the initial estimation, to this estimation with spectral whitening, and finally, to this estimation with the ideal postfiltering defined in eq. 4, in order to check the validity of the convolution assumption. Those results clearly show the necessity of the postfiltering stage, since it greatly enhances the readability of the first reflections.



(a) onset marking and segmentation



(b) exponential windowing and energy normalization

Figure 1: segmentation on the onsets performed over a speech sequence (the vertical strips correspond to the choosen segments)

2.3. Conclusion on this stage

Cepstral processing appears as a good means to perform the task of this stage, which is to provide an accurate low-level description of the early part of the room effect by estimating the first part of 'its' room impulse response.
Practical experiments showed that when the segmentation is properly performed, results match the theory quite well. However, a theoretical limitation, i.e. the loss of nearly all non minimum-phase information, has to be overcome, or at least, taken into account, since it entails the energy levels of early reflections to be inaccurately estimated in mixed-phase cases (the time position is correctly estimated). Thus, cues based on relative energy level of the early reflections part (such as *perceptual distance* or *envelopment*) may be biased.
As a further development, postfiltering could be improved, for example by trying to minimize correlation or even statistical dependence between the estimate of a room response with each estimate of the sources.

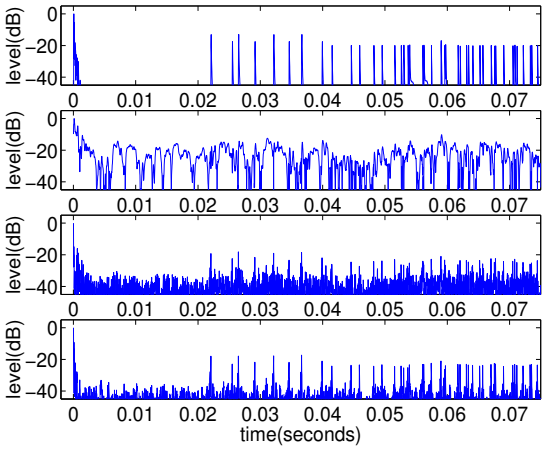


Figure 2: homomorphic deconvolution of a synthetic impulse response over 18 segments :
top : reflectogram of the original impulse response
second : reflectogram of the estimation without postfiltering
third : reflectogram of the estimation after spectral whitening
bottom : reflectogram of the estimation with ideal postfiltering

3. BINAURAL COINCIDENCE DETECTION

This stage aims at deriving high-level objective criteria related to auditory spatial cues, such as *localization* or *spatial impression*. It consists in simultaneous processing of left and right information, in our case the early parts of the binaural room impulse responses that were previously estimated.

3.1. Principle of cross-correlogram

The first (and simpliest) model of binaural interaction has been proposed by Jeffress [12]. It is composed of delays and EE (excitation-excitation) neurons, firing only for a specific delay between left and right channel. Since it is an efficient model for estimating source azimuth, it has been adapted to a signal processing formulation, as the **cross-correlogram** :

$$\Psi_{x,y}(t,\tau) = \int_{-\infty}^{+\infty} y^*(\nu).x(\nu+\tau).w(t-\nu).d\nu \quad (5)$$

where $w(t)$ is a window which models short-term memory. A normalized definition can be used, which makes the correlogram bounded by -1 and 1 :

$$\rho_{xy}(t,\tau) = \frac{\Psi_{x,y}(t,\tau)}{\sqrt{\Psi_{x,x}(t+\tau,0).\Psi_{y,y}(t,0)}} \quad (6)$$

The main drawback of equation 6 is that, contrary to the case of long-term correlation, for a given couple (t, τ), either x or y can be zero or nearly zero during all the window, thus the normalized correlation does not mean anything.
Correlograms have been applied to some auditory scene analysis problems, such as pitch estimation (in its mono formulation), source localization and separation, or room acoustics planning problems [5].
The idea here is to use the correlogram as a localization model of the direct sound and the first reflections, as well as a short-term correlation estimator between left and right :

- The delay τ corresponding to the correlation maximum for a given time provides an estimation of delay between both channels for a given event (direct sound or early reflection). This delay corresponds to the **Interaural Time Differences** (ITD) in binaural hearing, and thus is helpful for estimating source *azimuth*. It is also a useful indicator for isolating side reflections, which are partly responsible for *spaciousness*. Spaciousness can be estimated in medium and high frequencies using criteria like **Lateral efficiency** (LF) [2].

- The maximum level of correlation between 0 and 80 ms can be easily derived from the correlogram. It is usually defined as the **early Interaural Cross Correlation** (early IACC) [6] and can be related to *auditory source width.*

3.2. Pratical issues

3.2.1. Resynchronization

Before applying the cross correlogram on the estimates of the early part of the responses, they have to be resynchronized with respect to the time delay between them, which was lost during the cepstral processing. If not, all the relative delays between events will be biased, and in particular, the direct sound ITD will be artificially set to zero. Fortunately, it can be estimated from the recording [7] by applying the correlogram directly on it first. In fig 3 is shown the croiss-correlogram of a segment of the recording. Since the direct sound predominates in comparison with the remaining part of the room effect, the short-term autocorrelation pattern of the dry source signal is shifted with respect to the direct sound ITD between both channels.

3.2.2. Choosing the right window

The shape and length of the analysis window have to be carefully chosen, according to the signals and the features that have to be analyzed :

- Long and smooth windows (Hanning, cosine...) are suited for quasi-stationary signals : as a matter of fact, when the statistical properties of a signal vary slowly, the correlogram is an unbiased estimator of the statistical nonstationary correlation.

- Short and sharp windows are suited for transient analysis. The window we used in this case is the exponential window, i.e :

$$w(t) = \begin{cases} 0 & if\, t < 0 \\ e^{-\frac{t}{\tau}} & if\, t \geq 0 \end{cases}$$

  Thaks to its discontinuity at t=0, this window is able to represent fast transients without the blurring entailed by Hanning-like windows. It behaves as a short-term memory window, since it is causal, and gives more importance to recent information than to past information. For this reason, such a window has been often used in binaural modeling.

3.2.3. Example

In fig 4 (top) is shown the correlogram of the estimated binaural room impulse responses with ideal postfiltering defined in eq. 4. The actual impulse responses used here are the same that were used in part 2, and consist of : a direct sound (whose ITD equals to 0.22 ms), eight side reflections with an ITD of $\pm$0.35 ms and a time of arrival varying from 20 ms to 40 ms, a cluster of late reflections with a time of arrival varying from 40 ms to 120 ms, and finally, a reverberant tail, with a reverberation time of 1s for each frequency.
The direct sound is clearly visible as a thin horizontal line around $\tau \simeq 0.2\,ms$. The side reflections are represented by two groups of four spots for $\tau \simeq \pm 0.35\,ms$. The group of late uncorrelated reflections is represented by a global pattern for t>50 ms, with a loss of coherence ($C_{max} = 0.5$). In this case, an accurate time-space representation of the room effect is obtained. However, as written above, this ideal postfilter can not be computed in a blind situation.
With actual postfiltering, that is spectral whitening (fig 4 - bottom), the representation is less easily readable: the first reflections are still visible, but they can hardly be distinguished from the global noise as well as a large blur juste after the direct sound: there is still residual information related to the source, which is common to both channels.

3.3. Conclusion on this stage

Although the correlogram is a very simple and incomplete model of human binaural hearing, it provides useful information on time differences and cross-correlation between both channels. Since time differences are closely

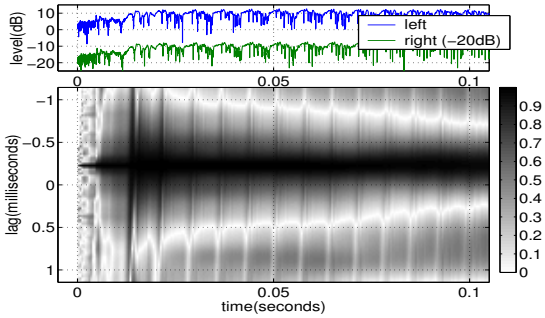
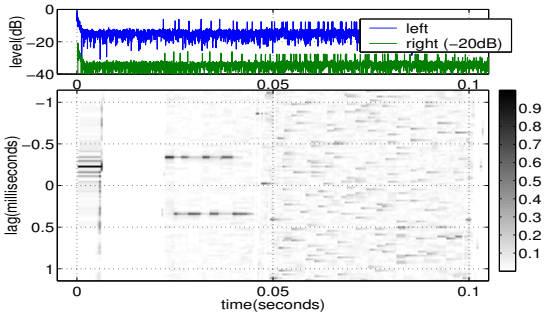
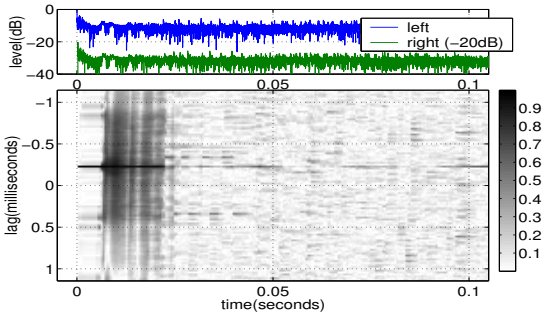


Figure 3: Correlogram computed directly on a segment of the binaural recording (top : signal level in dB - bottom : cross-correlogram). The direct sound ITD correspond to the maximum with respect to τ. Here, ITD$\simeq$0.2 ms



(a) estimated impulse responses with ideal postfiltering



(b) estimated impulse responses with spectral whitening

Figure 4: Correlograms of the estimated impulse responses with ideal and actual postfitlering (top : signal level in dB - bottom : cross-correlogram)

linked to the azimuth of the sound event, it can also be considered as a spatial map of the room effect.

In the future, this model should be replaced by more complete models [9, 14]. In particular, taking into account the precedence effect would allow us to consider ITD fluctuations (see part 1) as an extra cause of auditory spatial impression.

Another useful extension would be to work with peripheral processing, that is cochlear filterbanks, in order to be able to evaluate frequency dependent room acoustics perceptual cues. The main obstacle at present is that the previous stage (i.e. cepstral processing) is not able to estimate the spectral envelope of the response.

4. ANALYSIS OF REVERBERATION

According to section 1, the minute details of the reverberant tail are not perceptually relevant, whereas the envelope gives useful information on cues such as *reverberance* and diffuse-field spectral colour. Thus, this stage aims at deriving the (frequency-dependent) reverberation time from signal sudden silences, that are the only parts of the recording where the room resonates freely. This technique is sometimes used by acousticians to estimate reverberation time from a recording, or during a live concert.

4.1. Principle of the analysis

This analysis module uses the well-known principle of **energy decay curves**, proposed by Schroeder [20] as an efficient and robust method to derive reverberation time from a room impulse response. It can be applied to real recordings, since during free resonances, the room behaves the same way whatever the excitation was (impulse, music,...): the energy contained in it just before the sudden silence decays exponentially with respect to RT. As a matter-of-fact, let us consider the case of a dry source signal suddenly silent convolved by a quite simple model of reverberation, that is an exponentially-decaying white gaussian noise :

$$y(t) = (h * x)(t)$$

where x(t) is the source signal, which is set to zero for $t \geq 0$ (sudden silence), and h(t) is the impulse response :

$$h(t) = n(t).e^{-\delta.t}$$

where n(t) is a white gaussian noise with the power σ_n^2. This reverberation model corresponds to the case where reverberation time does not depend on frequency in the considered frequency band. The averaged backward integration of the output signal (i.e. its "energy decay curve") is defined by :

$$I_y(t) = \left\langle \int_t^{+\infty} y^2(\alpha).d\alpha \right\rangle = \int_t^{+\infty} \left\langle y^2(\alpha) \right\rangle .d\alpha$$

where $\langle . \rangle$ denotes statistical average. Deriving $\langle y^2(\alpha) \rangle$ for $\alpha \geq 0$ yields :

$$\begin{aligned} \langle y^2(\alpha) \rangle &= \sigma_n^2 . e^{-2\delta\alpha} . \left\{ \int_{-\infty}^{0} x^2(\tau) . e^{2\delta\tau} . d\tau \right\} \\ &= \sigma_n^2 . e^{-2\delta t} . C \end{aligned}$$

since $x(\tau) = 0$ for $\tau \geq 0$. The sum in curly brackets is constant, hence finally :

$$I_y(t) = C . \frac{\sigma_n^2}{2\delta} . e^{-2\delta t} \tag{7}$$

In other words, the energy decay curve of the output signal after a sudden silence decays exponentially with respect to reverberation time, in the same way as the energy decay curve of the impulse response. The idea developped here is then to find and isolate such sudden silences, and to compute an energy decay curve for each of them in each frequency band, to estimate the slope, and then to derive a mean estimation.

4.2. Practical issues

4.2.1. Limitations of a real case

In practice, the assumption of a sudden and definitive silence on the source signal is rarely satisfied : not only are *real* silences quite rare, even in speech, but silence durations are also never sufficient to make equation 7 a good approximation of the real decay curve, which is no longer exponentially decaying. Fig. 5 shows an example of a decay curve computed over a segment of relative silence in a reverberant speech recording. It can easily be seen that the curve is no longer linear, and that the remaining noise modifies the slope compared to the slope of the impulse response energy decay curve; thus, the reverberation time estimate is biased.

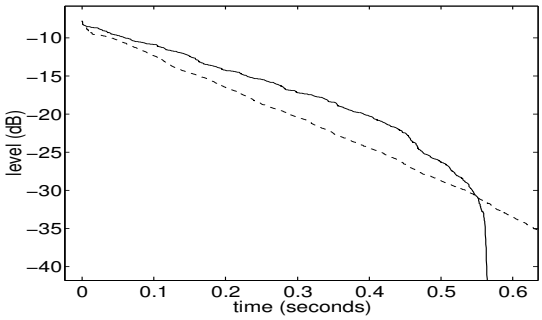


Figure 5: Typical energy decay curve computed over a segment of relative silence in reverberant speech
solid line : reverberant speech decay curve
dashed : impulse response decay curve

The estimation may be biased by two kinds of causes :

- If the beginning of the segment is set too early, the source could still emit some sound, and the RT is underestimated because of the resulting excessive energy. This problem makes the **early decay time** hard to estimate, since it entails distinguishing residual source energy from the early energy decay of the room.
- If the end of the segment is set too late, or if the sound source is not completely silent during the decay, the RT is overestimated (this is the case of fig. 5).

To overcome the second problem (assuming that the beginning of the segment is carefully set), the analysis is traditionaly performed over several segments, and the lowest RT estimation is chosen.

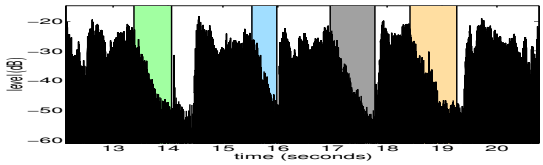
4.2.2. Architecture proposal

The analysis is performed in several steps :

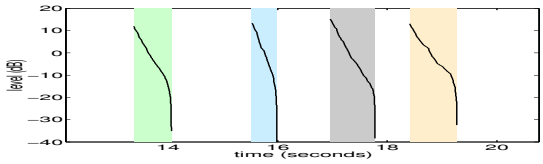
1. Location of sudden silences, thanks to a root-mean-square plot of the signal
2. Segmentation : the silence before the next onset has to be long enough (at least several hundreds of milliseconds), in order to obtain an accurate estimate
3. Band-pass filtering
4. linear or non linear regression analysis to derive reverberation time : the former is more robust, the latter is more accurate
5. Averaging over both channels
6. Elimination of estimates that are too far from the mean (non accurate estimations), and selection of the minimum RT estimation

4.2.3. Example

Again, the technique is applied on a test signal. The source signal is the same speech recording as in section 2.2.4. The impulse response is artificially generated, and provides a RT of 1s whatever the frequency. For this test, the decay curves were computed broadband. Eighteen segments of silences are set in the reverberated recording, as shown in fig. 6. Their average duration is approximately 0.5 s. Five segments were eliminated, since the estimation was too far from the mean (the relative difference between local estimation and mean estimation was more than 20%). The minimum estimation over the remaining segments is 0.91 s.
Although this method provides a good estimation in this case, it must be noticed that it is highly sensitive to the



(a) decay marking and segmentation



(b) backward integration segment by segment

Figure 6: segmentation on the decays performed over a speech sequence (the vertical strips correspond to the chosen segments)

location of the segments, in accordance with preliminary remarks of part 4.2.1. Moreover, it is not precise enough to take into account double slope decays, or to derive the early decay time.

4.3. Conclusion on this stage

Whereas this method works fine within the assumptions of the theory, that is with a definitive silences, it is much more delicate to apply in a more realistic case. Actually, its main limitation is that long and complete silences are hard to find in music. A possible approach to overcome this limitation could be to apply a time-dependent adaptative filter on the signal before processing it, in order to keep only the relevant information and free as possible from disturbances, i.e. noise from non-silent frequency bands, and onsets on (possibly) other bands. In this way, the signal to noise ratio should be raised, as well as the silence duration. Thus, we could meet again audition behaviour : we are able to estimate the size of the room thanks to sudden silences...or sudden pitch changes, which allow us to hear the decay in the frequency band that used to contain the old pitch.

Moreover, we are not able at present to make the difference between the resonance of the room and a resonance of a sound source, which also has an exponential shape most of the time. A way to surpass this drawback would be to evaluate the correlation during the decay between both channels, since the reverberation is diffuse, whereas the resonance of the source is well localized.

5. CONCLUSION

To summarize, the work presented here is an attempt to go further in computer-based audio scene analysis, which aims at estimating spatial cues of a scene directly from a recording. It uses both auditory models and non physiologically-based signal processing methods, and relies on both perceptual and physical features of room acoustics.

The paradigm assumed here (one fixed source) is quite restrictive, but it can be considered as a starting point, and could be compared to blind source separation techniques, which traditionaly adopt the opposite point of view (several sources, instantaneous mixtures) [18]. In the future, mixing both issues would be a great challenge ! Actually, the case of convolutive mixtures is already one of the main goals in blind source separation [15, 1], but since it is focused on source estimations, it does not provide enough accuracy (the number of frequency bands is limited) to estimate the room effect; moreover, it does not take into account its double nature (deterministic in the early part, and probabilistic in the late part).

One of the main problems that we have to face up to now is to link the results between the estimators, in order to obtain a coherent high-level representation of the room effect. In particular, the energy levels have to be carefully adjusted between the estimations of the early part and the reverberant tail, because they drive cues that are related to the complete impulse response, such as Dir/Rev ratio , early decay tim**e** (EDT), or clarity index (C80).

The architecture is not yet complete (in particular, the evaluation stage is lacking), and each stage has not been fully studied and developed, but the approach opted here seems promising and provides interesting adequacies with human audition.

REFERENCES

[1] J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *proceedings of the 2nd international workshop on independent component analysis and blind signal separation*, June 2000. 5

[2] M. Barron and A.H. Marshall. Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure. *J. Sound Vib.*, 77(2):211–232, 1981. 3.1

[3] A. Baskind and J.-D. Polack. Sound Power Radiated by Sources in Diffuse Field. In *proc. AES 108th convention*, February 2000. (document)

[4] D. Bees, M. Blostein, and P. Kabal. Reverberant speech enhancement using cepstral processing. In *Proceedings of ICASSP 91, vol. 2*, pages 977–980, 1991. 2.2.2

[5] J. Blauert, M. Bodden, and H. Lehnert. Binaural signal processing and room acoustics planning. *IEICE Trans. Fundamentals*, E75-A(11):1454–1459, november 1992. (document), 3.1

[6] J. Blauert and W. Cobben. Some consideration of binaural cross correlation analysis. *Acustica*, 39:96–104, 1978. 3.1

[7] Markus Bodden. Auditory models for spatial impression, envelopment and localization. In *Proceedings of the AES 15th international conference*, pages 150–156, 1998. 3.2.1

[8] D. Brillinger and R. Irizarry. An investigation of the second and higher-order spectra of music. *Signal Processing*, 39(161-179), 1998. 2.1.2

[9] N. I. Durlach. Equalization and cancellation theory of binaural masking level differences. *JASA*, 35:1206–1218, 1963. 3.3

[10] D. Griesinger. General overview of spatial impression, envelopment, localization, and externalisation. In *proceedings of the AES 15th international conference*, pages 136–148, 1999. 1.2

[11] James Robert Hopgood. *Nonstationary signal processing with application to reverberation cancellation in acoustics environments*. PhD thesis, University of Cambridge, Dpt of Engineering, Signal processing laboratory, April 2001. 2

[12] L. A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41:35–39, 1948. 3.1

[13] J.-M. Jot, L. Cerveau, and O. Warusfel. Analysis and synthesis of room reverberation based on a time-frequency model. In *AES 103rd convention preprint*. AES, September 1997. (document)

[14] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition i. simulation of lateralization for stationary signals ii. the law of the first wave front. *JASA*, 80(6):1608–1630, December 1986. 3.3

[15] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41:1–24, 2001. 5

[16] Oppenheim and Shafer. *Discrete-Time signal processing*. Prentice-Hall, second edition only edition, 1989. 2.1.1

[17] A.P. Petropulu and Subramaniam S. Cepstrum based deconvolution for speech dereverberation. In *proceedings of the ICASSP 94, vol. 1*, pages 9–13, 1994. 2.1.1

[18] Dinh-Tuan Pham and Jean-François Cardoso. Blind separation of instantaneous mixtures of non stationary sources. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland*, pages 187–192, June 2000. 5

[19] J.-D. Polack, H. Alrutz, and M. R. Schroeder. The modulation transfer function of music signal and its applications to reverberation measurement. *Acustica*, 54:257–265, 1984. (document)

[20] M. R. Schroeder. New method for measuring reverberation time. *JASA*, 37:409–412, 1965. 4.1

[21] M. Tohyama, R.H. Lyon, and T. Koike. Reverberant phase in a room and zeros in the complex frequency plane. *JASA*, 89(4):1701–1707, April 1991. 2.2.3

[22] Gabriel Weinreich. Directional Tone Color. *JASA*, 101(4):2338–2346, April 1997. 1.1

E

Publication 2

PITCH-TRACKING OF REVERBERANT SOUNDS, APPLICATION TO SPATIAL DESCRIPTION OF SOUND SCENES

ALEXIS BASKIND AND ALAIN DE CHEVEIGNÉ

IRCAM, 1 place Igor-Stravinsky, 75004 Paris, France
Alexis.Baskind@ircam.fr
Alain.de.Cheveigne@ircam.fr

Fundamental frequency (F0) is useful as a perceptually-relevant sound descriptor, and also as an ingredient for signal processing applied to analysis of sound scenes. Here, a recently proposed multiple-F0 algorithm is adapted to handle reverberation in monophonic or multichannel recordings; the information that is obtained from it is then applied to estimation of reverberation time from recorded musical signals.

INTRODUCTION

Fundamental frequency ("F0") estimation is an initial step in many systems for the analysis of complex sound scenes, such as speech recognition, score following, low-bitrate coding of musical signals, etc... Many algorithms have been developped for this purpose. The great majority of them rely on time-frequency or time-lag analysis. Most of those techniques assume a single periodic signal at each instant, and thus are designed for monodic signals. Their behavior in the presence of background noise depends mainly on the signal-to-noise ratio and the decorrelation between signal and noise, the latter being often assumed stationary. Among these techniques, a recent pitch-tracker, called YIN, has proven to be robust and efficient, and is also fast enough to be implemented in real-time [9].

The presence of reverberation makes F0 estimation task more difficult, as its spectral structure competes with that of the direct sound. Thus most pitch-tracking devices fail to accurately estimate the fundamental frequency of reverberant sounds, especially at transients. Figure 1 shows an example of this breakdown, for the very first seconds of a recording of Jean Sebastien Bach's"partita" for solo flute: at the top the estimate provided by YIN for the dry recording, and at bottom the estimate from a reverberant version of this recording, synthesized by convolution with an artificial impulse response (the reverberation time at low frequencies was approximately 1.5 second, and the clarity indexC_{80} was +6 dB). What can be seen is a blurring of the estimate, especially when notes are close to each other in time and/or frequency. This is undoubtedly due to the presence of both current direct sound and reverberation of the preceding notes. It is an obstacle for any kind of further analysis that requires an accurate estimation of running fundamental frequency.

A closer look at the reverberation gives a clue to overcome this problem: since the tail of the impulse response is made of the superimposition of partially coherent echoes

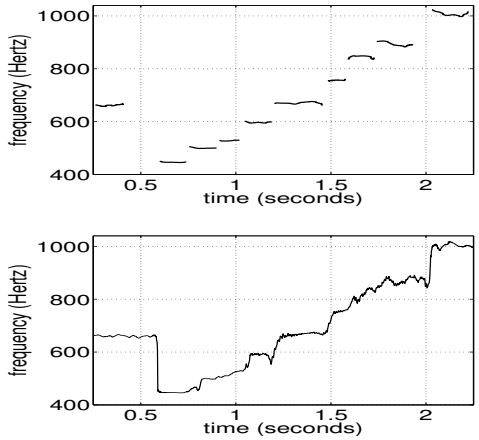


Figure 1: single-f0 estimation of monodic music, in dry (*top*) and reverberant (*bottom*) conditions

of the direct sound, the periodicity of the latter provides a strong constraint on the spectral content of the former. Although reverberation of a harmonic signal cannot be considered strictly harmonic[1], the assumption of approximate harmonicity appears reasonable (see figure 2).

The aim of this study is to provide a fundamental-frequency estimator suitable for reverberant monodic recordings, that takes into account the specific behavior of such signals. For that purpose, we adapt a recently-developped extension of YIN to multiple-F0 estimation, called MMM [8], to the task of estimating the F0 of the direct and reverberated parts of a monodic recording. MMM works by joint cancellation of the various harmonic sources present, by searching through a two-dimensional lag space. The coordinates of the minimum of the cancellation residual provide estimatesof the two periods in the signal. In our

[1] Actually, inharmonicity of the reverberant decay seems to be a relevant cue for estimating reverberation time[16].

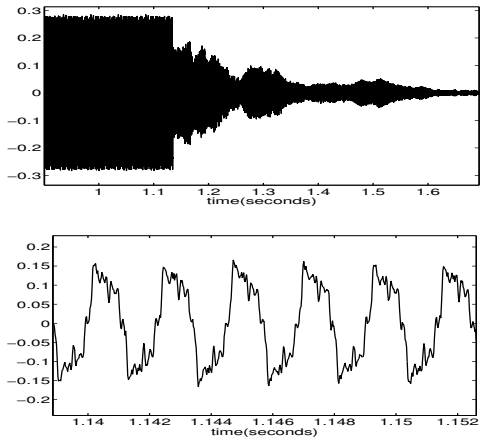


Figure 2: Global shape (top) and detail (bottom) of a reveberated square wave (f0=440 Hz)

case, one estimate is assigned to the direct sound, the other to the reverberation. To ease estimation, the algorithm takes into account the nature ofreverberated sound by constraining the parameter assigned to the reverberation to be consistent with previous estimates of the F0 of the direct sound. Estimation is made yet more reliable by working on both channels of a stereo recording.

F0 estimates, once obtained, are used to tune comb filters to isolate direct and reverberant streams one from another, in order to perform analysis of the spatial features of the scene. As a detailed example, a method for estimating reverberation time from musical signals is proposed here, which is based on the derivation of short-time pitch-synchronous spectra of such isolated streams. An analysis of the decay is performed on each channel of those spectra between time limits that are defined with the knowledge of interchannel pitch-synchronous short-time coherence.

Knowledge of reverberation characteristics, as well as other spatial features of the scene, is of use for applications such as production and post-production of multichannel sound (such as automated mixing, or cinema dubbing), or indexation of binaural and multichannel recordings in databases [5].

1. DOUBLE-F0 ESTIMATION FOR REVERBERANT MUSIC

1.1. Cancellation model for double-f0 estimation

YIN and MMM are related to a cancellation model of pitch perception [7] similar to Licklider's traditionnal autocorrelation model [13]. Autocorrelation and cancellation are both physiologically plausible, and have many similarities. However, the cancellation model has a useful feature for our purpose, in that the cancellation residuals can be used to judge the quality of the estimation.

The principle of double cancellation is illustrated in figure3.

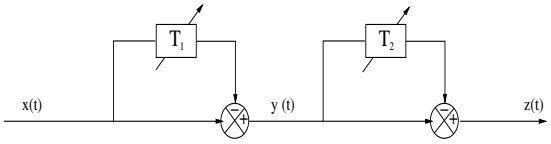


Figure 3: Double cancellation model

The principle is the following. Supposing that $x(t)$ is the signal to estimate, consider the signal

$$z(t) = x\,(t) - x\,(t - T_1) - x\,(t - T_2) + x\,(t - T_1 - T_2)$$

that results from comb-filtering $x(t)$ successively by lags T_1 and T_2. Denoting the instantaneous power of the cancellation residue $P_z(t)$:

$$P_z(t) = \sum_{i=t}^{t+W} z^2(i)$$

, the algorithm looks for the lags T_1 and T_2 that minimize this power. The value of the power as a function of T_1 and T_2 is called **double difference function** (DDF):

$$ddf\,(t, T_1, T_2) = P_z(t)$$

The DDF is thus a running bi-dimensional pattern. An example of this pattern is provided in figure4 The couple of lags that minimize the DDF constitute the estimates of the periods of the two harmonic sources that are assumed to be mixed. The significant number of possible minima (dark zones in figure 4) reveals an ambiguity that is inherent to the method: if the pair (T_1, T_2) cancels the difference function, then any pair $(m.T_1, n.T_2)$, with m and n being positive integers, will cancel it too. Thus, if this feature is not taken into account, the algorithm will tend to select sub-octaves of the note instead of the actual fundamental frequency. The method that is choosen to overcome this problem is to select the pair that minimizes efficiently the d.d.f., and at the same time provides the smallest lags.

The quality of the estimation, as well as the relevance of this assumption, can be evaluated thanks to the following ratios, called "aperiodicity measures", all bounded between 0 and 1:

$$ap(t) = P_z(t)/P_x(t)$$

$$ap_1(t) = P_z(t)/P_{y_1}(t)$$

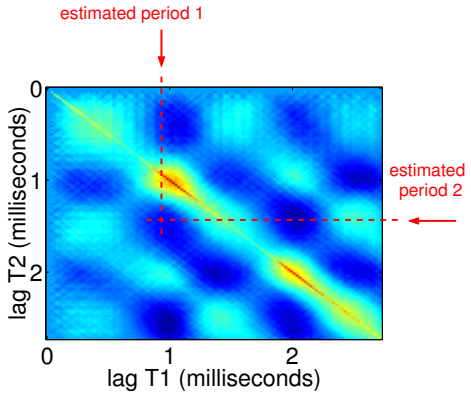


Figure 4: An example of the double difference function for a given time frame of the signal. Arrows indicate the two estimated periods. Dark zones correspond to lower values, whereas light zones correspond to higher values.

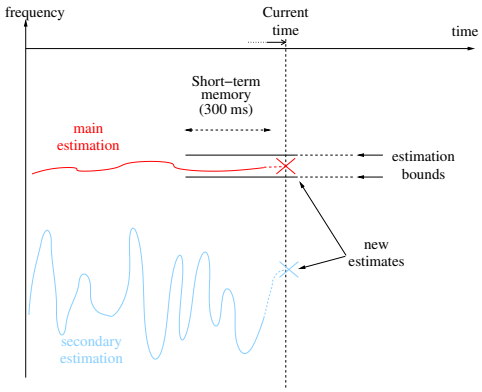


Figure 5: Double-F0 estimation for reverberant monodic sounds. Frequency bounds are determined by the prominent frequency in the last 300 ms.

$$ap_2(t) = P_z(t)/P_{y_2}(t)$$

...where $y_1(t)$ and $y_2(t)$ are the signals that result from a single cancellation, i.e. $y_1(t) = x(t) - x(t - T_1)$ and $y_2(t) = x(t) - x(t - T_2)$. The first ratio $ap(t)$ allows to evaluate the quality of the double-f0 model taken as a whole, whereas $ap_1(t)$ and $ap_2(t)$ are useful to compare the estimation to the single-f0 estimation that YIN provided. All this data, added to YIN data, is processed by a decision module, which decides which of the models (i.e. one or two harmonic sources) is the most accurate, and what is (or are) the fundamental frequency(ies) of the corresponding source(s).

1.2. The case of reverberant sounds

This algorithm is more general than necessary in the case that concerns us, and also somewhat fragile. As the reverberated sound is not perfectly periodic, the standard MMM algorithm may fail. It may for example split the reverberated sound into parts and incorrectly assign F0s to each part. In our case, there are strong constraints between the F0s of direct and reverberated sounds, and we modified the MMM algorithm to incorporate these constraints.

However, this problem can be overcome, taking into account the fact that the running fundamental frequency of the reverberant tail of a sound (assuming local harmonicity) is quite close to the actual fundamental frequency of the sound itself. Thus, by constraining one of the estimations to rely within bounds that would be determined by the main F0 estimation in the very near past (several hundreds of milliseconds), we expect the algorithm to detect the reverberant stream more efficiently (see figure 5).

As a practical example, is shown on figure 6 the result of the double-F0 estimation of the same short excerpt as in introduction, and comparison with single-f0 estimation: among others transients are precisely detected, thus most of notes can be distinguished anew as discrete events with a nearly constant fundamental frequency.

In the case of two-channel or multichannel recordings, we can also benefit from the redundant information on fundamental frequency over all recording channels that contain enough direct sound. Different approaches could be employed in order to use this additional information. Ours at present relies on the same overall principle as the single channel estimation, but whereas a separate single- and double-F0 estimation is performed on each channel, the decision module is common to all channels, providing a unique fundamental frequency estimation at a given time, corresponding to the lowest residual aperiodicity. It is also worth noticing that the cancellation model, when generalized to two or several channels, is suitable for estimating the delays between channels. Thus, a possible extension of this architecture would be for instance a binaural signal detector, which could at the same time estimate the localization of the source (or at least its lateralization) and its pitch.

2. APPLICATION TO RUNNING ESTIMATION OF REVERBERATION TIME

2.1. The problem of estimating reverberation time from music

The idea of deriving reverberation time from musical signals is not new, since it is one of the ways to solve the major problem of calculating RT in occupied halls. As a matter-of-fact, traditional impulse response measurements, using pseudo-random noises or short impulses such as gun shots, cannot be used in this case, since it cor-

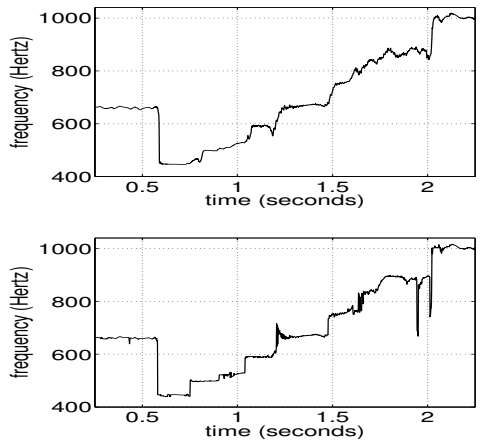


Figure 6: comparison between single-f0 (*top*) and double-f0 (*bottom*) estimation of monodic reverberant music

responds most of the time to the situation of a concert. As an alternative to extrapolating estimates from measurements made in the empty hall [3, 6], many acousticians tried to derive reverberation time directly from the music that is diffused during the concert. One approach for instance is to compare modulation transfer functions measured close to the musicians and in the audience [14]. Another is to observe the shape of the autocorrelation envelope [10], or to analyze the decay after"stop-chords" in the music [6]. This latter method, which does not need precise knowledge of the source signal, may be very useful, but also very sensitive to the "quality" of the silence [5]. Its main limitation is that it needs to focus on stop-chords, or, equivalently, any broadband signal burst that stops suddenly. Such moments are not frequent in the signal, and may be also hard to detect automatically without an analysis module that is consecrated to this specific purpose.

The solution that is proposed here is inspired from our everyday experienceof recorded music. Late reverberation is perceptible during complete silences, but also after sudden frequency changes when the source is narrowband or harmonic.

A useful application of the pitch-tracker presented above can thus be foreseen. Assuming that the fundamental frequency of the source signal is known with good accuracy, we can use it in at least in two different (and complementary) ways.First, the frequency bands in which reverberation actually occurs are known, since they correspond to the fundamental frequency of the present note and its harmonics; second, the fundamental frequencies of the previous and subsequent notes can be used to cancel them so that the decay can be measured. This latter operation, which could be performed using basic first-order comb filters, is a major help in our attempt to isolate the decay.

2.2. Pitch-synchronous time-frequency analysis of reverberant decays

Our method for estimating reverberation time relies on an accurate time-frequency analysis. The standard short-time Fourier transform with a constant number of frequency channels is convenient and allows the use of the well-known FFT optimized algorithm. However, since the instantaneous pitch of the direct sound is known, it is possible to achieve a better accuracy by using pitch-synchronous analysis. The principle remains the same, except that the number of frequency channels now depends on the main fundamental frequency of the actual note, in an attempt to make each channel matching the best the corresponding harmonic of the note.

Short-time Fourier spectrum

Short-time Fourier spectrum is often used as a basic ingredient for describing reverberation in narrow bands [11]. It has been chosen here mainly because it allows pitch-synchronous analysis, but may be replaced with success by other techniques, such as modified discrete cosine transform, as well as constant-Q or ERB filterbanks.

The short-time Fourier spectrum of the raw signal is difficult to exploit for reverberation analysis because of the overlapping of successive notes. For this reason, the signal is first preprocessed by applying one or more comb-filters with delays that match the periods of the preceding and following notes. On figure 7 is provided an example of the performance of this quite simple method: the considered reverberated note of this flute recording is the first 'A' (446 Hz), where as the previous one as well as the next ones are to be cancelled by this preprocessing stage. The comparison of the spectra at the fundamental frequency shows that despite a significant decrease (20 dB !) of the global level, the signal-to-noise ratio is greatly enhanced: cancelling allowed to reveal the onset of the note (arround 0.2 seconds), as well as the exponential decay (between 0.2 and 0.5 seconds) without ambiguity, with much greater dynamic and time ranges.

Linear regression for RT estimation

Inspiring from traditionnal methods for measuring RT from impulse responses, we estimate it in this case by finding the straight line that best approximates the decay. However (and as usual), the precision that may be achieved by directly applying a linear regression method over short-time Fourier spectrum segments is quite poor, especially at low frequencies. This is due to oscillations in the decay that correspond to the stochastic nature of reverberation. A well known method to reduce this variance is backward integration of instantaneous power. This technique, that has been proposed by Schroeder [15] in order to re-

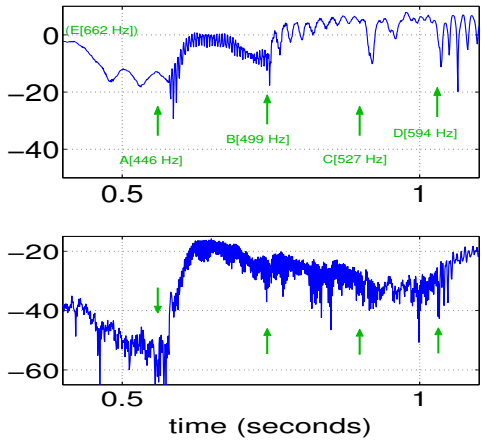


Figure 7: Example of the efficiency of pitch-synchronous comb-filtering: energy within the fundamental frequency band. Arrows indicate onset times of actual and following notes, with their corresponding name, as well as fundamental frequency in brackets. *Top*: without filtering preceding and following notes. *Bottom*: with filtering

duce the number of measurements that are needed to derive reverberation time, was designed for octave or third-octave band-filtered impulse responses, but has been applied with success to narrow-band representations [11, 4]. Applying it on segments of musical signals instead of impulse responses makes sense, but entails some additional difficulties that are related to two reasons: first, the probable presence of background noise during the estimation, which is most of time not stationary (contrary to noise in impulse response measurements) since it corresponds to other sources and to reverberation; second, the necessity to define accurately times for the beginning and the end of the analysis. Both problems can be handled efficiently thanks to an additional cue, short-time coherence.

Short-time interchannel coherence

The time-domain definition of coherence between two signals $x(m)$ and $y(m)$ for time frame m and frequency channel k may be defined as follows:

$$C_{xy}\,(m,k) = \frac{|\Phi_{xy}(m,k)|}{\sqrt{\Phi_{xx}(m,k).\Phi_{yy}(m,k)}}$$

In this equation, $\Phi_{xy}(m,k)$ represents the short-time cross-spectrum between the two channels, and $\Phi_{xx}(m,k)$ and $\Phi_{yy}(m,k)$ are the short-time autospectra. Short-time cross- and auto-spectra are defined by the following recursive equations:

$$\Phi_{xy}(m,k) = \lambda.\Phi_{xy}(m-1,k) + X(m,k).Y^*(m,k)$$

$$\Phi_{xx}(m,k) = \lambda.\Phi_{xx}(m-1,k) + |X(m,k)|^2$$

$$\Phi_{yy}(m,k) = \lambda.\Phi_{xy}(m-1,k) + |Y(m,k)|^2$$

where $X(m,k)$ and $Y(m,k)$ are the (complex) short-time Fourier transforms of signals $x(m)$ and $y(m)$, respectively. Those definitions of cross- and auto-spectra corresponds to the convolution of the instantaneous cross- or auto-spectra with an exponential window $w(m) = e^{-m/\tau}$ which damping coefficient τ equals $-1/ln(\lambda)$.

The use of interchannel coherence for the analysis of reverberant signal is based on the simple assumption that cross-correlation (and thus coherence) between two microphones in a room closely depends on time: when the direct sound or the early specular reflections reaches the microphones, the correlation is maximum and thus close to 1, since the signals in both microphones only differ in time and level[2], whereas as time passes, the reflections that reach the microphones are more and more numerous and diffuse, so that the interchannel coherence gradually falls to a minimum that roughly corresponds to the theoretical interchannel coherence in a diffuse field, that is

$$C_{th}(k) = \frac{sin\,(kd)}{kd}$$

,where k is the wave number for given frequency, and d is the distance between microphones [12].

Short-time interchannel coherence has been proposed by Allen et al. [1] as a way to estimate the direct-to-reverberant ratio in each frequency channel, in order to perform dereverberation of speech. Avendano et Jot [2] used it more recently to separate reverberation from direct sound, and send the former to rear channels in a 5.1 context.

Figure 8 shows an example of successful use of short-time Fourier spectrum and interchannel coherence to the estimation of reverberation time, on a binaural recording: interchannel coherence is, as expected, maximum on the signal onsets (three onsets are not visible in this channel, partly thanks to previous comb-filtering), and falls regularly to a minimum value of 0.4. On the basis of this representation, the start and stop limits for analysing the decay may be defined, by looking at the times where coherence falls below a given threshold (in that case, 0.95), and just before the next onset. If the time limits are adequately defined, the backward integration provides a quite linear shape, and linear regression may be performed with a good precision.

In order to prevent estimations over segments which are too corrupted by background noise, several quality requirements on spectrum and coherence have been defined, that allows to decide if a given narrow-band segment may be accepted or has to be rejected: first, the dynamic range

[2] This is not exactly the case in binaural listening, since head shadow entails differences in time and level that depend on frequency, when the incident wave is not frontal. However, this assumption remains reasonable, especially when comparing interaural coherence in early and late parts of the room effect.

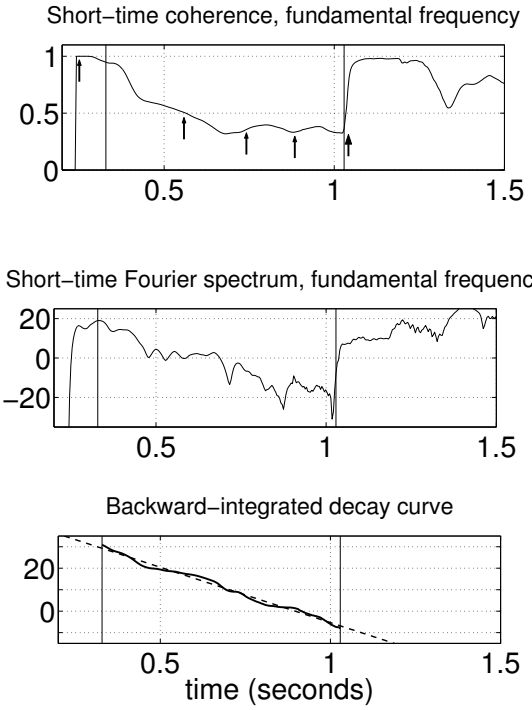


Figure 8: The principle of running estimation of reverberant decay on a specific example. All plots concern the frequency channel corresponding to fundamental frequency of the actual note (662 Hz). Vertical lines are calculated limits for regression. Arrows on top figure are onset times of actual and following notes (after comb-filtering). *Top*: short-time coherence. *Middle*: short-time Fourier spectrum. *Bottom*: backward integration of Fourier spectrum

of the spectrum during the decay must be sufficient; second, coherence must be high enough just after the onset, and low enough in the late part of the decay; third, the time window between that has been estimated for the analysis must be long enough.

It is worth noticing that, whereas parameters used in short-time Fourier transform are not critical towards the quality of the representation for our purpose, the accuracy of the estimation is very sensitive to the length of the averaging window that is used for the estimation of interchannel coherence: it was found from our experiments that an exponential window with a damping time of 100ms provides reliable results.

2.3. Practical example

This method is applied on a longer excerpt (22 seconds) of the example used above: the dry recording that has been used, which is a recording solo flute close to the musician, has been spatialized by convolution with a set of synthetic binaural room impulse responses, with the following objective measurements: clarity index C_{80} equals 6.05 dB, and in the frequency range of interest, mean early decay time EDT_{10} is 1 second, and mean RT_{30} is 1.5 seconds. The two-channel "binaural" recording that is obtained is processed first by the pitch-tracker, and then by the reverberation time estimation device. Among the 330 notes of this recording, 111 where choosen for their sufficient signal-to-noise ratio, providing 120 narrow-band segments that match the requirements on coherence and dynamic range.

All the estimates were gathered in third-octave bands. For each band, the mean value as well as the standard deviation is computed. The standard deviation provides an estimation of the "confidence interval" (this term is not fully accurate, since the estimation for each band is not gaussian). On figure 9 are shown the results of this analysis, as well as the actual reverberation time RT_{30} and early decay time EDT_{10}, derived directly in third-octave bands from the impulse responses.

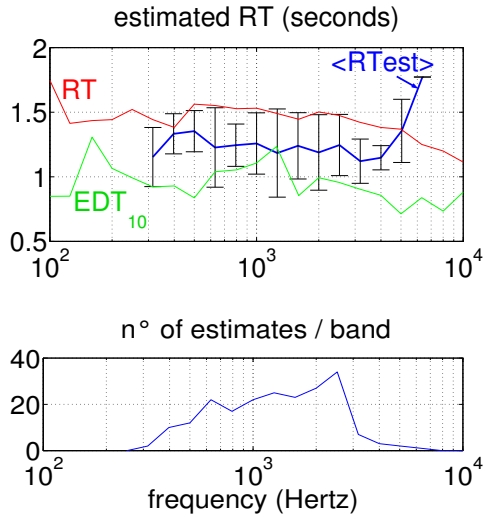


Figure 9: Results of the estimation over a 22 seconds excerpt. *Top*: mean estimation and "confidence intervals", compaired with actual RT_{30} and EDT_{10}. *Bottom*: number of estimates in each third-octave band

Not considering the two higher bands, which estimations are not at all correct, mainly because of the lack of estimations at those frequencies, it is obvious from this plot that most of the estimated decay times are comprised between early decay time and reverberation time. A closer look at the data showed us that the longer a given segment is, the better the estimation matches the reverberation time, and the shorter this segment is, the better the estimation matches the early decay time. Those results are consistent with physics and perception of reverberation: as a matter-of-fact, running reverberance, to which

EDT corresponds, is audible anytime during the signal as soon as it is not continuous, and late reverberance, which is better described by RT, is audible only during silences or sudden pitch changes.

CONCLUSION

This paper presents methods that aim at deriving two perceptually-relevant objective descriptors of a reverberant sound scene: the pitch, and the reverberation time. The pitch-tracker that is proposed here provides very encouraging results; the estimation itself, as well as the decision module are constantly improved in order to provide more accurate results, but it is at this stage anyway better suitable for reverberant signals than usual single-voice pitch-trackers. The method for deriving reverberation time that was developped shows remarkable adequacy to traditional objective measurements, even if it is not at present able to distinguish actual reverberation time from early decay time. Further work is achieved in that direction, which mainly consists in finding the adequate setup of analysis parameters for each purpose.
The use of the fundamental frequency information in deriving a spatial description is not limited to the estimation of reverberation time. Potentially it provides a useful source of information for a more complete analysis of spatial features of a sound scene, such as onset detection and localization.

REFERENCES

[1] J. Allen, D. Berkley, and J. Blauert. Multimicrophone signal processing technique to remove room reverberation of speech signals. *J. Acoust. Soc. Am.*, 62(2):912–915, 1977.

[2] C. Avendano and J.-M. Jot. Frequency domain techniques for stereo to multichannel upmix. In *Proc. AES 22nd international conference "Virtual, synthetic and entertainment audio", Espoo, Finland*, pages 121–130, June 2002.

[3] M. Barron. *Auditorium acoustics and Architectural Design*. E & FN Spon/Chapman & Hall, 1993.

[4] A. Baskind and J.-D. Polack. Sound Power Radiated by Sources in Diffuse Field. In *proc. AES 108th convention*, February 2000.

[5] A. Baskind and O. Warusfel. Methods for blind computational estimation of perceptual attributes of room acoustics. In *proc. AES 22nd international conference "Virtual, synthetic and entertainment audio", Espoo, Finland*, June 2002.

[6] L. Beranek. *Concert and opera halls : how they sound*. Acoustical Society of America, 1996.

[7] A. de Cheveigné. Cancellation model of pitch perception. *J. Acoust. Soc. Am.*, 103(3):1261–1271, march 1998.

[8] A. de Cheveigné and A. Baskind. F0 estimation of one or several voices. In *proc. Eurospeech (submitted)*, 2003.

[9] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, April 2002.

[10] M. Hansen. A method for calculating reverberation time from musical signals. Technical report, report from the Acoustics Laboratory, Technical University of Denmark, 1995. Report no 60, ISSN 0105-3027.

[11] J.-M. Jot, L. Cerveau, and O. Warusfel. Analysis and synthesis of room reverberation based on a time-frequency model. In *proc. AES 103rd convention*. AES, September 1997.

[12] H. Kuttruff. *Room Acoustics*. SPON Press, 4th edition, 2000.

[13] J.C.R Licklider. A duplex theory of pitch perception. *Experientia*, 7(4):128–132, 1951.

[14] J.-D. Polack, H. Alrutz, and M. R. Schroeder. The modulation transfer function of music signal and its applications to reverberation measurement. *Acustica*, 54:257–265, 1984.

[15] M. R. Schroeder. New method for measuring reverberation time. *J. Acoust. Soc. Am.*, 37:409–412, 1965.

[16] M. Wu and D. Wang. A one-microphone algorithm for reverberant speech enhancement. To be presented in ICASSP2003, Hong Kong, 2003.

Bibliographie

AFNOR. Acoustique : mesurage de la durée de réverbération des salles en référence à d'autres paramètres acoustiques (norme internationale NF EN ISO 3382 :2000), Mai 2000.

V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, et Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.*, 112(5, pt. 1) :2053–2064, 2002.

V. R. Algazi, R. O. Duda, R. P. Morrison, et D. M. Thompson. Structural composition and decomposition of HRTF's. In *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, Mohonk Mountain House, New Paltz, NY*, pages 99–102, Octobre 2001a.

V. R. Algazi, R. O. Duda, D. M. Thompson, et C. Avendano. The CIPIC HRTF database. In *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, Mohonk Mountain House*, pages 99–102, Octobre 2001b.

J. Allen, D. Berkley, et J. Blauert. Multimicrophone signal processing technique to remove room reverberation of speech signals. *J. Acoust. Soc. Am.*, 62(2) :912–915, 1977.

J. B. Allen et D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4) :943–950, 1979.

C. Avendano et J.-M. Jot. Frequency domain techniques for stereo to multichannel upmix. In *Proc. AES 22nd international conference "Virtual, synthetic and entertainment audio", Espoo, Finland*, pages 121–130, Juin 2002.

M. Barron. Objective measures of spatial impression in concert halls. In *proc. 11th International Congress on Acoustics*, volume 7, pages 105–108, 1983.

M. Barron. *Auditorium acoustics and Architectural Design*. E & FN Spon/Chapman & Hall, 1993.

M. Barron et A.H. Marshall. Spatial impression due to early lateral reflections in concert halls : the derivation of a physical measure. *J. Sound Vib.*, 77(2) :211–232, 1981.

Michael Barron. The subjective effects of first reflections in concert halls - The need for lateral reflections. *J. Sound Vib.*, 15(4) :475–494, 1971.

A. Baskind. Etude de la puissance acoustique rayonnée en champ diffus. mémoire de dea d'acoustique, traitement de signal et informatique appliqués à la musique, Université Aix-Marseille II, 1999.

A. Baskind et A. de Cheveigné. Pitch-Tracking of Reverberant Sounds, Application to Spatial Description of Sound Scenes. In *Proc. AES 24th conference "Multichannel Audio - The New Reality"*, Juin 2003.

A. Baskind et J.-D. Polack. Sound Power Radiated by Sources in Diffuse Field. In *proc. AES 108th convention*, Février 2000.

A. Baskind et O. Warusfel. Methods for blind computational estimation of perceptual attributes of room acoustics. In *proc. AES 22nd international conference "Virtual, synthetic and entertainment audio", Espoo, Finland*, Juin 2002.

D. Bees, M. Blostein, et P. Kabal. Reverberant speech enhancement using cepstral processing. In *Proceedings of ICASSP 91, vol. 2*, pages 977–980, 1991.

L. Beranek. *Music, Acoustics and Architecture*. John Wiley and Sons, 1962.

L. Beranek. *Concert and opera halls : how they sound*. Acoustical Society of America, 1996.

J. Berg et F. Rumsey. Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *proc. Audio Engineering Society 19th int. conf., Schloss Elmau, Germany*, Juin 2001.

J. Blauert. Localization and the law of the first wavefront in the median plane. *J. Acoust. Soc. Am.*, 50(2, pt. 2) :466–470, Août 1971.

J. Blauert. *Spatial hearing - the psychophysics of human sound*. MIT press, 1997.

J. Blauert et W. Cobben. Some consideration of binaural cross correlation analysis. *Acustica*, 39 :96–104, 1978.

J. Blauert et W. Lindemann. Auditory spatiousness : some further psychoacoustic analyses. *J. Acoust. Soc. Am.*, 80(2) :533–542, Août 1986a.

J. Blauert et W. Lindemann. Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *J. Acoust. Soc. Am.*, 79(3) :806–813, Mars 1986b.

Barry Blesser. An interdisciplinary synthesis of reverberation viewpoints. *JAES*, 49(10), Octobre 2001.

M. (1993) Bodden. Modeling Human Sound Source Localization and the Cocktail-Party-Effect. *Acta Acustica*, 1(1) :43–55., 1993.

J. S. Bradley et G. A. Soulodre. The influence of late arriving energy on spatial impression. *J. Acoust. Soc. Am.*, 97(4) :2263–2271, Avril 1995a.

J. S. Bradley et G. A. Soulodre. Objective measures of listener envelopment. *J. Acoust. Soc. Am.*, 98(5 pt. 1) :2591–2597, Novembre 1995b.

J.S. Bradley, R.D. Reich, et S.G. Norcross. On the combined effects of early- and late-arriving sound on spatial impression in concert halls. *J. Acoust. Soc. Am.*, 108(2) :651–661, Août 2000.

Dirk Jeroen Breebaart. *Modeling binaural signal detection*. Thèse de Doctorat, Eindhoven university of technology, Juin 2001.

C. P. Brown et R. O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Processing*, 6(5) :476–488, Septembre 1998.

G. Brown et M. Cooke. Computational auditory scene analysis. *Comput. Speech Lang.*, 8(4) : 297–336, 1994.

J. C. Chen, K. Yao, et R. E. Hudson. Source localization and beamforming. *IEEE signal processing magazine*, 19(2), 2002.

A. de Cheveigné. Separation of concurrent harmonic sounds : Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93(6) :3271–3290, Juin 1993.

A. de Cheveigné. Cancellation model of pitch perception. *J. Acoust. Soc. Am.*, 103(3) :1261–1271, march 1998.

A. de Cheveigné et A. Baskind. F0 estimation of one or several voices. In *proc. Eurospeech*, 2003.

A. de Cheveigné et H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4) :1917–1930, Avril 2002.

W. Chung, S. Carlile, et P. Leong. A performance adequate computational model for auditory localization. *J. Acoust. Soc. Am.*, 107(1) :432–445, Janvier 2000.

P. D. Coleman. Failure to localize the source distance of an unfamiliar sound. *J. Acoust. Soc. Am.*, 44(3) :345–346, Mars 1962.

R. K. Cook, R.V. Waterhouse, S. E. Berendt, et M. C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *J. Acoust. Soc. Am.*, 27(6) :1072–1077, Novembre 1955.

L. Cremer. Early reflections in some modern concert halls. *J. Acoust. Soc. Am.*, 85(3) :1213–1225, Mars 1989.

R. E. Crochiere et L. R. Rabiner. *Multirate digital signal processing*. Prentice-Hall, 1983.

J.F. Culling et Q. Summerfield. Perceptual separation of concurrent speech sounds : absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.*, 98(2 pt.1) : 785 – 797, 1995.

J. Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Thèse de Doctorat, Université Paris 6, Juin 2000.

N. I. Durlach. Equalization and cancellation theory of binaural masking level differences. *J. Acoust. Soc. Am.*, 35 :1206–1218, 1963.

Daniel P. W. Ellis. *Prediction-driven computational auditory scene analysis*. Thèse de Doctorat, MIT, 1996.

W. E. Feddersen, T. T. Sandel, D. C. Teas, et L. A. Jeffress. Localization of high-frequency tones. *J. Acoust. Soc. Am.*, 29(9) :988–991, Septembre 1957.

F. Freeland, L. Biscainho, et P. Diniz. Efficient HRTF Interpolation in 3D Moving Sound. In *proc. AES 22nd int. conf. "Virtual, synthetic and entertainment audio", Espoo, Finland*, Juin 2002.

W. Gaik. Combined evaluation of interaural time and intensity differneces : Psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94(1) :98–110, July 1993.

B. Gardner et K. D. Martin. HRTF Measurements of a KEMAR Dummy-Head Microphone. Technical Report technical report #280, MIT Media Laboratory Perceptual Computing, 1994. URL `http://sound.media.mit.edu/KEMAR.html`.

M. B. Gardner et R. S. Gardner. Problem of localization in the median plane : effect of pinnae cavity occlusion. *J. Acoust. Soc. Am.*, 53(2) :400–408, Février 1973.

M.B. Gardner. Some Monaural and Binaural Facets of Median Plane Localization. *J. Acoust. Soc. Am.*, 54(6) :1489–1495, Décembre 1973.

D. Griesinger. The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acta Acustica*, 83 :721–731, 1997.

D. Griesinger. Objective Measures of Spaciousness and Envelopment. In *Proceedings of the 16th international conference on Spatial Sound Reproduction*, pages 27–41, Avril 1999.

M. Hansen. A method for calculating reverberation time from musical signals. Technical report, report from the Acoustics Laboratory, Technical University of Denmark, 1995. Report no 60, ISSN 0105-3027.

W. M. Hartmann. Localization of sound in rooms. *J. Acoust. Soc. Am.*, 74(5) :1380–1391, Novembre 1983.

W. M. Hartmann. *Signals, sound and sensation*. Springer-Verlag, 1996.

G. B. Henning. Detectability of interaural delay in high-frequency complex waveforms. *J. Acoust. Soc. Am.*, 55(1) :84–90, Janvier 1974.

R. Hershkowitz et N. Durlach. Interaural Time and Amplitude JNDs for a 500-Hz Tone. *J. Acoust. Soc. Am.*, 46 :1464–1467, 1969.

Finn Jacobsen et Thibaut Roisin. The coherence of reverberant sound fields. *J. Acoust. Soc. Am.*, 108(1) :204–210, Juillet 2000.

L. A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, 41 :35–39, 1948.

J.-M. Jot. *Etude et réalisation d'un spatialisateur de sons par modèles physiques et perceptifs*. Thèse de Doctorat, École Nationale Supérieure des Télécommunications, Septembre 1992.

J.-M. Jot, L. Cerveau, et O. Warusfel. Analysis and synthesis of room reverberation based on a time-frequency model. In *proc. AES 103rd convention*. AES, Septembre 1997.

Eckhard Kahle. *Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras*. Thèse de Doctorat, Université du Maine, Le Mans, Juin 1995.

B. F. G. Katz. *Measurement and calculation of individual head-related transfer functions using a boundary element model including the measurement and effect of skin and hair impedance*. Thèse de Doctorat, Pennsylvania state university, Mai 1998.

Gary S. Kendall. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4) :71–87, 1995.

G.F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Am.*, 62(1) :157–167, 1977.

H. Kuttruff. *Room Acoustics*. SPON Press, 4th edition, 2000.

V. Larcher. *Techniques de spatialisation des sons pour la réalité virtuelle*. Thèse de Doctorat, Université Paris 6, Mai 2001.

D. M. Leakey, B. M. Sayers, et C. Cherry. Binaural fusion of low- and high- frequency sounds. *J. Acoust. Soc. Am.*, 30(3) :222, Mars 1958.

W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition I. simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6) :1608–1622, Décembre 1986a.

W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition II. the law of the first wave front. *J. Acoust. Soc. Am.*, 80(6) :1623–1630, Décembre 1986b.

R. Y. Litovsky, M. L. Hawley, et R. M. Dizon. Measurements of precedence phenomena in binaural and monaural conditions. *J. Acoust. Soc. Am.*, 101(5) :3083, 1997. (abstract).

R. Y. Litovsky et N. A. Macmillan. Sound localization precision under conditions of the precedence effect : effects of azimuth and standard stimuli. *J. Acoust. Soc. Am.*, 96(2 pt. 1) :752–758, Août 1994.

R.Y. Litovsky, H.S. Colburn, W.A. Yost, et S.J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106(4) :1633–1654, Octobre 1999.

Q.-G. Liu, B. Champagne, et P. Kabal. Room speech dereverberation via minimum-phase and all-pass component processing of multi-microphone signals. In *proceedings of the ICASSP 95*, pages 571–574, 1995.

R.H. Lyon. Range and frequency dependence of transfer function phase. *J. Acoust. Soc. Am.*, 76(5) :1433–1437, 1984.

E. A. Macpherson et J. C. Middlebrooks. Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited. *J. Acoust. Soc. Am.*, 111(5) :2219–2236, Mai 2002.

K.D. Martin. A computational model of spatial hearing. Mémoire de m. sc., MIT, Juin 1995.

J. Max et J.-L. Lacoume. *Méthodes et techniques de traitement du signal et applications aux mesures physiques*. Masson, 1996.

R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am.*, 79(3) :702–711, Mars 1986.

A. W. Mills. On the minimum audible angle. *J. Acoust. Soc. Am.*, 30(4) :237–246, Avril 1958.

James A. Moorer. About this reverberation business. *Computer Music Journal*, 3(32) :13–28, 1979.

M. Morimoto. The contribution of the two ears to the perception of vertical angle in sagittal planes. *J. Acoust. Soc. Am.*, 109(4) :1596–1603, Avril 2001.

M. Morimoto et Z. Maekawa. Effects of low frequency components on auditory spaciousness. *Acustica*, 66 :190–196, 1988.

A. D. Musicant et R. A. Butler. Influence of monaural spectral cues on binaural localization. *J. Acoust. Soc. Am.*, 77(1) :202–208, Janvier 1985.

R. Nicol. Implémentation et évaluation d'un traitement d'antenne acoustique. Master's thesis, Conservatoire National des Arts et Métiers, 1996.

R. D. Patterson, I. Nimmo-Smith, D. L. Weber, et R. Milroy. The deterioration of hearing with age : Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *J. Acoust. Soc. Am.*, 72(6) :1788–1803, 1982.

R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, et M. Allerhand. *Auditory physiology and perception*, chapter "Complex sounds and auditory images", pages 429–446. Oxford-Pergamon, 1991.

Renato S. Pellegrini. *A virtual reference listening room as an application of auditory virtual environments*. Thèse de Doctorat, Institute of Communications Acoustics, Ruhr - University of Bochum, Germany, 2001.

D. R. Perrot. Role of signal onset in sound localization. *J. Acoust. Soc. Am.*, 45(2) :436–445, Février 1969.

D. R. Perrot, R. Briggs, et S. Perrot. Binaural fusion : its limits as defined by signal duration and signal onset. *J. Acoust. Soc. Am.*, 47(2 part. 2) :565–568, Février 1970.

A. P. Petropulu et S. Subramaniam. Cepstrum based deconvolution for speech dereverberation. In *proceedings of the ICASSP 94, vol. 1*, pages 9–13, 1994.

J.-D. Polack. *Sur la fonction de transfert de la modulation des signaux musicaux*. Thèse de Doctorat, Université Georg-August (Göttingen), 1982.

J. D. Polack. *La transmission de l'energie sonore dans les salles*. Thèse de Doctorat, Université du Maine, 1988.

J. D. Polack. Modifying chambers to play billiards : the foundations of reverberation theory. *Acustica*, 76 :257–272, 1992.

J.-D. Polack. Eigenvalue distribution in rectangular rooms. 2001.

J.-D. Polack, H. Alrutz, et M. R. Schroeder. The modulation transfer function of music signals and its applications to reverberation measurement. *Acustica*, 54 :257–265, 1984.

Jan Potter. *On the binaural modelling of spaciousness in room acoustics*. Thèse de Doctorat, Delft University of Technology, Netherlands, Avril 1993.

V. Pulkki. Analyzing virtual sound source attributes using a binaural auditory model. *JAES*, 47(4) :203–217, April 1999.

B. Rakerd et W. M. Hartmann. Localization of sound in rooms II : the effects of a single reflecting surface. *J. Acoust. Soc. Am.*, 78(2) :524–533, Août 1985.

B. Rakerd et W. M. Hartmann. Localization of sound in rooms III : Onset and duration effects. *J. Acoust. Soc. Am.*, 80(6) :1695–1706, Décembre 1986.

B. Rakerd, J. Hsu, et W. M. Hartmann. The Haas effect with and without binaural differences. *J. Acoust. Soc. Am.*, 101(5) :3083, 1997. (abstract).

Lord Rayleigh. Our perception of the direction of a source of sound. *Nature*, XIV :32–33, 1876.

Lord Rayleigh. Acoustical observations IV. *Philosophical Magazine*, XIII :340–347, 1882.

Lord Rayleigh. On our perception of sound direction. *Philosophical magazine*, XIII :214–232, 1907.

X. Rodet et F. Jaillet. Detection and modeling of fast attack transients. In *proc ICMC'01, La Habana, Cuba*, Septembre 2001.

S. K. Roffler et R. A. Butler. Factors that influence the localization of sound in the vertical plane. *J. Acoust. Soc. Am.*, 43(6) :1255–1259, Juin 1968a.

S. K. Roffler et R. A. Butler. Localization of tonal stimuli in the vertical plane. *J. Acoust. Soc. Am.*, 43(6) :1260–1266, Juin 1968b.

F. Rumsey. *Spatial Audio*. Focal Press, 2001.

F. Rumsey. Spatial quality evaluation for reproduced sound : terminology, meaning and a scene-based paradigm. *JAES*, 50(9) :651–666, Septembre 2002.

W. C. Sabine. *Collected papers on acoustics*. Peninsula Publishing (réédité en 1993), 1922.

B. McA. Sayers et E. C. Cherry. Mechanism of binaural fusion in the hearing of speech. *J. Acoust. Soc. Am.*, 29(9) :973–987, Septembre 1957.

C. Schauer et P. Paschke. A spike-based model of binaural sound localization. *International Journal of Neural Systems*, 9(5), 1999.

M. R. Schroeder. New method for measuring reverberation time. *J. Acoust. Soc. Am.*, 37 : 409–412, 1965.

M. R. Schroeder. Integrated impulse method measuring sound decay without impulses. *J. Acoust. Soc. Am.*, 66 :497–500, 1979.

M. R. Schroeder. Modulation transfer function : definition and measurement. *Acustica*, 49 : 179–182, 1981.

M.R. Schroeder et J.L Hall. Model for Mechanical to Neutral Transduction in the Auditory Receptor. *J. Acoust. Soc. Am.*, 5 :1055 – 1060, Mai 1974.

E. D. Schubert et J. Wernick. Enveloppe versus microstructure in the fusion of dichotic signals. *J. Acoust. Soc. Am.*, 45(6) :844–849, Juin 1969.

C.L. Searle, L.D. Braida, M.F. Davis, et H.S. Colburn. Model for auditory localization. *J. Acoust. Soc. Am.*, 60(5) :1164–1175, Novembre 1976.

M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical report, Apple Computer - Perception Group, 1993.

M. Slaney et R.F. Lyon. *On the importance of time - A temporal representation of sound in "Visual Representations of Speech Signals"*, pages pp. 95–116. John Wiley and sons, 1993.

Malcolm Slaney. Lyon's cochlear model. Technical report, Apple, 1988.

S. S. Stevens et E. B. Newman. The localization of actual sources of sound. *Amer. J. Psychol.*, 48 :297–306, 1936.

B. Supper, T. Brookes, et F. Rumsey. A new approach to detecting auditory onsets within a binaural stream. In *proc. AES 114th convention*, Mars 2003.

G. Theile et G. Plenge. Localization of lateral phantom sources. *JAES*, 25 :196–200, 1977.

Harvey Thornburg et Fabien Gouyon. A flexible analysis-synthesis method for transients. In *proc. ICMC-00*, 2000.

M. Tohyama. *Encyclopedia of acoustics vol.2*, chapter 77, "Response statistics of rooms", pages 913–923. Malcolm J. Crocker, 1997.

M. Tohyama et R. H. Lyon. Zeros of a transfer function in a multi-degree-of-freedom vibrating system. *J. Acoust. Soc. Am.*, 86(5) :1854–1863, Novembre 1989.

M. Tohyama, R.H. Lyon, et T. Koike. Reverberant phase in a room and zeros in the complex frequency plane. *J. Acoust. Soc. Am.*, 89(4) :1701–1707, Avril 1991.

D. J. Tollin. Computational model of the lateralization of clicks and their echoes. In S. Greenberg et M. Slaney, editors, *Proceedings of the NATO advanced study institute on computational hearing*, pages 77–82, 1998.

F. E. Toole et B. McA. Sayers. Lateralization judgments and the nature of binaural acoustic images. *J. Acoust. Soc. Am.*, 37(2) :319–324, Février 1965.

C. Trahiotis et L. R. Bernstein. Lateralization of bands of noise and sinusoidally amplitude-modulated tones : effects of spectral locus and bandwidth. *J. Acoust. Soc. Am.*, 79(6) : 1950–1957, Juin 1986.

V. Välimäki et T. I. Laakso. Principles of fractional delay filters. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, 2000.

G. Vandernoot et E. Rio. The Listen HRTF database. Technical report, IRCAM/Projet Listen, 2003. URL `http://www.ircam.fr/equipes/salles/listen/`.

K. Varma. Time-delay-estimate based direction-of-arrival estimation for speech in reverberant environments. Mémoire de m. sc., Virginia Polytechnic Institute and State University, Octobre 2002.

F.L. Wightman et D.J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3) :1648 – 1661, 1992.

M. Wu et D. Wang. A one-microphone algorithm for reverberant speech enhancement. To be presented in ICASSP2003, Hong Kong, 2003.

Pavel Zahorik. Assessing auditory distance perception using virtual acoustics. *J. Acoust. Soc. Am.*, 114(4) :1832–1846, Avril 2002.

Index

Accord interrompu, méthode de l', 157
Autospectre à court-terme, 22

Bancs de filtres
  auditifs, 73
  temps-échelle, 73
  uniformes, 73
Bandes critiques, 74
Bandes rectangulaires équivalentes, 74
Biais de localisation, 32
Bruit interrompu, méthode du, 154

Cône de confusion, 36
Champ diffus, 17
Champ réverbérant, 18
  et ergodicité locale, 198
  et stationnarité locale, 18, 165, 198
Coefficient de corrélation
  définition, 17
Cohérence
  à court-terme, 22, 84
  définition, 17
  en champ réverbérant, 17, 20
Corrélation
  à court-terme, 22, 66
  définition, 17
  en champ réverbérant, 17
  interaurale précoce, 44
  interaurale tardive, 46
  normalisée, 17
  normalisée à court-terme, 67
Courbe de décroissance en énergie, voir Décroissance intégrée
Courbes iso-ILD, 126
Courbes iso-ITD, 126

Décroissance intégrée, méthode de la, 155
Détection
  définition, 52
  de réverbération, 52, 92, 154
  de source, 52, 92
  en bandes limitées, 70
Détection d'enveloppe, 75
Détection par égalisation et annulation
  erreur absolue, 62, 68
  erreur normalisée, 63, 68
  et pointage d'antennes, 55
  modèle auditif de, 55
Différences interaurales
  d'intensité, 34
  de temps, 35
Durée de décroissance initiale, 47

Effet de précédence, 37
  et masquage, 38
Élargissement de l'événement auditif, 39, 40
Élargissement de l'événement auditif, 28
Élévation, 36
Enveloppement de l'auditeur, 29, 45
ERB, voir Bandes rectangulaires équivalentes
Ergodicité, voir Champ réverbérant
  spatiale, 13, 198
Estimation de la fréquence fondamentale, 102, 149
Estimation du retard
  ambiguïté de phase en bande étroite, 76, 129
  méthodes usuelles, 65
  résolution limitée en basses fréquences, 78, 129
Événement auditif, 28
Événement sonore, 28

Flou de localisation, 32
Fréquence de Schroeder, 15
Fraction d'énergie latérale, 44

Impression spatiale, 28
  continue, 31
  d'arrière-plan, 31
  et masquage, 31
  précoce, 31
Incohérence
  spatiale, 14
  spectrale, 43
  temporelle, 14, 43
Indice de décision, 133
Indice de décision sur la direction, 130, 133
  après intégration fréquentielle, 136
  après intégration séquentielle, 144
Indice de détection, 64, 69, 133

Interspectre à court-terme, 22, 83
ITDG, voir Temps d'arrivée de la première réflexion

Latéralisation, 116

Modèle de coïncidence, 55

Pinna notch, 118
Puissance
à court-terme, 67

Réverbérance, 28
Relief de décroissance, 156
spectre courant futur, 156
spectrogramme cumulé, 156
Retard
d'enveloppe, 74
de groupe, 75
de phase, 75
de porteuse, 74
fractionnaire, 103

Seuil de détection, 63, 92, 97, 133
Seuil de pertinence de l'estimation de la direction, 133
Site, 36
Stationnarité, voir Champ réverbérant
Stop chord, voir Accord interrompu

Temps d'arrivée de la première réflexion, 45
Temps de mélange, 12
Temps de réverbération, 47, 154
Transformée de Fourier
à court-terme, 73, 80
gammatone, 74
Transformée en cosinus discrète modifiée, 73
Transformation de Gabor, voir transformée de Fourier à court-terme