Approximations non-linéaires pour l'analyse de signaux sonores

Rémi Gribonval

UNIVERSITÉ DE PARIS-IX DAUPHINE
U.F.R. MATHÉMATIQUES DE LA DÉCISION

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PARIS IX DAUPHINE

spécialité : Mathématiques Appliquées

par

Rémi GRIBONVAL

Approximations non-linéaires pour l'analyse des signaux sonores

Soutenue le 7 Septembre 1999 devant le jury composé de

MM.	Yali AMIT	rapporteur
	Jean-Pierre AUBIN	président
	Emmanuel BACRY	examinateur
	Donald GEMAN	examinateur
	Stéphane MALLAT	directeur de thèse
	Xavier RODET	examinateur
	Bruno TORRÉSANI	rapporteur

Remerciements

Je tiens à remercier Emmanuel Bacry et Stéphane Mallat, qui m'ont tous deux encadré lors de ce travail de thèse. Je leur suis redevable d'une somme considérable de temps et d'énergie.

Ma rencontre avec Emmanuel Bacry à l'Ecole Normale Supérieure, à l'occasion de son cours sur les aspects mathématiques et informatiques de la musique, a été décisive en bien des manières. C'est en effet à la suite de longues discussions avec lui que j'ai orienté mon travail dans la direction prise dans cette thèse. Lors de mon stage de DEA à l'IRCAM, et tout au long de ce travail de thèse au CMAP, il a profondément transformé ma façon de travailler et de concevoir l'interaction entre mathématiques, musique, et programmation informatique. Enfin c'est à lui que je dois la chance extraordinaire d'avoir fait la connaissance de Stéphane Mallat. Je suis donc très heureux qu'il ait accepté de participer au jury.

J'ai énormément appris au contact de Stéphane Mallat, et ma gratitude pour lui est immense. Il a fait preuve d'une exceptionnelle disponibilité, et j'ai bénéficié auprès de lui d'un apprentissage scientifique exceptionnel, grâce à la rigueur de son éthique scientifique et à sa grande inspiration.

J'ai été galvanisé par l'enthousiame que Xavier Rodet a insufflé à notre travail et l'excellent accueil qu'il m'a réservé dans l'équipe Analyse-Synthèse de l'IRCAM. Je dois à son expérience avisée d'être resté proche des applications sonores, et je le remercie d'avoir accepté de participer au jury.

J'adresse toute mon amitié à Philippe Depalle, qui m'avait encadré lors de mon stage de DEA. Sa compagnie a toujours été un grand plaisir et il a toujours prodigué avec beaucoup de gentillesse et de patience les conseils les plus judicieux.

C'est dans le cadre splendide du CIRM à Luminy que j'ai eu la chance de faire la connaissance de Yali Amit, dont les idées originales et brillantes m'étaient déjà connues par ses écrits. J'ai été très flatté qu'il s'intéresse à mon travail, et je le remercie vivement d'avoir accepté d'être l'un des rapporteurs de cette thèse.

J'ai eu le plaisir de découvrir la grande gentillesse et l'efficacité de Bruno Torrésani lorsqu'il m'a fait l'honneur d'être lui aussi rapporteur de cette thèse. J'espère vivement avoir de nouveau la chance de bénéficier de la précision et de la pertinence de ses commentaires.

J'ai beaucoup apprécié les nombreuses discussions passionnantes que j'ai eues avec Donald Geman, aussi bien lors de son séjour au CMAP qu'à l'occasion d'un second passage au CIRM. Je suis d'autant plus heureux qu'il ait accepté de participer au jury.

J'ai été très flatté de l'intérêt que Jean-Pierre Aubin a porté à ce travail, et je le remercie vivement de m'avoir fait l'honneur de présider le jury.

J'ai eu la chance de pouvoir profiter du savoir-faire et des moyens inégalés de l'IRCAM, dont je tiens à rendre hommage au directeur scientifique Hughes

Vinet. J'en profite pour saluer les nombreux membres de l'équipe Analyse-Synthèse, Geoffroy Peeters, Diemo Schwartz, Stefania Serafin, Christophe Vergez, Marcelo Wanderley, ... et tous ceux que j'ai pu croiser lors d'un de mes passages épisodiques. Une mention spéciale est dédiée à Laurent Ghys, administrateur système, pour le dévouement, la patience et l'efficacité avec lesquels il m'a aidé à résoudre tant de questions informatiques.

L'atmosphère chaleureuse qui règne au CMAP, le charisme et le talent des directeurs qui s'y sont succédés, Jean-Claude Nédélec, Pierre-Arnaud Raviart, et Vincent Giovangigli, l'efficacité et la gentillesse de Jeanne Bailleul, Geo Boleat, Liliane Doaré et Nathalie Limonta, les longues discussions entre collègues après le café me laisseront un excellent souvenir des années que j'y ai passées. La compétence d'Aldjia Mazari et de Pedro Ferreira pour régler mes difficultés informatiques m'a été d'un grand secours. J'ai beaucoup apprécié l'humour flegmatique d'Erwan Le Pennec, dont la compagnie a été un grand plaisir, et qui m'a rendu de fiers services. Je lui souhaite bonne chance pour les années qui viennent. Je suis très heureux d'avoir eu pour compagnons d'aventure Maureen Clerc, Christophe Bernard et Jérôme Kalifa. Nous avons partagé beaucoup d'expériences en quelques années, et j'espère que cela va durer. J'adresse en particulier tous mes voeux à Jerome Kalifa et Nadine ainsi qu'a Christophe Bernard et Rita.

Enfin, que Vérène soit infiniment remerciée pour la patience et l'endurance avec lesquelles elle m'a soutenu et supporté. Je lui dédie cette thèse ainsi qu'à Alice.

Table des matières

Notations employées

Notations générales

$\Im(z), \Re(z), \overline{z}$	Partie imaginaire, partie réelle et conjugué d'un nombre complexe z
$f = \mathcal{O}(g)$	f est dominée par g : il existe une constante C telle que $\|f\| \leq C \|g\|$
$f \sim g$	f est équivalente à g : $f = \mathcal{O}(g)$ et $g = \mathcal{O}(f)$
$\triangleq$	"est égal par définition"
$[a, b]$	Intervalle fermé de nombre réels compris entre les bornes a et b
$[\![n_1, n_2]\!]$	$[n_1, n_2] \cap \mathbb{Z}$

Probabilités

$X \sim \mathcal{P}$	La variable aléatoire X suit la loi $\mathcal{P}$
$\mathbb{E}\{.\}$	Espérance d'une variable aléatoire
$\mathcal{P}(A)$	Probabilité d'un événement A

Approximations non-linéaires

$\boldsymbol{H}$	Espace de Hilbert
$\langle ., . \rangle$	Produit scalaire
$\|.\|$	Norme $\boldsymbol{L}^2$
$\mathcal{D}$	Dictionnaire, et parfois plus précisément dictionnaire de Gabor
$\mathcal{D}^+$	Dictionnaire de Gabor chirpé
$R^m x$	Résidu d'un Matching Pursuit après m itérations
$f(t)$	Signal à temps continu
$\delta(t-u)$	Masse de Dirac au temps u
$f[n]$	Signal à temps discret
$\delta[n-p]$	Masse de Dirac au temps discret p
$\widehat{f}(\omega)$	Transformée de Fourier du signal f

Résumé

La classification de signaux en grande dimension rend nécessaire la sélection d'un petit nombre de structures caractéristiques pour représenter chaque signal. Les approximations non-linéaires donnent lieu à des représentations concises, parce qu'elles s'adaptent à la structure de chaque signal analysé. Leur emploi est prometteur.

Une première partie de ce travail définit des représentations adaptatives rapides de signaux comme combinaison linéaire d'atomes extraits d'un dictionnaire de vecteurs. A partir de l'algorithme de Matching Pursuit, plusieurs méthodes itératives sont proposées pour mettre en lumière les structures caractéristiques des signaux sonores. Le Matching Pursuit Harmonique décompose un signal en composantes harmoniques élémentaires. Le Matching Pursuit "Chirpé" extrait les variations de fréquence instantanée en tirant parti d'une analyse fine des *ridges* du dictionnaire de Gabor multi-échelle. Les approximations fournies par le Matching Pursuit Haute-résolution préservent les transitoires des signaux analysés, en imposant des contraintes de résolution temporelle. Nous accélérons ces techniques en employant des sous-dictionnaires de maxima locaux.

Notre travail est consacré dans un second temps à l'étude de l'"Analyse Discriminante Non-linéaire". Pour classifier des signaux, les méthodes d'Analyse Discriminante Linéaire réduisent la dimension en les projetant sur un sous-espace pré-déterminé. Une projection adaptative, en fonction du signal analysé, extrait de celui-ci des caractéristiques qui lui sont propres. Celles-ci le distinguent et permettent de le classifier efficacement. Nous déterminons la stratégie optimale de projection adaptative pour la classification de bruits gaussiens colorés. Afin de classifier des transitoires, nous explorons enfin une méthode utilisant les maxima du module de la transformée en ondelettes et des arbres de décision. Cette approche permet de surmonter les difficultés liées à l'invariance par translation des signaux à classifier.

Chapitre 1

Introduction

1.1 Réduire la dimension pour extraire de l'information

Le propre d'un signal, c'est de contenir de l'information. Qu'il s'agisse de l'enregistrement d'un séisme, qui traduit son parcours dans l'écorce terrestre, d'un son musical, dont le contenu est à la fois symbolique et subtil, ou bien d'une image où l'identité d'un visage est visible, on a souvent besoin d'extraire l'information qui nous intéresse. Les besoins de compression, de débruitage, de déconvolution, d'estimation de paramètres et de reconnaissance automatique de signaux rassemblent ainsi de façon féconde le Traitement du Signal et la Théorie de l'Information, nés il y a cinquante ans, sous l'impulsion notamment de Shannon et Gabor, pour modéliser la transmission d'information et les systèmes de communication. L'étude du codage, de la transmission et du décodage de l'information est loin d'être achevée. Mais aujourd'hui c'est aussi la nature qui transmet de l'information, et c'est à nous de la décoder.

Beaucoup des problèmes posés par ces besoins ont trouvé une réponse grâce aux apports de l'Analyse Harmonique, et en particulier des techniques récentes d'approximation adaptative de signaux. En effet pour extraire de l'information d'un signal, il faut connaître sa structure, afin d'en réduire la redondance pour ne garder que la "substantifique moelle". Aujourd'hui apparaissent en effet de nombreuses situations où on dispose de gigantesques bases de données (ex : analyse d'IRM en médecine, données des sondes spatiales, enregistrement de séismes ...). Elles sont constituées d'images ou de signaux qui vivent dans des espaces de grande dimension N : ainsi un son de qualité CD (*i.e.* échantillonné à 44.1 kHz) d'une durée de 1.5 seconde est un signal de $N \approx 65536 = 2^{16}$ échantillons, tandis qu'une image a couramment $N = 512 \times 512 = 2^{18}$ pixels.

L'extraction d'information nécessite donc de réduire fortement la dimension, ce qui peut se faire en projetant le signal x sur un sous-espace $\boldsymbol{V}_M$

de petite dimension $M \ll N$. En compression cela se traduit par l'utilisation d'un codage par transformée. En reconnaissance, il s'agit d'une forme d'Analyse Discriminante Linéaire.

Approximations linéaires

Les techniques d'approximation linéaire, telles que l'Analyse en Composantes Principales, fixent une fois pour toutes une base orthonormale $\{g_m\}_{m=1}^N$ (il s'agit dans le cas de l'Analyse en Composantes Principales de la base de Karhunen-Loève) et décomposent le signal sur les M premières composantes

$$P_{\boldsymbol{V}_M} x = \sum_{m=1}^{M} \langle x, g_m \rangle \, g_m$$

c'est-à-dire que le sous-espace $\boldsymbol{V}_M$ est indépendant de x.

Approximations non-linéaires

Au contraire, les techniques d'approximation non-linéaire choisissent le sous-espace $\boldsymbol{V}_M$ de manière *adaptative*, en fonction du signal x. Ainsi les approximations non-linéaires dans une base orthonormale sélectionnent les M plus grands coefficients

$$P_{\boldsymbol{V}_{M(x)}} x = \sum_{k=1}^{M} \langle x, g_{m_k} \rangle \, g_{m_k}$$

où la base (g_{m_k}) est classée par ordre décroissant des coefficients

$$|\langle x, g_{m_1} \rangle| \geq |\langle x, g_{m_2} \rangle| \geq \ldots \geq |\langle x, g_{m_k} \rangle| \, .$$

Pour que cette projection contienne bien l'information que l'on recherche, elle doit être adaptée au modèle de la nature dont on dispose. La force des approximations non-linéaires, c'est d'adapter également la projection au signal étudié. Nous allons nous attacher à employer cet atout non seulement pour approcher des signaux, mais également pour les classifier et pour en estimer des paramètres.

1.2 Mesure d'information : énergie, entropie et perception

Comment mesurer la quantité d'information que l'on a extrait d'un signal ? La réponse dépend bien sûr de ce que l'on compte faire de cette information.

Mesures d' énergie

Pour les applications de compression et de débruitage, il est naturel de mesurer l'information à l'aune de la dégradation que le signal a subie. Les techniques d'approximation de signaux mesurent cette dégradation à l'aide de critères métriques (rapport signal à bruit et taux de distorsion) liés à l'*énergie*. On mesure la qualité d'une approximation avec l'erreur quadratique

$$\varepsilon_M[x] = \left\| x - P_{\boldsymbol{V}_M} x \right\|_2^2 = \|x\|_2^2 - \left\| P_{\boldsymbol{V}_M} x \right\|_2^2 .$$

Minimiser cette erreur revient à maximiser l'énergie de la projection orthogonale

$$\left\| P_{\boldsymbol{V}_M} x \right\|_2^2$$

Si X est un processus aléatoire en dimension finie, les approximations linéaires définies avec la base de Karhunen-Loève[1] minimisent l'erreur quadratique moyenne

$$\mathbb{E}\{\varepsilon_M[X]\}$$

pour $1 \leq M \leq N$. La base de Karhunen-Loève fournit donc les meilleures approximations linéaires (*i.e.* non-adaptatives) d'un processus X. Cependant comme cette base est déterminée par les moments jusqu'à l'ordre 2 de la loi $\mathcal{P}(X)$ du processus, elle manque toutes les autres structures de celui-ci. Dès que X n'est pas gaussien, les performances des approximations non-linéaires sont meilleures.

Nous montrerons au chapitre 2 un exemple illustrant la supériorité des approximations non-linéaires sur les approximations linéaires. Nous rappellerons les principales techniques actuelles d'approximation non-linéaire, en insistant sur ce qu'apportent les stratégies utilisant la *redondance*, telles que les algorithmes de meilleure base [CW92] ou de poursuite [MZ93] [CD95], qui utilisent respectivement des bibliothèques de bases et des dictionnaires de vecteurs. Leur utilisation pour la compression [JN84] [VK95] [DeV98], le débruitage et la déconvolution [DJ94] [Kal99], pour traiter des signaux non-gaussiens et/ou non-stationnaires [DMvS97] [CM97], conduit à des algorithmes très performants. Nous verrons enfin qu'un aspect tout à fait non négligeable de ces techniques est qu'elle fournissent des algorithmes rapides, leur complexité de l'ordre de $\mathcal{O}(N)$ (transformée en ondelettes orthogonale) à $\mathcal{O}(N \log^2 N)$ (transformée en cosinus locaux [CM91]) étant à mettre en regard des $\mathcal{O}(N^2)$ que coûte un changement de base sans algorithme rapide associé.

Aux chapitres 3,4, 5, et 6, consacrés à l'analyse de signaux sonores, nous introduirons des algorithmes de poursuite, inspirés du Matching Pursuit

[1] formée des composantes principales définies par l' Analyse en Composantes Principales

[MZ93], développés pour s'adapter à certaines caractéristiques particulières des signaux acoustiques.

Nous commençons, au chapitre 3, par étendre la définition du Matching Pursuit, en introduisant la notion de Matching Pursuit "Moléculaire". A l'aide de cet outil nous développons le *Matching Pursuit Harmonique*, qui utilise comme briques élémentaires non pas des atomes temps-fréquence, mais des "molécules" associées aux structures harmoniques que l'on s'attend à trouver dans les signaux sonores. Nous insistons sur l'efficacité algorithmique en mettant au point un algorithme rapide. Celui-ci fournit des représentations temps-fréquence structurées où la présence de notes (durée, hauteur) ne se lit pas seulement visuellement, mais est explicitement présente dans la décomposition. Contrairement à beaucoup de techniques de détection de fréquence fondamentale, cet algorithme n'a aucune difficulté à détecter la présence simultanée de plusieurs fondamentales, dans le cas de la polyphonie.

Le chapitre 4 est consacré à l'accélération des techniques de poursuite. Nous y développons une technique de poursuite sur des sous-dictionnaires de maxima locaux, introduite par Bergeaud [Ber95] pour l'analyse d'images. Nous montrons qu'elle réduit la complexité de $\mathcal{O}(MN \log^2 N)$ à $\mathcal{O}(MN)$.

L'algorithme de *Matching Pursuit "Chirpé"* que nous introduisons au chapitre 5 est développé en vue de mettre en lumière les variations de fréquence instantanée des signaux sonores. Notre algorithme utilise un dictionnaire de *chirps* gaussiens et une version modifiée du Matching Pursuit pour obtenir une décomposition du signal en atomes chirpés avec une complexité algorithmique ($\mathcal{O}(MN \log^2 N)$). C'est bien plus faible que les $\mathcal{O}(MN^2 \log^2 N)$ normalement requis [Bul95] [Bul99] pour appliquer directement le Matching Pursuit sur ce dictionnaire. La décomposition atomique qu'il fournit permet non seulement de mesurer les variations de fréquence instantanée du signal, mais également de manipuler séparément les parties transitoires et les parties stationnaires. On peut également transposer la hauteur (sans changement de durée) en respectant finement la "phase" du signal.

Mesures "perceptives"

Les critères purement énergétiques peuvent créer des artefacts dans des conditions extrêmes (fort taux de compression, débruitage dans un bruit très intense, *etc.*) : les effets de blocs, les oscillations de Gibbs, la forme de l'ondelette qui devient visible sur une image compressée, voilà quelques exemples connus d'artefacts perceptivement gênants en traitement de l'image. Les codeurs audio employés commercialement (MUSICAM, Dolby^{TM} AC-3,...) emploient des modèles de masquage auditif, qui permettent de dégrader le signal dans des zones non-perceptibles, afin de restituer plus fidèlement les parties critiques.

Nous expliquons au chapitre 6 comment, en modifiant un critère initialement énergétique, nous avons réussi à éliminer des artefacts auditivement

gênants (tel que le pré-écho) du Matching Pursuit, définissant ce que nous avons appelé *Matching Pursuit Haute Résolution* [GBM+96] [GDR+96]. Il ne s'agit pas à proprement parler d'un critère perceptif, mais d'un critère non-linéaire de sélection d'atomes menant à une *super-résolution* temporelle, à la manière du critère $\boldsymbol{l}^1$ utilisé dans le Basis Pursuit [CD95].

Mesures d'entropie

Pour estimer un (ou des) *paramètre*(s) (ex : la fréquence instantanée, pour effectuer une dé-modulation FM), ou déterminer une *classe* (ex : pour reconnaître l'identité d'un visage sur une photo), l'énergie est une mesure d'information mal adaptée. Des critères non-linéaires, tels que le critère haute-résolution exposé au chapitre 6 peuvent parfois s'avérer adaptés. Les *statistiques d'ordre supérieur* [Men91] offrent également un certain nombre d'outils permettant de sortir du cadre restreint des signaux gaussiens [DT96].

La théorie de l'information fournit des outils pour mesurer directement la dépendance statistique entre la projection $P_{\boldsymbol{V}_M} x$ du signal en petite dimension et les grandeurs à estimer. On rappellera ainsi au chapitre 7 le principe de l'Analyse en Composantes Indépendantes [Com94], et sa supériorité sur l'Analyse en Composantes Principales. On fera le point sur les techniques récemment développées par Saito [Sai94] [SC94] [Sai98] et Liu et Ling [LL99] pour tirer parti de l'Analyse Harmonique et de ses algorithmes rapides afin d'obtenir des coordonnées informatives.

1.3 Analyse Discriminante Non-linéaire

Une projection sur un sous-espace $\boldsymbol{V}_M(x)$ adapté au signal x peut s'avérer utile en classification. Une telle "Analyse Discriminante Non-linéaire" peut en effet s'adapter aux caractéristiques propres du signal qui le distinguent et permettent de le classifier efficacement. Cela est illustré simplement avec l'exemple suivant, où il n'est pas particulièrement question de signaux, mais qui concerne plus généralement un problème de reconnaissance. Les 20 questions que peut poser le joueur sont l'analogue des M coordonnées associées à la projection $P_{\boldsymbol{V}_M} x$ d'un signal.

Exemple : le Jeu des 20 questions

Un joueur peut poser 20 *questions, qu'il peut choisir librement, pour identifier un personnage ou un objet. Il n'a manifestement aucun intérêt à demander systématiquement "s'il a des roues", car la réponse ne lui apportera aucune information s'il s'agit d'un personnage. Cependant, dès que le joueur sait qu'il s'agit d'un objet, cette question devient intéressante à poser. Il vaut donc mieux pour le joueur demander d'abord si c'est un objet (c'est-à-dire*

poser une question générale), puis raffiner ses demandes en fonction *de l'information qu'il a déjà acquise.*

Nous rappelons au chapitre 7 la différence entre Analyse Discriminante Linéaire (classification *passive*) et Analyse Discriminante Non-linéaire (classification *active*). Les Bases Discriminantes Locales [SC94], les bases "les moins statistiquement dépendantes" [Sai98], ou la technique de poursuite d'information de Liu [LL99], sont du ressort de l'Analyse Discriminante Linéaire [Fuk72]. Tout comme l'Analyse en Composantes Principales, ces techniques définissent en effet la projection $P_{\boldsymbol{V}_M} x$ *indépendamment* du signal x dont on veut extraire de l'information. Elles s'"adaptent", certes, mais seulement à la structure *globale* du processus X, et non à la réalisation *particulière* x qu'il faut traiter. Nous déterminons sur deux exemples la stratégie séquentielle optimale de projection adaptative. L'un des résultats les plus intéressants est que pour classifier des bruits gaussiens colorés, il est payant de s'adapter à la réalisation observée. Nous verrons cependant que l'Analyse Discriminante Non-linéaire, plus efficace en principe, pose des problèmes statistiques dans sa mise en pratique.

Afin de classifier des transitoires, nous explorons au chapitre 8 une méthode utilisant les extrema de la transformée en ondelettes et des arbres de décision [BFOS84]. Nous obtenons des performances de classification bien supérieures à celles de l'Analyse Discriminante Linéaire.

Première partie

Approximation non-linéaire

Chapitre 2

Approximations adaptatives de signaux sonores

La parole, qui transmet du sens, la musique, porteuse d'émotions, sont loin d'être les signaux périodiques purs décrits par les modèles de signal sonore les plus simples. D'abord, parce qu'on y trouve des *transitoires*, attaques instrumentales ou consonnes occlusives. Ensuite, parce que même les parties d'un son que l'on a coutume de qualifier de "stationnaires" sont loin d'être stationnaires : ainsi la *fréquence instantanée*, qui peut être définie dans les parties entretenues des sons instrumentaux (ou la résonance des notes) subit généralement des variations au cours du temps, comme dans le *vibrato* de la chanteuse. Ainsi, c'est en variant au cours du temps que les signaux sonores transmettent de l'information.

Un intérêt certain pour l'analyse des signaux non-stationnaires est apparu ces dernières années. Avec les développements théoriques, pratiques et technologiques liés à l'usage des ondelettes, de nombreuses méthodes d'approximation de ces signaux signaux ont ainsi vu le jour, dans le cadre très prolifique de l'Analyse Harmonique. Si des enjeux pratiques importants, tels que la compression et le débruitage de signaux, ont pu motiver cette ébullition, les méthodes développées s'avèrent également appréciables dans beaucoup d'applications où l'on doit extraire de l'information d'un signal[1].

Redondance, adaptativité et efficacité algorithmique

Les avancées majeures qui expliquent les succès pratiques de ces techniques sont d'une part l'emploi de représentations *adaptatives*, d'autre part l'utilisation de la *redondance*, le tout étant généralement regroupé sous l'appellation "approximations non-linéaires". Redondance et adaptativité ont permis des améliorations substantielles de *qualité d'approximation* (par com-

[1] On verra ainsi au chapitre 7 comment la nécessité d'approcher efficacement des signaux apparaît pour résoudre un problème de classification de signaux gaussiens.

paraison aux meilleures méthodes dites "linéaires"). L'autre facteur de ces succès tient à l'existence d'*algorithmes rapides*, qui les rend concrètement utilisables pour traiter de vrais problèmes, sur de vrais signaux, c'est-à-dire en grande dimension.

Nous rappellerons dans ce chapitre pourquoi les techniques non-linéaires sont plus efficaces que les meilleures techniques linéaires, puis nous ferons brièvement le point sur chacune d'entre elles.

2.1 Approximation linéaire à M termes

Si $(g_m)_{m=1}^{\infty}$ est une base orthonormale de l'espace des signaux, on appelle approximation linéaire à M termes d'un signal x la projection orthogonale

$$P_{\boldsymbol{V}_M} x = \sum_{m=1}^{M} \langle x, g_m \rangle \, g_m \tag{2.1}$$

de ce signal sur M vecteurs fixés de la base, que, pour simplifier, on suppose correspondre aux M premiers indices m. La qualité de l'approximation ainsi obtenue est mesurée, à M fixé, par l'erreur quadratique

$$\varepsilon_M[x] = \left\| x - P_{\boldsymbol{V}_M} x \right\|_2^2 = \|x\|_2^2 - \left\| P_{\boldsymbol{V}_M} \right\|_2^2 = \sum_{m=M+1}^{\infty} |\langle x, g_m \rangle|^2 . \tag{2.2}$$

Si les coefficients vérifient $|\langle x, g_m \rangle| \leq A m^{-s}$, la décroissance de l'erreur est

$$\varepsilon_M[x] = \mathcal{O}(M^{1-2s}) \tag{2.3}$$

2.1.1 Base de Karhunen-Loève

En dimension finie N, lorsque les signaux x à approcher sont des réalisations d'un processus X d'énergie finie, on peut définir la meilleure approximation linéaire à M termes à l'aide des projecteurs $P_{\boldsymbol{V}_M}$, $1 \leq M \leq N$, qui minimisent l'erreur quadratique moyenne

$$\varepsilon_M = \mathbb{E}\{\varepsilon_M[X]\} \tag{2.4}$$

De manière équivalente, ces projecteurs maximisent l'énergie

$$\mathbb{E}\left\{ \left\| P_{\boldsymbol{V}_M} X \right\|_2^2 \right\}. \tag{2.5}$$

Les projecteurs optimaux sont obtenus comme en (2.1) à partir d'une base orthogonale, dite *base de Karhunen-Loève*, constituée des vecteurs propres[2] de l'opérateur de covariance

$$\langle u, Kv \rangle = \mathbb{E}\{ \langle u, X \rangle \langle X, v \rangle \} \tag{2.6}$$

[2] il s'agit des *composantes principales* du processus X

du processus X. K est en effet diagonalisable dans une base orthonormale, car symétrique et défini positif. L'ordre des vecteurs (g_m) de la base est tel que les valeurs propres associées soient décroissantes

$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_N^2. \tag{2.7}$$

2.1.2 Avantage de l'adaptativité

La base de Karhunen-Loève ne tient compte que des moments d'ordre 1 et 2 du processus X pour minimiser $\varepsilon_M[X]$ *en moyenne*. Les seuls processus aléatoires X qui soient entièrement descriptibles par leurs moments d'ordre 1 et 2 sont les processus gaussiens. Montrons donc sur un exemple les limitations fondamentales de la base de Karhunen-Loève (et des approximations *linéaires*), et mettons en lumière les avantages apportés par l'*adaptativité*. Soit X un processus (en dimension finie N)

$$X[n] = A\delta[n - P] + W[n] \tag{2.8}$$

constitué d'un "dirac glissant" δ auquel on a ajouté un bruit W centré, gaussien et cyclo-stationnaire mais *non blanc* (*i.e.* d'opérateur de covariance $K_W \neq \lambda^2 Id$). L'amplitude A est tirée avec équiprobabilité dans $\{-\sqrt{N}, +\sqrt{N}\}$, et l'emplacement P est uniformément distribué sur $[\![0, N-1]\!]$. On suppose de plus que les trois variables aléatoires W,A et P sont indépendantes. Alors X est un bruit cyclo-stationnaire *non gaussien*, centré, d'opérateur de covariance

$$K_X = Id + K_W, \tag{2.9}$$

Comme W est cyclo-stationnaire, K_W est un opérateur de convolution circulaire. Il est donc diagonalisé dans la base de Fourier discrète $e_k, 1 \leq k \leq N$, si bien qu'il en est de même de K_X.

En supposant que les valeurs propres σ_k^2 de K_W sont classées par ordre décroissant (*i.e.* W est un bruit *basse fréquence*), celles de K_X sont

$$1 + \sigma_1^2 \geq 1 + \sigma_2^2 \geq \ldots \geq 1 + \sigma_N^2 \tag{2.10}$$

donc la meilleure approximation linéaire à M termes dans la base de Karhunen-Loève est une approximation basse fréquence caractérisée par l'erreur

$$\varepsilon_M = \sum_{m=M+1}^{N} (1 + \sigma_m^2) = (N - M) + \sum_{m=M+1}^{N} \sigma_m^2. \tag{2.11}$$

Dans le cas limite ou le bruit est presque blanc, $\sigma_m^2 \approx \sigma^2$ est presque constant et

$$\varepsilon_M \approx (N - M)(1 + \sigma^2) \tag{2.12}$$

On peut obtenir de meilleures approximations à M termes de X dans la base de diracs, à condition de choisir les M termes de façon *adaptative*, c'est-à-dire en fonction du signal x à approcher. En effet, soit p la valeur prise par la variable aléatoire P dans la réalisation x du processus X : l'approximation de x avec *un* vecteur $\delta[n-p]$ fournit une erreur d'approximation à 1 terme

$$\varepsilon_1[x] = \sum_{n \neq p} |x[n]|^2 = \sum_{n \neq p} |W[n]|^2 = \|W\|^2 - |W[p]|^2 \tag{2.13}$$

et à M termes

$$\varepsilon_M[x] = \sum_{n \notin \{p\} \cup I_{M-1}(x)} |W[n]|^2 \leq \sum_{n \notin \{p\} \cup I_{M-1}} |W[n]|^2 \tag{2.14}$$

où $I_{M-1}(x)$ est l'ensemble de $M-1$ indices (ne contenant pas p) qui permet de minimiser $\varepsilon_M[x]$, et I_{M-1} n'importe quel ensemble de $M-1$ indices ne contenant pas p. Selon que $p \in [\![1, M-1]\!]$ (ce qui se produit avec une probabilité $(M-1)/N$) ou non, on prend $I_{M-1} = [\![1, M]\!] \setminus \{p\}$ uo $I_{M-1} = [\![1, M-1]\!]$, et l'on obtient les majorations

$$\varepsilon_M[x] \leq \sum_{m=M}^{N} |W[n]|^2 - |W[M]|^2 \tag{2.15}$$

ou

$$\varepsilon_M[x] \leq \sum_{m=M}^{N} |W[n]|^2 - |W[p]|^2 . \tag{2.16}$$

Comme P est indépendant de W, $\mathbb{E}\left\{ |W[P]|^2 \right\} = (\sum_m \sigma_m^2)/N \approx \sigma^2$. En passant à l'espérance on a donc

$$\mathbb{E}\{\varepsilon_M[x]\} \leq (N-M)\sigma^2. \tag{2.17}$$

La qualité d'approximation non-linéaire (2.17) est bien meilleure que (2.11).

2.2 Approximation non-linéaire à M termes

Les approximations *non-linéaires* de signaux, sont potentiellement bien plus efficaces que les approximations linéaires. En outre, elles permettent d'extraire des caractéristiques *non-gaussiennes* des signaux, porteuses potentielles d'information[3], telles que le paramètre P dans l'exemple ci-dessus. Dans une base orthonormale $(g_m)_{m=1}^{\infty}$, une approximation non-linéaire à M termes d'un signal x s'écrit

$$P_{\boldsymbol{V}_M(x)} x = \sum_{m \in I_M(x)} \langle x, g_m \rangle \, g_m \tag{2.18}$$

[3] On verra au chapitre 7 que la base de Karhunen-Loève peut également être peu performante pour la classification de signaux.

où l'ensemble de M indices $I_M(x)$ dépend de x. Comme l'erreur quadratique vaut

$$\varepsilon_M[x] = \sum_{m \notin I_M(x)} |\langle x, g_m \rangle|^2 = \|x\|^2 - \sum_{m \in I_M(x)} |\langle x, g_m \rangle|^2 , \tag{2.19}$$

le choix optimal de $I_M(x)$ est obtenu en prenant les M indices associés aux plus grands coefficients, *i.e.* , en notant (g_{m_k}) la base classée dans l'ordre décroissant des coefficients $|\langle x, g_{m_k} \rangle|^2$,

$$I_M(x) = \{m_k, 1 \leq k \leq M\} . \tag{2.20}$$

Un signal x est d'autant mieux approché par une telle approximation non-linéaire que ses coefficients $\langle x, g_m \rangle$ sont plus concentrés sur quelques vecteurs de la base seulement. On peut mesurer cette concentration à l'aide de l'appartenance de la suite $\langle x, g_{m_k} \rangle$ à des espaces l^p faibles

$$|\langle x, g_{m_k} \rangle|^p \leq C k^{-1} \tag{2.21}$$

Les inégalités de Jackson et de Bernstein relient la plus petite valeur de $p < 1$ pour laquelle (2.21) est vraie et la vitesse de décroissance de $\varepsilon_M[x]$:

$$\varepsilon_M[x] = \mathcal{O}(M^{1-2/p}) \tag{2.22}$$

2.2.1 Complexité algorithmique de la projection adaptative

Lorsque la base orthogonale (g_m) est quelconque, les approximations linéaires à M termes nécessitent le calcul de $\langle x, g_m \rangle$, $1 \leq m \leq M$. Leur calcul a donc une complexité algorithmique de

$$\mathcal{O}(MN). \tag{2.23}$$

Pour obtenir la meilleure approximation non-linéaire, il faut connaître la valeur de *tous* les coefficients, si bien que le coût algorithmique est celui d'un changement de base

$$\mathcal{O}(N^2). \tag{2.24}$$

Cependant certaines bases orthogonales sont associées à des *algorithmes rapides* de changement de base. Ainsi la Transformée de Fourier Rapide FFT $\mathcal{O}(N \log N)$, la Transformée en Ondelettes Rapide FWT $\mathcal{O}(N)$ [Mal89] [BCR91], la transformée associée à une famille particulière de paquets d'ondelettes $\mathcal{O}(N \log N)$ ou de cosinus locaux $\mathcal{O}(N \log^2 N)$ [CM91], diminuent fortement la *complexité algorithmique* du changement de coordonnées. La projection adaptative sur les M plus grands cosinus locaux, par exemple, se

fait alors en trois étapes : changement de coordonnées ($\mathcal{O}(N \log^2 N)$), sélection des M plus grandes et mise à zéro des autres $\mathcal{O}(N)$, changement de coordonnées inverse $\mathcal{O}(N \log^2 N)$). Le coût total

$$\mathcal{O}(N \log^2 N), \tag{2.25}$$

toujours dominé par le changement de base, est bien plus faible que $\mathcal{O}(N^2)$. Le coût des approximations linéaires dans ces bases peut également être réduit, en utilisant aussi trois étapes (la deuxième étape n'est plus adaptative). Une approximation linéaire calculée par ce moyen coûte

$$\mathcal{O}(N \log^2 N), \tag{2.26}$$

ce qui est plus faible que $\mathcal{O}(MN)$ si M est grand devant $\log^2 N$.

2.2.2 Choix de la base

Dans le cas des approximations linéaires, avant d'effectuer la projection (2.1), il faut *calculer* la base de Karhunen-Loève. Pour cela on estime l'opérateur de covariance K et on le diagonalise. Comme K est associé à une matrice de taille $N \times N$, sa diagonalisation coûte $\mathcal{O}(N^3)$. Cependant lorsque le processus X est cyclo-stationnaire, sa base de Karhunen-Loève est la base de Fourier discrète, et l'on peut donc éviter ce calcul préalable.

Pour les approximations non-linéaires, on emploie souvent une base d'ondelettes, qui constitue une base inconditionnelle de nombreux espaces fonctionnels ($\boldsymbol{L}^p$, espaces de Besov, ...). La théorie de l'approximation établit les liens entre la régularité d'un signal x, sa norme dans ces espaces fonctionnels, et la vitesse de décroissance de ses coefficients d'ondelettes (2.21). Le lecteur intéressé par ces aspects pourra se référer à l'introduction aux approximations non-linéaires de De Vore [DeV98].

Une base orthogonale est d'autant plus appropriée pour approcher une classe de signaux que les coefficients des signaux sont concentrés sur peu de vecteurs. Le choix de la base orthogonale la plus appropriée dépend de la classe de signaux et donc de l'application envisagée. Ainsi pour l'analyse de signaux réguliers par morceaux, une base d'ondelettes de régularité suffisante, telle que les ondelettes à support compact de Daubechies [Dau88] est adaptée. Lorsque les signaux présentent des oscillations plutôt que des singularités temporelles, des paquets d'ondelettes, ou bien des cosinus locaux [CM91], ou encore des bases orthonormales d'ondelettes "chirpées" [BJ93a] sont sans doute plus appropriées.

2.3 Algorithme de meilleure base ("Best Basis")

L'analyse mathématique ne permet pas toujours de déterminer une base optimale pour un problème d'approximation donné. On peut avoir intérêt

à *adapter* également la base employée au signal x, de façon à *concentrer* autant que possible son énergie sur peu de coefficients. Les algorithmes de meilleure base [CW92] choisissent une base dans une *bibliothèque* $(\mathcal{B}^\lambda)_{\lambda\in\Lambda}$ de bases orthonormales $\mathcal{B}^\lambda = (g_m^\lambda)_{m=1}^N$, en minimisant une fonction de coût "additive"

$$\mathcal{C}(\mathcal{B}^\lambda, x) \triangleq \sum_{m=1}^{N} \Phi\left(\frac{\left|\langle x, g_m^\lambda\rangle\right|^2}{\|x\|^2}\right) \tag{2.27}$$

définie à partir d'une fonction concave arbitraire Φ (par exemple $\Phi(x) = x\log 1/x$).

Le théorème suivant, dont on trouvera une démonstration dans [Mal98], montre que la relation $\mathcal{C}(\mathcal{B}^\lambda, x) < \mathcal{C}(\mathcal{B}^\mu, x)$ entre le coût de deux bases est suffisante pour savoir que, pour tout M, $\mathcal{B}^\lambda$ concentre mieux l'énergie de x sur ses M composantes les plus fortes que ne le fait $\mathcal{B}^\mu$.

Théorème 1 (Hardy-Littlewood-Pòlya) *Soient $(x_m)_{m=1}^N$ et $(y_m)_{m=1}^N$ deux suites décroissantes de N réels de somme 1. Alors les deux propriétés suivantes sont équivalentes :*

(i) Pour tout M,

$$\sum_{m=1}^{M} x_m \geq \sum_{m=1}^{M} y_m$$

(ii) Pour toute fonction concave Φ,

$$\sum_{m=1}^{N} \Phi(x_m) \leq \sum_{m=1}^{N} \Phi(y_m)$$

Le coût reflète donc la capacité de la base à approcher x avec peu de vecteurs, si bien que la base sélectionnée selon

$$\mathcal{C}(\mathcal{B}^{\lambda_0}, x) = \min_\lambda \mathcal{C}(\mathcal{B}^\lambda, x) \tag{2.28}$$

est la plus *adaptée* au signal x.

Coifman et Wickerhauser [CW92] ont montré qu'en utilisant une bibliothèque de bases structurée en arbre binaires (comme la bibliothèque des paquets d'ondelettes ou celle des cosinus locaux [CM91]) on dispose d'un algorithme rapide qui, après calcul des divers coefficients $(\langle x, g_m^\lambda\rangle)_{1\leq m\leq N, \lambda\in\Lambda}$, sélectionne une meilleure base en $\mathcal{O}(N)$ opérations. Le coût total de la procédure est alors dominé par la décomposition du signal dans la bibliothèque. Dans le cas des paquets d'ondelettes, ce coût est de $\mathcal{O}(N\log N)$, tandis que pour les cosinus locaux il est de $\mathcal{O}(N\log^2 N)$.

2.4 Représentations *redondantes* et dictionnaires

Les signaux sonores (parole, musique, ...) sont non-stationnaires. Ils contiennent des structures à différentes échelles (transitoires de très courte durée, parties soutenues et résonances de notes qui durent, ...) et différentes fréquences (par exemple les différents *partiels*, ou "harmoniques" d'une même note) à des instants variés. Ces différents structures se superposent, dès que plusieurs locuteurs ou plusieurs instruments s'expriment simultanément. Ainsi, un signal qui présente simultanément des structures qui ne sont pas orthogonales, telles qu'une sinusoïde et un dirac superposés,

$$x(t) = \delta(t) + e^{i\omega t} \tag{2.29}$$

ne peut pas être représenté concisément comme somme de ces deux composantes dans une base orthonormale. Pour de tels signaux, l'efficacité des approximations à M termes dans une base orthogonale est donc limitée.

2.4.1 Extraction de ridges de transformées redondantes

Pour représenter correctement les signaux sonores, il est nécessaire d'introduire de la *redondance*, en ne se limitant plus à une famille orthogonale. Ainsi, pour analyser les variations de fréquence instantanée de signaux acoustiques, Delprat, Kronland-Martinet, *et.al.* [Del92] [DEG+92] [GKM96] [KMG96] extraient les *ridges* de représentations temps-fréquence ou temps-échelle redondantes, telles que la transformée de Fourier à court terme

$$\left\langle x, g\left(t-u\right) e^{i\xi(t-u)} \right\rangle \tag{2.30}$$

ou la transformée en ondelettes continue de Morlet

$$\left\langle x, \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{i\frac{\xi_0}{s}(t-u)} \right\rangle. \tag{2.31}$$

Toutefois, étant donné la présence simultanée d'*oscillations* et de *transitoires* dans les signaux sonores, il est souhaitable d'analyser indépendamment l'*échelle* s et la *fréquence* ξ des phénomènes mis en jeu. Cela n'est pas possible avec les outils temps-fréquence/temps-échelle classiques : la transformée de Fourier à court terme utilise une fenêtre d'analyse de taille fixée, tandis que l'ondelette d'analyse utilisée dans la transformée en ondelettes a une fréquence $\xi = \xi_0/s$ liée à son échelle. La transformée de Fourier multi-échelle [Pea91] utilisée par Pearson n'a pas cet inconvénient. Cependant elle ne fournit pas une décomposition du signal en structures élémentaires : elle le compare à un *dictionnaire* de formes d'ondes élémentaires, de différentes échelles, temps et fréquence.

2.4.2 Dictionnaire temps-fréquence multi-échelle de Gabor

On appelle *dictionnaire* une famille redondante

$$\mathcal{D} = \{g_\gamma, \gamma \in \Gamma\} \tag{2.32}$$

de vecteurs unitaires $\|g_\gamma\| = 1$, ou *atomes*. L'analyse des signaux sonores[4], nécessite un dictionnaire temps-fréquence multi-échelle, dont les atomes sont caractérisés par un indice

$$\gamma \stackrel{\Delta}{=} (s, u, \xi) \tag{2.33}$$

choisi dans un ensemble d'indices $\Gamma \subset \mathbb{R}^+ \times \mathbb{R}^2$.

Un tel dictionnaire s'obtient en réunissant les vecteurs des différentes bases de la bibliothèque de paquets d'ondelettes, ou de celle des cosinus locaux [CM91]. On s'intéresse ici au dictionnaire multi-échelle de Gabor [QC94] [MZ93][5], qui comprend de l'ordre de $\mathcal{O}(N \log N)$ atomes temps-fréquence.

Il est constitué de la collection des atomes temps-fréquence obtenus dilatation, translation et modulation d'une "fenêtre" $g(t)$. Une fenêtre est une fonction paire et positive, dont l'essentiel de l'énergie est localisée temporellement autour du temps 0 et, dans le domaine de Fourier, autour de la fréquence 0. En raison de ses propriétés optimales de localisation combinée temps/fréquence, au sens du principe d'incertitude de Heisenberg, on utilisera souvent une fenêtre gaussienne

$$g(t) = \frac{1}{\pi^{1/4}} \exp\left(-t^2/2\right). \tag{2.34}$$

L'atome temps-fréquence d'échelle s, de temps u et de fréquence ξ s'écrit

$$g_{(s,u,\xi)}(t) \stackrel{\Delta}{=} \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{i\xi(t-u)} \tag{2.35}$$

Le facteur de normalisation (au sens de la norme $\boldsymbol{L}^2$) $1/\sqrt{s}$ nous assure que les atomes sont tous d'énergie 1.

L'atome $g_{(s,u,\xi)}$ est centré autour du temps u avec une dispersion temporelle Δu d'énergie de l'ordre de s. Sa transformée de Fourier est localisée autour de la fréquence ξ, avec une dispersion $\Delta\xi$ de l'ordre de $1/s$. Sa transformée de Wigner-Ville [Fla93] (la figure 2.1 représente un atome chirpé gaussien et sa transformée de Wigner-Ville), qui définit sa répartition énergétique dans le plan temps-fréquence, se déduit de celle de la fenêtre de départ g par la relation

$$WV[g_{(s,u,\xi)}](t,\omega) = WV[g]\left(\frac{t-u}{s}, s(\omega-\xi)\right) \tag{2.36}$$

[4] Pour des applications spécifiques, il est possible de définir un dictionnaire adéquat [MC97], avec l'inconvénient cependant de ne pas avoir d'algorithme rapide.

[5] On en utilisera une extension, le dictionnaire de Gabor "chirpé", qui comprend $\mathcal{O}(N^2)$ atomes, au chapitre 5

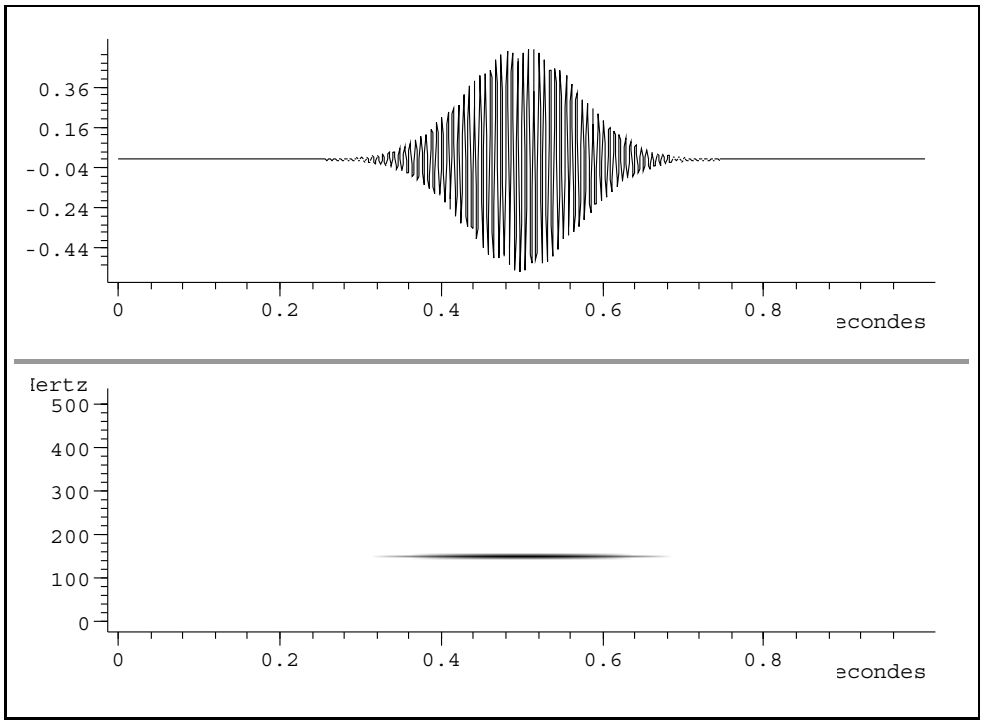

FIG. 2.1 – Un atome gaussien et sa transformée de Wigner-Ville.

Dans le cas particulier d'une fenêtre gaussienne, la transformée de Wigner-Ville d'un atome est donc une bosse gaussienne en deux dimensions

$$WV[g_{(s,u,\xi)}](t,\omega) = \frac{1}{\pi} e^{-\frac{(t-u)^2}{s^2} - s^2(\omega-\xi)^2} \tag{2.37}$$

essentiellement concentrée dans le rectangle

$$(t,\omega) \in [u-\Delta u, u+\Delta u] \times [\xi - \Delta\xi, \xi + \Delta\xi]\,.$$

Les paramètres sont échantillonnés selon

$$s = a^j, j \in \mathbb{Z} \tag{2.38}$$
$$u = n \times \Delta u(s), n \in \mathbb{Z}, \tag{2.39}$$
$$\xi = k \times \Delta\xi(s), k \in \mathbb{Z}, \tag{2.40}$$

où les pas d'échantillonnage

$$\Delta u(s) \triangleq s\Delta u(1) \tag{2.41}$$
$$\Delta\xi(s) \triangleq s^{-1}\Delta\xi(1) \tag{2.42}$$

sont invariants par changement d'échelle. Watson et Gilholm [WG98] justifient cet échantillonnage "critique" à l'aide des propriétés du tenseur associé

à la métrique Riemannienne avec laquelle on définit la distance

$$1 - |\langle g_{\gamma_1}, g_{\gamma_2} \rangle|^2$$

entre triplets (s, u, ξ) de $\mathbb{R}^3$. Pour un signal de N points, on doit donc considérer $\mathcal{O}(\log N)$ échelle, et $\mathcal{O}(N)$ couples (u, ξ) par échelle. Le dictionnaire de Gabor multi-échelle comprend donc $\mathcal{O}(N \log N)$ atomes.

2.5 Décomposition atomique dans un dictionnaire

A partir d'un dictionnaire $\mathcal{D}$ donné, on peut chercher à approcher le signal x à l'aide d'une *décomposition atomique*

$$x_M = \sum_{m=1}^{M} \alpha_m g_{\gamma_m} \tag{2.43}$$

avec M atomes $(g_{\gamma_m})_{m=1}^{M}$ choisis dans $\mathcal{D}$. Davis [Dav94] a montré que l'obtention d'une telle approximation x_M de x telle que $\|x_M - x\| \leq \varepsilon$ est un problème NP-complet. Il n'est donc pas question d'exiger la meilleure décomposition atomique, mais plutôt de chercher à s'en approcher : les techniques de "poursuite" contournent la NP-complétude en empruntant des stratégies sous-optimales pour obtenir de "bonnes" décomposition atomiques des signaux.

2.5.1 Poursuite de base ("Basis Pursuit")

Le Basis Pursuit [CD95] fait appel aux techniques de la programmation linéaire pour obtenir une décomposition atomique $x_M = x$ minimisant le critère $\boldsymbol{l}^1$

$$\|(\alpha_{\gamma_m})\|_1 = \sum_m |\alpha_{\gamma_m}| \,. \tag{2.44}$$

Il aboutit à la sélection d'une *base* (non nécessairement orthogonale) de N vecteurs, d'où son nom. En dépit de l'utilisation des techniques les plus récentes de programmation linéaires (notamment l'algorithme de point intérieur de Karmarkar [Shr98]), d'accélérations avec les algorithmes rapides liés au dictionnaire utilisé, le coût algorithmique du Basis Pursuit est de l'ordre de $\mathcal{O}(P^{3.5})$, où P est la taille du dictionnaire. Bien qu'il mène expérimentalement à des décompositions compactes des signaux, la complexité du Basis Pursuit est donc trop grande pour l'employer effectivement.

2.5.2 Poursuite adaptative ("Matching Pursuit")

Le Matching Pursuit [MZ93] (ou poursuite adaptative) est une technique itérative sous-optimale de sélection d'une approximation adaptative à M

termes d'un signal x. Étant donnée une approximation $x_m = \sum_1^m \alpha_n g_{\gamma_n}$ à m atomes, spécifiée par les coefficients et les indices $(\alpha_n, g_{\gamma_n})_{1 \leq n \leq m}$, la poursuite détermine une approximation à $m+1$ atomes de façon *gloutonne*, en étendant la précédente décomposition à l'aide du choix de l'atome $g_{\gamma_{m+1}}$ et de son coefficient α_{m+1}. Rappelons ici la définition du Matching Pursuit introduite par Mallat et Zhang [MZ93]. On commence par choisir un premier atome g_{γ_1} dans le dictionnaire $\mathcal{D}$ de façon à s'adapter au mieux au signal analysé x, selon une mesure de corrélation

$$C(x, g_\gamma) \stackrel{\Delta}{=} |\langle x, g_\gamma \rangle|^2 \tag{2.45}$$

Le carré du produit scalaire de x avec l'atome g_γ, $|\langle x, g_\gamma \rangle|^2$, représente l'*énergie* de x le long de la direction de g_γ. Le premier vecteur est donc choisi selon le critère

$$\gamma_1 = \arg\max_\gamma |\langle x, g_\gamma \rangle|^2 \tag{2.46}$$

et le premier *résidu* de x est défini par la projection orthogonale

$$R^1 x = x - \langle x, g_{\gamma_1} \rangle \, g_{\gamma_1}. \tag{2.47}$$

L'énergie du résidu est alors donnée par la relation

$$\left\| R^1 x \right\|^2 = \|x\|^2 - |\langle x, g_{\gamma_1} \rangle|^2 \tag{2.48}$$

En itérant cette procédure, on obtient par induction

$$\begin{aligned} g_{\gamma_{m+1}} &= \arg\max_\gamma |\langle R^m x, g_\gamma \rangle|^2 && (2.49) \\ R^{m+1} x &= R^m x - \left\langle R^m x, g_{\gamma_{m+1}} \right\rangle g_{\gamma_{m+1}} && (2.50) \\ \left\| R^{m+1} x \right\|^2 &= \|R^m x\|^2 - \left| \left\langle R^m x, g_{\gamma_{m+1}} \right\rangle \right|^2 && (2.51) \end{aligned}$$

et finalement, en notant $R^0 x = x$, on obtient la décomposition de x comme combinaison linéaire

$$x = \sum_{m=1}^{M} \left\langle R^{m-1} x, g_{\gamma_m} \right\rangle g_{\gamma_m} + R^M x \tag{2.52}$$

avec la conservation d'énergie

$$\|x\|^2 = \sum_{m=1}^{M} \left| \left\langle R^{m-1} x, g_{\gamma_m} \right\rangle \right|^2 + \left\| R^M x \right\|^2 \tag{2.53}$$

analogue à ce qu'on obtiendrait avec une décomposition dans une base orthonormale, et ceci bien que la famille de vecteurs sélectionnés ne soit en général absolument pas orthonormale. Un résultat de Jones [Jon87] sur le Projection

Pursuit de Huber [Hub85] prouve la convergence de cet algorithme : dès que le dictionnaire $\mathcal{D}$ est complet, le résidu $R^M x = x - x_M$ tend vers zéro et l'on dispose de la représentation

$$x = \sum_{m=1}^{\infty} \left\langle R^{m-1}x, g_{\gamma_m} \right\rangle g_{\gamma_m} \tag{2.54}$$

$$\|x\|^2 = \sum_{m=1}^{\infty} \left|\left\langle R^{m-1}x, g_{\gamma_m} \right\rangle\right|^2 \tag{2.55}$$

En dimension finie N, la convergence s'effectue à une vitesse exponentielle $\left\|R^M x\right\| \sim e^{-\lambda(\mathcal{D})M}$, caractéristique du dictionnaire. En dimension infinie, le lien entre la régularité du signal x et la vitesse de décroissance de $\left\|R^M x\right\| = \|x - x_M\|$ est pour l'instant beaucoup plus mal connu que pour l'approximation à M termes dans une base orthonormale d'ondelettes. Le lecteur intéressé pourra consulter Temlyakov [Tem98, Tem99b, Tem99a] ou De Vore [DeV98].

2.5.3 Matching Pursuit Orthogonal

Même en dimension finie, le Matching Pursuit nécessite une infinité d'itérations pour reconstruire x. Le Matching Pursuit Orthogonal, introduit par Zhang [Zha93], Davis [Dav94] [DMA97] et Pati *et al.* [PRK93] permet de s'assurer que la poursuite cesse après un nombre fini d'étapes. L'algorithme initial est modifié comme suit : une fois les m vecteurs $g_{\gamma_1}, \ldots, g_{\gamma_m}$ sélectionnés, on considère $P_{\boldsymbol{V}_m}$ le projecteur orthogonal sur le sous-espace

$$\boldsymbol{V}_m = Vect\left\{g_{\gamma_1}, \ldots, g_{\gamma_m}\right\}. \tag{2.56}$$

La meilleure approximation de x avec ces m vecteurs est $P_{\boldsymbol{V}_m} x$, Elle permet de définir le résidu comme

$$R^m x = x - P_{\boldsymbol{V}_m} x. \tag{2.57}$$

On peut alors itérer le procédé à l'aide de (2.49).

Cet algorithme nécessite le calcul de l'orthonormalisée de Gram-Schmidt de la famille $(g_{\gamma_m})_{m=1}^N$ et augmente assez sensiblement la complexité algorithmique de la poursuite. Pour un dictionnaire multi-échelle de Gabor elle est de l'ordre de $\mathcal{O}(MN \log^2 N)$.

2.5.4 Généralisations

Le principe du Matching Pursuit est souple. Suivant le dictionnaire et le critère de sélection d'atomes employés, il permet d'approcher efficacement différentes classes de signaux. On s'intéresse aux chapitres suivants à

des variantes du Matching Pursuit. Le chapitre 3 montrera comment définir un Matching Pursuit "moléculaire", en sélectionnant de façon adaptative des sous-espaces plutôt que des atomes. Ainsi pour décomposer un signal musical en structures harmoniques, on introduira le Matching Pursuit Harmonique. Le chapitre 4 est consacré à l'accélération du Matching Pursuit : le Matching Pursuit Rapide que nous avons développé réduit la complexité à $\mathcal{O}(MN)$. On définit au chapitre 5 une poursuite modifiée dans le dictionnaire de Gabor chirpé, avec une complexité de $\mathcal{O}(MN)$. Enfin on introduit au chapitre 6 le Matching Pursuit "Haute Résolution", qui sélectionne les atomes avec un critère différent du pur critère énergétique usuellement employé. Ce critère introduit une super-résolution temporelle, et améliore l'analyse des transitoires.

Chapitre 3

Matching Pursuit sur un dictionnaire de "molécules"

Nous définissons dans ce chapitre une extension naturelle du Matching Pursuit atomique, le Matching Pursuit "moléculaire". Au lieu de projections itératives sur les droites engendrées par des *atomes* g_γ choisis dans un dictionnaire atomique, on choisi des projections sur des *molécules*, c'est-à-dire des sous-espaces vectoriels $\boldsymbol{V}_\gamma$ de $\boldsymbol{H}$ de dimension plus grande que 1, choisis dans un *dictionnaire de molécules*.

La première section est consacrée à la définition "abstraite" de cet algorithme, à partir d'idées issues du Projection Pursuit de Huber [Hub85].

On s'intéresse dans un second temps à deux dictionnaires de molécules particuliers. Le dictionnaire de molécules "di-atomiques réelles" est le cadre naturel pour définir une poursuite avec des atomes à valeurs réelles, comme l'ont fait remarquer Bergeaud [Ber95] et Goodwin [Goo97]. Nous introduisons ensuite le dictionnaire de molécules *harmoniques*, afin de définir le Matching Pursuit Harmonique, destiné à décomposer les signaux sonores en *structures harmoniques*.

3.1 Matching Pursuit avec des dictionnaires de *molécules*

Un Matching Pursuit "moléculaire" diffère du Matching Pursuit "atomique" par le fait qu'à chaque itération on adapte au résidu un *sous-espace* $\boldsymbol{V}_\gamma$ de $\boldsymbol{H}$ qui n'est plus contraint à être une droite. Ce sous espace est choisi dans un dictionnaire de "molécules"

$$\mathcal{D}_{mol} = \{\boldsymbol{V}_\gamma, \gamma \in \Gamma_{mol}\}\,. \tag{3.1}$$

3.1.1 Principe

On commence donc par sélectionner la première molécule $\boldsymbol{V}_{\gamma_1}$ de la décomposition de manière à maximiser une mesure de corrélation

$$C(x, \boldsymbol{V}_\gamma) \stackrel{\Delta}{=} \left\| P_{\boldsymbol{V}_\gamma} x \right\|^2 \tag{3.2}$$

où $P_{\boldsymbol{V}_\gamma}$ est l'opérateur de projection orthogonale sur $\boldsymbol{V}_\gamma$. La grandeur $\left\| P_{\boldsymbol{V}_\gamma} x \right\|^2$ représente donc l'*énergie* de x dans la direction de la molécule $\boldsymbol{V}_\gamma$. Le choix du premier indice γ_1 est donc effectué selon le critère

$$\gamma_1 = \arg\max_\gamma \left\| P_{\boldsymbol{V}_\gamma} x \right\|^2 \tag{3.3}$$

et le premier *résidu* de x est calculé cette fois-ci à l'aide de la projection orthogonale

$$R^1 x = x - P_{\boldsymbol{V}_{\gamma_1}} x. \tag{3.4}$$

L'énergie du résidu est alors donnée par la relation

$$\left\| R^1 x \right\|^2 = \|x\|^2 - \left\| P_{\boldsymbol{V}_{\gamma_1}} x \right\|^2 \tag{3.5}$$

En itérant ce procédé on obtient par induction

$$\gamma_{m+1} = \arg\max_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^m x \right\|^2 \tag{3.6}$$

$$R^{m+1} x = R^m x - P_{\boldsymbol{V}_{\gamma_{m+1}}} R^m x \tag{3.7}$$

$$\left\| R^{m+1} x \right\|^2 = \|R^m x\|^2 - \left\| P_{\boldsymbol{V}_{\gamma_{m+1}}} R^m x \right\|^2 \tag{3.8}$$

et on peut finalement, en notant $R^0 x = x$, reconstruire x à partir des projections successives obtenues

$$x = \sum_{m=1}^{M} P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x + R^M x \tag{3.9}$$

avec la conservation d'énergie

$$\|x\|^2 = \sum_{m=1}^{M} \left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x \right\|^2 + \left\| R^M x \right\|^2 \tag{3.10}$$

3.1.2 Convergence

Le procédé itératif utilisé converge si le résidu vérifie

$$R^M x \longrightarrow 0. \tag{3.11}$$

On peut alors reconstruire le signal

$$x = \sum_{m=1}^{\infty} P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x \tag{3.12}$$

$$\|x\|^2 = \sum_{m=1}^{\infty} \left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x \right\|^2. \tag{3.13}$$

Un théorème de Jones [Jon87] sur la convergence du Projection Pursuit de Huber [Hub85] prouve la convergence du Matching Pursuit *atomique* [MZ93] dès que le dictionnaire atomique $\mathcal{D}$ utilisé est *complet*, *i.e.* lorsque l'adhérence de l'espace vectoriel $\boldsymbol{W} = Vect\{\mathcal{D}\}$ qu'il engendre est égale à l'espace $\boldsymbol{H}$ tout entier. Si ce n'est pas le cas, il y a toujours convergence, mais pas vers zéro

$$R^M x \longrightarrow P_{\overline{\boldsymbol{W}}^{\perp}} x \tag{3.14}$$

où $P_{\overline{\boldsymbol{W}}^{\perp}}$ est le projecteur orthogonal sur le complément orthogonal de $\overline{\boldsymbol{W}}$ dans $\boldsymbol{H}$. L'approximation

$$x_M = x + R^M x \longrightarrow P_{\overline{\boldsymbol{W}}} x \tag{3.15}$$

ne permet donc pas de reconstruire le signal. Un résultat de Rejtö et Walter [RW92] permet d'étendre le résultat de convergence à la poursuite moléculaire. Il suffit encore que le dictionnaire "moléculaire" $\mathcal{D}_{mol}$ de sous-espaces vectoriels utilisé engendre un sous-espace vectoriel dense de $\boldsymbol{H}$ pour être assuré de la convergence (3.11). Rejtö et Walter établissent de plus la convergence d'une forme faible de poursuite, définie par le choix, à chaque étape, non pas de la meilleure molécule (qui remplit la condition (3.6)) mais d'une "bonne" molécule $\boldsymbol{V}_{\gamma_m}$ vérifiant la condition plus faible

$$\left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x \right\|^2 \geq \rho \sup_{\gamma} \left\| P_{\boldsymbol{V}_{\gamma}} R^{m-1} x \right\|^2 \tag{3.16}$$

où $\rho > 0$ est un facteur de sous-optimalité indépendant[1] de m.

L'avantage de cette condition assouplie est que la recherche de la "meilleure" molécule peut s'effectuer à chaque étape m dans une sous-famille $\mathcal{D}_m$ du dictionnaire $\mathcal{D}$, dont le nombre d'éléments est beaucoup plus petit, ce qui peut accélérer cette recherche. On en verra une application au chapitre 4. La contrepartie est un affaiblissement de la vitesse de convergence.

[1] Des résultats récents de Temlyakov [Tem99b] permettent de traiter le cas où ce facteur varie avec m, à condition que $\sum_m \frac{\sqrt{\rho_m}}{m} = \infty$.

3.1.3 Vitesse de convergence en dimension finie

En dimension finie, il y a convergence à vitesse exponentielle. Pour un dictionnaire $\mathcal{D}$, on peut en effet définir

$$\beta(\mathcal{D}) \triangleq \inf_{x \in \boldsymbol{H}} \sup_{\boldsymbol{V}_\gamma \in \mathcal{D}} \frac{\left\| P_{\boldsymbol{V}_\gamma} x \right\|^2}{\|x\|^2}. \tag{3.17}$$

Comme on est en dimension finie, la sphère unité est compacte, donc $x \mapsto \sup_\gamma \left\| P_{\boldsymbol{V}_\gamma} x \right\|^2 / \|x\|^2$ atteint son infimum qui est strictement positif car $\mathcal{D}$ est complet et contient donc au moins une base. On a donc $\beta(\mathcal{D}) > 0$. L'équation (3.8) nous donne alors à chaque étape

$$\frac{\|R^m x\|^2}{\|R^{m-1} x\|^2} = 1 - \frac{\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2}{\|R^{m-1} x\|^2} \leq 1 - \beta(\mathcal{D}) \tag{3.18}$$

Pour tout $M \geq 1$ on a donc

$$\left\| R^M x \right\|^2 \leq \|x\|^2 \left(1 - \beta\left(\mathcal{D}\right)\right)^M = \|x\|^2 e^{-\lambda(\mathcal{D}) M} \tag{3.19}$$

où $\lambda(\mathcal{D}) = -\log(1 - \beta(\mathcal{D}))$ est une borne inférieure sur le taux de convergence. Lorsque la dimension N est grande et que $\mathcal{D}$ n'est pas trop redondant, $\beta \ll 1$, si bien que $\lambda \approx \beta$. Lorsque le choix de molécule est fait avec la condition affaiblie (3.16), on sait que la convergence est toujours garantie, mais la borne sur la vitesse asymptotique de décroissance de l'énergie du résidu est affaiblie d'un facteur ρ, *i.e.* $\lambda = -\log(1 - \rho\beta) \approx \rho\beta$.

Dans ce chapitre on s'intéresse à deux dictionnaires de molécules. Le dictionnaire $\mathcal{D}_r$ de molécules "di-atomiques" est le cadre naturel pour analyser des signaux à valeurs réelles avec une poursuite sur des atomes réels. Par ailleurs, dans le cadre de l'analyse de signaux musicaux, le dictionnaire de molécules *harmoniques* $\mathcal{D}_h$ permet de décomposer un signal en structures harmoniques.

3.2 Matching Pursuit avec des atomes réels

Le Matching Pursuit moléculaire est le bon cadre pour définir la poursuite avec des atomes temps-fréquence à valeurs réelles

$$g_{(s,u,\xi,\phi)} = K_{(s,u,\xi,\phi)} g\left(\frac{t-u}{s}\right) \cos\left(\xi(t-u) + \phi\right) \tag{3.20}$$

où $K_{s,u,\xi,\phi}$ est un facteur de normalisation $\boldsymbol{L}^2$. En effet, la procédure *ad hoc* de sélection d'un "bon" atome réel suggérée par Mallat et Zhang [MZ93] n'est pas optimale. Elle consiste à choisir le meilleur atome complexe (voir (2.35))

$$\gamma_m = (s_m, u_m, \xi_m) = \arg\max_\gamma \left| \left\langle R^{m-1} x, g_\gamma \right\rangle \right| \tag{3.21}$$

et à utiliser comme phase l'*argument* de son produit scalaire avec le résidu

$$e^{i\phi_{arg,m}} = \frac{\langle R^{m-1}x, g_{\gamma_m}\rangle}{|\langle R^{m-1}x, g_\gamma\rangle|}. \tag{3.22}$$

Le but est de sélectionner un atome réel g_{γ_m,ϕ_m} sans balayer le paramètre de phase ϕ, afin de limiter la complexité de la poursuite. Le formalisme du Matching Pursuit moléculaire permet d'atteindre ce but en fournissant l'atome réel optimal.

3.2.1 Molécules "di-atomiques" réelles

Comme l'ont fait remarquer Bergeaud [Ber95] et Goodwin [Goo97], chaque atome réel $g_{\gamma,\phi}$ est associé à un atome complexe g_γ et à son conjugué $\overline{g_\gamma}$. Il vérifie

$$g_{\gamma,\phi} = \frac{K_{\gamma,\phi}}{2}\left(e^{i\phi}g_\gamma + e^{-i\phi}\overline{g_\gamma}\right), \tag{3.23}$$

où $K_{\gamma,\phi}$ est un facteur de normalisation $\boldsymbol{L}^2$. L'ensemble des vecteurs $g_{\gamma,\phi}$, lorsque ϕ varie, engendre donc l'espace engendré par g_γ et $\overline{g_\gamma}$

$$\boldsymbol{V}_\gamma \stackrel{\Delta}{=} Vect\,\{g_\gamma, \overline{g_\gamma}\} \tag{3.24}$$

La projection orthogonale $P_{\boldsymbol{V}_\gamma}R^{m-1}x$ du résidu $R^{m-1}x$ sur $\boldsymbol{V}_\gamma$ est un vecteur dont la direction est l'atome réel $g_{\gamma,\phi}$ de phase optimale

$$\sup_{\gamma,\phi}\left|\langle R^{m-1}x, g_{\gamma,\phi}\rangle\right|^2 = \sup_\gamma \sup_\phi \left|\langle R^{m-1}x, g_{\gamma,\phi}\rangle\right|^2 = \sup_\gamma \left\|P_{\boldsymbol{V}_\gamma}R^{m-1}x\right\|^2. \tag{3.25}$$

La sélection du meilleur atome temps-fréquence réel $g_{\gamma,\phi}$ est donc équivalente à celle de la meilleure molécule $\boldsymbol{V}_\gamma$ en fonction de l'énergie $\left\|P_{\boldsymbol{V}_\gamma}R^{m-1}x\right\|^2$ de la projection sur ce sous-espace[2] de dimension 2.

3.2.2 Complétude du dictionnaire de molécules di-atomiques

Le dictionnaire $\mathcal{D}_r$ de molécules "di-atomiques" de Gabor (3.24) est complet, car il engendre le même sous-espace vectoriel de $\boldsymbol{L}^2(\mathbb{R})$ que le dictionnaire des atomes de Gabor complexes $\mathcal{D}_c$. En effet $\mathcal{D}_c$ est stable par

[2] En filant la métaphore physique associée à la définition d'*atomes*, on va appeler ce sous-espace une *molécule di-atomique*. De façon tout à fait analogue, dans le monde physique, les *molécules di-atomiques* d'oxygène sont faites d'*atomes* d'oxygène réunis en paires. Il en est de même des molécules d'hydrogène, de chlore,...Le même état de fait se retrouve dans le monde des atomes appelés à représenter un signal réel : chaque atome complexe fait partie d'une paire en étant associé à son conjugué. L'analyse de signaux à valeurs réelles fait donc plutôt intervenir des *molécules di-atomiques* que des atomes, c'est-à-dire des plans complexes plutôt que des droites.

passage au conjugué, puisque $\overline{g_{(s,u,\xi)}} = g_{(s,u,-\xi)}$. Comme $\boldsymbol{V}_\gamma = Vect\{g_\gamma, \overline{g_\gamma}\}$ et $\mathcal{D}_c = \{g_\gamma, \gamma \in \Gamma_{atom}\} = \{\overline{g_\gamma}, \gamma \in \Gamma_{atom}\}$, on a bien

$$\boldsymbol{W} = Vect\{g_\gamma, \gamma \in \Gamma_{atom}\} = Vect\{\boldsymbol{V}_\gamma, \gamma \in \Gamma_{atom}\}. \tag{3.26}$$

En vertu de cette complétude, la poursuite sur des molécules di-atomiques réelles est donc convergente.

3.2.3 Projection orthogonale sur une molécule di-atomique

La projection orthogonale sur une molécule di-atomique $\boldsymbol{V}_\gamma$ se calcule sans problème car on connaît une base $g_\gamma, \overline{g_\gamma}$ de $\boldsymbol{V}_\gamma$, et sa base bi-orthogonale[3] $\widetilde{g_\gamma}, \widetilde{\overline{g_\gamma}}$:

$$\widetilde{g_\gamma} = \frac{1}{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2} \{g_\gamma - \langle g_\gamma, \overline{g_\gamma} \rangle \overline{g_\gamma}\} \tag{3.27}$$

$$\widetilde{\overline{g_\gamma}} = \overline{\widetilde{g_\gamma}} \tag{3.28}$$

si bien que

$$P_{\boldsymbol{V}_\gamma} R^{m-1} x = \left\langle R^{m-1} x, g_\gamma \right\rangle \widetilde{g_\gamma} + \left\langle R^{m-1} x, \overline{g_\gamma} \right\rangle \widetilde{\overline{g_\gamma}}, \tag{3.29}$$

et

$$\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2 = \frac{2\Re \left\{ \left| \left\langle R^{m-1} x, g_\gamma \right\rangle \right|^2 - \langle g_\gamma, \overline{g_\gamma} \rangle \left\langle R^{m-1} x, g_\gamma \right\rangle^2 \right\}}{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2} \tag{3.30}$$

Cas particulier des atomes temps-fréquence symétriques

Pour des atomes temps-fréquence g_γ construits à partir d'une fenêtre *symétrique* $g(t)$ (ce qui est le cas des atomes gaussiens), le produit scalaire $\langle g_\gamma, \overline{g_\gamma} \rangle$ est un nombre *réel*. L'équation (3.30) se simplifie alors en

$$\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2 = 2 \frac{1 - \langle g_\gamma, \overline{g_\gamma} \rangle \cos 2\phi_{arg}}{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2} \left| \left\langle R^{m-1} x, g_\gamma \right\rangle \right|^2 \tag{3.31}$$

où ϕ_{arg} est l'argument de $\left\langle R^{m-1} x, g_\gamma \right\rangle$[4]. L'atome réel *optimal* g_{γ_m, ϕ_m} vérifie donc

$$P_{\boldsymbol{V}_\gamma} R^{m-1} x = \left\langle R^{m-1} x, g_{\gamma_m, \phi_m} \right\rangle g_{\gamma_m, \phi_m} = \left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\| g_{\gamma_m, \phi_m} \tag{3.32}$$

[3] Un traitement particulier intervient lorsque g_γ est *déjà* un atome réel, auquel cas il est égal à son conjugué et $\boldsymbol{V}_\gamma$ est de dimension 1 au lieu de 2. Aucune notion de phase n'intervient alors, et l'on a tout simplement $P_{\boldsymbol{V}_\gamma} R^{m-1} x = \left\langle R^{m-1} x, g_\gamma \right\rangle g_\gamma$ et $\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2 = \left| \left\langle R^{m-1} x, g_\gamma \right\rangle \right|^2$

[4] On peut cependant avoir besoin de l'équation (3.30) lorsque la fenêtre n'est pas symétrique : c'est le cas du dictionnaire de sinusoïdes amorties employé par Goodwin [Goo97] ; de même au chapitre 5, comme on ajoute un paramètre de *chirp* dans la définition des atomes, on doit faire appel à l'expression (3.30).

et a pour phase

$$e^{i\phi_m} = \frac{\left\langle R^{m-1}x, \widetilde{g_\gamma}\right\rangle}{\left|\left\langle R^{m-1}x, \widetilde{g_\gamma}\right\rangle\right|} \tag{3.33}$$

$$= \frac{\left\langle R^{m-1}x, g_\gamma\right\rangle - \left\langle \overline{g_\gamma}, g_\gamma\right\rangle \left\langle R^{m-1}x, \overline{g_\gamma}\right\rangle}{\left|\left\langle R^{m-1}x, g\right\rangle - \left\langle \overline{g_\gamma}, g_\gamma\right\rangle \left\langle R^{m-1}x, \overline{g_\gamma}\right\rangle\right|} \tag{3.34}$$

Pour le sélectionner, il suffit de choisir $\gamma_m = (s_m, u_m, \xi_m)$ qui rend maximale la corrélation

$$\gamma_m = \arg\max_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\| \tag{3.35}$$

calculée à l'aide des produits scalaires $\langle R^m x, g_\gamma\rangle$ avec les atomes complexes. Pour l'indice sélectionné, *et pour celui-là seulement*, il reste à calculer la phase ϕ_m d'après l'équation (3.33). Cette procédure *exacte* est manifestement différente de la procédure *approximative* proposée par Mallat et Zhang, puisque la phase exacte ϕ_m vérifie l'équation (3.33), alors que la phase *ad hoc* $\phi_{arg,m}$ vérifie (3.22), et que l'indice γ_m est choisi avec le critère (3.35) au lieu de (3.21).

La procédure optimale de calcul de la phase n'augmente pas la complexité des calculs d'un facteur mesurable, comme on le verra au chapitre 4. Par contre elle augmente sensiblement la vitesse de décroissance de l'énergie du résidu $\left\|R^M x\right\|^2$, améliorant ainsi la qualité d'approximation lorsque le nombre d'itérations M est fixé. C'est cette amélioration que nous étudions maintenant.

3.2.4 Amélioration de l'approximation à M atomes réels

Nous comparons ici la procédure *ad hoc* de choix du meilleur atome réel avec le choix optimal, dans le cadre du dictionnaire de Gabor. A partir des définitions (3.22) et (3.33), comme $\tan\phi = \Im(e^{i\phi})/\Re(e^{i\phi})$, on établit la relation

$$\tan\phi = \frac{1 + \left\langle \overline{g_\gamma}, g_\gamma\right\rangle}{1 - \left\langle \overline{g_\gamma}, g_\gamma\right\rangle} \tan\phi_{arg} \tag{3.36}$$

entre la phase *optimale* ϕ_{opt} et l'argument ϕ_{arg} du produit scalaire $\left\langle R^{m-1}x, g_\gamma\right\rangle$. Les deux phases sont donc quasiment identiques lorsque $\left\langle \overline{g_\gamma}, g_\gamma\right\rangle \approx 0$.

Par ailleurs d'après (3.22) et (3.23), on a

$$\left|\left\langle R^{m-1}x, g_{\gamma,\phi_{arg}}\right\rangle\right|^2 = K^2_{\gamma,\phi_{arg}} \left|\left\langle R^{m-1}x, g_\gamma\right\rangle\right|^2. \tag{3.37}$$

et comme le facteur de normalisation utilisé en (3.23) vaut

$$K^2_{\gamma,\phi_{arg}} = \frac{2}{1 + \Re\left(e^{-2i\phi_{arg}} \left\langle \overline{g_\gamma}, g_\gamma\right\rangle\right)} = \frac{2}{1 + \left\langle \overline{g_\gamma}, g_\gamma\right\rangle \cos 2\phi_{arg}}, \tag{3.38}$$

on aboutit finalement, en utilisant (3.31), à la relation

$$\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2 = \frac{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2 \cos^2 2\phi_{arg}}{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2} \left| \left\langle R^{m-1} x, g_{\gamma,\phi_{arg}} \right\rangle \right|^2. \quad (3.39)$$

La perte engendrée par le choix de phase *ad hoc* est mesurée par

$$\rho\left(R^{m-1}x, g_\gamma\right) \triangleq \frac{\left|\left\langle R^{m-1}x, g_\gamma \right\rangle\right|^2}{\left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\|^2} = \frac{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2}{1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2 \cos^2 2\phi_{arg}}. \quad (3.40)$$

Elle est comprise entre $1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2$ et 1. Or dans le dictionnaire de Gabor, grâce à la relation (3.73), on a

$$\langle \overline{g_\gamma}, g_\gamma \rangle = \widehat{g^2}(-2s\xi) \quad (3.41)$$

Perte dans le dictionnaire de Gabor

Pour la plupart des atomes temps-fréquence du dictionnaire multi-échelle de Gabor, on a donc $\langle g_\gamma, \overline{g_\gamma} \rangle \approx 0$ car[5] $\xi \gg 1/s$, si bien que $\rho(R^{m-1}x, g_\gamma) \approx 1$, c'est-à-dire qu'il n'y a pas de perte.

Les atomes g_γ de fréquence nulle sont égaux à leur conjugué[6], et dans leur cas $\boldsymbol{V}_\gamma$ est de dimension 1 et $\rho(R^{m-1}x, g_\gamma) = 1$.

Dans le cas limite où la fréquence[7] ξ est petite devant la résolution fréquentielle $1/s$, l'atome et son conjugué interagissent, et $\rho(R^{m-1}x, g_\gamma)$ peut, au pire, descendre jusqu'à la valeur $1 - |\langle \overline{g_\gamma}, g_\gamma \rangle|^2$.

Comparaison théorique des vitesses de décroissance

Il est *a priori* difficile de comparer directement les vitesses de convergence des algorithmes de poursuite réelle avec phase optimale ou *ad hoc*, car les séquences d'atomes $g_{\gamma_{arg,m}}$ et $g_{\gamma_{opt,m}}$ qu'ils produisent sont distinctes. La relation (3.40), qui mesure la perte sur un atome particulier, ne permet donc pas de comparer de façon *déterministe* les énergies $\left\| P_{\boldsymbol{V}_{\gamma_{opt,m}}} R_{opt}^{m-1} x \right\|^2$ et $\left| \left\langle R_{arg}^{m-1} x, g_{\gamma_{arg,m}} \right\rangle \right|^2$, si bien qu'il est impossible de prédire exactement l'écart qui va exister entre les énergies résiduelles $\left\| R_{arg}^m x \right\|^2$ et $\left\| R_{opt}^m x \right\|^2$.

On peut cependant pressentir que la différence entre les deux algorithmes, en terme de vitesse de décroissance de l'énergie du résidu, se traduira par une perte asymptotique moyenne. Celle-ci dépendra à la fois de la *fréquence*

[5] Cela correspond simplement au fait que les supports fréquentiels de g_γ et $\overline{g_\gamma}$ sont essentiellement disjoints.

[6] Lorsque l'on travaille sur des signaux discrets, la même situation se produit pour les atomes g_γ à la fréquence de Nyquist.

[7] Ou sa différence avec la fréquence de Nyquist, dans le cas de signaux discrets.

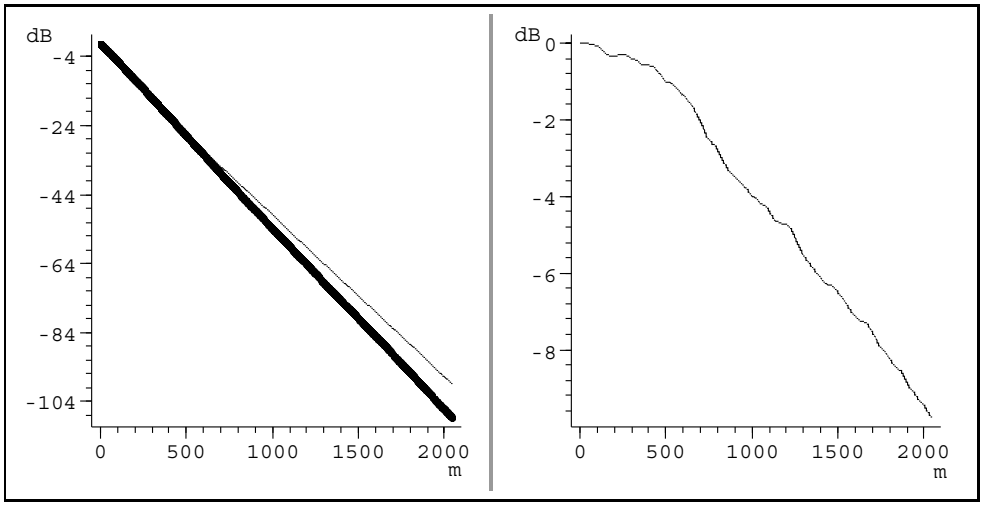

FIG. 3.1 – Décroissances de l'énergie (en décibels) du résidu $\|R^m x\|$ d'un Matching Pursuit Réel effectué sur un bruit blanc gaussien, en fonction du nombre d'itérations m. -A gauche : en gras, avec la phase optimale ; en traits simples, avec la phase *ad hoc*. -A droite : évolution de la différence entre les deux courbes du cadre de gauche. Après 2000 itérations on atteint quasiment $10dB$ de différence en faveur de la phase optimale.

d'apparition[8] des atomes où une perte ($\rho < 1$) est possible, et de la valeur effective ρ de la perte alors engendrée.

Comparaison numérique des décroissances asymptotiques

C'est numériquement que nous illustrons maintenant la perte asymptotique engendrée par la procédure *ad hoc*. On peut observer sur la figure 3.1 la différence de comportement entre les deux algorithmes de poursuite réelle : on a analysé un bruit blanc gaussien de 1024 points avec chacune des méthodes de calcul de la phase, en utilisant un dictionnaire temps-fréquence multi-échelle de Gabor ; les courbes de gauche représentent la décroissance de l'énergie du résidu, en décibels, en fonction du nombre d'itérations M : la courbe du haut (en trait fin) correspond à l'utilisation de la phase *ad hoc*, celle du bas (en trait gras) à la phase optimale. La courbe de droite montre la différence entre les deux courbes de gauche : après 2000 itérations, on atteint quasiment $10dB$ de différence en faveur de la phase optimale.

[8] Le *bruit de dictionnaire* décrit par Davis [Dav94] [DMA] modélise le comportement asymptotique du résidu $R^M x$, et ses propriétés pourraient permettre d'estimer cette fréquence d'apparition.

3.2.5 Représentation temps-fréquence associée

On peut construire une représentation temps-fréquence [MZ93]

$$E_{complexe}[x](t,\omega) \triangleq \sum_{m=1}^{\infty} \left|\left\langle R^{m-1}x, g_{\gamma_m}\right\rangle\right|^2 WV[g_{\gamma_m}](t,\omega) \tag{3.42}$$

à partir de la décomposition atomique (2.54) d'un signal x. Cette représentation est exempte des termes oscillants qui apparaîssent dans les représentations bi-linéaires telles que la transformée de Wigner-Ville [Fla93].

Comme les atomes de Gabor réels $g_{\gamma,\phi}$ sont combinaison linéaire (3.23) d'atomes de Gabor complexes leur représentation temps-fréquence est

$$\begin{aligned} E_{complexe}[g_{\gamma,\phi}](t,\omega) &= \frac{K_{\gamma,\phi}^2}{4}\left\{WV[g_\gamma](t,\omega) + WV[\overline{g_\gamma}](t,\omega)\right\} \\ &= \frac{K_{\gamma,\phi}^2}{4}\left\{WV[g_\gamma](t,\omega) + WV[g_\gamma](t,-\omega)\right\} \end{aligned} \tag{3.43}$$

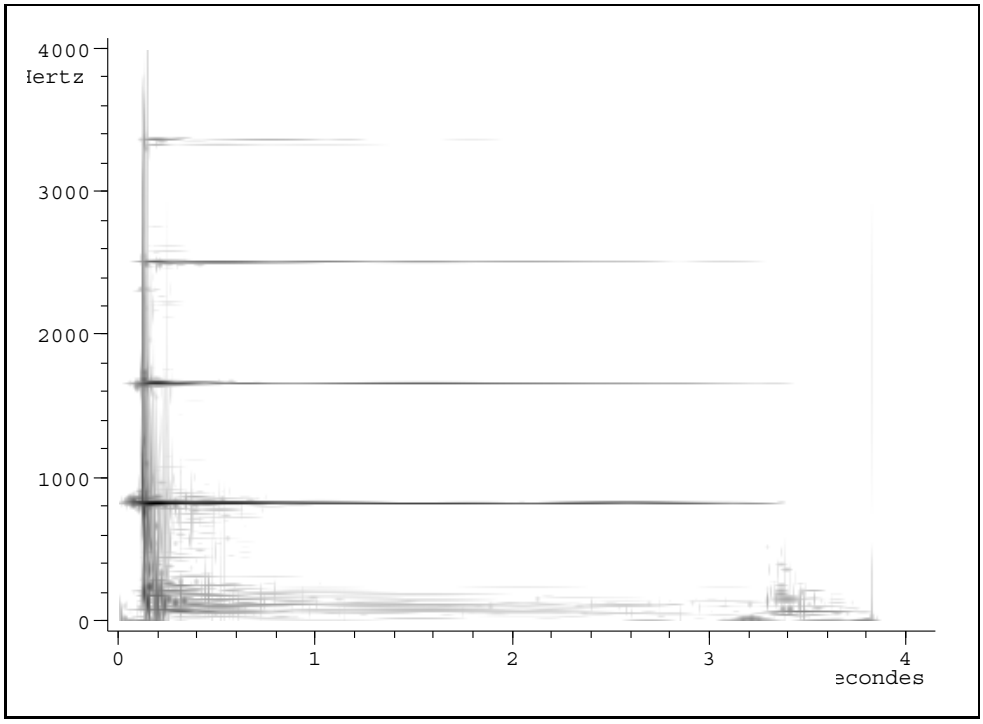

FIG. 3.2 – Représentation temps-fréquence d'un son de piano, obtenue à l'aide d'un Matching Pursuit avec un dictionnaire multi-échelle de Gabor gaussien réel (1000 atomes). On peut y lire la présence simultanée de transitoires et de structures fréquentielles quasi-harmoniques.

A partir de la décomposition

$$x = \sum_{m=1}^{\infty} \left| \left\langle R^{m-1} x, g_{\gamma_m,\phi_m} \right\rangle \right|^2 g_{\gamma_m,\phi_m} \tag{3.44}$$

d'un signal x sur le dictionnaire de Gabor réel, on définit

$$E_{reel}[x] = \sum_{m=1}^{\infty} \left| \left\langle R^{m-1} x, g_{\gamma_m,\phi_m} \right\rangle \right|^2 E_{complexe}[g_{\gamma_m,\phi_m}](t,\omega) \tag{3.45}$$

Un exemple d'une telle représentation est montré sur la figure 3.2, où l'on observe la représentation temps-fréquence d'une note de piano, obtenue à partir des 1000 premiers atomes temps-fréquence de Gabor réels sélectionnés par un Matching Pursuit. On peut y lire la présence simultanée de structures temps-fréquence à différentes échelles : la structure la plus visible, sans doute, est la partie quasi-harmonique de la note, de fréquence fondamentale $820Hz$, représentée par les lignes horizontales. Chacune est associée à quelques atomes à grande échelle, bien localisés en fréquence. On repère, au début et à la fin de la note, des structures verticales, associées à des atomes de petite échelle. Elles sont adaptées à la présence à ces instants de parties transitoires de faible durée : l'attaque et la chute des étouffoirs sur la corde du piano. Enfin, en observant de près la représentation, on peut également constater qu'une structure harmonique à grande échelle est présente en-dessous de $100Hz$, avec une fréquence fondamentale de l'ordre de $20Hz$: elle correspond à la résonance de la table d'harmonie du piano.

Les structures harmoniques, telles que la résonance de la note ou celle de la table dans ce son de piano, sont omniprésentes dans les sons musicaux. Le Matching Pursuit Harmonique que nous introduisons ci-après permet de décomposer un son en structures harmoniques élémentaires au lieu d'atomes temps-fréquence élémentaires.

3.3 Matching Pursuit Harmonique

Le Matching Pursuit Harmonique est un Matching Pursuit moléculaire effectué sur un dictionnaire $\mathcal{D}_h$ de molécules harmoniques que l'on va maintenant définir.

3.3.1 Molécules harmoniques

Le dictionnaire de molécules harmoniques est constitué de sous-espaces associés à des *structures harmoniques*. Pour des fréquences

$$0 < \xi_1 < \ldots < \xi_K$$

on note $\vec{\xi} = (\xi_k)_{1 \leq k \leq K}$ et on définit la molécule harmonique $\boldsymbol{V}_{(s,u,\vec{\xi})}$

$$\boldsymbol{V}_{(s,u,\vec{\xi})} \stackrel{\Delta}{=} Vect\left\{g_{(s,u,\xi_k)}, \overline{g_{(s,u,\xi_k)}}, k \in [\![1, K]\!]\right\}. \tag{3.46}$$

Tout signal de ce sous-espace est combinaison linéaire de $2K$ atomes complexes. Chaque molécule harmonique est donc de dimension au plus $2K$, et peut être représentée par l'indice

$$\gamma \stackrel{\Delta}{=} (s, u, \vec{\xi}). \tag{3.47}$$

Les signaux *réels* de $\boldsymbol{V}_\gamma$ sont combinaison de K atomes *réels* $g_{(s,u,\xi_k,\phi_k)}$, appelés *partiels*. L'ensemble de molécules harmoniques que l'on va considérer est caractérisé par la loi des partiels $\xi_k(\xi_1)$ et le domaine de fréquences fondamentales, qui constituent des *contraintes* rendant compte des informations de haut niveau et des modélisations *a priori* du signal dont on dispose.

3.3.2 Loi des partiels

En première approximation, le k-ème partiel ξ_k est relié à la fréquence fondamentale ξ_1 (souvent notée f_0 dans la littérature [Dov94]) par la relation d'harmonicité

$$\xi_k \approx k\xi_1. \tag{3.48}$$

Bien que des modélisations fines [Fle62] de la production physique du signal puissent préciser des écarts à l'harmonicité, nous nous contenterons ici de la loi harmonique approximative[9]. L'étalement spectral de $g_{(s,u,\xi_k)}$ étant de l'ordre de $1/s$, on impose au k-ème partiel d'appartenir à l'intervalle fréquentiel

$$\xi_k \in I_k(s, \xi_1) \stackrel{\Delta}{=} \left[k\xi_1 - \frac{\mu}{2s}, k\xi_1 + \frac{\mu}{2s}\right] \tag{3.49}$$

où μ est un paramètre de tolérance qui autorise de plus ou moins grandes déviations par rapport à l'harmonicité stricte. Pour respecter l'ordre des partiels $\xi_1 < \ldots < \xi_k \ldots < \xi_K$, il suffira que la fréquence fondamentale vérifie

$$\xi_1 > \mu/s \tag{3.50}$$

3.3.3 Domaine de fréquences fondamentales

On ne veut chercher que des structures harmoniques significatives[10]. Il faut donc restreindre autant que possible le domaine

$$I_1(u, s) \stackrel{\Delta}{=} [\xi_1^{min}(s, u), \xi_1^{max}(s, u)] \tag{3.51}$$

[9] On pourrait adapter la loi des partiels aux connaissances *a priori* sur l'inharmonicité du signal analysé (*e. g.* le type d'instrument joué).

[10] Pour limiter les erreurs "d'octave" [Dov94] sur le choix de leur fréquence fondamentale.

dans lequel peut varier la fréquence fondamentale ξ_1 à chercher. Cette information de haut niveau peut venir de la *tessiture* du (des) instrument(s) joués, d'un *pré-traitement* (détection de fréquence fondamentale [Dov94]) voire, dans le cadre du suivi de partition, d'informations *a priori* sur les *instants d'arrivée*, les *durées* probables et les *hauteurs* des notes attendues. Le domaine I_1 peut donc bien dépendre de s et u, comme exprimé en (3.51).

3.3.4 Complétude du dictionnaire de molécules harmoniques

Tout dictionnaire qui *contient* un dictionnaire complet est complet. Pour nous assurer la complétude du dictionnaire de molécules employé dans la poursuite harmonique (et donc la convergence de la poursuite), il suffit donc d'y inclure le dictionnaire d'atomes de Gabor réels, qui est complet. Le dictionnaire moléculaire $\mathcal{D}_h$ utilisé dans la poursuite harmonique est donc défini comme la *réunion* du dictionnaire de Gabor réel et de l'ensemble des molécules harmoniques

$$\mathcal{D}_h \stackrel{\Delta}{=} \mathcal{D}_r \cup \{\boldsymbol{V}_\gamma, \gamma \in \Gamma_h\}. \tag{3.52}$$

3.3.5 Choix approché de la meilleure molécule harmonique

A chaque itération de poursuite, il faut choisir une meilleure molécule harmonique, ce qui nécessite de calculer $\left\|P_{\boldsymbol{V}_\gamma} R^{m-1}x\right\|^2$. A partir d'une base orthonormale $(g_k)_{1\leq|k|\leq K}$ de $\boldsymbol{V}_\gamma$, ce calcul est immédiat

$$\left\|P_{\boldsymbol{V}_\gamma} R^{m-1}x\right\|^2 = \sum_k \left|\left\langle R^{m-1}x, g_k\right\rangle\right|^2. \tag{3.53}$$

de même que celui de la projection orthogonale

$$P_{\boldsymbol{V}_\gamma} R^{m-1}x = \sum_k \left\langle R^{m-1}x, g_k\right\rangle g_k. \tag{3.54}$$

Quand on ne dispose pas d'une telle base orthonormale, on peut effectuer une sélection *approximative* de la meilleure molécule harmonique $\boldsymbol{V}_{\gamma_m}$, à partir des corrélations avec les atomes réels optimaux. Soit en effet une molécule harmonique de dimension $2K$

$$\boldsymbol{V}_\gamma = Vect\left\{g_{\gamma_k}, \overline{g_{\gamma_k}},\ 1\leq k\leq K\right\}. \tag{3.55}$$

Elle est somme des K sous-espaces de dimension 2

$$\boldsymbol{W}_k = Vect\left\{g_{\gamma_k}, \overline{g_{\gamma_k}}\right\}. \tag{3.56}$$

Or l'énergie du résidu dans la direction de $\boldsymbol{W}_k$

$$\left\|P_{\boldsymbol{W}_k} R^{m-1}x\right\|^2 = \sup_{\phi_k} \left|\left\langle R^{m-1}x, g_{\gamma_k,\phi_k}\right\rangle\right|^2 \tag{3.57}$$

est connue grâce aux résultats de la section 3.2.3. Définissons alors la corrélation approchée entre le résidu et une molécule

$$C(R^{m-1}x, \boldsymbol{V}_\gamma) \triangleq \sum_{k=1}^{K} \left\| P_{\boldsymbol{W}_k} R^{m-1}x \right\|^2 \tag{3.58}$$

$$= \left\langle R^{m-1}x, \sum_k P_{\boldsymbol{W}_k} R^{m-1}x \right\rangle . \tag{3.59}$$

La corrélation $C(u, \boldsymbol{V}_\gamma)$, restreinte aux vecteurs u de $\boldsymbol{V}_\gamma$, est la forme quadratique définie positive associée à l'opérateur $L_\gamma = \sum_k P_{\boldsymbol{W}_k}$. Elle vérifie donc l'encadrement

$$A_\gamma \|u\|^2 \leq C(u, \boldsymbol{V}_\gamma) \leq B_\gamma \|u\|^2 \tag{3.60}$$

où les bornes optimales A_γ et B_γ sont respectivement la plus petite et la plus grande des valeurs propres de cet opérateur.

Théorème 2 *Si*

$$\rho \triangleq \frac{\inf_\gamma A_\gamma}{\sup_\gamma B_\gamma} > 0 \tag{3.61}$$

alors la poursuite effectuée en remplaçant la sélection (3.6) de la meilleure molécule par le critère approché[11] *(3.58)*

$$\gamma_m = \arg\max_\gamma C(R^{m-1}x, \boldsymbol{V}_\gamma) \tag{3.62}$$

est convergente.

Preuve

D'abord, comme $\boldsymbol{W}_k \subset \boldsymbol{V}_\gamma$, on a $P_{\boldsymbol{W}_k} P_{\boldsymbol{V}_\gamma} = P_{\boldsymbol{W}_k}$, et donc

$$C(R^{m-1}x, \boldsymbol{V}_\gamma) = C(P_{\boldsymbol{V}_\gamma} R^{m-1}x, \boldsymbol{V}_\gamma). \tag{3.63}$$

[11] Remarque : on pourrait également raisonner à l'aide de la corrélation "renormalisée"

$$C'(x, \boldsymbol{V}_\gamma) \triangleq \frac{1}{A_\gamma} C(x, \boldsymbol{V}_\gamma)$$

auquel cas on aboutit à

$$\left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1}x \right\|^2 \geq \frac{A_{\gamma_m}}{B_{\gamma_m}} \sup_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2$$

et il suffit que

$$\rho' \triangleq \inf_\gamma \frac{A_\gamma}{B_\gamma} > 0$$

pour prouver la convergence.

D'après l'encadrement (3.60), on obtient donc

$$\frac{1}{B_\gamma} C(R^{m-1}x, \boldsymbol{V}_\gamma) \leq \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2 \leq \frac{1}{A_\gamma} C(R^{m-1}x, \boldsymbol{V}_\gamma) \tag{3.64}$$

Par conséquent

$$\begin{aligned}
\sup_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2 &\leq \sup_\gamma \left\{ \frac{1}{A_\gamma} C(R^{m-1}x, \boldsymbol{V}_\gamma) \right\} && (3.65)\\
&\leq \frac{1}{\inf_\gamma A_\gamma} \sup_\gamma C(R^{m-1}x, \boldsymbol{V}_\gamma) && (3.66)\\
&\leq \frac{1}{\inf_\gamma A_\gamma} C(R^{m-1}x, \boldsymbol{V}_{\gamma_m}) && (3.67)\\
&\leq \frac{B_{\gamma_m}}{\inf_\gamma A_\gamma} \left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1}x \right\|^2 && (3.68)
\end{aligned}$$

et donc

$$\left\| P_{\boldsymbol{V}_{\gamma_m}} R^{m-1}x \right\|^2 \geq \frac{\inf_\gamma A_\gamma}{B_{\gamma_m}} \sup_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2 \geq \rho \sup_\gamma \left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2 \tag{3.69}$$

En vertu de la convergence de la version *faible* (3.16) du Matching Pursuit moléculaire, si $\rho > 0$, l'algorithme est convergent.

□.

3.3.6 Quasi-orthogonalité des partiels

Nous allons maintenant établir la borne uniforme (3.61) lorsque la base $\{g_{\gamma_k}, \overline{g_{\gamma_k}}\}$ vérifie une condition de *quasi-orthogonalité* uniforme. On note pour cela $g_k = g_{\gamma_k}$ et $g_{-k} = \overline{g_{\gamma_k}}$:

Théorème 3 *Si, pour toute molécule, la base est telle que*

$$|\langle g_k, g_l \rangle| \leq \varepsilon, \ \forall k \neq l, |k|, |l| \in [\![1, K]\!] \tag{3.70}$$

avec

$$\varepsilon < 1/K$$

alors la poursuite harmonique converge.

Preuve

On commence par établir l'encadrement

$$\frac{2}{1+\varepsilon} \sum_{k=1}^{K} |\langle x, g_k \rangle|^2 \leq C(x, \boldsymbol{V}_\gamma) \leq \frac{2}{1-\varepsilon} \sum_{k=1}^{K} |\langle x, g_k \rangle|^2 \tag{3.71}$$

pour tout signal réel $x \in \boldsymbol{V}_\gamma$. On établit ensuite un encadrement des valeurs propres de l'opérateur associé à la forme quadratique $x \mapsto \sum_k |\langle x, g_k\rangle|^2$. On en déduit une borne uniforme pour $C(x, \boldsymbol{V}_\gamma)$ permettant d'appliquer le théorème 2.

- Pour tout k, on a d'après (3.30)

$$\frac{2}{1+|\langle g_{\gamma_k}, \overline{g_{\gamma_k}}\rangle|}|\langle x, g_k\rangle|^2 \leq \left\|P_{\boldsymbol{V}_{\gamma_k}} x\right\|^2 \leq \frac{2}{1-|\langle g_{\gamma_k}, \overline{g_{\gamma_k}}\rangle|}|\langle x, g_k\rangle|^2 \tag{3.72}$$

  Comme $|\langle g_{\gamma_k}, \overline{g_{\gamma_k}}\rangle| \leq \varepsilon$, on en déduit l'encadrement (3.71).
- Les valeurs propres de l'opérateur associé à la forme quadratique

$$x \mapsto \sum_k |\langle x, g_k\rangle|^2$$

  sont celles de la matrice de Gram $G = (\langle g_k, g_l\rangle)_{k,l}$ de la famille $(g_k)_{k=1}^K$. D'après l'inégalité de Cauchy-Schwartz, si L_k est le k-ème vecteur ligne de $G - I$, et $\|U\| = 1$

$$\begin{aligned}\|(G-I)U\|^2 &= \sum_k \langle L_k, U\rangle^2 \leq \sum_k \|L_k\|^2 \|U\|^2 = \sum_k \|L_k\|^2 = \sum_{k\neq l} \langle g_k, g_l\rangle^2 \\ &\leq (K^2-K)\varepsilon^2\end{aligned}$$

  donc les valeurs propres de $G - I$ vérifient $|\lambda_k| \leq K\varepsilon$. Les valeurs propres $1 + \lambda_k$ de G sont donc dans l'intervalle

$$[1 - K\varepsilon, 1 + K\varepsilon].$$

- Les valeurs propres de $C(., \boldsymbol{V}_\gamma)$ sont donc dans l'intervalle

$$\left[\frac{1-K\varepsilon}{1+\varepsilon}, \frac{1+K\varepsilon}{1-\varepsilon}\right].$$

  Le rapport ρ entre la plus petite et la plus grande vérifie donc

$$\rho \geq \frac{1-K\varepsilon}{1+K\varepsilon}\frac{1-\varepsilon}{1+\varepsilon} > 0$$

  dès que $K\varepsilon < 1$.

□.

3.3.7 Quasi-orthogonalité dans le dictionnaire de Gabor

La condition de *quasi-orthogonalité* s'exprime dans le dictionnaire de Gabor à l'aide des produits scalaires

$$\begin{aligned}\left\langle g_{(s,u,\xi_k)}, g_{(s,u,\xi_l)}\right\rangle &= \int_{-\infty}^{+\infty} \frac{1}{s} g^2\left(\frac{t-u}{s}\right) e^{-i(\xi_l-\xi_k)t} dt \\ &= \int_{-\infty}^{+\infty} g^2(t) e^{-is(\xi_l-\xi_k)t} dt \\ &= \widehat{g^2}\left(s(\xi_l-\xi_k)\right). \end{aligned} \tag{3.73}$$

Comme $g^2(t)$ est concentrée fréquentiellement autour de 0, la condition de quasi-orthogonalité (3.70) devient

$$\inf_{k \neq l} |\xi_k - \xi_l| \geq \mu_g(\varepsilon)/s \tag{3.74}$$

où $\mu_g(\varepsilon)$ est la plus petite valeur telle que

$$|\omega| \geq \mu_g(\varepsilon) \Longrightarrow \widehat{g^2}(\omega) \leq \varepsilon. \tag{3.75}$$

D'après la loi des partiels (3.49), comme $\xi_{-k} = \xi_k$, on a

$$\begin{aligned} \inf_{k \neq l} |\xi_k - \xi_l| &= \min\left(\inf_{k \geq 1}(\xi_{k+1} - \xi_k), \xi_1 - \xi_{-1}\right) \\ &\geq \min\left(\xi_1 - \mu/s, 2\xi_1\right) = \xi_1 - \mu/s. \end{aligned}$$

Il suffit donc que

$$\xi_1 \geq \left(\mu + \mu_g(\varepsilon)\right)/s \tag{3.76}$$

ce qui est un peu plus contraignant que (3.50). Le domaine de fréquences fondamentales est donc limité par la condition

$$\xi_1^{min}(s,u) > (\mu + \mu_g(\varepsilon))/s. \tag{3.77}$$

3.3.8 Recherche rapide de la molécule la plus corrélée

Pour rechercher la molécule $\boldsymbol{V}_\gamma$ la plus corrélée au résidu, il faut *a priori* parcourir tous les indices $\gamma = (s, u, \vec{\xi}) \in \Gamma_h$ et calculer les corrélations associées. Cependant l'expression (3.58) de la corrélation qu'on utilise nous permet d'effectuer la recherche de l'optimum de façon efficace, en deux temps, et de réduire ainsi le coût de cette sélection.

Recherche rapide des paramètres (s, u, ξ_1) optimaux

Grâce à la forme de la corrélation utilisée, à s, u et $\xi_1 \in I_1(s,u)$ fixés, on peut optimiser *indépendamment* chaque partiel ξ_k

$$\sup_{\xi_k \in I_k(s,\xi_1)} \left\| P_{\boldsymbol{W}_{(s,u,\xi_k)}} R^{m-1} x \right\|^2, k = 2..K \tag{3.78}$$

et obtenir par sommation

$$\begin{aligned} \sup_{\substack{\xi_k \in I_k(s,\xi_1), \\ k=2..K}} C(R^{m-1}x, \boldsymbol{V}_\gamma) &= \left\| P_{\boldsymbol{W}_{(s,u,\xi_1)}} R^{m-1} x \right\|^2 \\ &\quad + \sum_{k=2}^{K} \sup_{\xi_k \in I_k(s,\xi_1)} \left\| P_{\boldsymbol{W}_{(s,u,\xi_k)}} R^{m-1} x \right\|^2. \end{aligned} \tag{3.79}$$

Cela nécessite le calcul du maximum local de $\xi \mapsto \left\| P_{\boldsymbol{W}_{(s,u,\xi)}} R^{m-1} x \right\|^2$ sur les $K-1$ intervalles $I_k(s,\xi_1)$. La longueur de ces intervalles étant de l'ordre du pas d'échantillonnage fréquentiel $1/s$ à l'échelle s, le coût de ce calcul est de l'ordre de

$$\mathcal{O}(K). \tag{3.80}$$

Détermination fine des partiels $\xi_k, k \geq 2$

Une fois que la localisation *grossière* (s_m, u_m, ξ_1^m) de la meilleure molécule, au sens de (3.58), a été déterminée rapidement avec la stratégie que l'on vient de décrire, il reste à déterminer *précisément* la position de ses partiels ξ_k^m. Il suffit, pour cela, de déterminer pour chaque $k = 2..K$

$$\xi_k^m = \arg \max_{\xi_k \in I_k(s_m, \xi_1^m)} \left\| P_{\boldsymbol{W}_{(s_m,u_m,\xi_k)}} R^{m-1} x \right\|^2. \tag{3.81}$$

Afin de ne pas être limité en résolution fréquentielle par le pas de discrétisation fréquentielle du dictionnaire employé numériquement, on effectue une interpolation parabolique du spectre autour du maximum discret trouvé.

3.3.9 Projection sur la molécule sélectionnée

Une fois la molécule $\boldsymbol{V}_{\gamma_m}$ sélectionnée, il faut calculer $P_{\boldsymbol{V}_{\gamma_m}} R^{m-1} x$ pour terminer l'itération de poursuite en cours. On sait effectuer exactement ce calcul dans les molécules di-atomiques, ou lorsque l'on a une base orthogonale de $\boldsymbol{V}_\gamma$. Lorsque la base "naturelle" de $\boldsymbol{V}_\gamma$ dont on dispose n'est pas orthogonale et que la dimension $2K$ est grande, on peut utiliser un des algorithmes itératifs connus de reconstruction dans un frame [Mal98]. Pour calculer la projection avec une erreur relative ε, le nombre d'itérations nécessaires est

- avec l'algorithme de Richardson extrapolé :

$$n_{ER} \approx \frac{1}{2\rho_{\gamma_m}} \log_e \frac{1}{\varepsilon} \tag{3.82}$$

- avec la descente de gradient :

$$n_{GR} \approx \frac{1}{2\sqrt{\rho_{\gamma_m}}} \log_e \frac{2}{\varepsilon} \approx \sqrt{\rho_{\gamma_m}} n_{ER} \tag{3.83}$$

où $\rho_{\gamma_m} = A_{\gamma_m}/B_{\gamma_m}$ mesure l'*étroitesse* du frame de $\boldsymbol{V}_{\gamma_m}$ utilisé. Quand on se place dans la condition suffisante de convergence (3.61), on a $\rho_{\gamma_m} \geq \rho > 0$. La poursuite sera donc d'autant plus efficace, en termes de qualité d'approximation (c'est-à-dire de vitesse de convergence du résidu vers 0) et de complexité algorithmique, que la borne ρ définie en (3.61) sera grande, c'est-à-dire proche de 1.

Avec la condition de quasi-orthogonalité (3.70), si $\varepsilon \ll 1/K$, on a $\rho \approx 1$ et il suffit donc d'une itération pour obtenir

$$P_{\boldsymbol{V}_{\gamma_m}} R^{m-1}x \approx \sum_{k=1}^{K} \left\langle R^{m-1}x, g_{(s_m,u_m,\xi_k^m,\phi_k^m)} \right\rangle g_{(s_m,u_m,\xi_k^m,\phi_k^m)}. \tag{3.84}$$

avec une très faible erreur relative.

3.3.10 Résumé de l'algorithme

Résumons maintenant les grandes lignes d'une itération de Matching Pursuit Harmonique :

1. Calcul des corrélations du résidu avec les atomes complexes
2. Calcul des corrélations avec les "meilleurs" atomes réels
3. Calcul des corrélations "approchées" avec les molécules "grossières"
4. Sélection de la meilleure molécule "grossière"
5. Détermination fine des partiels de la molécule sélectionnée
6. Projection orthogonale sur la molécule "fine" sélectionnée
7. Mise à jour du résidu

3.3.11 Représentation temps-fréquence associée

Comme tout signal $x \in \boldsymbol{V}_{(s,u,\vec{\xi})}$ est combinaison linéaire des atomes réels engendrant la molécule

$$x(t) = \sum_{k=1}^{K} \alpha_k g_{(s,u,\xi_k,\phi_k)}, \tag{3.85}$$

sa représentation temps-fréquence est

$$E_{reel}[x](t,\omega) = \sum_{k=1}^{K} |\alpha_k|^2 \, E_{complexe}[g_{(s,u,\xi_k,\phi_k)}](t,\omega). \tag{3.86}$$

Un exemple d'une telle représentation temps-fréquence est donné à la figure 3.3. Le vecteur x est choisi dans la molécule harmonique de durée $s = 0.5$ seconde, située au temps $u = 0.3$ et dont la fréquence fondamentale est $\xi_1 = 50$ Hertz. Cette molécule est de dimension 6, et le coefficient affecté au troisième partiel dans le vecteur x est nul. La représentation temps-fréquence d'un signal x décomposé en somme de vecteurs $x = \sum_m x_{\gamma_m}, x_{\gamma_m} \in \boldsymbol{V}_{\gamma_m}$ à l'aide du Matching Pursuit Harmonique est

$$E_{harm}[x](t,\omega) \triangleq \sum_{m} E_{reel}[x_m](t,\omega). \tag{3.87}$$

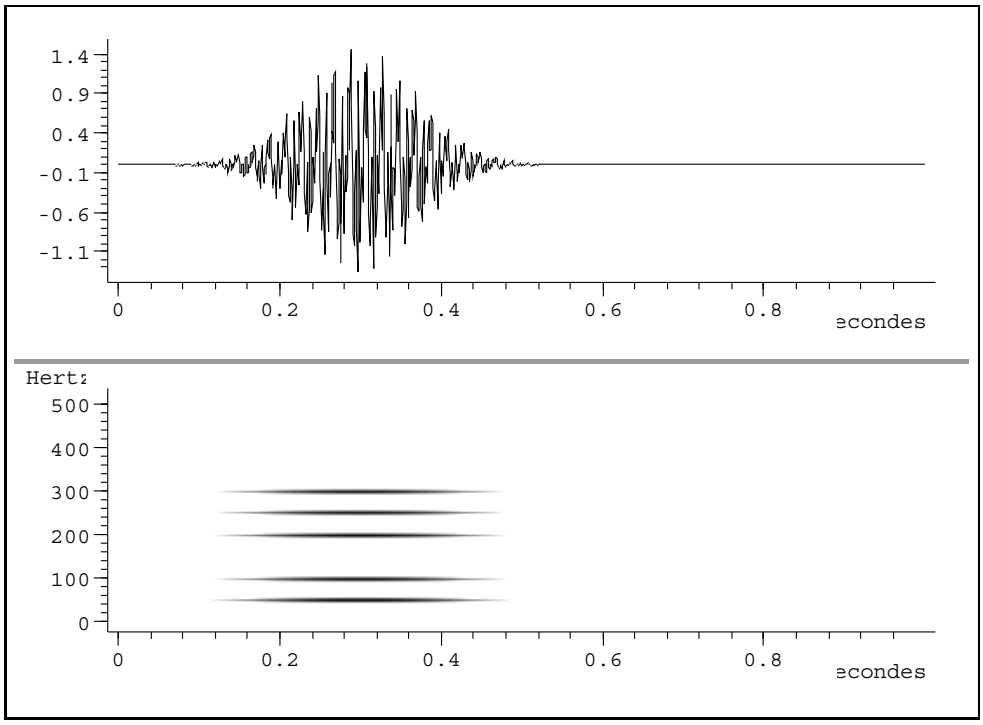

FIG. 3.3 – Représentation temps-fréquence d'un signal x choisi dans une molécule harmonique $\boldsymbol{V}_{(s,u,\vec{\xi})}$ de dimension 6. Le coefficient du troisième partiel est nul.

La figure 3.4 représente une phrase de clarinette, extraite de la pièce *Dialogue de l'ombre double*, de P. Boulez [Bou91] et sa décomposition avec une poursuite harmonique. On y repère la succession des notes qui constituent la phrase. Leur durée se traduit par l'échelle des molécules harmoniques sélectionnées, et leur hauteur par la fréquence fondamentale de celles-ci. La réverbération de la salle "prolonge" chaque note alors que la note suivante a déjà été jouée par l'instrumentiste. Ce phénomène est visible sur la représentation temps-fréquence obtenue, sous la forme du "tuilage" des structures harmoniques qu'on observe entre la première et la deuxième note par exemple. Cela montre que notre méthode peut détecter simultanément plusieurs fondamentales, ce qui lui ouvre les portes de l'analyse de sons polyphoniques. Par ailleurs on peut observer quelques atomes à petite échelle : celui marquant le début de la deuxième note est représenté par une tache verticale à l'instant $t = 0.3$. De tels atomes caractérisent la présence de transitoires. Enfin on peut remarquer sur cette analyse que la fondamentale de chaque structure harmonique est beaucoup plus forte que ses partiels d'ordre supérieur, au point que ceux-ci sont à peine visibles. Ils sont en effet plus de $20dB$ en dessous du fondamental.

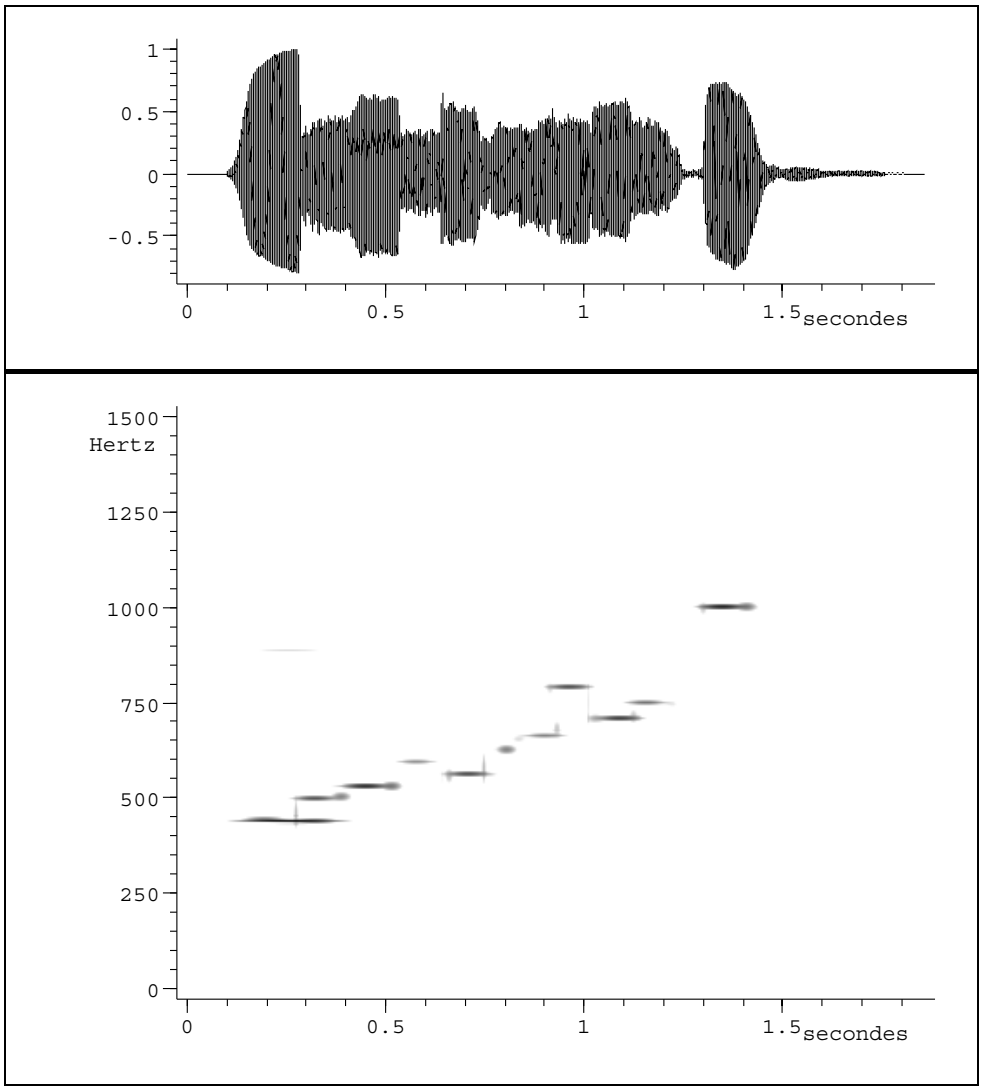

FIG. 3.4 – Décomposition en molécules harmoniques d'une phrase de clarinette (Extrait de *Dialogue de l'ombre double*, de P. Boulez [Bou91]).On y repère la succession des notes qui constituent la phrase. Leur durée se traduit par l'échelle des molécules harmoniques sélectionnées, et leur hauteur par la fréquence fondamentale de celles-ci. Des atomes ou des molécules à petite échelle (comme à l'instant $t = 0.3$) repèrent les transitoires.

Chapitre 4

Matching Pursuit Rapide

Nous nous intéressons dans ce chapitre à la complexité des algorithmes de poursuite, et nous développons une technique d'accélération valable pour le Matching Pursuit dans un dictionnaire de Gabor d'atomes réels et le Matching Pursuit Harmonique que nous venons d'introduire.

Nous commençons par rappeler que la complexité de la poursuite "standard" sur le dictionnaire de Gabor (complexe ou réel) est de l'ordre de

$$\mathcal{O}\left(MN\log^2 N\right), \tag{4.1}$$

tandis que celle du Matching Pursuit Harmonique est

$$\mathcal{O}\left(MN\log N(K+\log N)\right), \tag{4.2}$$

où K est le nombre de partiels d'une molécule harmonique.

Cette complexité est grande devant la complexité $\mathcal{O}(MN)$ de la projection linéaire sur un sous-espace de dimension M déterminé par une base orthogonale. Il est donc intéressant d'envisager les moyens d'accélérer la décomposition.

Nous détaillons alors la méthode d'accélération de la poursuite que nous avons mise au point. Le Matching Pursuit Rapide utilise les maxima locaux du dictionnaire de Gabor et permet d'obtenir une complexité

$$\mathcal{O}(MN). \tag{4.3}$$

De façon analogue, la complexité du Matching Pursuit Harmonique Rapide n'est que de

$$\mathcal{O}(KMN). \tag{4.4}$$

Bien que la décomposition atomique (respectivement moléculaire) fournie par l'algorithme accéléré diffère de celle fournie par l'algorithme standard, nous verrons sur quelques exemples que les structures qu'elle extrait du signal sont néanmoins assez similaires.

4.1 Complexité initiale du Matching Pursuit

Le coût d'une itération de Matching Pursuit, que ce soit avec un dictionnaire d'atomes complexes, réels ou de molécules harmoniques de Gabor, dépend essentiellement du coût algorithmique de calcul de l'ensemble des produits scalaires $\left\langle R^{m-1}x, g_\gamma \right\rangle$ du résidu avec tous les atomes du dictionnaire de Gabor complexe, ainsi que, le cas échéant, des corrélations avec les molécules. Il est utile en effet de rappeler qu'une itération de Matching Pursuit se présente ainsi

1. Calcul des produits scalaires avec les atomes complexes $\left\langle R^{m-1}x, g_\gamma \right\rangle$.
2. Calcul des corrélations $\sup_\phi \left|\left\langle R^{m-1}x, g_{\gamma,\phi} \right\rangle\right|^2$ avec les meilleurs atomes réels.

2-*bis* *Matching Pursuit Harmonique* :
Calcul des corrélations avec les meilleures molécules harmoniques.

3. Sélection du meilleur atome ou de la meilleure molécule
   - *Matching Pursuit Atomique* :
   $$(\gamma_m, \phi_m) \stackrel{\Delta}{=} \arg\max \left|\left\langle R^{m-1}x, g_{\gamma,\phi} \right\rangle\right|$$
   - *Matching Pursuit Harmonique* :
   $$\gamma_m \stackrel{\Delta}{=} \arg\max C(R^{m-1}x, \boldsymbol{V}_\gamma)$$
4. Mise à jour du résidu
   - *Matching Pursuit Atomique* :
   $$R^m x = R^{m-1}x - \left\langle R^{m-1}x, g_{\gamma_m,\phi_m} \right\rangle \tag{4.5}$$
   - *Matching Pursuit Harmonique* : Utilisation de la formule approchée (3.84) de projection sur la molécule choisie
   $$\begin{aligned} R^m x &= R^{m-1}x - P_{\boldsymbol{V}_{\gamma_m}} R^{m-1}x & (4.6)\\ &= R^{m-1}x - \sum_{k=1}^{K} \left\langle R^{m-1}x, g_{(s_m,u_m,\xi_k^m,\phi_k^m)} \right\rangle g_{(s_m,u_m,\xi_k^m,\phi_k^m)} & (4.7) \end{aligned}$$

Nous allons ci-dessous évaluer successivement la complexité de chacune des étapes mises en jeu.

4.1.1 Calcul des produits scalaires avec les atomes complexes

La première étape se traduit par le calcul de P_{atom} produits scalaires, où $P_{atom} \geq N$ est la taille du dictionnaire. Le coût est d'*au moins* 1 opération(s) par atome. Le dictionnaire de Gabor multi-échelle $\mathcal{D}$ échantillonné de manière critique est de taille $P_{atom} = \mathcal{O}(N \log N)$, et la Transformée de

Fourier Rapide permet d'effectuer le calcul de produits scalaires en question, qui correspond au calcul de $\log N$ spectrogrammes, avec des tailles de fenêtre $s = 2^j$ allant de 1 à N de manière dyadique. Sa complexité algorithmique est donc $\log N$ fois celle du calcul d'un spectrogramme, soit

$$\mathcal{O}(N \log^2 N) = \mathcal{O}(P_{atom} \log N). \tag{4.8}$$

ce qui correspond à une coût par atome de $\mathcal{O}(\log N)$.

4.1.2 Calcul des corrélations avec les atomes réels

Le paramètre de phase ϕ, dont l'ajout caractérise le passage du dictionnaire complexe au dictionnaire réel, n'augmente pas la complexité de la poursuite. En effet, en vertu de l'équation (3.25) et de la formule (3.30), on peut calculer $\sup_\phi \left|\left\langle R^{m-1}x, g_{\gamma,\phi}\right\rangle\right|^2$ sans parcourir les différentes valeurs du paramètre de phase, ni même calculer le paramètre de phase optimum : il suffit de disposer des produits scalaires $\left\langle R^{m-1}x, g_\gamma\right\rangle$ et $\langle \overline{g_\gamma}, g_\gamma\rangle$. Or $\left\langle R^{m-1}x, g_\gamma\right\rangle$ est connu grâce à la première étape, et comme $\langle \overline{g_\gamma}, g_\gamma\rangle$ ne dépend pas du résidu, le calcul de $\sup_\phi \left|\left\langle R^{m-1}x, g_{\gamma,\phi}\right\rangle\right|^2$ coûte $\mathcal{O}(1)$ pour chaque γ. Cette seconde étape coûte donc

$$\mathcal{O}(P_{atom}). \tag{4.9}$$

4.1.3 Calcul des corrélations avec les molécules

On a montré en 3.3.8 que ce calcul s'effectue de façon rapide (en $\mathcal{O}(K)$ par molécule "grossière" (s, u, ξ_1)) à l'aide de l'expression (3.79). Or, en échantillonnant le triplet (s, u, ξ_1) comme dans le dictionnaire atomique, aux limitations (3.51) près, on définit $P_{mol} \leq P_{atom} = \mathcal{O}(N \log N)$ localisations "grossières" de molécules. La complexité de cette étape est donc

$$\mathcal{O}(P_{mol} K) \tag{4.10}$$

4.1.4 Sélection du meilleur atome ou de la meilleure molécule

Il suffit de parcourir les P_{atom} atomes pour sélectionner le meilleur en

$$\mathcal{O}(P_{atom}) \tag{4.11}$$

La sélection de la meilleure molécule se fait, elle, en deux temps, comme on l'a vu en 3.3.8 : on parcourt les P_{mol} molécules grossières disponibles afin de sélectionner la meilleure (s_m, u_m, ξ_1^m) en $\mathcal{O}(P_{mol})$, puis on détermine ensuite finement ses K partiels à l'aide de K interpolations de Newton, en $\mathcal{O}(K)$. Cette sélection coûte donc

$$\mathcal{O}(P_{mol} + K) \tag{4.12}$$

4.1.5 Mise à jour du résidu

La projection du résidu sur l'atome ou la molécule sélectionnée se fait en $\mathcal{O}(2^j)$ (respectivement $\mathcal{O}(K2^j)$ pour une molécule de dimension K) si le support temporel de l'atome ou de la molécule est de taille $s = 2^j$. Comme 2^j peut atteindre N, le coût de la mise à jour du résidu est

$$\mathcal{O}(N) \tag{4.13}$$

avec des atomes et

$$\mathcal{O}(KN) \tag{4.14}$$

avec des molécules.

4.1.6 Formules rapides de mise à jour des corrélations

Étant donnée les expressions (4.5) et (4.6) de mise à jour du résidu, on dispose de formules de mise à jour des produits scalaires

$$\langle R^m x, g_\gamma \rangle = \langle R^{m-1} x, g_\gamma \rangle - \langle R^{m-1} x, g_{\gamma_m, \phi_m} \rangle \langle g_{\gamma_m, \phi_m}, g_\gamma \rangle . \tag{4.15}$$

ou

$$\langle R^m x, g_\gamma \rangle = \langle R^{m-1} x, g_\gamma \rangle - \sum_{k=1}^{K} \left\langle R^{m-1} x, g_{(s_m, u_m, \xi_k^m, \phi_k^m)} \right\rangle \left\langle g_{(s_m, u_m, \xi_k^m, \phi_k^m)}, g_\gamma \right\rangle . \tag{4.16}$$

Le produit scalaire est donc inchangé pour tous les atomes g_γ orthogonaux à l'atome choisi (ou à chacun des partiels de la molécule harmonique choisie) !

On peut donc se contenter de mettre à jour les seuls produits scalaires qui sont modifiés. Comme le dictionnaire numériquement employé est constitué d'atomes *à support compact*, on ne met à jour que les produits scalaires du résidu avec les atomes g_γ *dont le support temporel intersecte* celui de l'atome ou de la molécule choisi(e). Cette optimisation est sensible lorsque l'échelle s_m est petite, car peu de produits scalaires doivent être mis à jour. Au contraire, lorsque l'échelle est grande (de l'ordre de N), tous les produits scalaires doivent être recalculés.

4.1.7 Complexité totale

Le coût d'une itération de Matching Pursuit atomique est en définitive

$$\mathcal{O}\left(P_{atom} \log N + P_{atom} + P_{atom} + N\right) = \mathcal{O}\left(P_{atom} \log N\right) \tag{4.17}$$

tandis que celui d'une itération de Matching Pursuit Harmonique vaut

$$\mathcal{O}\left(P_{atom} \log N + P_{atom} + P_{mol} K + P_{mol} + K + KN\right) = \mathcal{O}\left(P_{atom}(K + \log N)\right) . \tag{4.18}$$

La complexité de M itérations du Matching Pursuit atomique est donc bien dominée par la complexité du calcul des produits scalaires avec les atomes complexes

$$\mathcal{O}\left(MN\log^2 N\right). \tag{4.19}$$

Celle de M itérations du Matching Pursuit Harmonique

$$\mathcal{O}\left(MN\log N(K+\log N)\right) \tag{4.20}$$

met aussi en jeu le calcul des corrélations avec les molécules "grossières".

4.2 Poursuite dans des sous-dictionnaire adaptés

La complexité de la poursuite est élevée comparée au coût $\mathcal{O}(MN)$ ($\mathcal{O}(KMN)$) de la reconstruction d'un signal comme combinaison linéaire de M atomes (de KM partiels). De plus, les atomes successivement sélectionnés ne sont pas orthogonaux les uns aux autres, bien que l'énergie soit conservée (3.13). Il est donc possible que

$$\langle x, g_{\gamma_m}\rangle = 0$$

alors que, par définition, $\left|\left\langle R^{m-1}x, g_{\gamma_m}\right\rangle\right| > 0$. Cela signifie que les atomes sélectionnés ne reflètent pas nécessairement des caractéristiques du *signal*, mais peuvent être des "artefacts" de l'algorithme de poursuite.

Les *maxima locaux* de la "carte d'énergie" du signal dans les coordonnées du dictionnaire $\mathcal{D}$ peuvent être considérés comme des caractéristiques intrinsèques du signal[1]. La sélection des plus grands d'entre eux mène à peu d'atomes, par rapport au nombre total d'atomes P_{atom} du dictionnaire. Comme Bergeaud [Ber95] l'a fait pour l'analyse d'images avec le Matching Pursuit [BM96], nous avons eu l'idée [Gri95] [Gri96] d'effectuer la poursuite dans des *sous-dictionnaires* $\mathcal{D}_m$ de $\mathcal{D}$, adaptés au fur et à mesure des itérations, et ne contenant que de tels maxima locaux. Grâce à cela la poursuite est plus rapide, et les atomes sélectionnés plus représentatifs du signal[2].

4.2.1 Sous-dictionnaire de maxima locaux

A partir du résidu $R^{m-1}x$, on définit le sous-dictionnaire de maxima locaux $\mathcal{D}(R^{m-1}x, \varepsilon) \subset \mathcal{D}$ comme l'ensemble des atomes/molécules où la fonction de corrélation est supérieure au seuil ε et admet un maximum local dans la direction du temps ou de la fréquence.

[1] Chen et Donoho [CD95] ont fait remarquer que le Basis Pursuit effectué dans un dictionnaire d'ondelettes dyadiques (*cf* chapitre 8) semblait sélectionner les extrema locaux de cette transformée.

[2] Excepté dans le régime asymptotique où toutes les caractéristiques saillantes du signal ont déjà été ôtées, et ou le résidu atteint le comportement de *bruit de dictionnaire* défini par Davis [Dav94].

Maxima locaux du dictionnaire de Gabor réel

L'atome $g_\gamma = g_{(s,u,\xi)}$ est un maximum local si l'une des fonctions partielles

$$u \mapsto \left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\| \tag{4.21}$$

$$\xi \mapsto \left\| P_{\boldsymbol{V}_\gamma} R^{m-1} x \right\| \tag{4.22}$$

y est localement maximale. Les maxima locaux *temporels* sont localisés aux abords des *singularités* de $R^{m-1}x$ et caractérisent donc ses transitoires, tandis que les maxima locaux *fréquentiels* sont placés sur les *ridges* et repèrent sa fréquence instantanée.

Maxima locaux du dictionnaire de molécules harmoniques

Une molécule harmonique $\boldsymbol{V}_{(s,u,\vec{\xi})}$ est un maximum local fréquentiel si la fonction

$$\xi_1 \mapsto \sup_{\xi_k \in \mathrm{I}_k(\xi_1), k=2..K} C(R^{m-1}x, \boldsymbol{V}_{(s,u,\vec{\xi})}) \tag{4.23}$$

admet un maximum local en ξ_1, et si les partiels sont optimaux pour cette fondamentale

$$(\xi_2, \ldots, \xi_K) = \arg \max_{\xi_k \in \mathrm{I}_k(\xi_1), k=2..K} C(R^{m-1}x, \boldsymbol{V}_{(s,u,\vec{\xi})}) \tag{4.24}$$

On définit de même les maxima locaux temporels.

4.2.2 Construction "périodique" de sous-dictionnaires

La détermination de $\mathcal{D}(R^{m-1}x, \varepsilon_m)$ nécessite le calcul des corrélations avec *tou(te)s* les atomes (les molécules) de $\mathcal{D}$ pour y détecter des maxima. Afin de réduire la complexité algorithmique, le sous-dictionnaire $\mathcal{D}_m$ n'est de la forme $\mathcal{D}(R^{m-1}x, \varepsilon_m)$ que pour certaines itérations $(m_p)_{p\geq 1}$

$$\mathcal{D}_{m_p} \stackrel{\Delta}{=} \mathcal{D}(R^{m_p-1}x, \varepsilon_p). \tag{4.25}$$

réparties plus ou moins régulièrement et aussi peu fréquentes que possible (on verra plus tard comment elles sont déterminées). Le seuil ε_p est déterminé de façon à réduire effectivement la complexité. La taille

$$P_{m_p}(\varepsilon_p) \stackrel{\Delta}{=} \#\mathcal{D}(R^{m_p-1}x, \varepsilon_p) \tag{4.26}$$

du sous-dictionnaire doit être convenablement choisie, afin que la recherche de son meilleur élément et la mise à jour des corrélations, soient peu coûteuses. Nous laissons en suspens quelques instants encore les questions du choix efficace de ε_p et de la valeur optimale de P_{m_p}.

4.2.3 Itérations dans un sous-dictionnaire

Entre les itérations m_p et $m_{p+1}-1$, on utilise des sous-dictionnaires

$$\mathcal{D}_{m_p} \supset \mathcal{D}_{m_p+1} \supset \ldots \supset \mathcal{D}_{m_{p+1}-1} \tag{4.27}$$

extraits itérativement de $\mathcal{D}_{m_p}$ en ne conservant que les atomes dont la corrélation avec le nouveau résidu dépasse encore le seuil ε_p :

$$\mathcal{D}_{m+1} \triangleq \left\{ g_\gamma \in \mathcal{D}_m, \left\| P_{\boldsymbol{V}_\gamma} R^m x \right\|^2 \geq \varepsilon_p \right\}, m_p \leq m < m_{p+1} \tag{4.28}$$

Comme on a $\left\| P_{\boldsymbol{V}_{\gamma_m}} R^m x \right\| = 0$, le cardinal du sous-dictionnaire décroît strictement à chaque itération, si bien qu'il arrive un moment où il est vide. Il est alors nécessaire de reconstruire un "vrai" sous-dictionnaire de maxima, et c'est ainsi que l'on défini l'instant m_{p+1}. La suite m_p est donc reliée à la taille $P_{m_p}(\varepsilon_p)$ du sous-dictionnaire $\mathcal{D}_{m_p}$ par la relation

$$m_{p+1} - m_p \leq P_{m_p}(\varepsilon_p) \tag{4.29}$$

4.2.4 Mise à jour rapide des produits scalaires

Pour les itérations m comprises entre m_p et m_{p+1}, on utilise les formules de mise à jour (4.15) et (4.16), non seulement pour *déterminer quels* produits scalaires changent, mais aussi pour *calculer* leur nouvelle valeur. On dispose à ce effet de la formule analytique (A.2) de calcul du produit scalaire entre deux atomes gaussiens à temps continu, ainsi que (A.8) pour le produit scalaire entre atomes gaussiens discrets.

La formule (A.8) pour les atomes discrets, démontrée et discutée en annexe A, n'est pas utilisable telle quelle en pratique car elle fait intervenir une somme infinie. On montre toutefois dans la même annexe qu'une version approchée consistant à effectuer la somme partielle de très peu de termes permet de calculer de façon rapide, en $\mathcal{O}(1)$, et avec une précision relative de 10^{-5}, les produits scalaires $\langle g_{\gamma_m}, g_\gamma \rangle$ nécessaires pour calculer les nouveaux produits scalaires et les nouvelles corrélations.

4.2.5 Détermination rapide du seuil ε_p

Supposons que la taille désirée P_{m_p} du sous–dictionnaire $\mathcal{D}_{m_p}$ est fixée : pour construire $\mathcal{D}_{m_p}$, il suffit de *trier* par ordre décroissant les maxima locaux de $\mathcal{D}(R^{m_p-1}x, 0)$ et de sélectionner les P_{m_p} plus grands. Le seuil ε_p tel que $\mathcal{D}_{m_p} = \mathcal{D}(R^{m_p-1}x, \varepsilon_p)$ est alors déterminé par la valeur de la corrélation du P_{m_p}-ème maximum local.

On utilise un algorithme de tri rapide [Knu98], dont la complexité pour trier P objets est $\mathcal{O}(P \log P)$ fois celle de la comparaison élémentaire entre deux objets. Comme le nombre total de maxima est majoré par la taille P du

dictionnaire, une borne supérieure sur le coût de la détermination du seuil est

$$\mathcal{O}(P \log P) = \mathcal{O}(N \log^2 N \log\log N) \tag{4.30}$$

4.2.6 Résumé de l'algorithme

- Lorsque $m = m_p$:
  1. Calcul de toutes les corrélations nécessaires (avec les atomes réels optimaux, avec les molécules "grossières").
  2. Construction du dictionnaire de *tous* les maxima locaux $\mathcal{D}(R^{m-1}x, 0)$, en parcourant le dictionnaire $\mathcal{D}$ pour y détecter les maxima locaux. On obtient $P_{m_p}(0)$ maxima locaux.
  3. Détermination du seuil ε_p tel que le dictionnaire $\mathcal{D}_{m_p} \stackrel{\Delta}{=} \mathcal{D}(R^{m_p-1}x, \varepsilon_p)$ ait le nombre de vecteurs P_{obj} que l'on s'est fixé comme objectif
  $$P_{m_p}(\varepsilon_p) = \#\mathcal{D}(R^{m_p-1}x, \varepsilon_p) = P_{obj} \tag{4.31}$$
- Pour $m \in [\![m_p, m_{p+1} - 1]\!]$:
  1. Sélection du meilleur atome g_{γ_m} (ou de la meilleur molécule) en parcourant les $P_m \leq P_{m_p} - (m - m_p)$ qui sont dans le sous-dictionnaire $\mathcal{D}_m \subset \mathcal{D}_{m_p}$
  2. Mise à jour des P_m corrélations par la formule rapide et élimination des atomes dont la corrélation est passée sous le seuil ε_p.
  3. Mise à jour du résidu.

4.2.7 Convergence de l'algorithme accéléré

Cet algorithme accéléré est l'analogue du Matching Pursuit Accéléré introduit par Bergeaud [Ber95] pour l'analyse d'images. Il en a prouvé la convergence lorsque le nombre d'itérations dans un sous-dictionnaire est uniformément borné.

4.2.8 Complexité du Matching Pursuit Rapide

Pour estimer la complexité de cet algorithme rapide et le nombre de vecteurs P_{obj} à inclure dans les sous-dictionnaires $\mathcal{D}_{m_p}$, on fera une seule hypothèse : à chaque étape, le nombre d'itérations nécessaires pour vider le dictionnaire $\mathcal{D}_{m_p}$ est de l'ordre de la taille initiale du dictionnaire, *i.e.*

$$m_{p+1} - m_p = \mathcal{O}(P_{m_p}) = \mathcal{O}(P_{obj}), \tag{4.32}$$

On vérifiera la validité pratique de cette hypothèse dans les exemples numériques.

Complexité des constructions de sous-dictionnaires

Avec cette hypothèse, on sait que pour choisir M atomes (respectivement M molécules), on a eu $\mathcal{O}(M/P_{obj})$ sous-dictionnaires de taille $P_{m_p} = P_{obj}$ à construire.

Lors de chaque construction, on a dû calculer toutes les corrélations nécessaires, ce qui a coûté $\mathcal{O}(P_{atom} \log N)$ (respectivement $\mathcal{O}\left(P_{atom}(K + \log N)\right)$). On a ensuite détecté les maxima locaux en parcourant les P_{atom} atomes (respectivement les $P_{mol} = \mathcal{O}(P_{atom})$ molécules), et l'on en a conservé P_{obj} en sélectionnant le seuil ε_{m_p} en $\mathcal{O}(P_{atom} \log P_{atom})$.

Le coût total de chaque construction d'un dictionnaire de maxima locaux est donc

$$\mathcal{O}(N \log^2 N) \tag{4.33}$$

Complexité des itérations dans les sous-dictionnaires

Par ailleurs, on a effectué M itérations dans des sous-dictionnaires, dont le coût individuel se décompose comme suit

- sélection du meilleur atome de $\mathcal{D}_m$ (respectivement de la meilleure molécule) en parcourant les $\mathcal{O}(P_m)$ disponibles.
- mise à jour du résidu en $\mathcal{O}(N)$ (respectivement $\mathcal{O}(KN)$).
- mise à jour des produits scalaires du résidu avec les P_m atomes, à l'aide de la formule (4.15) (respectivement avec les KP_m partiels des P_m molécules, à l'aide de la formule (4.16)), ce qui coûte $\mathcal{O}(1)$ pour chaque atome (respectivement $\mathcal{O}(K)$ pour chaque partiel) ;
- mise à jour des P_m corrélations avec les atomes réels (respectivement les P_m molécules) du sous-dictionnaire $\mathcal{D}_m$, ce qui coûte $\mathcal{O}(1)$ pour chacun (respectivement $\mathcal{O}(K)$ additions pour chacune) ;
- élimination des atomes (respectivement des molécules) dont la corrélation est passée sous le seuil ε_p.

Le coût individuel d'une itération dans un sous dictionnaire, pour un Matching Pursuit atomique accéléré, est donc de

$$\mathcal{O}(P_{obj} + N) \tag{4.34}$$

tandis qu'il vaut

$$\mathcal{O}(K^2 P_{obj} + KN) \tag{4.35}$$

pour un Matching Pursuit Harmonique accéléré.

Complexité totale

Le coût total de M itérations de la poursuite accélérée est donc

$$\mathcal{O}\left(MN\left(\frac{\log^2 N}{P_{obj}} + \frac{P_{obj}}{N} + 1\right)\right) \tag{4.36}$$

tandis que pour M itérations de Matching Pursuit Harmonique accéléré on obtient

$$\mathcal{O}\left(MN\left(\frac{\log^2 N}{P_{obj}} + \frac{K\log N}{P_{obj}} + K^2\frac{P_{obj}}{N} + K\right)\right) \tag{4.37}$$

Choix de la taille des sous-dictionnaires

L'ordre de grandeur du coût du Matching Pursuit atomique accéléré est minimisé lorsque P_{obj} est de l'ordre de

$$P_{obj} \propto \sqrt{N}\log N. \tag{4.38}$$

La complexité atteinte avec ce choix est alors de l'ordre de

$$\mathcal{O}\left(MN\left(1 + \frac{\log N}{\sqrt{N}}\right)\right) = \mathcal{O}(MN). \tag{4.39}$$

Dans le cas du Matching Pursuit Harmonique accéléré, si $K = \mathcal{O}(\log N)$, il faut cette fois choisir

$$P_{obj} \propto \frac{\sqrt{N}}{K}\sqrt{\log N(K + \log N)} \approx \frac{\sqrt{N}}{K}\log N \tag{4.40}$$

et l'on aboutit donc a une complexité

$$\mathcal{O}\left(MN\left(K + K\frac{\log N}{\sqrt{N}}\right)\right) = \mathcal{O}\left(KMN\right). \tag{4.41}$$

Le gain de complexité par rapport à la poursuite normale est donc, dès que N est grand, de l'ordre de $\log^2 N$. On peut donc parler de Matching Pursuit *Rapide*.

4.2.9 Résultats numériques

La représentation temps-fréquence affichée sur la figure 3.2 concerne un signal musical de $N = 32000 \approx 2^{15}$ points. Nous l'avons obtenue par un Matching Pursuit "standard" avec 1000 itérations. Le temps de calcul, sur une station de travail Dec^{TM} Alpha 600, était de l'ordre de cinq heures, soit 300 minutes. Nous avons effectué [Gri95], [Gri96] la même analyse en utilisant la poursuite accélérée, avec 600 maxima locaux : il a suffit de 3 minutes pour parvenir au résultat, que l'on représente sur la figure 4.1. Le gain en temps de calcul était donc de l'ordre de 100. On peut constater que les représentations temps-fréquence obtenues avec les deux algorithmes, pour le même son de piano, sont similaires.

La comparaison du gain observé (100) avec le gain théoriquement prévu n'a pas grand sens, car des constantes interviennent dans les définitions des ordres de grandeur. Nous avons établi pour quelques valeurs de N, avec

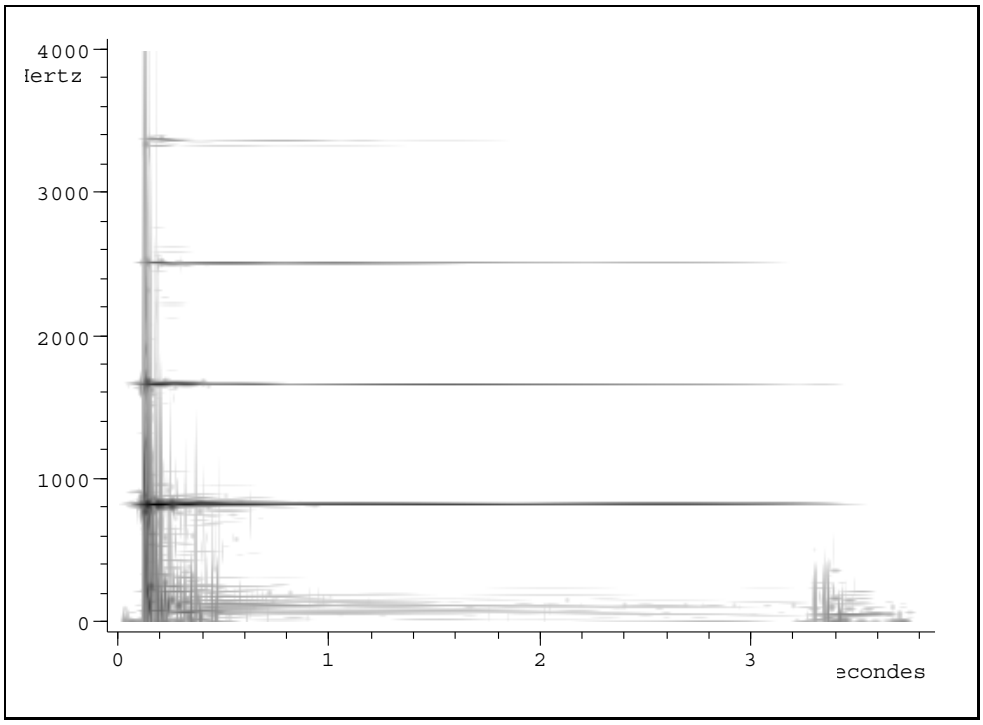

FIG. 4.1 – Représentation temps-fréquence d'un son de piano, obtenue à l'aide d'un Matching Pursuit Rapide avec un dictionnaire multi-échelle de Gabor gaussien ($M = 1000$ atomes, $P_{obj} = 600$ maxima locaux par sous-dictionnaire).

$P_{obj} = \sqrt{N} \log N$, les temps de calcul avec et sans accélération, ainsi que les gains de vitesse correspondants. Ils sont résumés dans le tableau 4.1. Par ailleurs, afin de valider l'hypothèse (4.32), nous avons également mesuré, pour différentes valeurs de N, le nombre moyen Δm d'itérations effectuées dans un sous-dictionnaire. Les résultats sont rassemblés dans le tableau 4.2. On y constate que ce nombre est bien du même ordre de grandeur que P_{obj}.

N	P_{obj}	MP	MPR	Gain
128	24	32	6	5,3
256	39	51	9	5,7
512	61	90	12	7,5
1024	96	157	18	8,7
2048	150	316	24	13,2
4096	231	666	33	20,2

TAB. 4.1 – Temps de calcul nécessaire, en secondes, pour effectuer un Matching Pursuit standard (MP) et un Matching Pursuit Rapide (MPR), pour différentes valeurs de la taille de signal N. On indique le nombre de maxima locaux P_{obj} utilisés dans l'algorithme rapide. La dernière colonne indique le gain en vitesse correspondant. Les calculs ont été effectués sur un ordinateur de type PC, muni d'un processeur "Celeron" à 300 Mhz.

N	P_{obj}	Δm	$\Delta m / P_{obj}$
128	24	7.7	0.32
256	39	11.6	0.30
512	61	17.7	0.29
1024	96	27.7	0.29
2048	150	42.7	0.28
4096	231	68.3	0.29

TAB. 4.2 – Nombre moyen d'itérations Δm effectué dans chaque sous-dictionnaire $\mathcal{D}_{m_p}$ lors d'un Matching Pursuit Rapide, pour différentes valeurs de la taille de signal N. On le compare au nombre $P_{obj} = \sqrt{N}\log(N)$ de maxima locaux que contient chacun de ces sous-dictionnaires.

Chapitre 5

"Matching Pursuit" Rapide avec un dictionnaire d'atomes modulés en fréquence

Nous proposons dans ce chapitre un algorithme de Matching Pursuit modifié permettant d'effectuer une décomposition atomique *rapide* d'un signal dans un dictionnaire $\mathcal{D}^+$ multi-échelle de *chirps*[1] [Bul95] [Bul99] [MH95]

$$g_{(s,u,\xi,c)}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{i\left(\xi(t-u)+\frac{c}{2}(t-u)^2\right)} \tag{5.1}$$

dont la fréquence instantanée $\omega = \xi + c(t-u)$ varie linéairement avec le temps. Le dictionnaire $\mathcal{D}^+$ est une extension du dictionnaire temps-fréquence multi-échelle $\mathcal{D}$ de Gabor [QC94].

Complexité

L'ajout d'un paramètre de chirp c aux trois paramètres d'échelle s, de temps u et de fréquence ξ du dictionnaire de Gabor $\mathcal{D}$ fait du dictionnaire "chirpé" $\mathcal{D}^+$ un ensemble de très grande taille $\mathcal{O}(N^2)$. Le Matching Pursuit "brutal" dans un tel dictionnaire est de complexité

$$\mathcal{O}(M N^2 \log N). \tag{5.2}$$

Sa mise en œuvre par Bultan [Bul95] [Bul99] requiert une grande puissance de calcul. Elle est donc limitée à l'analyse de petits signaux (*e.g.* $N = 256$ points) avec peu d'itérations (*e.g.* $M = 10$).

[1] Le terme *chirp* désigne, en anglais, l'onomatopée caractérisant le cri ("cui-cui") des oiseaux. Il désignera ici aussi bien les atomes *chirpés* de $\mathcal{D}^+$, les signaux dont la fréquence varie linéairement avec le temps et, lorsque l'ambiguité ne sera pas possible, la mesure de la *pente* c de cette variation linéaire, dont l'unité est le Hertz par seconde ($Hz.s^{-1}$).

Matching Pursuit Chirpé à complexité réduite

Bultan a proposé, pour réduire la complexité, une solution *ad hoc* qui consiste à limiter la résolution du paramètre de chirp. Nous souhaitons analyser des signaux réels de grande taille, en suffisamment d'itérations pour en obtenir de bonnes approximations, *sans limiter la résolution* ni avoir recours à une puissance de calcul démesurée. Pour cela, nous introduisons ici un algorithme substantiellement modifié, le Matching Pursuit "de ridges". Dans un dictionnaire $\mathcal{D}^+$ à enveloppes *gaussiennes*, nous obtenons une complexité

$$\mathcal{O}(MN\log^2 N), \tag{5.3}$$

identique à celle du Matching Pursuit sur le dictionnaire de Gabor simple.

Pour parvenir à ce niveau de réduction de la complexité, on a dû faire appel à deux idées. D'une part, comme $\mathcal{D}^+$ est une extension du dictionnaire de Gabor $\mathcal{D}$, on cherche le "meilleur atome chirpé" $g_{(s_m,u_m,\xi_m,c_m)} \in \mathcal{D}^+$ en deux temps. On commence par déterminer le meilleur atome de Gabor non-chirpé, puis on optimise ses paramètres d'échelle et de chirp pour augmenter la corrélation avec le résidu. D'autre part, comme la *recherche exhaustive* des meilleurs paramètres nécessite de balayer toutes les valeurs possibles du chirp c, elle peut coûter encore très cher ($\mathcal{O}(N^2)$ par itération). Pour atteindre la complexité annoncée, on la remplace par une *estimation* des meilleurs paramètres. Pour construire un estimateur local rapide, en $\mathcal{O}(1)$, nous utilisons un théorème de "ridges" du dictionnaire de Gabor gaussien que nous établissons pour l'occasion.

Dans ce chapitre on commence donc par définir le dictionnaire temps-fréquence chirpé $\mathcal{D}^+$ et mettre en évidence la complexité numérique qu'il impose. On établit ensuite deux théorèmes de "ridges" du dictionnaire de Gabor gaussien multi-échelle, à l'aide desquels on analyse la sélection du "meilleur atome chirpé local". On présente enfin les résultats obtenus avec notre algorithme modifié, sur différents signaux.

5.1 Dictionnaire temps-fréquence d'atomes chirpés

Un atome chirpé (5.1), repéré à l'aide de son indice (s, u, ξ, c), est centré autour du temps u avec une dispersion temporelle de l'énergie de l'ordre de s. Sa transformée de Wigner-Ville (représentée sur la figure 5.1 pour un atome chirpé gaussien), qui définit sa répartition énergétique dans le plan temps-fréquence, est concentrée autour de la droite $\omega = \xi + c(t-u)$. Sa dispersion est de l'ordre de $1/s$ dans la direction de ω. En effet, d'après les propriétés de la transformée de Wigner-Ville [Fla93][Mal98], on a

$$WV[g_{(s,u,\xi,c)}]\,(t,\omega) = WV[g_{(s,0,0,0)}]\,(t-u, \omega-\xi-c(t-u))\,.$$

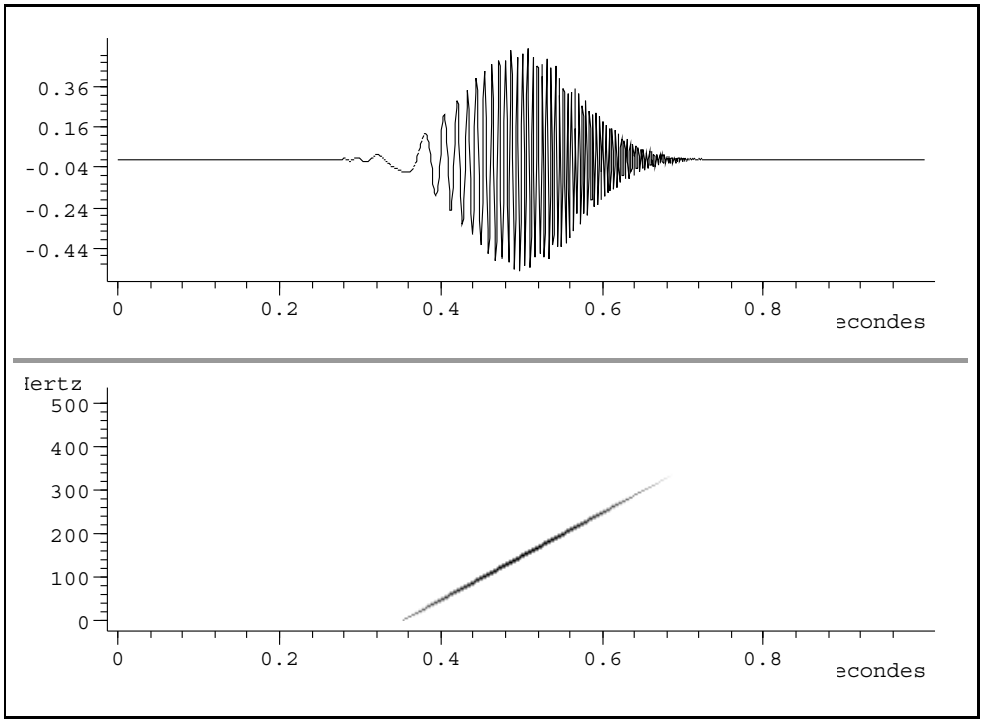

FIG. 5.1 – Un atome chirpé gaussien et sa transformée de Wigner-Ville.

5.1.1 Discrétisation du dictionnaire

L'indice (s, u, ξ, c) prend ses valeurs dans $\Gamma^+ \subset \Gamma \times \mathbb{R}$, où $\Gamma \subset \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}$ est l'ensemble des valeurs prises par l'indice $\gamma = (s, u, \xi)$ des atomes du dictionnaire temps-fréquence de Gabor. L'échantillonnage de s,u, et ξ est donc celui utilisé dans le dictionnaire de Gabor multi-échelle [QC94], *i.e.* comme en (2.38)-(2.42) . L'analyse de Watson et Gilholm [WG98] permet d'échantillonner le chirp c

$$c = l \times \Delta c(s), l \in \mathbb{Z}, \tag{5.4}$$

$$\Delta c(s) \stackrel{\Delta}{=} s^{-2} \Delta c(1) \tag{5.5}$$

5.1.2 Échantillonnage "critique" du chirp

Expliquons en quoi cet échantillonnage du chirp c est "critique". Pour distinguer deux atomes $g_{(s,u,\xi,c)}$ et $g_{(s,u,\xi,c')}$, il faut que $\Delta c = |c - c'|$ soit suffisamment grand pour que les supports de l'énergie temps-fréquence des deux atomes diffèrent. Comme le support temporel des deux atomes est identique, on peut les distinguer grâce à leurs fréquences instantanées $\xi + c(t-u)$ et $\xi + c'(t-u)$. Celles-ci atteignent leur différence maximale, de l'ordre de $\Delta c\, s$, aux "extrémités" du support des atomes. Pour distinguer les deux atomes, il faut donc que $\Delta c\, s$ soit légèrement plus grand que leur

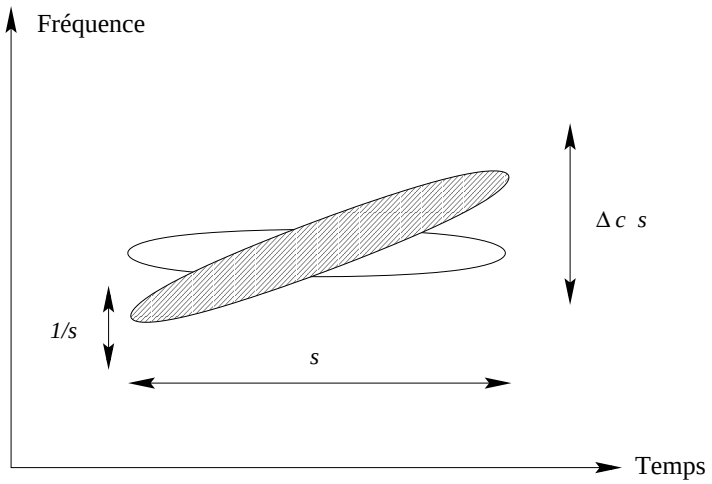

FIG. 5.2 – Échantillonnage du paramètre de chirp.

dispersion fréquentielle, qui est de l'ordre de $1/s$ (voire la figure 5.2), *i.e.* $\Delta c(s) \times s^2$ doit être de l'ordre de 1. Cela conduit bien à la condition (5.5).

5.1.3 Taille du dictionnaire discret

La taille du dictionnaire chirpé $\mathcal{D}^+$ est directement liée aux pas d'échantillonnage $a, \Delta u(1)$, $\Delta\xi(1)$ et $\Delta c(1)$, ainsi qu'aux bornes délimitant les paramètre s, u, ξ, c admissibles.

Pour analyser un signal discret de N points, on considère des échelles $s = a^j$ entre 1 et N, ce qui fait au total

$$\mathcal{O}(\log N) \tag{5.6}$$

échelles. Pour chacun des indices j, $u \in [0, N-1]$ peut prendre

$$\mathcal{O}(N/a^j) \tag{5.7}$$

valeurs. Par ailleurs la fréquence instantanée $\omega(t) = \xi + c(t-u)$ de chaque atome doit vérifier

$$\xi + c(t-u) \in [0, \pi], \ \forall t \in [u - s/2, u + s/2]. \tag{5.8}$$

En effet la borne supérieure π de l'intervalle (5.8)) traduit la condition d'échantillonnage de Nyquist, tandis que sa borne inférieure 0 contraint l'atome à être analytique, en évitant le repliement des fréquences négatives dans la partie positive du spectre, comme illustré sur la figure 5.3. Ces contraintes se traduisent par

$$\xi \in [0, \pi] \tag{5.9}$$

$$|c|\, s/2 \leq \min(\pi - \xi, \xi) \leq \pi/2. \tag{5.10}$$

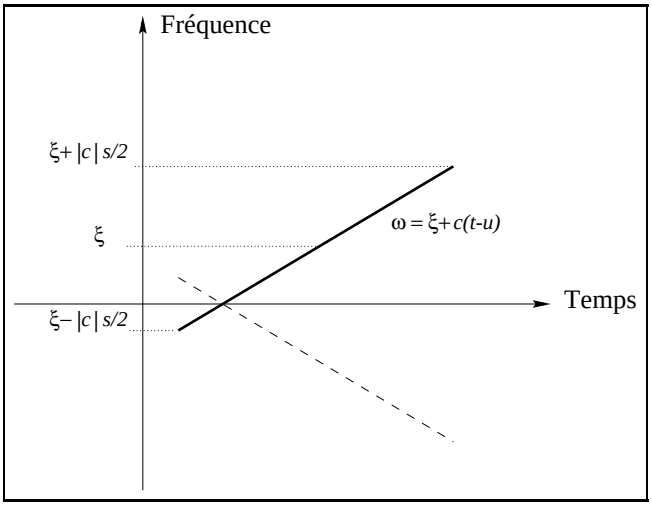

FIG. 5.3 – Condition de non-repliement pour un atome chirpé discret : si des fréquences négatives $\omega = \xi + c(t-u)$ apparaîssent, elles se replient dans la partie positive du spectre.

Pour une échelle $s = a^j$ fixée, ξ prend donc

$$\mathcal{O}(a^j) \tag{5.11}$$

valeurs. Le chirp c peut, lui, prendre en moyenne

$$\mathcal{O}(a^j) \tag{5.12}$$

valeurs, de l'unique valeur $c = 0$ possible si $\xi = 0$ ou $\xi = \pi$, aux $\mathcal{O}(a^j)$ possibles pour $\xi = \pi/2$. Au total, $\mathcal{D}^+$ comprend donc de l'ordre de

$$\sum_{j=0}^{\log N} \mathcal{O}(N/a^j) \times \mathcal{O}(a^j) \times \mathcal{O}(a^j) = \mathcal{O}(N^2) \tag{5.13}$$

atomes. La taille du dictionnaire est due avant tout au grand nombre de valeurs possibles du paramètre de chirp c á grande échelle.

Exemple

Un signal musical d'une durée de 1.5 secondes, échantillonné à 44.1 kHz, a une taille d'environ $N = 2^{16} = 65536$ échantillons. Le dictionnaire chirpé critique comprendra donc $2^{32} \approx 4.10^9$ atomes lorsque les échelles sont choisies de manière dyadique ($a = 2$).

5.1.4 Coût du calcul des produits scalaires

Étant donnée la taille $\mathcal{O}(N^2)$ du dictionnaire, le calcul de *tous* les produits scalaires $\langle x, g_{(s,u,\xi,c)} \rangle$ avec les atomes de $\mathcal{D}^+$ ne peut coûter moins de

$\mathcal{O}(N^2)$ opérations. On peut, de fait, l'effectuer en $\mathcal{O}(N^2 \log N)$ en utilisant des algorithmes de FFT avec des fenêtres appropriées, comme l'ont fait remarquer Bultan [Bul95] [Bul99], Watson et Gilholm [WG98]. En effet, à s et c fixés, les atomes

$$g_{(s,u,\xi,c)}(t) = g_{(s,0,0,c)}(t-u)e^{i\xi(t-u)}$$

se déduisent par translation et modulation fréquentielle de la fenêtre chirpée

$$g_{s,c}(t) = g_{(s,0,0,c)}(t).$$

Leurs produits scalaires se calculent donc comme une transformée de Fourier à court terme avec la fenêtre $g_{s,c}$. Chaque transformée de Fourier à court terme coûte $\mathcal{O}(N \log s)$ car elle nécessite $\mathcal{O}(N/s)$ FFT, de coût unitaire $\mathcal{O}(s \log s)$. Comme il y a $\mathcal{O}(a^j)$ chirps à l'échelle $s = a^j$, le coût total est donc

$$\sum_{j=0}^{\log N} \mathcal{O}(N \times j) \times \mathcal{O}(a^j) = \mathcal{O}(N^2 \log N) \tag{5.14}$$

5.1.5 Complexité du Matching Pursuit Chirpé "brutal"

Une analyse de complexité en tout point analogue à celle effectuée au chapitre 4 montrerait donc qu'il faut

$$\mathcal{O}(MN^2 \log N) \tag{5.15}$$

opérations pour effectuer M itérations de poursuite sur un signal de N points avec un tel dictionnaire.

Exemple

Pour analyser le signal musical du précédent exemple, le calcul des produits scalaires coûtera environ $2^{32} \times 16 \approx 64.10^9$ opérations à chaque itération. Si la puissance de calcul d'un calculateur d'aujourd'hui est de 10^8 à 10^9 opérations par secondes, il faut entre 1 et 10 minutes pour chaque itération de poursuite. Cela est à mettre en regard de la durée de 1.5 seconde du signal. On peut noter que Bultan [Bul95] [Bul99] a réalisé un tel Matching Pursuit "brutal" sur un dictionnaire de chirps : il a dû se contenter de travailler sur de petits signaux (typiquement $N = 256$), et ne présente en général que des résultats obtenus avec une dizaine d'itérations.

5.2 Matching Pursuit *de ridges*

Pour réduire la complexité de la poursuite sur un dictionnaire de chirps, Bultan [Bul95] [Bul99] limite la résolution du paramètre de chirp c aux

grandes échelles.

$$\Delta c(s) = \max\left(s^{-2}\Delta c(1), \Delta c_{min}\right). \tag{5.16}$$

Cette solution *ad hoc* ne permet pas à la poursuite de représenter de façon concise (avec peu d'atomes) des chirps $a(t)e^{i\phi(t)}$ de grande échelle, dès lors que leur pente c *n'est pas* sur la grille $\Delta c_{min}\mathbb{Z}$ trop grossière. Nous suggérons donc ici un autre angle d'attaque pour réduire la complexité.

Choix approché en deux temps

On ne peut pas se permettre de calculer la corrélation $\left\langle R^{m-1}x, g_{(s,u,\xi,c)}\right\rangle$ de *chaque* atome de $\mathcal{D}^+$ avec le résidu. On doit donc effectuer un choix *approché* du "meilleur" atome $g_{(s_m,u_m,\xi_m,c_m)}$, en faisant appel à une méthode de coût aussi faible que possible. On aimerait que celle-ci nous fournisse un atome à peine moins bon que celui qu'on aurait obtenu au prix fort, avec la poursuite "brutale".

Comme le dictionnaire de Gabor $\mathcal{D}$, dont $\mathcal{D}^+$ est une extension, est complet, les produits scalaires $\left\langle R^{m-1}x, g_\gamma\right\rangle, g_\gamma \in \mathcal{D}$ contiennent toute l'information disponible dans le signal. Nous allons montrer, à l'aide d'un théorème *de ridge* de ce dictionnaire, que le *sous-dictionnaire* de ses *maxima locaux* [Ber95] [Gri95] [Gri96] [WG98] (voir également le chapitre 4) contient l'information relative à la fréquence instantanée et à ses variations. On peut ainsi obtenir un "bon" atome chirpé $g_{(s_m,u_m,\xi_m,c_m)}$ (à défaut du "meilleur") à partir des maxima locaux de $\mathcal{D}$.

Cela mène à une poursuite en deux temps : une première passe consiste à sélectionner le meilleur atome "simple"

$$g_{(s_m^{ridge},u_m^{ridge},\xi_m^{ridge})} \stackrel{\Delta}{=} \arg\max_{g_{(s,u,\xi)}\in\mathcal{D}} \left|\left\langle R^{m-1}x, g_{(s,u,\xi)}\right\rangle\right| \tag{5.17}$$

qui est donc sur un *ridge* de $\mathcal{D}$. Dans un second temps on explore son *voisinage* dans le dictionnaire $\mathcal{D}^+$ afin de trouver *un bon* atome chirpé

$$g_{(s_m,u_m,\xi_m,c_m)} \stackrel{\Delta}{=} \arg\max_{g_{(s,u_m^{ridge},\xi_m^{ridge},c)}} \left|\left\langle R^{m-1}x, g_{(s,u_m^{ridge},\xi_m^{ridge},c)}\right\rangle\right| \tag{5.18}$$

en optimisant localement les paramètres de chirp c et d'échelle s.

Analogie : choix d'un atome de Gabor à partir des ridges de la transformée en ondelettes

Supposons un instant que la recherche du meilleur atome de Gabor soit trop coûteuse. On peut alors le sélectionner avec une recherche en deux temps, à partir des crêtes (ou *ridges*) de son sous-dictionnaire constitué des ondelettes de Morlet.

Soit $x(t) = a(t)e^{i\phi(t)}$ et $\langle x, \psi_{(s,u)} \rangle$ sa transformée en ondelettes, calculée avec une ondelette de Morlet

$$\psi_{(s,u)}(t) = \frac{1}{\sqrt{s}} g((t-u)/s) e^{i\omega_0 t/s}. \tag{5.19}$$

Les *ridges* de cette transformée sont les maxima locaux $s(u)$ de $s \mapsto \left|\langle x, \psi_{(s,u)} \rangle\right|$. Delprat, Escudié, Guillemain, Kronland-Martinet, Tchamitchian et Torrésani [Del92] [DEG+92] ont montré, à l'aide d'arguments de phase stationnaire, pour des signaux asymptotiques en temps, que ces ridges suivent la fréquence instantanée

$$\omega_0/s(u) \approx \phi'(u) \tag{5.20}$$

dès que l'échelle s est petite devant les variations d'amplitude et de fréquence instantanée. Or l'ondelette $\psi_{(s,u)}$ est un atome temps-fréquence particulier $g_{(s,u,\omega_0/s,0)}$. Le maximum d'amplitude (s^{ridge}, u^{ridge}) de $\langle x, \psi_{(s,u)} \rangle$ correspond donc au meilleur atome d'un certain sous-dictionnaire (invariant par translation et dilatation) de $\mathcal{D}$.

Plutôt que d'effectuer le calcul de tous les produits scalaires de x avec les atomes de $\mathcal{D}$ afin de sélectionner le plus grand, on peut tirer parti de l'information dont on dispose avec les ridges de la transformée en ondelettes. Un "bon" atome temps-fréquence est en effet

$$\psi_{(s^{ridge}, u^{ridge})} = g_{(s^{ridge}, u^{ridge}, \omega_0/s^{ridge})}. \tag{5.21}$$

Cependant, les atomes temps-fréquence ont un degré de liberté de plus que les ondelettes, car s et ξ peuvent varier indépendamment. En jouant sur le paramètre d'échelle s, on peut donc raisonnablement chercher un gain de corrélation. Dans ce cadre, une recherche en deux temps du meilleur atome de Gabor consisterait donc à :

1. repérer le maximum de la transformée en ondelettes de $R^{m-1}x$:

$$(s_m^{ridge}, u_m^{ridge}) = \arg\max_{(s,u)} \left|\langle R^{m-1}x, \psi_{(s,u)} \rangle\right|. \tag{5.22}$$

   L'atome $g_{(s_m^{ridge}, u_m^{ridge}, \omega_0/s_m^{ridge})}$ est candidat à la maximisation de $\left|\langle R^{m-1}x, g_{(s,u,\xi)} \rangle\right|$.

2. optimiser l'échelle s pour augmenter la corrélation :

$$g_{(s_m, u_m, \omega_0/s_m^{ridge})} = \arg \max_{g_{(s, u_m^{ridge}, \omega_0/s_m^{ridge})}} \left|\left\langle R^{m-1}x, g_{(s, u_m^{ridge}, \omega_0/s_m^{ridge})} \right\rangle\right|. \tag{5.23}$$

Estimation du chirp et de l'échelle localement optimaux

Dans le cas qui nous intéresse, après avoir sélectionné le meilleur atome de Gabor, le deuxième temps (5.18) nécessite un *balayage exhaustif* du voisinage de l'atome choisi, afin d'optimiser le paramètre de chirp c et l'échelle. La recherche en deux temps n'est donc pas suffisante pour réduire la complexité, car ce balayage est encore très coûteux. On le remplace par une *estimation rapide* de l'échelle s_m et du chirp c_m localement optimaux. Le *Matching Pursuit de ridges* ainsi défini a une complexité $\mathcal{O}(MN\log^2 N)$, identique à celle du Matching Pursuit sur le dictionnaire de Gabor $\mathcal{D}$.

C'est un théorème de *ridge* à l'ordre supérieur qui nous permet de comprendre le comportement local de $(s,u,\xi,c) \mapsto \left\langle R^{m-1}x, g_{(s,u,\xi,c)}\right\rangle$ au voisinage de ses maxima locaux et d'en extraire l'information locale qui nous intéresse sur s_m et c_m.

5.2.1 "Ridges" du dictionnaire de Gabor continu

On se place désormais dans un modèle de signal analytique

$$R^{m-1}x = a(t)e^{i\phi(t)} \tag{5.24}$$

où l'on peut définir une fréquence instantanée

$$\xi(t) = \phi'(t) \tag{5.25}$$

et un chirp instantané

$$c(t) = \phi''(t). \tag{5.26}$$

L'objet du théorème de *ridge* que nous allons établir est de prouver que, sous certaines hypothèses relatives aux variations des fonctions a et ϕ, le résidu $R^{m-1}x$, vu "à travers" un atome de Gabor gaussien $g_{(s,u,\xi)} \in \mathcal{D}$, ressemble à un atome de Gabor gaussien chirpé, *i.e.*

$$\left\langle R^{m-1}x, g_{(s,u,\xi)}\right\rangle \approx Ae^{i\Phi}\left\langle g_{s(u),u,\xi(u),c(u)}, g_{(s,u,\xi)}\right\rangle. \tag{5.27}$$

Ce résultat nous permet alors d'interpréter les maxima locaux (ou *ridges*) de la fonction $(s,u,\xi) \mapsto \left|\left\langle R^{m-1}x, g_{(s,u,\xi)}\right\rangle\right|$ en termes de fréquence instantanée et de chirp instantané.

Ridges du dictionnaire de Gabor

Il est connu [Mal98] que les maxima locaux $\xi(s,u)$ de $\xi \mapsto \left|\left\langle x, g_{(s,u,\xi)}\right\rangle\right|$ permettent de localiser la fréquence instantanée

$$\xi(s,u) \approx \phi'(u). \tag{5.28}$$

dès que l'échelle s est petite devant les variations d'amplitude et de fréquence instantanée.

Mais ce qui est plus intéressant à étudier est l'information qu'apporte maintenant l'échelle optimale : le maximum absolu de $\left|\left\langle x, g_{(s,u,\xi)}\right\rangle\right|$ est à la fois maximum local selon s, u et ξ. Le théorème suivant, que nous démontrons en annexe B, nous permet de relier les variations de phase et d'amplitude du signal à la localisation du "ridge" du dictionnaire de Gabor gaussien.

Théorème 4 *Soit $x(t) = a(t)e^{i\phi(t)}$ un signal analytique. On suppose que l'amplitude a , sa dérivée et la dérivée troisième de la phase ϕ sont bornées. Soit $g_{(s,u,\xi)}$ un atome temps-fréquence gaussien. Alors*

$$\left\langle x, g_{(s,u,\xi)}\right\rangle = a(u)e^{i\phi(u)}\left(\left\langle e^{i(\phi'(u)(t-u)+\phi''(u)(t-u)^2/2)}, g_{(s,u,\xi)}\right\rangle + \sqrt{s}\epsilon(s,u,\xi)\right) \tag{5.29}$$

où $|\epsilon(s,u,\xi)|$ est majoré par

$$\begin{aligned}\epsilon_{max}(s,u) &\triangleq s\sigma_1\frac{\|a'(u)\|_\infty}{a(u)} + \frac{s^3\sigma_3^3}{6}\left\|\phi'''\right\|_\infty e^{1/6} \\ &\quad +2\left(\left\|\phi'''\right\|_\infty s^3\right)^{1/3} e^{-\frac{1}{2\left(\|\phi'''\|_\infty s^3\right)^{2/3}}}\end{aligned} \tag{5.30}$$

avec

$$\sigma_k^k \triangleq \int |t|^k\, g(t)dt.$$

Cette approximation du produit scalaire mesure la ressemblance entre le signal analytique x et un chirp pur $e^{i(\phi'(u)(t-u)+\phi''(u)(t-u)^2/2)}$, vue "à travers" un atome d'analyse $g_{(s,u,\xi)}$. En effet, il exprime le fait que

$$\left\langle x, g_{(s,u,\xi)}\right\rangle \approx a(u)e^{i\phi(u)}\left\langle e^{i(\phi'(u)(t-u)+\phi''(u)(t-u)^2/2)}, g_{(s,u,\xi)}\right\rangle.$$

Si le terme d'erreur est assez petit, alors on voit que $\phi'(u)$ est estimé en prenant le maximum le long de la fréquence ξ. Plus précisément, comme

$$\left\langle e^{i(\phi'(u)(t-u)+\phi''(u)(t-u)^2/2)}, g_{(s,u,\xi)}\right\rangle = \widehat{g}_{(s,0,0,\phi''(u))}(\xi-\phi'(u)) \tag{5.31}$$

est la transformée de Fourier d'un atome gaussien chirpé, on en connaît une expression analytique (voir annexe A ou [Pap87]). On connaît en particulier son maximum à u fixé : en l'absence de terme d'erreur, le maximum local selon s et ξ, ou *ridge* serait situé en

$$\xi^{ridge} = \phi'(u) \tag{5.32}$$

et

$$s^{ridge} = \frac{1}{\sqrt{\phi''(u)}}. \tag{5.33}$$

et la valeur prise par le premier terme de (5.29) serait

$$\widehat{g}_{(s,0,0,\phi''(u))}(\xi - \phi'(u)) = \left(2\pi / \left|\phi''(u)\right|\right)^{1/4} = (2\pi)^{1/4}\sqrt{s^{ridge}}. \tag{5.34}$$

Les corollaires qui suivent quantifient l'erreur faite en estimant la fréquence et le chirp instantanés à partir de s^{ridge} et ξ^{ridge}. Ils sont démontrés en annexe B.

Estimation de la fréquence instantanée à partir du ridge

Le corollaire 1 montre que, si le terme d'erreur $\sqrt{s}\epsilon(s,u,\xi)$ est petit, on peut mesurer la fréquence instantanée à l'aide de la position du maximum selon ξ. Il quantifie l'erreur de mesure.

Corollaire 1 *Soit x un signal remplissant les hypothèse du théorème 4. Soient s et u tels que le terme d'erreur dans l'approximation (5.29) vérifie*

$$\epsilon_{max}(s,u) \leq \left(\frac{\pi}{4\left(1 + s^4\left|\phi''(u)\right|^2\right)}\right)^{1/4} \tag{5.35}$$

Alors le maximum absolu $\xi(s,u)$ de la fonction $\xi \mapsto \left|\left\langle x, g_{(s,u,\xi)}\right\rangle\right|$ vérifie

$$\left|\xi(s,u) - \phi'(u)\right| \leq \delta\xi(s,u) \tag{5.36}$$

et

$$\left|\left\langle x, g_{(s,u,\xi(s,u))}\right\rangle\right| = a(u)\left(\left(\frac{4\pi s^2}{1 + s^4\left|\phi''(u)\right|^2}\right)^{1/4} + \sqrt{s}\epsilon(s,u)\right) \tag{5.37}$$

où $|\epsilon(s,u)| \leq \epsilon_{max}(s,u)$ et où

$$\delta\xi(s,u) = \sqrt{\frac{2\left(1 + s^4\left|\phi''(u)\right|^2\right)}{s^2}\log\left(1 - \epsilon_{max}(s,u)\left(\frac{4\left(1 + s^4\left|\phi''(u)\right|^2\right)}{\pi}\right)^{1/4}\right)^{-1}}. \tag{5.38}$$

D'après sa définition (5.30), ϵ_{max} est petit dès que l'échelle s est petite. Il peut donc bien remplir la condition (5.35). La relation (5.36) montre alors que le pic du spectre local $\xi \mapsto \left\langle x, g_{(s,u,\xi)}\right\rangle$ permet d'estimer la fréquence instantanée $\phi'(u)$. La précision $\delta\xi(s,u)$ de l'estimation dépend, comme le montre l'expression (5.38), des valeurs relatives de s et du chirp instantané $\phi''(u)$. Lorsque s est très petite, la précision est faible à cause de l'étalement spectral de l'atome d'analyse, qui tend vers un dirac. Lorsque s est grande, la mesure de fréquence instantanée est imprécise parce que moyennée sur le support de l'atome qui est trop grand. C'est donc à une échelle intermédiaire que la précision sera la meilleure.

Estimation du chirp instantané à partir du ridge

On peut également estimer $\phi''(u)$ à l'aide du maximum dans la direction de s. Le produit scalaire sur le ridge (5.37) est en effet maximal (en négligeant le terme d'erreur) pour $s = 1/\sqrt{|\phi''(u)|}$. Le corollaire 2 donne les conditions de cette estimation et quantifie l'erreur de mesure de $\phi''(u)$.

Corollaire 2 *Soit x un signal vérifiant les hypothèses du théorème 4. Soit u un instant tel que $\phi''(u) \neq 0$, et $s_0 = 1/\sqrt{|\phi''(u)|}$. Soit $\lambda > 1$ un réel tel que la majoration (5.30) de l'erreur vérifie*

$$\epsilon_{max}(\lambda s_0, u) \leq \frac{(2\pi)^{1/4}}{2\sqrt{\lambda}}(1 - 2^{1/4}/\sqrt{\lambda}). \tag{5.39}$$

Alors la fonction $s \mapsto \left|\left\langle x, g_{(s,u,\xi(s,u))}\right\rangle\right|$ admet sur l'intervalle $]0, \lambda s_0]$ au moins un maximum local, et le plus grand $s(u)$ de ses maxima locaux sur cet intervalle vérifie

$$|\log s(u)/s_0| \leq \underline{\beta}(\lambda s_0, u)/2 \tag{5.40}$$

et

$$\left|\left\langle x, g_{(s(u),u,\xi(s(u),u))}\right\rangle\right| = a(u)\left(\frac{2\pi}{|\phi''(u)|}\right)^{1/4}(1 + \eta(u)) \tag{5.41}$$

où $\eta(u)$ est majoré par

$$\eta_{max}(\lambda s_0, u) \stackrel{\Delta}{=} \sqrt{\lambda}\epsilon_{max}(\lambda s_0, u)/(2\pi)^{1/4} \tag{5.42}$$

et

$$\underline{\beta}(\lambda s_0, u) \stackrel{\Delta}{=} \arg\cosh\left(1 - 2\eta_{max}(\lambda s_0, u)\right)^{-4}. \tag{5.43}$$

Plus ϵ_{max} est petit, plus on peut trouver une grande valeur λ_0 vérifiant (5.39). La relation (5.40) montre qu'alors au moins un maximum local de $s \mapsto \left|\left\langle x, g_{(s,u,\xi(s,u))}\right\rangle\right|$ est proche de l'échelle "idéale" (5.33) que l'on aurait obtenue en l'absence du terme d'erreur. Plus λ_0 est grand, plus on a des chances que ce maximum local soit un maximum absolu, car il est le maximum absolu sur un grand intervalle. Cependant on ne peut pas contrôler les maxima locaux hors de l'intervalle $]0, \lambda_0 s_0]$, car $\epsilon_{max}(s, u)$ devient grand lorsque s devient grande. Par exemple, si l'on veut être sûr de trouver un maximum local dans $]0, 2s_0]$, il suffit que

$$\epsilon_{max}(2s_0, u) \leq 0.089$$

Une fois contrôlée l'*existence* d'un maximum local, on peut lire en (5.43) la précision avec laquelle ce maximum local permet de mesurer le chirp instantané $\phi''(u)$.

Conditions d'utilisation

Les bornes uniformes $\|.\|_\infty$ exigées dans le théorème 4 sont essentiellement techniques. En pratique, $\epsilon(s,u,\xi) \ll 1$ pour les "petites" échelles, caractérisées par

$$s \ll \left|a(u)/a'(u)\right| \tag{5.44}$$

et

$$s \ll 1/\left|\phi'''(u)\right|^{1/3}. \tag{5.45}$$

Pour que le terme d'erreur soit petit au voisinage de l'échelle $s \approx 1/\sqrt{|\phi''(u)|}$ qui nous intéresse, il suffit donc que

$$\left|a'(u)/a(u)\right| \ll \left|\phi''(u)\right|^{1/2} \tag{5.46}$$

et

$$\left|\phi'''(u)\right| \ll \left|\phi''(u)\right|^{3/2}. \tag{5.47}$$

L'approximation (5.49) n'est donc valable que si le terme de chirp *linéaire* est dominant devant les variations d'amplitude et les termes de chirp d'ordre supérieur.

Parmi tous les maxima locaux d'énergie (s_i, u_i, ξ_i) de $\mathcal{D}$, seuls ceux pour lesquels le terme d'erreur est petit permettent de mesurer la fréquence et le chirp instantanés

$$\phi'(u_i) \approx \xi_i \tag{5.48}$$

et

$$\phi''(u_i) \approx \pm\frac{1}{s_i^2} \tag{5.49}$$

avec des perturbations

$$\delta\xi(s_i,u_i) \approx \frac{\sqrt{8\,|\phi''(u_i)|}}{(2\pi)^{1/8}}\sqrt{\epsilon_{max}(s_i,u_i)} \tag{5.50}$$

et

$$\underline{\beta}(\lambda s_0(u_i),u_i) \approx \frac{4\lambda^{1/4}}{(2\pi)^{1/8}}\sqrt{\epsilon_{max}(\lambda s_0(u_i),u_i)} \tag{5.51}$$

que l'on obtient en développant (5.38) et (5.43) pour $\epsilon \ll 1$.

5.2.2 Recherche *locale* du meilleur atome chirpé

Lors de la m-ème itération de la poursuite, la localisation (5.17) du meilleur atome de $\mathcal{D}$ fournit *deux* "bons" candidats

$$g_{(s_m^{ridge},u_m^{ridge},\xi_m^{ridge},\pm 1/(s_m^{ridge})^2)}$$

à la maximisation de $\left\langle R^{m-1}x, g_{(s,u,\xi,c)}\right\rangle$. On peut ensuite jouer sur l'échelle s pour augmenter encore la corrélation. Cependant l'estimation du chirp optimal ainsi obtenue pose deux problèmes. Un premier problème, certes mineur, est l'indétermination du signe de $c_m = \pm 1/(s_m^{ridge})^2$, qu'il faut lever en calculant $\left\langle R^{m-1}x, g_{(s_m^{ridge},u_m^{ridge},\xi_m^{ridge},c_m)}\right\rangle$ pour chacun des signes et en sélectionnant le meilleur. Le principal problème vient de l'imprécision de cet estimateur lorsque l'échelle $s = 2^j$ est quantifiée grossièrement, ce qui est le cas dans le dictionnaire de Gabor $\mathcal{D}$, généralement employé. Il faut donc chercher également le meilleur paramètre de chirp.

Par rapport à une poursuite simple sur $\mathcal{D}$, on ajoute donc les deux étapes suivantes, dont nous calculons le coût :

1. $[\mathcal{O}((s_m^{ridge})^2)]$-Optimisation du chirp par "balayage" des $\mathcal{O}(s_m^{ridge})$ valeurs discrètes possibles

$$c_m = \arg\max_c \left|\left\langle R^{m-1}x, g_{(s_m^{ridge},u_m^{ridge},\xi_m^{ridge},c)}\right\rangle\right|. \tag{5.52}$$

   Chaque calcul de produit scalaire coûte $\mathcal{O}(s_m^{ridge})$.

2. $[\mathcal{O}(N \log N)]$-Optimisation de échelle par "balayage" des $\mathcal{O}(\log N)$ échelles a^j discrètes

$$s_m = \arg\max_s \left|\left\langle R^{m-1}x, g_{(s,u_m^{ridge},\xi_m^{ridge},c_m)}\right\rangle\right|. \tag{5.53}$$

   Chaque produit scalaire coûte $\mathcal{O}(a^j)$.

L'essentiel du sur-coût, par rapport à un Matching Pursuit avec un dictionnaire temps-fréquence "simple", provient de la nécessité de "balayer" les chirps possibles. En effet, comme s_m^{ridge} peut atteindre N, le coût de ce balayage peut atteindre $\mathcal{O}(N^2)$.

Il faut faire une analyse plus fine du comportement local de $(s,u,\xi) \mapsto \left\langle R^{m-1}x, g_{(s,u,\xi)}\right\rangle$ pour estimer c_m avec un coût raisonnable. Nous développons dans ce but une version plus fine du théorème 4.

5.2.3 Un théorème de ridge à l'ordre supérieur

Pour éviter le balayage "brutal" (5.52), on remplace la recherche *exhaustive* par une *estimation rapide* des bons paramètres. L'observation visuelle d'un spectrogramme (par exemple sur la figure 5.9) montre que l'information de pente fréquentielle est présente *localement* dans le spectre du signal, ce que le théorème 4 exprime mathématiquement. Pour construire un estimateur local des paramètres s et c optimaux, nous proposons une approche qui s'exprime en trois points :

- Sous certaines conditions de régularité, le résidu $R^{m-1}x = a(t)e^{i\phi(t)}$, vu "a travers" un atome $g_{(s,u,\xi,c)}$, est proche d'un atome chirpé

$$\left\langle R^{m-1}x, g_{(s,u,\xi,c)}\right\rangle \approx A(u)e^{i\Phi(u)} \left\langle g_{\gamma^+(u)}, g_{(s,u,\xi,c)}\right\rangle \tag{5.54}$$

- A u fixé, l'optimisation de l'énergie en fonction de s, ξ, c, est alors équivalente à l'estimation des paramètres de cet atome chirpé

$$\xi_m \approx \phi'(u) \tag{5.55}$$
$$c_m \approx \phi''(u) \tag{5.56}$$

et si $\alpha'(u) \ll \sqrt{\alpha''(u)}$, où $\alpha(t) = -\log a(t)$,

$$s_m^2 \approx 1/\alpha''(u) \tag{5.57}$$

- L'observation *locale* de $\xi \mapsto \left\langle R^{m-1}x, g_{(s,u,\xi)} \right\rangle$ permet d'estimer les paramètres α'', ϕ' et ϕ'', et donc les indices optimaux s_m, ξ_m, et c_m.

On construit ainsi un *estimateur local rapide* (de coût $\mathcal{O}(1)$) du chirp et de l'échelle qui maximisent la corrélation. Il nous faudra cependant prendre garde à *tester* la validité de l'approximation (5.54). Le théorème suivant, que nous démontrons en annexe B, est une version "à l'ordre supérieur" du théorème 4.

Théorème 5 *Soit $x(t) = a(t)e^{i\phi(t)}$ un signal analytique. On suppose que $\|a\|_\infty < \infty$, $\|\phi'''\|_\infty < \infty$ et $\|\alpha'''\|_\infty < \infty$, où $\alpha(t) \stackrel{\Delta}{=} -\log a(t)$. Soit u un instant où $\alpha''(u) > 0$, et $g_{(s,u,\xi,c)}$ un atome temps-fréquence gaussien chirpé. Alors*

$$\left\langle x, g_{(s,u,\xi,c)} \right\rangle = \frac{a(u)e^{i\phi(u)}}{(\alpha'')^{1/4}} e^{\frac{(\alpha')^2}{2\alpha''}} e^{-i\phi'\frac{\alpha'}{\alpha''} + i\frac{\phi''}{2}\left(\frac{\alpha'}{\alpha''}\right)^2} \left(\left\langle g_{\gamma^+(u)}, g_{(s,u,\xi,c)} \right\rangle + \epsilon(s,u,\xi,c) \right) \tag{5.58}$$

où

$$\gamma^+(u) = \left(\frac{1}{\sqrt{\alpha''(u)}}, u - \frac{\alpha'(u)}{\alpha''(u)}, \phi'(u) - \phi''(u)\frac{\alpha'(u)}{\alpha''(u)}, \phi''(u) \right) \tag{5.59}$$

et $|\epsilon(s,u,\xi,c)|$ est majoré par

$$\epsilon_{max}(s,u) = (\alpha'' s^2)^{1/4} \left(\frac{4\|a\|_\infty}{a(u)} (Ks^3)^{1/3} e^{-\frac{1}{2(Ks^3)^{2/3}}} + \frac{Ks^3\sigma_3^3}{6} e^{1/6} \right) \tag{5.60}$$

avec

$$K \stackrel{\Delta}{=} \left\|\alpha'''\right\|_\infty + \left\|\phi'''\right\|_\infty \tag{5.61}$$

et

$$\sigma_3^3 \stackrel{\Delta}{=} \int |t|^3 g(t) dt. \tag{5.62}$$

Cette approximation de $\left\langle x, g_{(s,u,\xi,c)} \right\rangle$ nous montre que x, observé "à travers" un atome de $\mathcal{D}^+$, *ressemble* à un atome $g_{\gamma^+(u)}$ de $\mathcal{D}^+$ dont la fréquence instantanée au temps u est $\phi'(u)$.

Estimation de paramètres à partir du ridge

Un raisonnement analogue à celui menant aux corollaires 1 et 2 permettrait de montrer que, si le terme d'erreur (5.60) est suffisamment petit[2], les valeurs des paramètres s, ξ, c sur le ridge permettent d'estimer la fréquence et le chirp instantanés

$$\xi^{ridge} = \arg\max_\xi \left|\left\langle x, g_{(s,u,\xi)}\right\rangle\right| \approx \phi'(u) \tag{5.63}$$

$$c^{ridge} = \arg\max_c \left|\left\langle x, g_{(s,u,\xi^{ridge},c)}\right\rangle\right| \approx \phi''(u). \tag{5.64}$$

De plus si $|\alpha'(u)/\alpha''(u)| \ll 1/\sqrt{\alpha''(u)}$, *i.e.* $|\alpha'(u)| \ll \sqrt{\alpha''(u)}$, alors

$$s^{ridge} = \arg\max_s \left|\left\langle x, g_{(s,u,\xi^{ridge},c^{ridge})}\right\rangle\right| \approx 1/\sqrt{\alpha''(u)}. \tag{5.65}$$

Conditions de validité de l'approximation

Là encore, les majorations uniformes de a, α''' et ϕ''' requises au théorème 5 sont essentiellement techniques. En pratique $\epsilon(s,u,\xi,c) \ll 1$ pour les "petites" échelles d'analyse s, c'est-à-dire dès que

$$s \ll 1/\left|\alpha'''(u)\right|^{1/3} \tag{5.66}$$

et

$$s \ll 1/\left|\phi'''(u)\right|^{1/3}. \tag{5.67}$$

Pour que le terme d'erreur soit petit au voisinage de l'échelle $s^{ridge} \approx 1/\sqrt{|\phi''(u)|}$ sélectionnée en (5.17), il suffit donc que

$$\left|\phi'''(u)\right| \ll \left|\phi''(u)\right|^{3/2} \tag{5.68}$$

$$\left|\alpha'''(u)\right| \ll \left|\phi''(u)\right|^{3/2}. \tag{5.69}$$

Par ailleurs l'hypothèse $\alpha''(u) > 0$ correspond à la condition

$$\frac{a''(u)}{a(u)} < \left(\frac{a'(u)}{a(u)}\right)^2 \tag{5.70}$$

qui est vérifiée dès que, par exemple, $a''(u) \leq 0$, c'est-à-dire sur les parties *concaves* de l'amplitude a, et en particulier au voisinage de ses maxima locaux suffisamment réguliers.

[2] La relation (5.65) n'est vraie que si le terme d'erreur (5.60) est petit à l'échelle $1/\sqrt{\alpha''(u)}$, c'est-à-dire si

$$\left|\alpha'''(u)\right| \ll \left|\alpha''(u)\right|^{3/2}$$

$$\left|\phi'''(u)\right| \ll \left|\alpha''(u)\right|^{3/2}$$

Il est donc bien équivalent[3] d'*optimiser* les paramètres s, ξ, c (en vue de maximiser l'énergie) et d'*estimer* les variations locales de phase et d'amplitude du signal modélisé par (5.24). Cependant, on veut éviter le balayage coûteux du paramètre c : cela est possible car on peut estimer $\phi''(u)$ à l'aide du *comportement local* de $\xi \mapsto \left\langle R^{m-1}x, g_{(s,u,\xi)} \right\rangle$ au voisinage du ridge, à l'échelle s_m^{ridge}.

5.2.4 Recherche locale rapide du meilleur atome chirpé

L'estimation spectrale des paramètres d'un atome gaussien chirpé fait appel à l'expression analytique (A.2) du produit scalaire entre deux atomes chirpés gaussiens. Lorsque l'échelle s_m^{ridge} et le temps u_m^{ridge} sont fixés par (5.17), l'observation *locale* de $\xi \mapsto \left\langle R^{m-1}x, g_{(s_m^{ridge}, u_m^{ridge}, \xi)} \right\rangle$, c'est-à-dire du spectrogramme à l'échelle s_m^{ridge} autour du temps u_m^{ridge}, permet d'estimer les paramètres s_m et c_m, grâce à la propriété suivante, démontrée en annexe B.

Proposition 1 *Soit $x(t)$ un signal et $g_{(s,u,\xi)}$ un atome gaussien non chirpé tels que l'approximation (5.58) soit vérifiée avec $\epsilon_{max} \ll 1$. Soit*

$$\left\langle x, g_{(s,u,\xi)} \right\rangle = A e^{i\Phi}$$

leur produit scalaire. Lorsque s et u sont fixés, $\xi \mapsto \log A(\xi)$ et $\xi \mapsto \Phi(\xi)$ sont des polynômes de second degré en ξ. Soient $(\log A)''(\xi)$ et $\Phi''(\xi)$ les "courbures" de ces deux paraboles. Alors

$$\phi''(u) = -\frac{\Phi''(\xi)}{((\log A)''(\xi))^2 + (\Phi''(\xi))^2} \tag{5.71}$$

et

$$\alpha''(u) + \frac{1}{s^2} = \frac{-\log A''(\xi)}{((\log A'')(\xi))^2 + (\Phi''(\xi))^2} \tag{5.72}$$

Grâce à cette propriété, on estime s_m et c_m sans "balayer" toutes les valeurs possibles de ces paramètres. On peut ensuite *ré-estimer* la fréquence ξ_m de l'atome. En effet ξ_m est maintenant l'emplacement du maximum des deux paraboles définies par $\log A(\xi)$ et $\Phi(\xi)$.

Ce type d'estimation a été utilisé par Marques et Almeida [MA89, MA86] pour l'analyse des non-stationarités dans le cadre du traitement de la parole. Cependant, travaillant dans le cadre d'une analyse de Fourier à fenêtre, ces auteurs n'utilisent que l'estimation de ϕ'' et ne tirent pas parti de l'*échelle locale* définie par $1/\sqrt{\alpha''}$. Leur méthode d'analyse a donc des difficultés avec les transitoires. La poursuite que nous utilisons est un outil multi-échelle et ne recontre pas ces problèmes.

[3] C'est pourquoi l'estimation de c_m par le biais de l'échelle sur le ridge (5.49) a un sens.

Validation du modèle

Pour *tester* la validité de l'approximation du résidu par un atome chirpé, on utilise la propriété suivante : lorsque l'approximation est valable, on a

$$\left|\Phi''(\xi)\right| \leq \frac{s^2}{2} \tag{5.73}$$

$$0 > \log A''(\xi) \geq -s^2. \tag{5.74}$$

Nous établissons ces inégalités en annexe A comme corollaire de la démonstration de la proposition 1. Pour être certain que le modèle n'est pas valide, il suffit donc que ces conditions ne soient pas respectées.

5.2.5 Estimation numérique par *interpolation*

On estime les "courbures" des paraboles $\log A(\xi)$ et $\Phi(\xi)$ à l'aide d'une interpolation parabolique[4] : on mesure pour cela le spectrogramme complexe à l'échelle s_m^{ridge} en trois points

$$\left\langle R^{m-1}x, g_\varepsilon \right\rangle = A_\varepsilon e^{i\Phi_\varepsilon},\ \varepsilon \in \{-1, 0, +1\}\,, \tag{5.75}$$

associés aux atomes temps-fréquence gaussiens

$$g_\varepsilon = g_{(s_m^{ridge}, u_m^{ridge}, \xi_m^{ridge} + \varepsilon\Delta\xi(s_m^{ridge}))}.$$

L'interpolation parabolique de trois valeurs $y[-1], y[0], y[1]$ par $y[k] = \frac{\alpha}{2}k^2 + \beta k + \gamma$ donne

$$\begin{aligned} \alpha &= y[-1] - 2y[0] + y[1] \\ \beta &= \frac{y[1] - y[-1]}{2}. \end{aligned}$$

La position de l'extrémum est alors $k_{ext} = -\frac{\beta}{\alpha}$ et la "courbure" α. En tenant compte du pas d'échantillonnage en fréquence $\Delta\xi(s_m^{ridge})$, la dérivée seconde le long de la fréquence se déduit de la dérivée "numérique" par

$$\alpha_\Phi = \Phi''(\xi)(\Delta\xi_s)^2 \tag{5.76}$$

$$\alpha_{\log A} = (\log A)''(\xi)(\Delta\xi_s)^2 \tag{5.77}$$

La figure 5.4 représente les "centres" M_{-1}, M_0 et M_1, des atomes g_ε, de coordonnées

$$(u_m^{ridge}, \xi_m^{ridge} + \varepsilon\Delta\xi(s_m^{ridge}))$$

ainsi qu'une ellipse grisée. Les points localisent l'endroit où les mesures du spectrogramme doivent être effectuées, tandis que l'ellipse symbolise la localisation temps-fréquence de l'atome chirpé de l'approximation (5.58).

[4] Pour une fenêtre quelconque de forme connue, sans chirp, McIntyre et Dermott ont montré [MD92] que la *régression* sur l'amplitude est plus robuste que l'interpolation. Avec des fenêtres gaussiennes, interpolation et régression sur trois points coïncident. La régression devient utile lorsque l'on mesure plus de points, mais alors ils sont plus loin du ridge et donc plus perturbés.

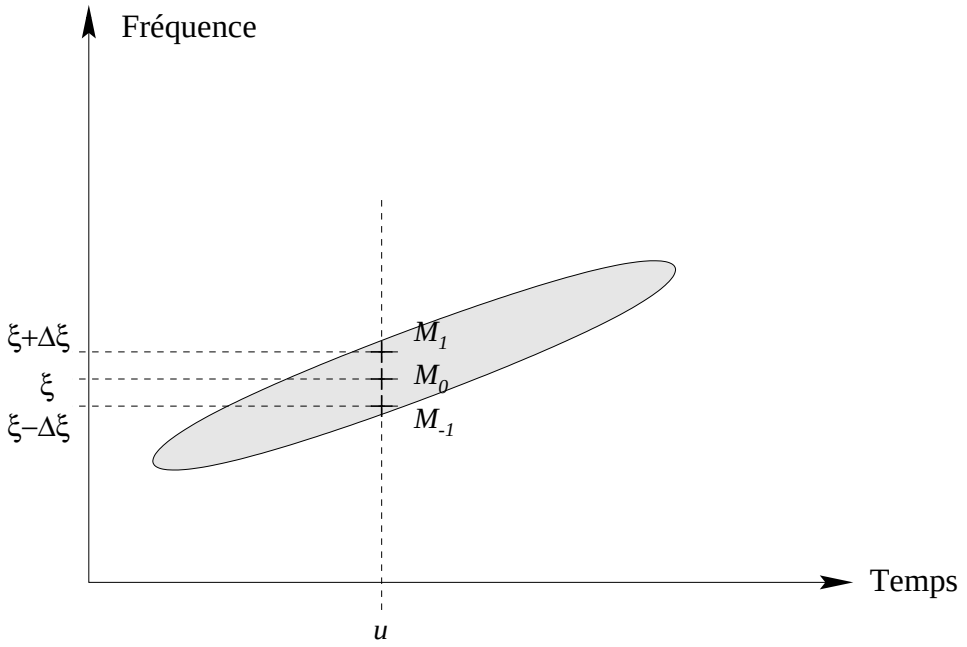

FIG. 5.4 – Estimation du chirp c_m et de l'échelle s_m à partir de trois points d'un spectrogramme.

Ambiguïté du déroulement de phase

Pour estimer α_Φ, un problème de "déroulement" se pose. Les phases Φ_ε n'étant définies qu'à 2π près, α_Φ n'est lui-même défini qu'à 2π près. Cependant, un nombre limité des valeurs de $\alpha_\Phi + 2n\pi$ est compatible avec la condition (5.73). Les seules valeurs de α_Φ qui ont un sens vérifient donc

$$|\alpha_\Phi| = \left|\Phi''(\xi)\right| \left(\Delta\xi(s_m^{ridge})\right)^2 \leq \frac{(s_m^{ridge})^2}{2}(\Delta\xi(s_m^{ridge}))^2 \tag{5.78}$$

Pour lever l'ambiguï té sur α_Φ, il suffit que l'intervalle dans lequel α_Φ a un sens soit de longueur strictement inférieure à 2π, c'est-à-dire que les mesures de Φ soient suffisamment rapprochées, avec un pas d'échantillonnage fréquentiel

$$\Delta\xi(s_m^{ridge}) < \frac{\sqrt{2\pi}}{s} \tag{5.79}$$

Si l'on utilisait l'analogue temporel de la proposition 1, un raisonnement analogue montrerait qu'il suffit que

$$\Delta u(s_m^{ridge}) < \sqrt{2\pi}s. \tag{5.80}$$

pour disposer d'un déroulement unique[5] de la phase $\Phi(u)$.

Complexité

Comme les trois produits scalaires complexes $\left\langle R^{m-1}x, g_\varepsilon \right\rangle$ ont *déjà* été calculés pour sélectionner le meilleur atome de Gabor non chirpé (5.17), la complexité de l'estimation des paramètres optimaux c_m et s_m ne dépend pas de N, et vaut $\mathcal{O}(1)$.[6]

5.3 Matching Pursuit Chirpé Réel Rapide

5.3.1 Résumé de l'algorithme et complexité

Chaque itération de la poursuite *de ridges* sur un dictionnaire de chirps gaussiens à valeurs réelles se décompose en un certain nombre d'étapes. Afin de sélectionner l'atome chirpé réel

$$g_{(s_m,u_m,\xi_m,c_m,\phi_m)},$$

on est amené successivement à effectuer

1. Le calcul des corrélations $\left\| P_{\boldsymbol{V}_\gamma} R^{m-1}x \right\|^2$ du résidu avec tous les atomes réels du dictionnaire de Gabor *non-chirpé* $\mathcal{D}$ (*cf* section 3.2.3).
2. La sélection du meilleur atome non-chirpé *réel*

$$(s_m^{ridge}, u_m^{ridge}, \xi_m^{ridge}, \phi_m^{ridge}) = \arg \max_{(s,u,\xi,\phi)} \left| \left\langle R^{m-1}x, g_{(s,u,\xi,\phi)} \right\rangle \right|. \quad (5.81)$$

3. L'estimation des paramètres optimaux s_m, ξ_m et c_m par une interpolation parabolique, à l'aide des produits scalaires avec les atomes complexes situés en

$$(s_m^{ridge}, u_m^{ridge}, \xi_m^{ridge} + \varepsilon\Delta\xi(s_m^{ridge})).$$

4. Le calcul du produit scalaire de l'atome chirpé complexe (s_m, u_m, ξ_m, c_m) sélectionné, puis de la phase *optimale* ϕ_m et de la corrélation du meilleur atome réel associé, selon les formules (3.33) et (3.30).
5. La mise à jour du résidu.

[5] Les conditions (5.79) et (5.80) ont donc pour conséquence que

$$\Delta u(s_m^{ridge}) \times \Delta\xi(s_m^{ridge}) < 2\pi,$$

c'est-à-dire [Dau92] que le réseau de gaussiennes à l'échelle s_m^{ridge} utilisé doit constituer un *frame* de $\boldsymbol{L}^2(\mathbb{R})$.

[6] Dans le domaine de la dé-modulation FM numérique en environnement atténué ("fading environment"), une technique aujourd'hui utilisée [KJ96] est la comparaison du signal à diverses fenêtres chirpées. Notre technique d'estimation rapide du chirp instantané pourrait permettre de réduire la complexité.

Complexité du Matching Pursuit Chirpé Réel

On a vu à la section 4.1 que les deux premières étapes coûtent $\mathcal{O}(N \log^2 N)$. Par ailleurs, on vient de voir que la troisième étape a un coût négligeable $\mathcal{O}(1)$. Enfin la quatrième étape demande de calculer *un* produit scalaire, et coûte donc au maximum $\mathcal{O}(N)$. Le coût additionnel du calcul de phase et d'énergie de l'atome réel est lui aussi négligeable $\mathcal{O}(1)$, il nécessite seulement de calculer $\left\langle \overline{g_{(s_m,u_m,\xi_m,c_m)}}, g_{(s_m,u_m,\xi_m,c_m)} \right\rangle$, ce que l'on fait avec une précision arbitraire en $\mathcal{O}(1)$ grâce à la formule analytique (A.2).

La complexité totale d'une itération de Matching Pursuit Chirpé Réel vaut donc $\mathcal{O}(N \log^2 N)$, et celle de M étapes est

$$\mathcal{O}(MN \log^2 N). \tag{5.82}$$

5.3.2 Poursuite avec des maxima locaux

En utilisant les idées que nous avons développées au chapitre 4, on peut encore réduire la complexité du Matching Pursuit Chirpé. On procède comme suit :

1. Détermination du sous-dictionnaire $\mathcal{D}_m$ de maxima locaux du dictionnaire de Gabor non chirpé $\mathcal{D}$.
2. Calcul des chirps optimaux pour chacun des atomes de ce sous-dictionnaire, à l'aide de l'estimation locale. On a alors un sous-dictionnaire $\mathcal{D}_m^+$ de "maxima locaux" du dictionnaire chirpé $\mathcal{D}^+$.
3. Poursuite "normale" dans le sous-dictionnaire, jusqu'à épuisement de celui-ci.

On obtient alors une complexité

$$\mathcal{O}(MN). \tag{5.83}$$

5.3.3 Sous-optimalité

La poursuite en deux temps (5.17)-(5.18) que nous avons proposée a un inconvénient : elle nous force à renoncer a *l'optimalité locale* du Matching Pursuit. En effet $g_{(s_m,u_m,\xi_m,c_m)}$ n'est pas forcément le meilleur atome de $\mathcal{D}^+$. La perte engendrée, en termes d'énergie capturée par l'atome choisi, est caractérisée par le rapport

$$\mu(R^{m-1}x) \triangleq \frac{\left|\left\langle R^{m-1}x, g_{(s_m,u_m,\xi_m,c_m)} \right\rangle\right|}{\sup_{(s,u,\xi,c)} \left|\left\langle R^{m-1}x, g_{(s,u,\xi,c)} \right\rangle\right|} \tag{5.84}$$

Illustrons cela sur un exemple : soient $g_{(s_1,u_1,\xi_1)} \in \mathcal{D}$ et $g_{(s_2,u_2,\xi_2,c_2)} \in \mathcal{D}^+$ deux atomes. On suppose que

$$\left\langle g_{(s_1,u_1,\xi_1)}, g_{(s_2,u_2,\xi_2,c_2)} \right\rangle \approx 0.$$

On analyse le signal

$$x = \alpha g_{(s_1,u_1,\xi_1)} + g_{(s_2,u_2,\xi_2,c_2)} \tag{5.85}$$

représenté sur la figure 5.5 avec sa distribution temps-fréquence définie par la décomposition atomique naturelle 5.85. On peut montrer à l'aide de l'ex-

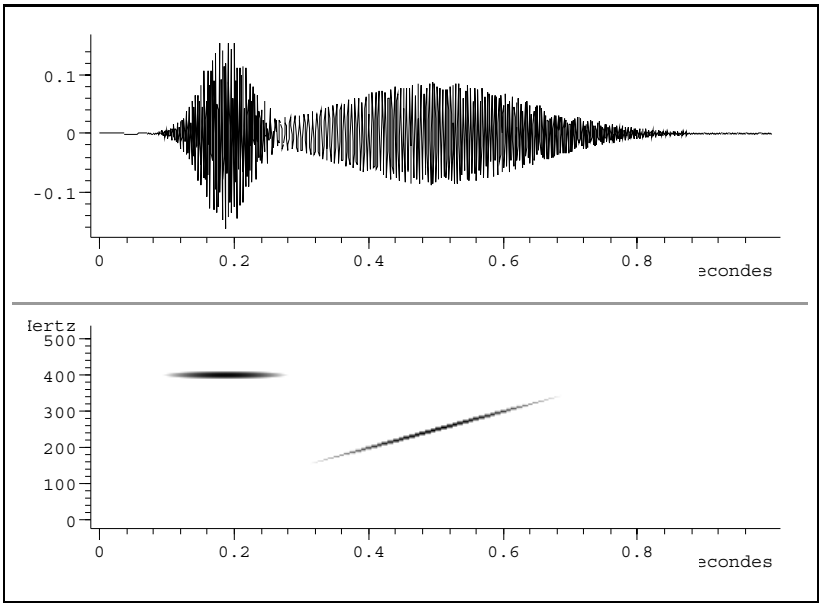

FIG. 5.5 – Représentation temps-fréquence d'un signal où le Matching Pursuit de ridges est sous-optimal.

pression analytique (A.2) que l'atome de $\mathcal{D}$ le plus corrélé à $g_{(s_2,u_2,\xi_2,c_2)}$ ne peut avoir avec lui un produit scalaire plus grand que

$$\mu(x) = \mu(g_{(s_2,u_2,\xi_2,c_2)}) = \sup_{g_{(s,u,\xi)} \in \mathcal{D}} \left| \left\langle g_{(s_2,u_2,\xi_2,c_2)}, g_{(s,u,\xi)} \right\rangle \right| = \left(\frac{2}{1 + \sqrt{1 + c_2^2 s_2^4}} \right)^{1/4} < 1. \tag{5.86}$$

Dès que $\alpha \in]\beta, 1[$, l'atome de $\mathcal{D}$ sélectionné dans un premier temps est $g_{(s_1,u_1,\xi_1)}$ bien que le meilleur atome chirpé soit $g_{(s_2,u_2,\xi_2,c_2)}$. Le second atome sélectionné est $g_{(s_2,u_2,\xi_2,c_2)}$, puis l'énergie du résidu devient négligeable.

La poursuite dans un sous-dictionnaire, que nous avons introduit dans le but d'accélérer l'algorithme, présente (outre l'accélération) un effet positif supplémentaire : la sélection d'un ensemble de maxima locaux du dictionnaire non chirpé augmente les chances d'y trouver la trace du meilleur atome chirpé. La sous-optimalité est donc atténuée.

5.4 Résultats numériques

Pour illustrer les perspectives offertes par notre outil d'analyse, nous avons effectué quelques analyses de signaux, d'abord sur des signaux artificiels de référence, puis sur quelques signaux réels "chirpés" connus.

5.4.1 Analyse d'un chirp hyperbolique

Nous avons d'abord analysé un chirp hyperbolique modulé en amplitude

$$x(t) \triangleq a(t) \cos(2\pi\omega \log t) \tag{5.87}$$

dont la fréquence et le chirp instantanés sont

$$\xi(t) = \omega/t \tag{5.88}$$
$$c(t) = -\omega/t^2. \tag{5.89}$$

La dérivée troisième de la phase étant $\phi'''(t) = 2\omega/t^3$, la condition (5.68) est indépendante de t et s'écrit

$$\omega \gg 4 \tag{5.90}$$

On représente à la figure 5.6 les résultats d'analyse du signal obtenu avec $\omega = 100$, échantillonné sur $N = 8192$ points. Sa fréquence instantanée dépasse, au voisinage de 0, la fréquence de Nyquist et donne lieu à un repliement spectral. La représentation du milieu est obtenue à l'aide de 1000 atomes chirpés, déterminés par un Matching Pursuit Chirpé accéléré avec des sous-dictionnaires de maxima locaux. On remarque que le repliement est détecté, et que l'énergie est bien concentrée autour de la fréquence instantanée. Le spectrogramme du signal, représenté en bas, est calculé avec une fenêtre gaussienne de 256 points. Au voisinage de l'origine, où la fréquence instantanée varie plus vite, la concentration de l'énergie autour de la celle-ci se dégrade au point que le spectrogramme finit par la perdre totalement.

5.4.2 Analyse d'un cri de chauve-souris

Pour repérer leur proies dans l'espace, certains animaux, tels les chauves-souris [CHT95a, CHT95b] ou les dauphins [WG98], émettent des ultrasons, modulés en fréquence, c'est-à-dire des *chirps*. En balayant ainsi le spectre en un intervalle de temps très court (la durée du chirp d'une chauve-souris *Eptesicus Fuscus* est d'environ 2.5 milli-secondes), ils parcourent une large gamme de longueurs d'onde et donc de dimensions d'objets. L'analyse des échos qui leurs parviennent leur permet de repérer, à la manière d'un sonar, la taille (liée à la fréquence de l'écho), la vitesse relative (par effet Doppler) des objets qui les entourent, et la distance de ceux-ci (liée au délai de retour de l'écho).

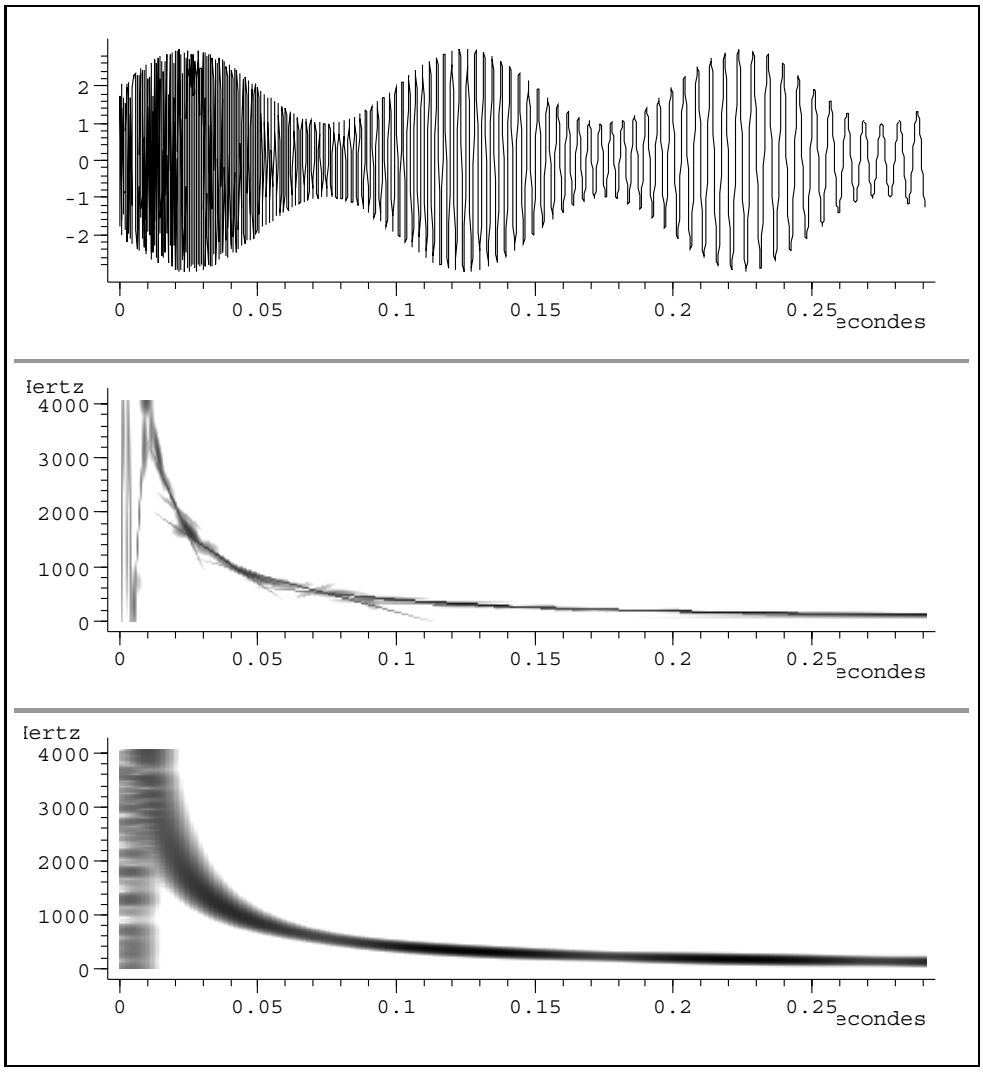

FIG. 5.6 – Représentations temps-fréquence d'un chirp hyperbolique. En haut : le signal analysé ($N = 8192$ points). Au milieu : représentation obtenue à l'aide de 1000 atomes déterminés avec la version accélérée du Matching Pursuit Chirpé de ridges. La fréquence instantanée est bien "suivie", y compris lors du repliement spectral au voisinage du temps $t = 0$. En bas : le spectrogramme du signal, calculé avec une fenêtre gaussienne de 256 points. La fréquence instantanée est perdue au voisinage de $t = 0$.

Nous avons ici analysé un tel chirp de 400 échantillons[7], échantillonné toutes les 7 micro-secondes. La figure 5.7 compare les représentations temps-fréquence de ce signal, obtenues par diverses méthodes.

(a) Le spectrogramme est effectué avec une fenêtre gaussienne de 64 points, choisie pour optimiser la concentration de l'énergie autour des chirps observés. Il possède de mauvaises caractéristiques de localisation temps-fréquence. Il est en effet obtenu par lissage de la transformée de Wigner-Ville avec un noyau lié à la fenêtre d'analyse.

(b) La transformée de Wigner-Ville [Fla93] est idéalement concentrée en temps-fréquence mais sa lisibilité est limitée par les termes oscillants d'interférence entre les différents chirps.

(c) Baraniuk et Jones [BJ93c, BJ93b, JB95] ont défini une représentation temps-fréquence adaptative en lissant la transformée de Wigner-Ville du signal avec un noyau optimal adapté à celui-ci. Elle fait disparaître les oscillations parasites, mais le dernier harmonique du chirp a ainsi tendance à disparaître.

(d) Le Matching Pursuit avec le dictionnaire de Gabor ordinaire fournit une décomposition assez morcelée des partiels chirpés, mais adapte néanmoins l'échelle des atomes qu'il sélectionne. Il concentre ainsi l'énergie autour de la fréquence instantanée du partiel.

(e) Le Matching Pursuit Chirpé de ridges représente le chirp de façon plus compacte : on peut observer que pour représenter chaque harmonique du chirp, il lui faut environ 5 atomes au lieu d'une dizaine pour le Matching Pursuit usuel. La décroissance de l'énergie du résidu, qui mesure la qualité de l'approximation, quantifie cela plus précisément. La figure 5.8 représente cette décroissance, en décibels, pour chaque type de poursuite.

5.4.3 Analyse du vibrato d'une voix chantée

La voix d'une chanteuse présente un vibrato, variation périodique de la fréquence instantanée, si caractéristique du timbre de la voix chantée. Il est absolument nécessaire de la reproduire lorsque l'on veut effectuer une synthèse réaliste de celle-ci [Rod80].

Schématiquement, dans un *vibrato*, la fréquence instantanée s'exprime comme

$$\phi'(t) = \cos(2\pi\omega t)$$

où ω est de l'ordre de 5 à 10 Hertz. Le chirp instantané ϕ'' admet donc des maxima locaux en $v_k = (k + 1/2)/\omega$. En ces points, $\phi'''(v_k) = 0$ tandis que $\phi''(v_k)$ est maximum, si bien que les conditions sont réunies pour mesurer le

[7] Je tiens à remercier Curtis Condon, Ken White, et Al Feng du Beckman Institute de l'Université de l'Illinois pour ce signal de chauve-souris et pour la permission de l'utiliser dans cette thèse.

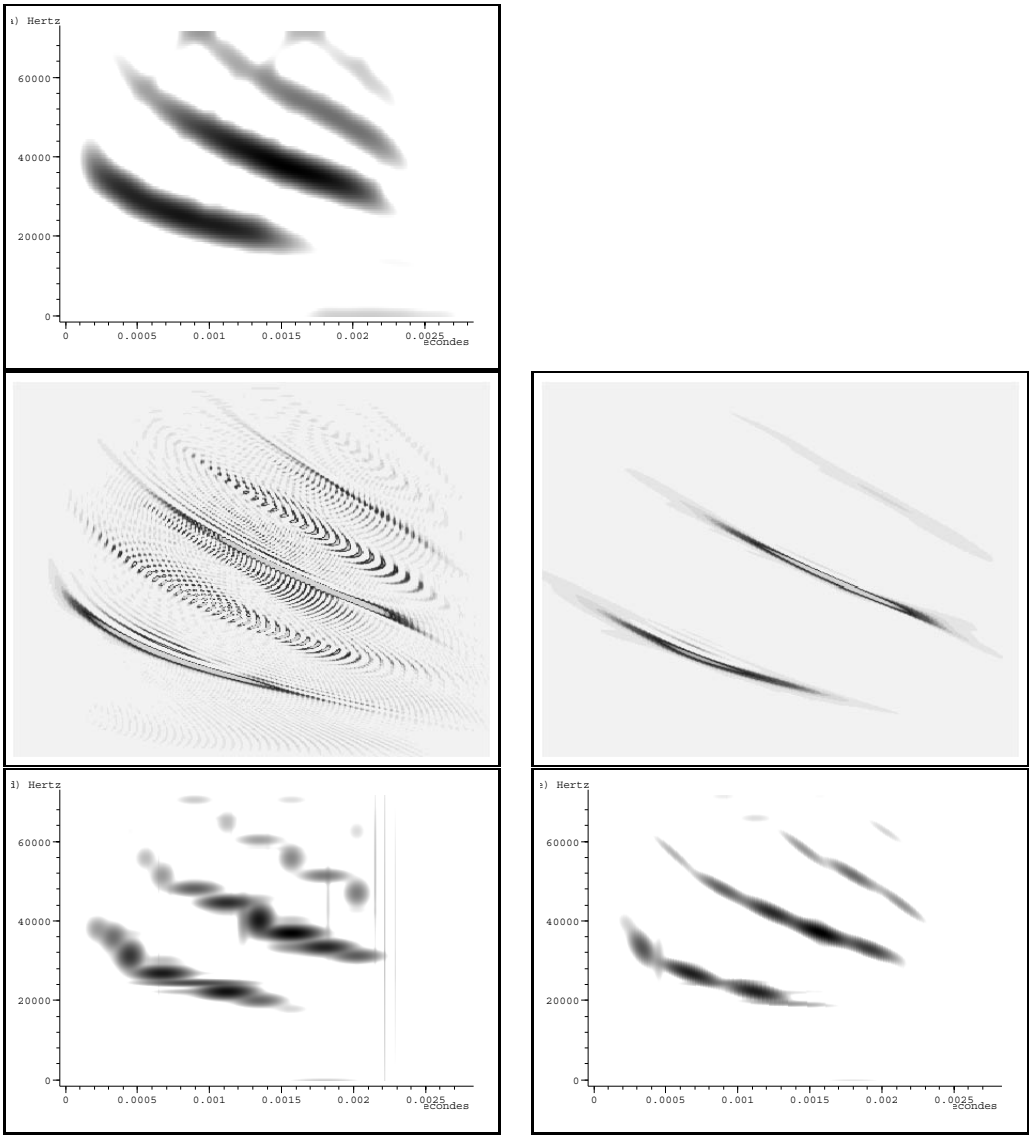

FIG. 5.7 – Différentes représentations temps-fréquence d'un chirp de chauve-souris. De haut en bas et de gauche à droite : (a) Spectrogramme, avec une fenêtre gaussienne de 64 points. La localisation temps-fréquence est imprécise. (b) Transformée de Wigner-Ville. Les termes oscillants gênent la lecture. (c) Lissage adaptatif de la transformée de Wigner-Ville avec un noyau optimal. Le quatrième harmonique disparaît. (d) Matching Pursuit avec un dictionnaire de Gabor (500 atomes). La fréquence instantanée est suivie, mais chaque chirp est morcelé en beaucoup d'atomes. (e) Matching Pursuit Chirpé (500 atomes). Chaque chirp est représenté par peu d'atomes, bien localisés autour de la fréquence instantanée.

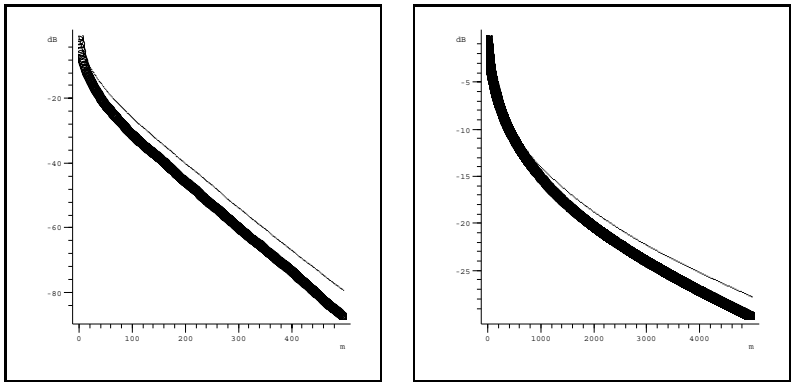

FIG. 5.8 – Décroissance, en décibels, de l'énergie du résidu d'un Matching Pursuit en fonction du nombre d'itérations. En gras, la courbe associée au Matching Pursuit Chirpé Rapide ; en fin, la décroissance obtenue avec le Matching Pursuit usuel. A gauche, le signal analysé est un chirp de chauve-souris, de 400 points. A droite, il s'agit d'un vibrato de chanteuse.

chirp à partir de l'échelle du ridge, comme proposé en (5.49). Cela ne serait par contre pas possible au voisinage des extrema locaux $u_k = k/\omega$ de la fréquence instantanée ϕ', où $\phi''(u_k) \approx 0$. En ces points, le corollaire 2 ne pourrait pas s'appliquer. En effet, l'échelle $s(u_k, \xi(u_k))$ des atomes de Gabor optimaux ne caractérise pas le chirp instantané en u_k : comme on l'a vu en (5.65), elle caractérise plutôt les variations locales d'amplitude.

On compare sur la figure 5.9 les représentations temps-fréquence obtenues par trois méthodes :

- le spectrogramme, où la fenêtre gaussienne de 512 points a été optimisée manuellement en vue de la lisibilité du vibrato,
- le Matching Pursuit dans le dictionnaire de Gabor non chirpé $\mathcal{D}$,
- le Matching Pursuit Chirpé de ridges.

Le signal, un extrait musical avec voix chantée et orchestre, est constitué d'environ $N = 30000$ échantillons, et les poursuites ont été effectuées avec $M = 5000$ itérations.

La fenêtre de grande échelle utilisée dans le spectrogramme lisse les transitoires des percussions. Au contraire, les deux représentations à base de poursuite permettent de lire ces attaques sous la forme de barres verticales associées à des atomes de petite échelle. De même les résonances des notes de l'orchestre, représentées par de fins traits horizontaux par les poursuites, sont mieux localisées en fréquence que dans le spectrogramme. Elles sont en effet représentées par des atomes dont l'échelle est plus grande que celle de la fenêtre d'analyse employée dans celui-ci. Enfin, le *vibrato* de la voix de la chanteuse est visible sur les trois représentations. Dans le dictionnaire de Gabor non chirpé $\mathcal{D}$, il est représenté par une multitude d'atomes à fréquence constante placés sur le "trajet" de la fréquence instantanée. Au

contraire, dans le dictionnaire chirpé $\mathcal{D}^+$, il est représenté comme un succession d'atomes chirpés "montants" et "descendants".

Enfin, on observe sur la figure de droite de la figure 5.8 que la décroissance, en décibels, de l'énergie du résidu est plus rapide avec la poursuite chirpée qu'avec la poursuite ordinaire. Pour obtenir une même qualité d'approximation, il faut donc moins d'atomes chirpés.

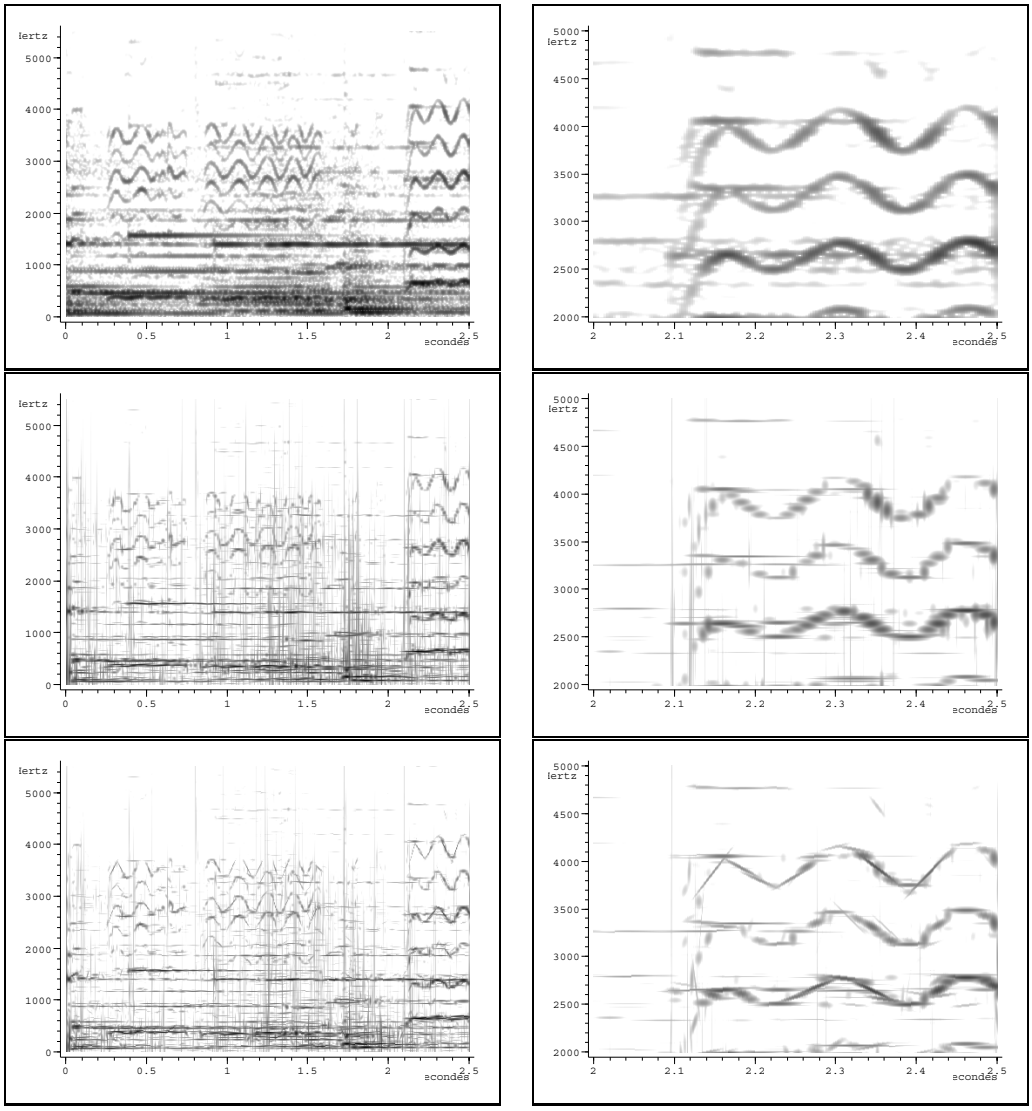

FIG. 5.9 – Représentations temps-fréquence d'un extrait musical avec voix chantée et orchestre (Extrait de *Seuils*, de M.-A. Dalbavie [Dal93]). Le signal comporte $N = 30000$ points. A droite : focalisation sur la représentation temps-fréquence de gauche. (a) Spectrogramme, avec une fenêtre gaussienne de 512 points. (b) Matching Pursuit dans le dictionnaire de Gabor $\mathcal{D}$ (5000 atomes). (c) Matching Pursuit Chirpé de ridges (5000 atomes). Les transitoires sont bien localisés par les poursuites, sous forme de lignes verticales associées à des atomes de petite échelle (aux temps $t = 0.8$ et $t = 2.1$ par exemple). Le vibrato, décomposé par la poursuite usuelle en une multitude d'atomes sur le trajet de la fréquence instantanée, est représenté à grands traits par le Matching Pursuit Chirpé, sous forme d'atomes chirpés "montants" et "descendants".

Chapitre 6

Matching Pursuit Haute Résolution

Le Matching Pursuit avec un dictionnaire temps-fréquence multi-échelle est très efficace pour analyser les signaux dans lesquels on trouve des structures à différentes échelles. Nous avons pu observer aux chapitres précédents que cet outil décompose bien les signaux sonores en transitoires (à petite échelle), parties entretenues et résonances de notes (à grande échelle). Cependant la résolution temporelle du Matching Pursuit n'est pas optimale. Par ailleurs, la reconstruction d'un signal sonore à partir des M premiers termes d'une décomposition atomique obtenue par un Matching Pursuit peut engendrer, si M n'est pas assez grand, un effet de pré-écho légèrement audible. Nous introduisons dans ce chapitre un critère "haute-résolution" de sélection d'atomes temps-fréquence. Il permet au Matching Pursuit "Haute Résolution" ainsi défini de surmonter ces problèmes.

6.1 Limitations de la poursuite

Le Matching Pursuit est un algorithme glouton, au sens où il optimise *à chaque itération* la quantité d'énergie (2.49) qu'il ôte au signal. Les atomes qu'il sélectionne sont ainsi adaptés à la représentation des structures globales du signal, mais pas forcément de ses structures locales.

6.1.1 Résolution temporelle

Considérons, par exemple, un signal composé de deux bosses modulées par la même sinusoïde

$$x(t) = \{g((t-u_1)/s) + g((t-u_2)/s)\}\, e^{i\xi t}. \tag{6.1}$$

Il admet une décomposition atomique "naturelle" en somme de deux atomes à l'échelle s. La figure 6.1-(a) représente un tel signal, et 6.1-(b) est la représentation temps-fréquence associée à sa décomposition atomique naturelle.

Si $|u_2 - u_1|$ est trop petit comparé à s, la poursuite décompose ce signal comme sur la figure 6.1-(c). Un "grand" atome (la ligne horizontale du milieu sur la figure 6.1-(c)) est d'abord sélectionné, à la fréquence ξ et à une échelle $s_1 \approx s + |u_2 - u_1| \geq s$. Il recouvre le support temporel des deux bosses. Ensuite, pour enlever l'énergie "créée" entre les deux bosses par le premier atome, la poursuite choisit deux atomes de la même taille que le premier, mais situés aux fréquences $\xi + \Delta\xi$ (la ligne du dessus) et $\xi - \Delta\xi$ (celle du dessous).

6.1.2 Pré-écho

On peut également observer que le Matching Pursuit ne conserve pas la localisation des attaques. La figure 6.2-(a) représente une "attaque" synthétique

$$x(t) = \chi(t \geq u) e^{-\alpha t} \cos(2\pi\omega t). \tag{6.2}$$

où $\chi(t \geq u)$ est l'échelon unité situé en u.

On observe sur la figure 6.2-(b) les atomes sélectionnés par la poursuite. Le premier atome choisi (une longue tache horizontale) est à grande échelle. Son support temporel s'étend au delà de l'instant u. Alors que le signal n'avait pas d'énergie avant l'instant u, le résidu $R^1 x$ en a, ainsi que tous les résidus d'ordre supérieur $R^m x$. Comme la reconstruction

$$x_M = \sum_{m=1}^{M} \left\langle R^{m-1} x, g_{\gamma_m} \right\rangle g_{\gamma_m} = x - R^M x \tag{6.3}$$

est effectuée avec un nombre fini M d'atomes, un léger pré-écho peut apparaître lors de la reconstruction. La figure 6.3-(b) montre le résidu de 100 itérations de poursuite effectuées sur le signal représenté en 6.3-(a). Le pré-écho s'y manifeste clairement.

6.1.3 Diagnostic

Le manque de résolution temporelle de la poursuite est dû au critère énergétique (2.49) utilisé lors de la sélection d'atomes. Celui-ci permet, pour ainsi dire, la "création" d'énergie dans le résidu là où le signal n'en avait pas : après quelques itérations, on peut avoir $\langle R^m x, g_\gamma \rangle \neq 0$ alors qu'initialement on avait $\langle x, g_\gamma \rangle = 0$. Ainsi dans l'exemple consacré au pré-écho, après le choix du premier atome, le résidu a de l'énergie *avant* l'instant u de l'attaque alors que le signal n'en avait pas. Après un certain nombre d'itérations, la poursuite sera donc amenée à choisir des atomes placés avant l'instant u, afin de capturer l'énergie du résidu qui y est présente.

Nous allons modifier la poursuite de façon à éliminer ce type de problèmes.

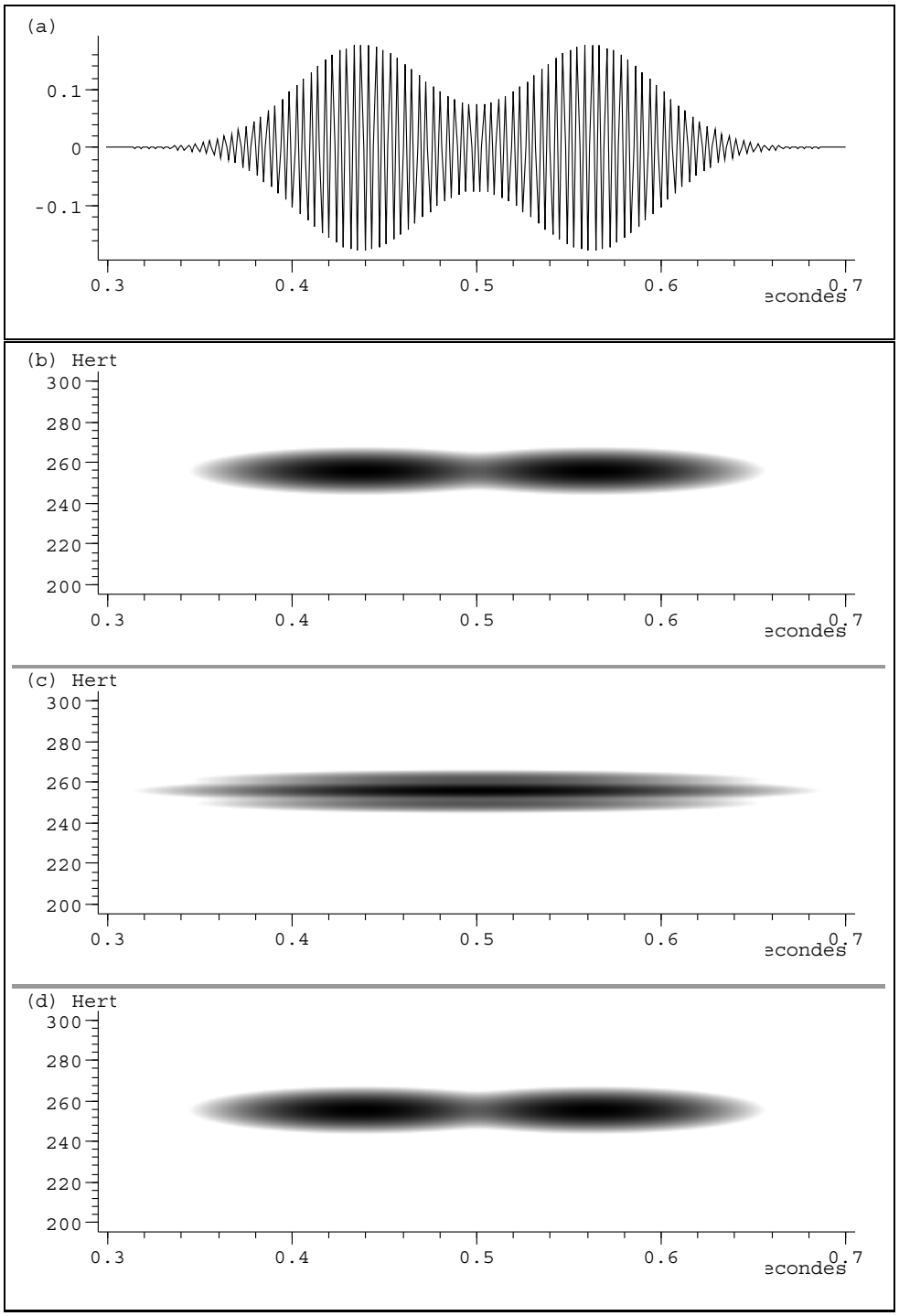

FIG. 6.1 – Un signal, constitué de la superposition de deux bosses (a), sa représentation temps-fréquence "idéale" (b), celles obtenues avec la poursuite ordinaire (c) et le Matching Pursuit Haute Résolution (d).

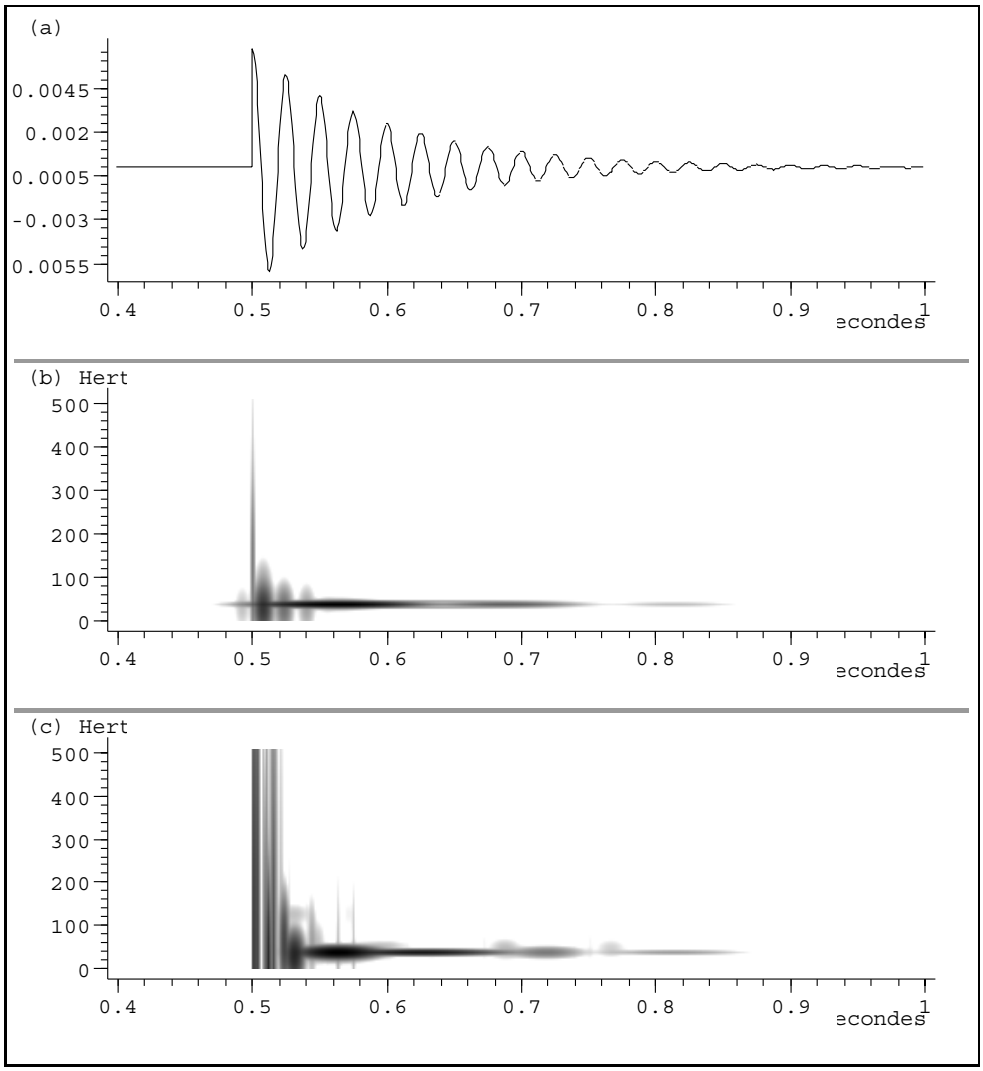

FIG. 6.2 – Une "attaque" synthétique (a) et ses représentations temps-fréquence obtenues avec le Matching Pursuit (b) et le Matching Pursuit Haute Résolution (c).

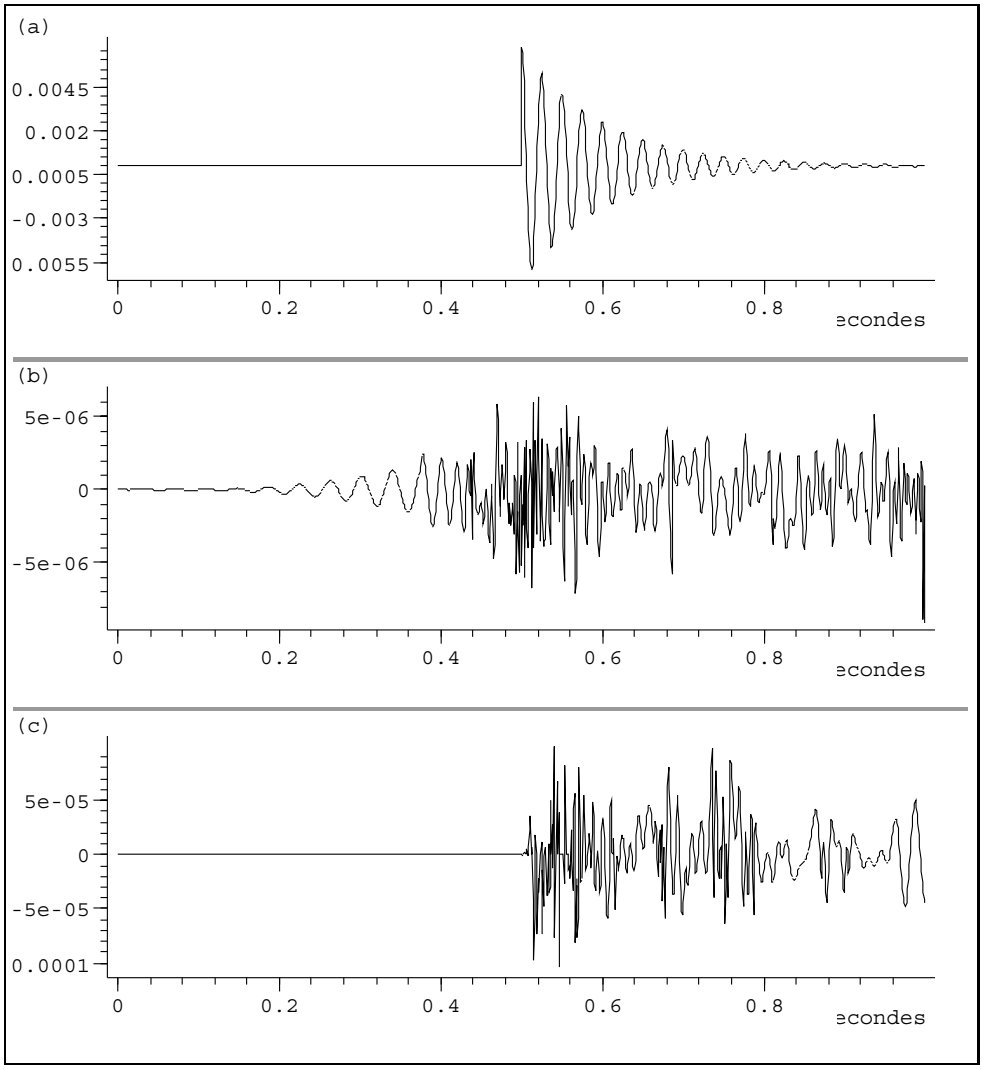

FIG. 6.3 – Phénomène de pré-écho pour une "attaque" synthétique. (a)-signal analysé ; (b)-résidu après 100 itérations de poursuite ordinaire ; (c)-résidu après 100 itérations de poursuite haute-résolution.

6.2 Critère haute résolution

Le manque de résolution est un problème commun à toutes les représentations linéaires du signal, à cause de la limite de résolution de Rayleigh. Le phénomène de pré-écho apparaît dans les techniques d'analyse-synthèse à fenêtre glissante [Moo78] [BCG94]. Il est dû au lissage des transitoires par la fenêtre d'analyse [MB96]. Le Basis Pursuit de Chen et Donoho [CD95] obtient une super-résolution en minimisant un critère $\boldsymbol{l}^1$ (non-linéaire), mais son coût algorithmique est beaucoup trop élevé. Le Matching Pursuit, technique non-linéaire et multi-échelle, n'est pas intrinsèquement limité en résolution. Pour améliorer sa résolution, on peut remplacer le produit scalaire, en tant que fonction de corrélation (2.45) par un autre critère de sélection d'atomes $C(x, g_\gamma)$. L'important est que la convergence de la poursuite soit toujours assurée. C'est ce que l'on a fait pour définir le Matching Pursuit Harmonique au chapitre 3, à partir d'une mesure de corrélation *équivalente* au critère énergétique. C'est également la stratégie employée par McClure et Carin [MC97] pour sélectionner des atomes dans un dictionnaire de forme d'onde définies par la physique du problème traité.

Nous avons défini un critère "haute-résolution" pour effectuer une poursuite sur un dictionnaire temps-fréquence multi-échelle [GBM+96, GDR+96, Gri95]. Parallèlement, Jaggi *et al.* [JCMW95] ont travaillé sur ce critère dans le cadre d'un dictionnaire de splines multi-échelle (non modulées en fréquence).

Nous commençons par introduire la notion de *sous-atome* d'un atome temps-fréquence. On définit alors le critère (ou corrélation) haute résolution. Cette nouvelle fonction de corrélation privilégie l'adaptation *locale* de l'atome g_γ au signal x sur son adaptation globale. Il n'est en effet pas suffisant que x ait beaucoup d'énergie dans la direction de g_γ : encore faut-il que x ne puisse pas être mieux décomposé à partir d'atomes à plus petite échelle que g_γ, judicieusement placés, comme sur l'exemple des deux bosses.

6.2.1 Sous-atomes

Pour chaque atome $g_\gamma = g_{(s,u,\xi)}$, introduisons un ensemble I_γ d'indices de "sous-atomes". L'ensemble I_γ correspond à des atomes $g_{\gamma_i}, \gamma_i \in I_\gamma$ à plus petite échelle que g_γ, dont le support temporel intersecte celui de g_γ, et modulés à la même fréquence

$$g_{\gamma_i} = g_{(s',u_i,\xi)},\ s' \leq s. \tag{6.4}$$

La figure 6.4 représente un atome (en pointillés) et une famille de cinq sous-atomes. Les sous-atomes vont nous permettre de mesurer l'adaptation *locale* de l'atome g_γ au signal x. On dira que g_γ est adapté *localement*, si les produits scalaires $\langle x, g_{\gamma_i} \rangle$ se comportent comme $\langle g_\gamma, g_{\gamma_i} \rangle$.

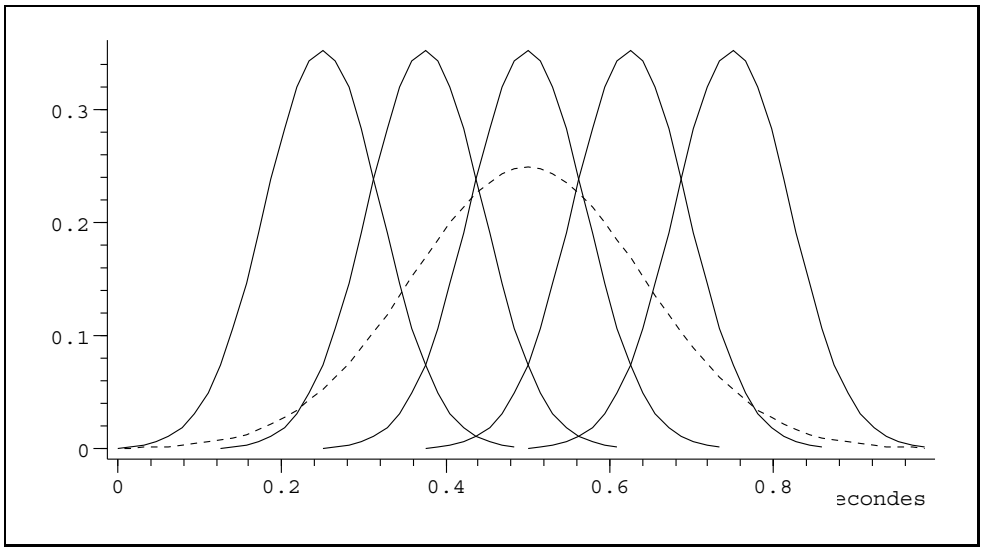

FIG. 6.4 – Un atome (en pointillés) et une famille de cinq sous-atomes à l'échelle dyadique inférieure.

Le choix de la famille de sous-atomes, et en particulier de leur échelle, est important pour définir l'adaptation locale plus précisément. Afin de ne pas introduire une échelle de référence, l'échelle des sous-atomes est *relative* à celle de g_γ. L'expérience nous a montré qu'en prenant les sous-atomes à l'échelle dyadique immédiatement plus petite que celle de g_γ,

$$s' = s/2 \tag{6.5}$$

on obtenait de bons résultats.

6.2.2 Corrélation haute-résolution

Supposons que l'atome g_γ est le premier choisi dans une poursuite. Si $C(x, g_\gamma)$ est le coefficient de corrélation qui lui est affecté, le résidu est alors

$$R^1 x = x - C(x, g_\gamma) g_\gamma.$$

Pour tout sous-atome $\gamma_i \in I_\gamma$, l'"énergie" de $R^1 x$ dans la direction de g_{γ_i} est mesurée par $\langle R^1 x, g_{\gamma_i} \rangle$. Pour qu'il n'y ait pas de "création" d'énergie, cette quantité doit être plus petite que l'énergie $\langle x, g_{\gamma_i} \rangle$ du signal dans la même direction. De plus, la quantité d'énergie "absorbée" $\langle C(x, g_\gamma) g_\gamma, g_{\gamma_i} \rangle$ ne doit pas non plus excéder l'énergie initiale $\langle x, g_{\gamma_i} \rangle$. Ces contraintes se traduisent sur le coefficient $C(x, g_\gamma)$ par

$$|\langle x, g_{\gamma_i} \rangle - C(x, g_\gamma) \langle g_\gamma, g_{\gamma_i} \rangle| \leq |\langle x, g_{\gamma_i} \rangle| \tag{6.6}$$

$$|\langle C(x, g_\gamma) g_\gamma, g_{\gamma_i} \rangle| \leq |\langle x, g_{\gamma_i} \rangle| . \tag{6.7}$$

Afin d'imposer une bonne résolution temporelle[1], on impose ces contraintes pour tous les sous-atomes de g_γ.

A chaque itération de poursuite, on cherche donc à ôter le plus d'énergie possible au résidu $R^{m-1}x$, tout en respectant les contraintes (6.6) et (6.7). La fonction de corrélation $C(x, g_\gamma)$ qui remplit ce rôle prend la forme suivante

$$C(x, g_\gamma) = \varepsilon \min_{\gamma_i \in I_\gamma} \left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| \tag{6.8}$$

avec

$$\varepsilon = \begin{cases} -1 \text{ si } \langle x, g_{\gamma_i} \rangle / \langle g_\gamma, g_{\gamma_i} \rangle \text{ est négatif pour tout } i \\ +1 \text{ si } \langle x, g_{\gamma_i} \rangle / \langle g_\gamma, g_{\gamma_i} \rangle \text{ est positif pour tout } i \\ 0 \text{ si } \langle x, g_{\gamma_i} \rangle / \langle g_\gamma, g_{\gamma_i} \rangle \text{ n'est pas de signe constant.} \end{cases} \tag{6.9}$$

Démonstration

- Commençons par traiter le cas des sous-atomes tels que $\langle g_\gamma, g_{\gamma_i} \rangle = 0$. Pour ceux-là, les contraintes n'imposent rien au coefficient $C(x, g_\gamma)$.
- Pour les autres sous-atomes, les contraintes prennent la forme

$$\begin{aligned} \left| C(x, g_\gamma) - \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| &\leq \left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| \\ |C(x, g_\gamma)| &\leq \left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right|. \end{aligned}$$

  Comme on travaille dans le cadre de signaux et d'atomes à valeurs réelles, on peut considérer le signe ϵ_i de $\langle x, g_{\gamma_i} \rangle / \langle g_\gamma, g_{\gamma_i} \rangle$. La première contrainte est donc

$$\left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| \geq \left| C(x, g_\gamma) - \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| = \left| \epsilon_i C(x, g_\gamma) - \left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right| \right|$$

  si bien que

$$0 \leq \epsilon_i C(x, g_\gamma) \leq \left| \frac{\langle x, g_{\gamma_i} \rangle}{\langle g_\gamma, g_{\gamma_i} \rangle} \right|$$

  pour tout i. La conclusion est alors immédiate.

6.2.3 Matching Pursuit Haute Résolution

On définit alors simplement la poursuite haute résolution comme une poursuite où le choix du meilleur atome est effectué, à chaque itération, selon le critère haute résolution (6.8) au lieu du produit scalaire. La mise à jour du résidu

$$R^m x = R^{m-1}x - C(R^{m-1}x, g_{\gamma_m}) g_{\gamma_m} \tag{6.10}$$

à chaque itération n'est alors plus une projection orthogonale dans la direction de l'atome sélectionné. L'énergie n'est donc plus conservée.

[1] Si l'on voulait améliorer la résolution fréquentielle, il suffirait de travailler avec une famille I_γ d'atomes adaptés, par exemple $g_{\gamma_i} = g_{(2s, u, \xi_i)}$.

6.2.4 Convergence

Si g_γ fait partie de l'ensemble I_γ de ses sous-atomes on est sûr, d'après la définition du critère haute résolution, qu'à chaque itération $C(R^{m-1}x, g_{\gamma_m})$ est du même signe que $\langle R^{m-1}x, g_{\gamma_m}\rangle$ et que

$$\left|C(R^{m-1}x, g_{\gamma_m})\right| \leq \left|\langle R^{m-1}x, g_{\gamma_m}\rangle\right|.$$

L'énergie du résidu décroît donc à chaque itération d'un facteur

$$\begin{aligned} \left\|R^{m-1}x\right\|^2 - \left\|R^m x\right\|^2 &= 2C(R^{m-1}x, g_{\gamma_m})\langle R^{m-1}x, g_{\gamma_m}\rangle - C(R^{m-1}x, g_{\gamma_m})^2 \\ &\in [0, \left|\langle R^{m-1}x, g_{\gamma_m}\rangle\right|^2] \end{aligned} \tag{6.11}$$

En dimension finie, on n'associe pas de sous-atome aux diracs $\delta[n]$ car ils sont déjà à la résolution temporelle la plus fine. Leur corrélation "haute-résolution" avec un signal est donc simplement leur produit scalaire. Par conséquent on a à chaque itération

$$\sup_{\gamma\in\Gamma} C(R^{m-1}x, g_\gamma)^2 \geq \sup_n \left|\langle R^{m-1}x, \delta[n]\rangle\right|^2 \geq \left\|R^{m-1}x\right\|^2 / N, \tag{6.12}$$

ce qui assure la convergence de la poursuite à une vitesse exponentielle $(1 - 1/\sqrt{N})^M$. De fait, à partir d'une certain nombre d'itérations, les atomes sélectionnés sont essentiellement des diracs, et la vitesse de la convergence est beaucoup plus lente que pour la poursuite usuelle, puisqu'à chaque itération on ôte moins d'énergie (6.11) au résidu. On peut comparer sur la figure 6.5 les vitesses de décroissance de l'énergie du résidu, en décibels, pour chacune des deux poursuites. Le signal analysé est l'attaque synthétique de la figure 6.2-(a), et l'on a effectué 200 itérations.

6.3 Résultats

6.3.1 Résolution temporelle améliorée

Dans la poursuite usuelle, le produit scalaire, utilisé comme fonction de corrélation, ne tenait pas compte de la présence ou de l'absence d'énergie dans le signal sur le support temps-fréquence de l'atome g_{γ_m} sélectionné. Au contraire, la nouvelle fonction de corrélation évite de "créer" de l'énergie à des instants où le signal n'en a pas. Elle permet ainsi de distinguer des structures temporelles proches, telles que les deux bosses de la figure 6.1-(a). On observe ainsi sur la figure 6.1-(d) la décomposition atomique fournie par la poursuite haute-résolution. Elle est composée exactement des atomes que l'on a employés pour définir le signal, dont la figure 6.1-(b) est la représentation. Le Matching Pursuit Haute Résolution a donc mieux extrait l'information présente dans le signal que la poursuite usuelle.

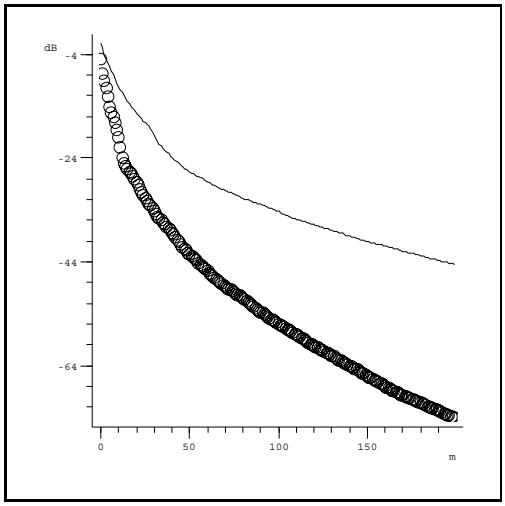

FIG. 6.5 – Décroissance, en décibels, de l'énergie du résidu $\left\|R^{m-1}x\right\|$ en fonction du nombre m d'itérations. En gras, avec la poursuite usuelle. En traits simples, avec la poursuite "haute-résolution".

A cause de la nouvelle fonction de corrélation, les atomes sélectionnés pour la décomposition ont un support temporel plus court qu'avec la poursuite usuelle. En vertu de l'inégalité de Heisenberg, ils ont donc un support fréquentiel plus large. La résolution fréquentielle du Matching Pursuit est donc diminuée par le critère haute-résolution, mais sa résolution temporelle est meilleure que celle de la poursuite usuelle.

6.3.2 Élimination du pré-écho

La poursuite haute résolution évite également l'effet de pré-écho. Ainsi, sur la figure 6.2-(c) on observe que le Matching Pursuit Haute-résolution n'introduit pas d'atomes dont le support temporel dépasse l'instant de l'attaque, contrairement au Matching Pursuit usuel. Cela se traduit sur la figure 6.3-(c) par l'absence de "création" d'énergie dans le résidu avant cet instant d'attaque. Le signal reconstruit, même avec un nombre limité d'atomes, ne présente donc pas de pré-écho. Cela est très important pour l'analyse-synthèse de signaux sonores, car l'oreille est très sensible aux transitoires. La seule contrepartie est une diminution de la vitesse de convergence : il faut donc plus d'atomes temps-fréquence pour atteindre une même qualité *métrique* d'approximation. On peut l'observer en comparant les échelles des figures 6.3-(b) et 6.3-(c) : le résidu de la poursuite usuelle est d'amplitude plus faible après 100 itérations que celui de la poursuite haute-résolution. Comme l'amélioration de la qualité perceptive compense cela en partie, la poursuite haute-résolution est sans doute à même de fournir une représentation du signal dans laquelle est présente l'information caractérisant les attaques, et plus généralement les transitoires.

Deuxième partie

Classification active de signaux

Chapitre 7

Sélection de caractéristiques

Les enjeux économiques et technologiques de la classification de signaux sont importants. Si la reconnaissance automatique de la parole continue, grande réussite de ces dernières années, a aujourd'hui rejoint la reconnaissance optique de caractère [AG97, AGW97] au stade de la commercialisation d'outils logiciels, la reconnaissance de locuteurs et l'identification de visages n'en sont pas encore là. Leurs applications potentielles, pour définir des *signatures* vocales ou visuelles, en font un enjeu dans le domaine de la sécurité bancaire ou de la preuve juridique. Dans le cadre de l'analyse de signaux musicaux, la reconnaissance automatique d'instruments de musique est encore un champ de recherches assez ouvert. Elle est nécessaire pour l'extraction automatique de partitions.

La classification de signaux consiste à associer à un *signal* x (une image, un son, *etc.*) une *classe* y (le nom du caractère, le mot prononcé, l'identité du locuteur, *etc*). Cette classe représente l'"identité" qu'un être humain, le *superviseur*, lui attribuerait naturellement. Dans un cadre probabiliste, x et y sont les réalisations respectives d'un processus X (à valeur dans $\mathbb{R}^N$, où N est très grand, voire infini) et d'une variable aléatoire Y, tirés avec une loi jointe

$$\mathcal{P}(X = x, Y = y) = \mathcal{P}(X = x|Y = y)\mathcal{P}(Y = y). \tag{7.1}$$

Le problème est donc de construire un estimateur $\widehat{Y}(X)$, appelé classificateur, qui minimise la probabilité d'erreur de classification

$$\mathcal{P}_e(\widehat{Y}) = \mathcal{P}\left(\widehat{Y}(X) \neq Y\right). \tag{7.2}$$

L'estimateur optimal $\widehat{Y}_{opt}$, qui est meilleur que tout autre estimateur $\widehat{Y}'$

$$\mathcal{P}_e(\widehat{Y}') \geq \mathcal{P}_e(\widehat{Y}). \tag{7.3}$$

est l'estimateur *bayesien*

$$\widehat{Y_B}(x) \stackrel{\Delta}{=} \arg\max_y \mathcal{P}(Y = y|X = x). \tag{7.4}$$

Pour construire le classificateur bayesien, on doit *estimer* la loi jointe $\mathcal{P}(x, y)$ à partir d'un ensemble d'échantillons

$$\mathcal{L} = \{(x_l, y_l),\ 1 \leq l \leq L\} \tag{7.5}$$

tirés selon cette loi. Cette phase d'*apprentissage* est particulièrement délicate en grande dimension, car l'estimation d'une densité de probabilité est un problème mal posé bien connu [Sco92], d'autant plus difficile que la dimension de la variable (x, y) est grande[1]. Cependant si l'on peut choisir une famille de $M \ll N$ *caractéristiques*

$$Q_1(x), \ldots, Q_M(x) \tag{7.6}$$

qui contient suffisamment d'information sur la classe Y, alors en utilisant les outils statistiques de classification [BFOS84] [CBB+97] on peut construire un classificateur

$$\widehat{Y}\left(Q_1(X), \ldots, Q_M(X)\right). \tag{7.7}$$

L'*apprentissage* sera facilité grâce à la réduction de dimension ainsi opérée. Le problème est alors déplacé vers un nouvel enjeu : la *sélection automatique de caractéristiques* dans des signaux de grande dimension.

L'objet de ce chapitre est de faire le point sur les stratégies aujourd'hui employées dans ce domaine. Par analogie avec les notions d'approximations linéaires et non-linéaires, on distinguera en particulier la sélection *passive* de la sélection *active* des caractéristiques $Q_1, \ldots, Q_M$ à observer.

7.1 Critère de sélection de caractéristiques

Pour sélectionner M caractéristiques dans un *dictionnaire*

$$\mathcal{Q} = \left\{Q_\gamma\ : x \in \mathbb{R}^N \mapsto Q_\gamma(x) \in \mathbb{R},\ \gamma \in \Gamma\right\} \tag{7.8}$$

de caractéristiques, on doit faire appel à un critère mesurant les qualités de la famille $Q_1, \ldots, Q_M$ en vue de la construction d'un classificateur (7.7).

7.1.1 Énergie

Si $(g_m)_{m=1}^M$ est une famille de $\mathbb{R}^N$ qui forme une base d'un sous-espace $\boldsymbol{V}_M$, et si $(\widetilde{g_m})$ est sa base duale[2], alors les caractéristiques

$$Q_m(x) = \langle x, g_m \rangle \tag{7.9}$$

[1] Lorsque l'échantillon d'apprentissage $\mathcal{L}$ est petit, l'estimation risque de trop s'adapter à celui-ci. Les performances de classification, apparemment bonnes, se dégraderont alors si on les mesure sur d'autre signaux tirés selon la même loi. Les théories statistiques de l'apprentissage, telles que le principe de "Minimum Description Length" [Ris83] [RY96] la Minimisation Structurelle du Risque [Vap95] [Vap98] ou la Sélection de modèles [BM97], traitent traitent ce genre de problèmes.

[2] Définie, rappelons-le, par $\langle g_m, \widetilde{g_n} \rangle = \delta[m - n]$

déterminent les composantes de la projection orthogonale

$$P_{\boldsymbol{V}_M} x = \sum_m \langle x, g_m \rangle \widetilde{g_m} = \sum_m \langle x, \widetilde{g_m} \rangle g_m \tag{7.10}$$

de x sur $\boldsymbol{V}_M$. La maximisation du critère énergétique

$$\mathbb{E}\left\{ \left\| P_{\boldsymbol{V}_M} X \right\|_2^2 \right\} \tag{7.11}$$

correspond alors simplement à minimiser l'erreur quadratique moyenne d'approximation (2.4), comme on l'a rappelé au chapitre 2. La meilleure base selon ce critère est la base de Karhunen-Loève , obtenue par l'Analyse en Composantes Principales . L'exemple suivant montre cependant que ce critère n'est pas adapté à la classification de signaux.

7.1.2 Insuffisance du critère énergétique

Soient y_0 et y_1 deux classes de processus gaussiens de lois respectives $\mathcal{N}(0, K_0)$ et $\mathcal{N}(0, K_1)$, où K_0 et K_1 sont les matrices diagonales

$$\begin{aligned} K_0 &= diag(\sigma_1^2, \ldots, \sigma_{M_0}^2, \sigma_{M+1}^2, \ldots, \sigma_N^2) \\ K_1 &= diag(\sigma_1^2, \ldots, \sigma_{M_0}^2, 0, \ldots, 0) \end{aligned}$$

de valeurs propres

$$\sigma_1^2 \geq \ldots \geq \sigma_N^2.$$

Soit à classifier un signal x issu, avec équiprobabilité, de l'une des deux classes

$$\mathcal{P}(Y = y_i) = 1/2.$$

Sa base de Karhunen-Loève est tout simplement la base canonique. Pour $M < M_0$, les lois marginales de X, sous chacune des deux classes, coïncident sur les M composantes principales les plus énergétiques. La vraisemblance

$$\mathcal{P}\left(P_{\boldsymbol{V}_M} X = P_{\boldsymbol{V}_M} x | Y = y_i \right) \tag{7.12}$$

de l'observation effectuée ne dépend donc pas de l'hypothèse $Y = y_i$, si bien que la relation de Bayes

$$\mathcal{P}\left(Y = y_i | P_{\boldsymbol{V}_M} X = P_{\boldsymbol{V}_M} x \right) = \mathcal{P}\left(P_{\boldsymbol{V}_M} X = P_{\boldsymbol{V}_M} x | Y = y_i \right) \frac{\mathcal{P}\left(Y = y_i\right)}{\mathcal{P}\left(P_{\boldsymbol{V}_M} X = P_{\boldsymbol{V}_M} x \right)} \tag{7.13}$$

ne permet pas de départager les deux classes à l'aide de l'observation de ces composantes. Au contraire, la composante la *moins* énergétique permet de prendre la bonne décision en testant la nullité de la N-ème composante.

7.1.3 Entropie, information mutuelle et entropie relative

Les critères fournis par la théorie de l'information [CT91] sont bien plus satisfaisants. L'entropie

$$H(Z) = \mathbb{E}_{\mathcal{P}(z)} \{-\log \mathcal{P}(Z)\} \tag{7.14}$$

d'une variable aléatoire Z mesure l'incertitude sur le résultat du tirage de cette variable. L'entropie conditionnelle de Y par rapport à X

$$H(Y|X) = \mathbb{E}_{\mathcal{P}(x,y)} \{-\log \mathcal{P}(Y|X)\} \tag{7.15}$$

est liée à la probabilité d'erreur de classification par l'inégalité de Fano

$$H\left(\mathcal{P}_e(\widehat{Y})\right) + \mathcal{P}_e(\widehat{Y}) \log\left(\#\mathcal{Y} - 1\right) \geq H(Y|X) \tag{7.16}$$

où l'on note $H(p) = -p\log p - (1-p)\log(1-p)$ et où Y prend ses valeurs dans l'ensemble $\mathcal{Y}$. Pour que la probabilité d'erreur d'un classificateur (7.7) soit faible, il faut donc que $H(Y|Q_1(X), \ldots, Q_M(X))$ soit faible, c'est-à-dire que l'information mutuelle

$$\begin{aligned} I\left(Q_1(X), \ldots, Q_M(X); Y\right) &= H(Y) - H(Y|Q_1(X), \ldots, Q_M(X)) \\ &= H(Q_1(X), \ldots, Q_M(X)) - H(Q_1(X), \ldots, Q_M(X)|Y) \\ &= \mathbb{E}_{\mathcal{P}(x,y)} \left\{ \log \frac{\mathcal{P}(Q_1(X), \ldots, Q_M(X), Y)}{\mathcal{P}(Q_1(X), \ldots, Q_M(X))\mathcal{P}(Y)} \right\} \end{aligned} \tag{7.17}$$

soit grande.

7.2 Sélection passive de caractéristiques

Nous parlerons de sélection *passive* de caractéristiques lorsque celles-ci sont sélectionnées *indépendamment* du signal x à classifier. La sélection passive d'une base (g_m) où les coordonnées sont informatives n'est alors rien d'autre que de l'Analyse Discriminante Linéaire [Fuk72]. Son efficacité pour la classification dépend alors du critère de sélection utilisé. Nous dressons ici un bref état des lieux des approches prometteuses dans ce domaine.

7.2.1 Analyse en Composantes Indépendantes

L'Analyse en Composantes Indépendantes [Com94], suppose que le processus X s'écrit

$$X = \sum_{m=1}^{N} \eta_m g_m \tag{7.18}$$

où les variables aléatoires η_m sont *indépendantes*. Elle a pour but de retrouver "la"[3] base (g_m) en faisant appel au résultat suivant [CT91]

Théorème 6 *Pour tout M-uplet de variables aléatoires $Z_1, \ldots, Z_M$, on a*

$$H(Z_1, \ldots, Z_M) \leq \sum_m H(Z_m) \tag{7.19}$$

avec égalité si, et seulement si, les M variables sont indépendantes.

Comme les caractéristiques $Q_m(X) = \langle X, \widetilde{g_m} \rangle$ déterminent X, on a

$$H(Q_1(X), \ldots, Q_N(X)) = H(X)$$

indépendamment de la base g_m. Les composantes $Q_1, \ldots Q_N$ sont donc statistiquement indépendantes si, et seulement si, elles minimisent

$$\sum_m H(Q_m(X)). \tag{7.20}$$

Toute sous-famille de M caractéristiques extraites d'une base de composantes indépendante vérifie

$$H(Q_{m_1}(X), \ldots, Q_{m_M}(X)|Y) \leq \sum_{k=1}^{M} H(Q_{m_k}(X)|Y), \tag{7.21}$$

et

$$H(Q_{m_1}(X), \ldots, Q_{m_M}(X)) = \sum_{k=1}^{M} H(Q_{m_k}(X)), \tag{7.22}$$

si bien que l'information mutuelle qu'elle apporte est

$$I(Q_{m_1}(X), \ldots, Q_{m_M}(X); Y) \geq \sum_{k=1}^{M} I(Q_{m_k}(X); Y). \tag{7.23}$$

En choisissant les M coordonnées les plus informatives

$$I(Q_{m_1}(X); Y) \geq \ldots \geq I(Q_{m_M}(X); Y) \tag{7.24}$$

on s'assure une borne inférieure aussi grande que possible dans (7.23), mais on ne garantit en aucun cas la sélection de l'optimum de (7.17).

[3] Si au moins deux composantes η_{m_1} et η_{m_2} sont gaussiennes, alors il n'y a pas unicité de l'écriture sous forme de composantes indépendantes, car la composante de X dans le sous-espace engendré par $\widetilde{g_{m_1}}$ et $\widetilde{g_{m_2}}$ est gaussienne, et cet espace admet une infinité de bases de composantes dé-corrélées donc indépendantes. On peut cependant choisir une base particulière de ce sous-espace à l'aide de l'Analyse en Composantes Principales .

7.2.2 Différence avec l'Analyse en Composantes Principales

Avec l'hypothèse (7.18), et l'indépendance des η_m, l'opérateur de covariance K du processus X s'écrit

$$\langle u, Kv \rangle = \mathbb{E}\{\langle u, X \rangle \langle X, v \rangle\} = \sum_{m=1}^{N} \sigma_m^2 \langle u, g_m \rangle \langle g_m, v \rangle \tag{7.25}$$

où σ_m^2 est la variance de η_m. Comme la base de Karhunen-Loève, qui diagonalise K, est orthogonale, elle ne coïncide pas nécessairement avec la base de composantes indépendantes (g_m) qui n'a, elle, aucune raison d'être orthonormale. L'Analyse en Composantes Indépendantes est donc distincte de l'Analyse en Composantes Principales .

Alors que la base de Karhunen-Loève est déterminée par diagonalisation de K, Bell et Sejnowski [BS95] utilisent des réseaux de neurones [NP94][DO96] pour maximiser l'information et déterminer les composantes indépendantes.

7.2.3 Base orthogonale "la moins statistiquement dépendante"

Toutefois un processus X ne se décompose pas nécessairement en composantes indépendantes (7.18), si bien que "la" base (g_m) n'existe pas forcément. Par contre on peut toujours déterminer une base qui minimise le critère (7.20). Saito [Sai98] propose ainsi de sélectionner la base *orthogonale* "la moins statistiquement dépendante", parmi une bibliothèque de bases orthonormales (paquets d'ondelettes ou cosinus locaux). Comme le critère (7.20) à minimiser est additif, il peut utiliser pour cela l'algorithme rapide de sélection d'une meilleure base [CW92] [Sai94] [SC94] [Wic91] de Coifman et Wickerhauser.

7.2.4 Poursuite passive d'information

Liu et Ling [LL99] proposent de s'inspirer du Matching Pursuit [MZ93] pour sélectionner *séquentiellement* les vecteurs (g_m) dans un dictionnaire $\mathcal{D}$, de façon *gloutonne*. Leur stratégie est précisément la suivante : le premier vecteur maximise

$$g_1 = \arg\max_{g \in \mathcal{D}} I(Q_g(X); Y) \tag{7.26}$$

où $Q_g(x) = \langle x, g \rangle$. On définit alors le processus résidu

$$R^1 X = X - \langle X, g_1 \rangle g_1. \tag{7.27}$$

En supposant que l'on a défini le processus résidu $R^m X$, à l'ordre m, on obtient par induction

$$g_{m+1} = \arg\max_{g \in \mathcal{D}} I\left(Q_{g_1}(X), \ldots, Q_{g_m}(R^{m-1}X), Q_g(R^m X); Y\right) \tag{7.28}$$

$$R^{m+1}X = R^m X - \langle R^m X, g_1 \rangle g_1. \tag{7.29}$$

Ils montrent numériquement la supériorité de leur technique sur l'Analyse en Composantes Principales, en termes de taux d'erreur de classification. Cette stratégie *passive*, ou sans mémoire, ne s'adapte pas au signal x à classifier. En effet, les M vecteurs $g_1, \ldots, g_M$ sont déterminés indépendamment de x, et la classification de x est effectuée à partir des mesures

$$\left\langle R^{m-1}x, g_m \right\rangle, \ 1 \leq m \leq M \tag{7.30}$$

qui sont des fonctions linéaires de x. Il s'agit donc bien ici d'une forme d'analyse discriminante linéaire [Fuk72].

7.3 Sélection active de caractéristiques

On a fait observer au chapitre 2 que les approximations *non-linéaires* étaient plus efficaces que les approximations *linéaires*, en raison de leur capacité à *s'adapter* au signal x à approcher. De façon tout à fait analogue, la sélection *active* de caractéristiques, qui *adapte* les caractéristiques observées en fonction de l'information déjà acquise sur le signal x, est potentiellement plus puissante que les techniques de sélection passive que l'on a présentées à la section précédente. Nous rappelons ci-dessous le principe de la réduction graduelle de l'incertitude [GJ96] [AG97], qui consiste à déterminer *activement* une séquence de caractéristiques à observer pour classifier le plus vite possible une réalisation x d'un processus X.

7.3.1 Choix actif/choix passif

Lorsqu'aucune observation n'a encore été effectuée sur x, toute l'information dont on dispose est constituée des probabilités *a priori*

$$p_y = \mathcal{P}(Y = y) \tag{7.31}$$

et des distributions initiales des caractéristiques sous chaque classe :

$$\mathcal{P}_y[Q](q) = \mathcal{P}(Q(X) = q | Y = y). \tag{7.32}$$

La meilleure *première* caractéristique Q_1 est donc

$$Q_1 = \arg\max_{Q \in \mathcal{Q}} I(Q(X); Y). \tag{7.33}$$

Une fois observée, elle apporte l'information $Q_1(X) = Q_1(x)$. La meilleure *deuxième* caractéristique maximise le critère *actif*

$$\arg\max_{Q\in\mathcal{Q}} I(Q(X); Y|Q_1(X) \;=\; Q_1(x)), \tag{7.34}$$

au lieu du critère *passif*

$$\begin{aligned}\arg\max_{Q\in\mathcal{Q}} I(Q(X), Q_1(X); Y) &= \arg\max_{Q\in\mathcal{Q}} \left(I(Q_1(X); Y) + I(Q(X); Y|Q_1(X))\right) \\ &= \arg\max_{Q\in\mathcal{Q}} I(Q(X); Y|Q_1(X))\end{aligned} \tag{7.35}$$

et peut donc *dépendre* de la réalisation x du processus X que l'on est en train de classifier.

7.3.2 Réduction graduelle de l'incertitude

Pour indiquer clairement que la m-ème caractéristique observée *dépend* de la réalisation x, on note

$$\mathcal{Q} = \{Q_\gamma,\ \gamma \in \Gamma\} \tag{7.36}$$

le dictionnaire de caractéristiques, et $Q_{\gamma_m(x)}$ la m-ème caractéristique, caractérisée par son indice $\gamma_m(x)$ qui dépend de la réalisation x. L'observation de $Q_{\gamma_m(x)}$ mène à la mesure

$$Q_{\gamma_m(x)}(X) = Q_{\gamma_m(x)}(x). \tag{7.37}$$

Conditionnement par rapport à l'information déjà acquise

Lorsque m caractéristiques $Q_{\gamma_k(x)}$ ont été sélectionnées et mesurées, l'information dont on dispose sur x est entièrement contenue dans la suite des mesures

$$Q_{\gamma_k(x)}(x),\ 1 \le k \le m. \tag{7.38}$$

L'ensemble des signaux menant à ces mesures est associé à la réalisation de l'événement

$$B_m(x) = \left\{Q_{\gamma_k(x)}(X) = Q_{\gamma_k(x)}(x), 1 \le k \le m\right\}. \tag{7.39}$$

Les lois *a posteriori* , conditionnées par cet événement, *dépendent* de $B_m(x)$ et donc de x. On note avec un indice m les lois conditionnées par cet événement, selon l'exemple

$$\mathcal{P}_m(Z = z) = \mathcal{P}(Z = z|B_m(x)). \tag{7.40}$$

Ainsi la probabilité *a posteriori* des classes est

$$p_{m,y} = \mathcal{P}_m(Y = y) = \mathcal{P}(Y = y|B_m(x)) \tag{7.41}$$

et la loi *a posteriori* de $Q(X)$ est le mélange

$$\mathcal{P}_m[Q] = \sum_y p_{m,y}\mathcal{P}_{m,y}[Q] \tag{7.42}$$

des lois *a posteriori* de $Q(X)$ sous chaque classe.

Meilleure $m+1$-ème caractéristique

Après avoir sélectionné et observé m caractéristiques, on choisit la $m+1$-ème en maximisant l'information mutuelle *conditionnelle*

$$Q_{\gamma_{m+1}(x)} = \arg\max_{Q\in\mathcal{Q}} I\left(Q(X); Y|B_m(x)\right). \tag{7.43}$$

On itère ensuite le procédé. Le choix actif (7.34) apporte, à chaque itération, plus d'information en moyenne que le choix passif (7.35), car

$$\begin{aligned}
\sup_{Q\in\mathcal{Q}} I\left(Q(X); Y|Q_{\gamma_k(x)}(X)\ 1 \le k \le m\right) &= \sup_{Q\in\mathcal{Q}} \mathbb{E}\{I\left(Q(X); Y|B_m(x)\right)\} \\
&\le \mathbb{E}\left\{\sup_{Q\in\mathcal{Q}} I\left(Q(X); Y|B_m(x)\right)\right\} \\
&\le \mathbb{E}\left\{I\left(Q_{\gamma_{m+1}(x)}(X); Y|B_m(x)\right)\right\}
\end{aligned}$$

Cependant, comme il s'agit d'une stratégie *gloutonne*, aucune optimalité globale n'est à attendre.

7.3.3 Arbres de décision

Les Arbres de Décision [BFOS84] de Breiman *et al.* constituent l'archétype de la classification active. Tous les signaux se voient en effet poser la même première question, disposée à la racine de l'arbre. Au fur et à mesure des réponses aux questions posées, les signaux sont mieux connus si bien qu'on leur pose des questions plus adaptées. On observe donc des caractéristiques qui leur sont plus spécifiques. Un telle approche a en outre l'avantage de mettre en lumière la structure des données (ce qui regroupe/ce qui distingue).

7.3.4 Problèmes d'ordre statistique

Le choix d'une stratégie active de sélection de caractéristiques pose des problèmes d'ordre statistique. En effet le critère qui sert à la sélection doit être *estimé* sur les données d'un ensemble d'apprentissage. Cette estimation

devient délicate lorsque les données sont rares [GMSV98]. C'est le cas par exemple de l'ensemble des signaux qui passent par un noeud donné d'un arbre de décision : dès que la profondeur dans l'arbre devient grande, ce sous-ensemble d'apprentissage n'est rapidement plus statistiquement significatif. L'estimation du critère peut alors être très bruitée, voire très biaisée, et mener à un mauvais choix de caractéristiques. C'est le problème classique de la trop grande adaptation aux données (*overfitting*), que les techniques d'arbres de décision traitent par *élagage* [EMS97].

Par ailleurs, indépendamment de la taille de l'ensemble d'apprentissage, l'entropie d'une variable X continue est difficile à estimer, car elle nécessite l'estimation d'une densité de probabilité [Sco92]. Viola [Vio95] [VSS95] propose pour l'estimer une méthode de noyaux, non paramétrique.

L'estimation est moins difficile pour l'entropie d'une caractéristique $Q(X)$ à valeurs discrètes. Ainsi, l'estimation de l'information mutuelle entre une classe Y et une question binaire $Q(X)$ n'est limitée que par la taille de l'échantillon d'apprentissage.

A partir d'un modèle de signal, défini par quelques paramètres, Geman et Jedynak [GJ96] proposent une solution élégante. Afin de détecter des autoroutes sur des images du satellite SPOT, ils construisent un classificateur *actif* sous la forme d'un arbre de décision avec une infinité de branches. Seul un nombre fini de branches est exploré, le bon choix étant fait *en ligne* en fonction du signal observé. Les difficultés liées à l'estimation de l'information mutuelle sont levées grâce à la modélisation du signal. L'estimation est rendue fiable par le fait qu'elle ne concerne que quelques paramètres du modèle, qui peuvent être estimés à l'avance, de façon *globale*, sur *tout* l'échantillon d'apprentissage.

Dans la section qui suit, on s'intéresse à la détermination *explicite* de la stratégie de sélection séquentielle active de caractéristiques, dans deux situations académiques. On compare en particulier les séquences obtenues avec celles fournies par la sélection passive.

7.4 Poursuite active d'information sur des classes gaussiennes

Dans le cadre de deux classes gaussiennes

$$y : x \sim \mathcal{N}(f_y, K_y), \tag{7.44}$$

on observe une réalisation x dont on veut déterminer la classe y. On sélectionne pour cela des caractéristiques linéaires

$$Q_g(x) = \langle x, g \rangle \tag{7.45}$$

où g est un *atome* d'un *dictionnaire* $\mathcal{D}$ d'atomes (voir chapitre 2). On établit les séquences

$$(g_m(x))_{m=1}^N \tag{7.46}$$

d'atomes optimaux obtenus par réduction graduelle de l'incertitude (c'est-à-dire par poursuite *active* d'information). On les compare à la famille

$$(g_m)_{m=1}^N \tag{7.47}$$

déterminée par la poursuite *passive* de Liu et Ling [LL99] (rappelée au paragraphe 7.2.4).

Dans une première situation, qui consiste à détecter un signal connu dans un bruit gaussien fixé, on établit des liens entre la poursuite active d'information et le Matching Pursuit Orthogonal. La meilleure stratégie active est alors ... la stratégie passive ! On étudie ensuite une situation d'identification de bruit gaussien coloré, où l'on montre qu'il est payant d'être actif.

7.4.1 Mélange de deux gaussiennes de même covariance

Dans le cas de *deux* classes y_0 et y_1 gaussiennes de *même couleur*

$$K_1 = K_0$$

centrées sur deux vecteurs f_0 et f_1, on peut déterminer explicitement la séquence optimale $g_m(x)$. Elle est indépendante de x, comme l'énonce le théorème suivant, démontré en annexe C.4.

Théorème 7 *Soit X un processus, mélange de deux classes gaussiennes de même opérateur de covariance K, centrées en f_0 et f_1. Soit x une réalisation de ce processus. Alors la stratégie* active *de sélection d'une séquence optimale $(g_m(x))$ d'atomes dans un dictionnaire $\mathcal{D}$ coïncide avec le* Matching Pursuit Orthogonal *sur le signal $K^{-1}(f_1 - f_0)$, où l'orthogonalité est relative au produit scalaire $\langle ., . \rangle_K$ induit par l'opérateur de covariance K.*

Les atomes sont donc choisis *indépendamment* de la réalisation x, puisque le Matching Pursuit Orthogonal est effectué sur un vecteur $K^{-1}(f_1 - f_0)$ qui ne dépend pas des observations déjà effectuées sur x. La séquence optimale

$$g_m, 1 \leq m \leq N$$

peut donc être déterminée à l'avance. Dans ce cas, la stratégie *active* coïncide donc avec la stratégie passive de Liu et Ling [LL99], à ceci près que la poursuite effectuée est *orthogonale*.

Commentaires

Supposons que $f_0 = 0$. La classification que l'on est en train d'effectuer est simplement la détection de f_1 dans un bruit gaussien. Si le dictionnaire $\mathcal{D}$ contient le *filtre adapté* [Pap86] au signal "déconvolé"[4] $K^{-1}(f_1 - f_0)$, alors

[4] Si les bruits sont stationnaires, K est en effet un opérateur de convolution

g_1 est le filtre adapté, et les observations suivantes sont alors inutiles car elles n'apportent plus aucune information.

7.4.2 Mélange de deux gaussiennes centrées

On s'intéresse maintenant au cas de *deux* classes associées à deux processus *centrés*, mais dont les opérateurs de covariance sont différents : la séquence optimale $g_m(x) = u_{k_m(x)}$ est alors extraite d'une base u_k connue à l'avance, mais son *ordre* $k_m(x)$ n'est plus pré-calculable. Dans ce cas, la meilleure stratégie active est différente de la stratégie passive. C'est ce que l'on établit ici à l'aide de quelques lemmes techniques. Le lemme suivant (démontré en annexe C.5) relie $(g_m(x))$ aux matrices de covariance K_0 et K_1

Lemme 1 *Pour la classification d'une réalisation x d'un mélange de deux classes gaussiennes centrées, de matrices de covariance K_0 et K_1, la séquence optimale $g_m(x)$ est constituée de vecteurs propres*[5] *de la matrice $K_0^{-1}K_1$*

$$g_m(x) = u_{k_m(x)} \tag{7.48}$$

où

$$K_0^{-1}K_1 u_k = \lambda_k^2 u_k \tag{7.49}$$

Considérons par exemple deux bruits stationnaires. La base de Fourier, qui diagonalise leur matrice de covariance K_i, diagonalise aussi $K_0^{-1}K_1$. L'identification de la "couleur" du bruit se fait alors par observation du spectre de x en des fréquences $\omega_m(x)$ bien choisies. Cependant on ne sait pas dans quel *ordre* ces observations doivent être effectuées.

Dans le cas général la base (u_k) de diagonalisation de $K_0^{-1}K_1$ n'est pas orthonormale. Elle ne l'est que si $K_0^{-1}K_1$ est symétrique, c'est-à-dire (puisque K_0 et K_1 sont symétriques) si K_0 et K_1 *commutent*. L' algorithme de sélection d'une base orthonormale "la moins statistiquement dépendante" proposé par Saito [Sai98] n'est donc en général pas capable de trouver une telle base.

Le lemme que l'on vient d'énoncer a laissé dans le flou le choix de l'*ordre* $k_m(x)$ d'observation des vecteurs : il n'a pas précisé si celui-ci dépendait de x. Le lemme suivant (démontré en annexe C.5) précise en partie ce point.

Lemme 2 *Le vecteur propre $g_{m+1}(x) = u_{k_{m+1}(x)}$ optimal à la $m+1$-ème itération est associé à l'une des deux valeurs propres extrémales de $K_0^{-1}K_1$ encore "disponibles"*

$$\underline{\lambda}_m^2(x) = \min_{k \notin \{k_l(x), 1 \le l \le m\}} \lambda_k^2, \tag{7.50}$$

$$\overline{\lambda}_m^2(x) = \max_{k \notin \{k_l(x), 1 \le l \le m\}} \lambda_k^2. \tag{7.51}$$

[5] La matrice $K_0^{-1}K_1$ est bien diagonalisable, bien qu'en général non symétrique, car elle est semblable à la matrice symétrique $\sqrt{K_0}^{-1} K_1 \sqrt{K_0}^{-1}$.

De plus,

- *si* $\overline{\lambda}_m^2(x) \geq \underline{\lambda}_m^2(x) \geq 1$ *alors*

$$\lambda_{k_{m+1}(x)}^2 = \overline{\lambda}_m^2(x);$$

- *si* $\underline{\lambda}_m^2(x) \leq \overline{\lambda}_m^2(x) \leq 1$ *alors*

$$\lambda_{k_{m+1}(x)}^2 = \underline{\lambda}_m^2(x).$$

- *si* $\overline{\lambda}_m^2(x) \geq 1 \geq \underline{\lambda}_m^2(x)$*, alors le choix entre ces deux valeurs propres extrémales peut dépendre de* $p_{m,0}$ *et* $p_{m,1} = 1 - p_{m,0}$*, car il faut déterminer*

$$\max \left\{ I\left(0, \overline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right), I\left(0, \underline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right) \right\}. \qquad (7.52)$$

Lors de la sélection du premier atome $u_{k_1(x)}$, les valeurs propres extrémales $\underline{\lambda}_0^2$ et $\overline{\lambda}_0^2$ ne dépendent pas de x. Comme les probabilités *a priori* p_0 et $p_1 = 1 - p_0$ n'en dépendent pas non plus, k_1 est indépendant de x. Les valeurs propres extrémales $\underline{\lambda}_1^2$ et $\overline{\lambda}_1^2$ ne dépendent donc pas non plus de x. Cependant cette fois-ci les probabilités *a posteriori* $p_{1,0}$ et $p_{1,1} = 1 - p_{1,0}$ en dépendent, car la première observation a apporté de l'information sur la classe y.

Le choix de $k_2(x)$ peut alors effectivement dépendre de x. En dépend-il vraiment ? Le lemme 2 montre que, dans le cas où toutes les valeurs propres sont plus grandes (respectivement plus petites) que 1, le choix de k_2 est, en fait, indépendant de x. Cependant ce lemme ne règle pas ce qui se passe si on trouve à la fois des valeurs propres plus petites que 1 et d'autres plus grandes que 1.

De façon générale l'ordre $k_m(x)$ est actif. On va le montrer sur un exemple précis, à l'aide du lemme suivant, démontré en annexe C.5.

Lemme 3 *Si* $\underline{\lambda}_m^2(x) = 1/\overline{\lambda}_m^2(x)$ *et* $\overline{\lambda}_m^2(x) \in]1, 3/2[$*, alors pour* p *suffisamment proche de* 1

$$I(0, \overline{\lambda}_m^2(x), p, 1-p) > I(0, \underline{\lambda}_m^2(x), p, 1-p) \qquad (7.53)$$

et

$$I(0, \overline{\lambda}_m^2(x), 1-p, p) < I(0, \underline{\lambda}_m^2(x), 1-p, p). \qquad (7.54)$$

Suivant la valeur de la probabilité *a posteriori* $p_{m,0}$, c'est-à-dire suivant la réalisation x observée, la meilleure observation change donc : elle est tantôt associée à la valeur propre $\underline{\lambda}_m^2$, tantôt à $\overline{\lambda}_m^2$. L'ordre $k_m(x)$ des vecteurs à observer n'est donc pas généralement calculable à l'avance : la stratégie active diffère de la stratégie passive.

Exemple

Considérons pour fixer les idées un exemple simple. Les deux classes sont caractérisées par

$$\begin{aligned} y_0 &: x = \eta_0 f_0 + \eta_1 f_1 + w \\ y_1 &: x = \eta_2 f_2 + w \end{aligned}$$

où $\eta_i \sim \mathcal{N}(0, \alpha_i^2)$ et $w \sim \mathcal{N}(0, I)$ sont des variables aléatoires gaussiennes indépendantes, et f_0, f_1 et f_2 des vecteurs unitaires deux à deux orthogonaux. Alors les opérateurs de covariance s'écrivent

$$\begin{aligned} K_0 &= I + \alpha_0^2 P_0 + \alpha_1^2 P_1, \\ K_1 &= I + \alpha_2^2 P_2, \end{aligned}$$

où P_i est le projecteur orthogonal sur $Vect\{f_i\}$. On obtient facilement $K_0^{-1} = \left(I - \frac{\alpha_0^2}{1+\alpha_0^2} P_0 - \frac{\alpha_1^2}{1+\alpha_1^2} P_1 \right)$, puis, comme f_0, f_1 et f_2 sont orthogonaux

$$K_0^{-1} K_1 = I - \frac{\alpha_0^2}{1+\alpha_0^2} P_0 - \frac{\alpha_1^2}{1+\alpha_1^2} P_1 + \alpha_2^2 P_2$$

Par conséquent les vecteurs propres de $K_0^{-1} K_1$ sont

- f_0, associé à la valeur propre $\lambda_0^2 = \dfrac{1}{1+\alpha_0^2}$;
- f_1, associé à la valeur propre $\lambda_1^2 = \dfrac{1}{1+\alpha_1^2}$;
- f_2, associé à la valeur propre $\lambda_2^2 = 1 + \alpha_2^2$;
- Tous les vecteurs u orthogonaux à f_0, f_1 et f_2, associés à la valeur propre $\lambda^2 = 1$.

Supposons par exemple que

$$\alpha_0^2 \geq \alpha_1^2 = \alpha_2^2$$

Si $p_0 = 1/2$, le premier vecteur choisi est, d'après le lemme 2,

$$g_1(x) = f_0.$$

Suite à l'observation de $\langle x, f_0 \rangle$ on établit les probabilités *a posteriori* $p_{1,0}$ et $p_{1,1} = 1 - p_{1,0}$. Le second vecteur observé est soit f_1, soit f_2. On a ici

$$\underline{\lambda}_m^2 = \lambda_1^2 \tag{7.55}$$

$$\overline{\lambda}_m^2 = \lambda_2^2 = 1/\underline{\lambda}_m^2 \tag{7.56}$$

Supposons que $\lambda_2^2 \in]1, 3/2[$, *i.e.* $\alpha_1^2 = \alpha_2^2 < 1/2$. D'après le lemme 3

– si $p_{1,0}$ est suffisamment proche de 1,

$$g_2(x) = f_2,$$

– si $p_{1,0}$ est suffisamment proche de 0,

$$g_2(x) = f_1.$$

Le deuxième vecteur observé *dépend* donc de x. Le troisième vecteur observé est celui, parmi f_1 et f_2, qui n'a pas encore été observé. Ensuite aucune observation n'apporte plus aucune information.

La stratégie séquentielle optimale de reconnaissance est donc bien ici *active*. Dans le cadre de la classification de transitoires, on va préciser au chapitre suivant un *algorithme* actif de classification.

Chapitre 8

Classification de singularités à l'aide d'arbres de décision

De nombreuses études psycho-acoustiques ont établi des liens entre des grandeurs *perceptives* telles que la hauteur, le timbre, à l'aide desquelles un être humain peut déterminer l'identité d'un son, et des caractéristiques telles que la fréquence instantanée, l'enveloppe spectrale, la position des formants, la fréquence fondamentale, *etc.* Pour effectuer une reconnaissance automatique de notes ou d'instruments dans un enregistrement musical, il faut donc mesurer ce type de caractéristiques. On connaît par contre mal aujourd'hui les caractéristiques physiques associées à d'autres grandeurs perceptives, telles que le "mordant" de l'attaque d'un violon. Il joue pourtant un grand rôle dans l'identification de cet instrument par un auditeur humain. Dans un enregistrement musical, les transitoires sont ainsi porteurs de beaucoup d'information [Gre75]. Un expérience classique de psycho-acoustique le montre bien : à partir de deux enregistrements, l'un de flûte et l'autre de violon, on génère deux sons "hybrides", constitués de l'attaque de l'un des instruments suivie de la partie entretenue de l'autre instrument. Les tests d'écoute effectués montrent alors que l'instrument identifié dans un tel son hybride est celui dont on a gardé l'attaque.

Les techniques que l'on a développées aux chapitres 3, 4, 5 et 6, sont bien adaptées, on l'a vu, pour caractériser simultanément les phénomènes transitoires et les parties oscillantes des signaux sonores. Elles permettent en outre de traiter *séparément* ces différentes parties, puisqu'elles décomposent les signaux en structures à différentes échelles. Ce type d'analyse est donc un bon outil pour mieux comprendre l'information présente dans les transitoires des signaux sonores.

L'analyse des relations entre les structures extraites d'un signal par les techniques de poursuite devrait permettre d'extraire l'information présente dans les attaques. Toutefois, en raison du nombre de paramètres (échelle, temps, fréquence, phase, ...) des dictionnaires temps-fréquence, il nous a

semblé nécessaire d'effectuer une première étude sur un dictionnaire plus simple et caractérisant néanmoins bien les transitoires. Une telle approche permet de cerner les difficultés et de construire les outils appropriés dans un paradigme simplifié, avant de s'attaquer au problème dans son ensemble.

La transformée en ondelettes $\langle x, \psi_{(s,u)} \rangle$, avec

$$\psi_{(s,u)}(t) \triangleq \frac{1}{\sqrt{s}} \psi \left(\frac{t-u}{s} \right) \tag{8.1}$$

constitue ce dictionnaire plus simple, et permet de détecter les singularités d'un signal x et de caractériser finement leur régularité. Comme on le rappellera dans ce chapitre, le comportement de la transformée en ondelettes mesure en effet, sous certaines conditions, la *force* des singularités du signal. On définit celle-ci à l'aide de l'exposant de Hölder local h du signal à l'instant v où est située la singularité. Il s'agit du *sup* de l'ensemble des réels $\alpha > 0$ pour lesquels x a la régularité Lipschitz-α en v, c'est-à-dire

$$|x(t) - p_v(t)| \leq K \, |t - v|^{\alpha} \text{ sur un voisinage de } v \tag{8.2}$$

où $p_v(t)$ est un polynôme (de degré $\leq h$). La transformée en ondelettes constitue donc l'outil adapté pour procéder à une caractérisation des transitoires d'un signal.

Comme les signaux auxquels on est confrontés en pratique ne sont pas nécessairement "calés" temporellement, c'est-à-dire que l'on ne connaît pas forcément le temps précis d'arrivée du transitoire, on a également besoin d'un outil invariant par translation. C'est la cas de la transformée en ondelettes continue, et de l'ensemble $\mathcal{E}(x)$ de ses extrema locaux, ou *squelette*.

C'est à partir du squelette $\mathcal{E}(x)$ que l'on développe dans ce chapitre une méthode de classification de signaux basée sur les *relations spatiales* et la force des singularités du signal. On utilise, pour effectuer la classification, des *arbres de décision* T, en faisant appel à un *dictionnaire* $\mathcal{Q}$ *de questions binaires* qui portent sur les relations spatiales et les forces de ces extrema. Il s'agit bien d'une classification *adaptative* : d'une part, les emplacements u_i des maxima locaux de $u \mapsto \left| \langle x, \psi_{(s,u)} \rangle \right|$ *dépendent* du signal x, donc la classification est faite à partir de caractéristiques adaptées au signal analysé ; d'autre part l'utilisation d'arbres de décision est un moyen d'adapter la stratégie de reconnaissance à l'information déjà extraite du signal, au fur et à mesure de la progression vers les feuilles de l'arbre.

8.1 Caractérisation de singularités avec la transformée en ondelettes

Deux caractéristiques de l'ondelette d'analyse ψ permettent de déterminer le comportement de la transformée en ondelettes $\langle x, \psi_{(s,u)} \rangle$ en fonction

de la régularité du signal. Il s'agit de la régularité de l'ondelette et du nombre n de ses moments nuls. La régularité est mesurée par son appartenance à l'ensemble C^k des fonctions k fois continûment dérivables. On dit que l'ondelette a n moments nuls si, et seulement si,

$$\left\langle t^k, \psi_{(s,u)} \right\rangle = 0, \text{ pour tout } 0 \leq k < n, \tag{8.3}$$

ce qui caractérise son orthogonalité avec la famille des polynômes de degré au plus $n-1$. Lorsqu'une ondelette a n moments nuls, elle ne "voit" pas le polynôme p_v qui intervient dans la définition de l'exposant de Hölder h (8.2), dès lors qu'il est de degré au plus $n-1$. On a alors

$$\langle x, \psi_{(s,u)} \rangle = \langle x - p_v, \psi_{(s,u)} \rangle \tag{8.4}$$

et l'exposant de Hölder local peut, comme on va le voir, être mesuré à partir du comportement de $\langle x, \psi_{(s,u)} \rangle$ pour u au voisinage de v [Jaf91] [HT91].

8.1.1 Caractérisation de l'exposant de Hölder local

Le théorème suivant permet de relier le comportement de la transformée en ondelettes aux petites échelles $s \to 0$ avec l'exposant de Hölder local.

Théorème 8 (Jaffard, Holschneider,Tchamitchian) *Soit ψ une ondelette de classe $\boldsymbol{C}^n$ ayant n moments nuls et un support compact. Soit $h \leq n$. Si x est Lipschitz-α en v, alors il existe A et $\eta > 0$ tels que*

$$\left|\langle x, \psi_{(s,u)} \rangle\right| \leq A\sqrt{s}\,(s^\alpha + |u-v|^\alpha) \tag{8.5}$$

pour tout $s > 0$ et tout u tel que $|u-v| \leq \eta$.

Réciproquement soit $\alpha < n$ une valeur non entière. S'il existe $\eta > 0$, $\beta > 0$ et A tels que

$$\left|\langle x, \psi_{(s,u)} \rangle\right| \leq A\sqrt{s}\left(s^\alpha + \frac{|u-v|^\alpha}{|\log_e |u-v||}\right) \tag{8.6}$$

$$\left|\langle x, \psi_{(s,u)} \rangle\right| \leq A s^{\beta+\frac{1}{2}} \tag{8.7}$$

pour tout $s > 0$ et tout u tel que $|u-v| \leq \eta$, alors la fonction $x(t)$ est Lipschitz-α régulière en v.

La condition (8.6) équivaut à ce que x soit uniformément Lipschitz-β.

8.1.2 Extrema locaux de la transformée en ondelettes

Pour mesurer l'exposant h à l'aide du théorème précédent, il n'est pas nécessaire de considérer toute la transformée en ondelettes du signal, mais seulement ses extrema locaux à s fixé, puisque ce sont les points pour lesquels la majoration (8.5) est la plus contraignante. Au chapitre 5, nous avons

défini les *ridges* de la transformée en ondelettes comme les maxima locaux de $s \mapsto \left|\left\langle x, \psi_{(s,u)}\right\rangle\right|$, à u fixé. Nous avons alors vu que l'emplacement $s(u)$ de ces ridges caractérisait la fréquence instantanée $\xi \approx \xi_0/s(u)$. Ici nous nous intéressons aux maxima locaux de

$$u \mapsto \left|\left\langle x, \psi_{(s,u)}\right\rangle\right|. \tag{8.8}$$

Le signal x a une *singularité isolée* en v si son exposant de Hölder y est plus petit que sur tout un voisinage de v. Une telle singularité isolée de x donne lieu à une (ou des) *ligne(s)* d'extrema $(s, u(s))$ se propageant [MH91] [HM89] [YP86] jusqu'aux échelles les plus fines en convergeant vers v. Si toutes ces lignes sont situées dans le *cône d'influence*

$$|u - v| \leq C_\psi s \tag{8.9}$$

de la singularité, où C_ψ est fonction de la taille du support de l'ondelette, c'est que celle-ci n'est pas *oscillante*[1]. D'après le théorème précédent, la valeur de la transformée en ondelettes le long de chaque ligne décroît alors selon la loi

$$\left|\left\langle x, \psi_{(s,u(s))}\right\rangle\right| \leq A' s^{h+1/2}. \tag{8.10}$$

On peut donc estimer l'exposant de Hölder en v en mesurant la pente maximale atteinte par la fonction $s \mapsto \log_2 \left|\left\langle x, \psi_{(s,u(s))}\right\rangle\right|$ le long d'une ligne convergeant vers v.

Le squelette

$$\mathcal{E}(x) \triangleq \left\{ \left((s_i, u_i), \left\langle x, \psi_{(s_i,u_i)}\right\rangle\right), \frac{\partial}{\partial_u} \left\langle x, \psi_{(s,u)}\right\rangle_{|(s,u)=(s_i,u_i)} = 0 \right\} \tag{8.11}$$

de x contient donc presque toute l'information du signal, sauf éventuellement ses singularités oscillantes. En effet l'algorithme de projection itérative de Mallat et Zhong [MZ92] [AO95], ou même une simple descente de gradient [Mal98], permet de "presque" [Mey94] reconstruire le signal à partir de son squelette.

Notation

On appellera *extremum* et l'on notera $e_i = ((s_i, u_i), a_i)$ tout extremum local de la transformée en ondelettes d'un signal, caractérisé par sa position (s_i, u_i) et son coefficient (ou amplitude) $a_i = \left\langle x, \psi_{(s_i,u_i)}\right\rangle$.

8.1.3 Invariance par translation

La représentation d'un signal sous forme de ses extrema remplit bien la condition d'invariance par translation dont on a besoin à cause de l'indétermination du temps d'arrivée d'un signal. Soit en effet τ_u l'opérateur de

[1] Un exemple de singularité oscillante est $\sin 1/x$.

translation, qui agit sur les signaux par

$$\tau_u x(t) \triangleq x(t-u). \tag{8.12}$$

Comme x admet un extremum local $e_i = ((s_i, u_i), \langle x, \psi_{(s_i,u_i)} \rangle)$ si, et seulement si, $\tau_u x$ admet un extremum local translaté de u

$$((s_i, u_i + u), \langle \tau_u x, \psi_{(s_i,u_i+u)} \rangle) = ((s_i, u_i + u), \langle x, \psi_{(s_i,u_i)} \rangle). \tag{8.13}$$

on peut définir l'opérateur de translation sur les extrema $e_i = ((s_i, u_i), a_i)$ par

$$\tau_u((s_i, u_i), a_i) \triangleq ((s_i, u_i + u), a_i). \tag{8.14}$$

D'après (8.13), le squelette est bien invariant par translation, car

$$\mathcal{E}(\tau_u x) = \{\tau_u e_i, e_i \in \mathcal{E}(x)\} \triangleq \tau_u \mathcal{E}(x) \tag{8.15}$$

8.2 Dictionnaire de questions binaires sur les extrema

L'information apportée par les singularités présentes dans un signal peut prendre deux formes : d'une part la *force* de chacune des singularités, d'autre part l'*organisation spatiale* de ces singularités. Toutes ces informations sont présentes dans la représentation $\mathcal{E}(x)$. Pour construire un classificateur à l'aide d'arbres de décision binaires, on va donc définir un *dictionnaire*

$$\mathcal{Q} \subset \{q \ : \mathcal{E}(x) \mapsto q\left(\mathcal{E}(x)\right) \in \{0, 1\}\} \tag{8.16}$$

de questions binaires, qui sont fonctions du signal *via* sa représentation en extrema. Le dictionnaire $\mathcal{Q}$ doit permettre de faire ressortir de la représentation $\mathcal{E}(x)$ du signal l'information qui y est présente. C'est pourquoi on le construit de façon à ce que les questions mesurent précisément les relations spatiales et les forces des singularités du signal.

8.2.1 Forme générale d'une question

Chaque question q a pour but de détecter dans le signal une certaine structure "élémentaire", et se pose donc sous la forme : "Y a-t-il dans le signal une telle structure ?". La présence d'une structure est caractérisée par l'existence de k extrema $e_1, \ldots, e_k \in \mathcal{E}(x)$ vérifiant une certaine relation

$$\mathcal{R}_q(e_1, \ldots, e_k). \tag{8.17}$$

Une question q quelconque prend donc la forme

$$q\left(\mathcal{E}(x)\right) \triangleq \begin{cases} 1 & \text{si } \exists (e_1, \ldots, e_k) \in \left(\mathcal{E}(x)\right)^k, \mathcal{R}_q(e_1, \ldots, e_k) = 1, \\ 0 & \text{sinon.} \end{cases} \tag{8.18}$$

De plus, afin de respecter l'invariance par translation, chaque question doit réagir de façon identique à deux signaux translatés, c'est-à-dire qu'il faut

$$\forall u \in \mathbb{R}\ q\left(\tau_u \mathcal{E}(x)\right) = q\left(\mathcal{E}(x)\right). \tag{8.19}$$

Cela se traduit en contrainte au niveau des relations $\mathcal{R}_q$

$$\forall u \in \mathbb{R}\ \mathcal{R}_q\left(\tau_u e_1, \ldots, \tau_u e_k\right) = \mathcal{R}_q\left(e_1, \ldots, e_k\right). \tag{8.20}$$

On utilisera donc des relations qui ne dépendent pas des *positions absolues* des extrema, mais plutôt de leurs *distances relatives*.

Occurrence(s) d'une question

On appellera *occurrence* d'une question q dans une représentation $\mathcal{E}(x)$ en extrema d'un signal x tout k-uplet $(e_1, \ldots, e_k) \in \left(\mathcal{E}(x)\right)^k$ d'extrema tel que $\mathcal{R}_q(e_1, \ldots, e_k) = 1$.

8.2.2 Relations élémentaires entre paires d'extrema

On veut construire un dictionnaire $\mathcal{Q}$ de questions qui rende compte de la force, des relations spatiales ainsi que de l'amplitude des singularités. On définit pour cela des relations élémentaires entre *paires* d'extrema, qui effectuent les mesures adéquates. On construira ensuite des relations plus complexes en combinant de telles relations élémentaires par conjonction.

Relations de distance

Les relations purement spatiales entre les singularités du signal se mesurent à l'aide de relations de distance entre les extrema. On ne considérera ici que des relations spatiales entre des singularités qui sont simultanément observables à certaines échelles. Elles se traduiront par la présence de deux extrema e_1 et e_2 à l'échelle s tels que

$$d_{min} \leq |u_2 - u_1| \leq d_{max} \tag{8.21}$$

Par ailleurs, pour s'assurer de la pertinence des extrema ainsi détectés, il faut que leurs amplitudes soient suffisamment fortes. Cela est paramétré par un *seuil* θ

$$\forall i,\ \ |a_i| \geq \theta \tag{8.22}$$

Enfin, on peut le cas échéant comparer les *signes* ϵ_1, ϵ_2 des deux singularités observées, à l'aide de la relation $\epsilon \stackrel{\Delta}{=} \epsilon_1\epsilon_2 \in \{-1, 1, \text{" } \pm 1\text{"}\}$ entre eux.

Une relation de distance est donc caractérisée par l'échelle s d'observation, l'intervalle de distance $[d_{min}, d_{max}]$, la relation ϵ et le seuil θ

$$\mathcal{R}[s, d_{min}, d_{max}, \epsilon, \theta]. \tag{8.23}$$

Relation de propagation inter-échelle

La force de chaque singularité se mesure, au contraire des relations spatiales, directement sur l'amplitude des extrema, au moyen de la mesure de la propagation inter-échelle de la singularité. Une relation de propagation inter-échelle, avec un exposant de Hölder compris entre h_{min} et h_{max}, existe donc entre deux extrema $e_1 = ((s_1, u_1), a_1)$ et $e_2 = ((s_2, u_2), a_2)$ si leurs deux échelles sont suffisamment proches et si la propagation s'effectue de façon compatible avec le cône d'influence et avec la loi de propagation (8.10)

$$h_{min} + 1/2 \quad \leq \quad \frac{\log_2(a_1/a_2)}{\log_2(s_1/s_2)} \leq h_{max} + 1/2 \tag{8.24}$$

$$|u_2 - u_1| \quad \leq \quad C_\psi s_2. \tag{8.25}$$

Lorsque l'on effectue le calcul *numérique* de la transformée en ondelettes d'un signal x, on discrétise bien sûr l'échelle et le temps. En particulier on n'observe cette transformée en ondelettes que pour $s = a^j$ (pour la transformée en ondelettes *dyadique*, $a = 2$). La proximité entre les échelles s_1 et s_2 se traduira donc simplement par le fait que ces échelles sont *consécutives*, *i.e.* $s_1 = a^{j_1}$, $s_2 = a^{j_2}$, $|j_1 - j_2| = 1$. Une relation de propagation inter-échelle

$$\mathcal{R}[s, h_{min}, h_{max}, C_\psi] \tag{8.26}$$

est donc caractérisée par la plus grande des deux échelles $s = \max(s_1, s_2)$, l'intervalle $[h_{min}, h_{max}]$, et la taille relative C_ψ du cône d'influence.

8.2.3 Dictionnaire de questions élémentaires

On définit un *dictionnaire de questions élémentaires* $\mathcal{Q}_{elem}$: c'est une partie de l'ensemble des questions associées à une relation de distance (8.23) sur des *paires* d'extrema

$$\mathcal{Q}_{elem} \stackrel{\Delta}{=} \{q \mid \mathcal{R}_q = \mathcal{R}[s, d_{min}, d_{max}, \epsilon, \theta], (s, d_{min}, d_{max}, \epsilon, \theta) \in \Gamma\}. \tag{8.27}$$

Il est caractérisé par l'échantillonnage Γ des paramètres des relations de distance. En pratique, les échelles

$$s = a^j \tag{8.28}$$

possibles pour les extrema d'un signal x sont déterminées par la transformée en ondelettes employée. On fera la plupart du temps appel à la transformée en ondelettes *dyadique*, *i.e.* avec $a = 2$. Par ailleurs, lorsque l'échelle s est fixée, il n'est pas utile d'avoir recours à une précision extrême sur les relations de distance $[d_{min}, d_{max}]$, car les emplacements u_i des extrema sont sensibles au bruit éventuellement ajouté au signal, et peuvent subir des fluctuations de l'ordre de s. Par conséquent les bornes de l'intervalle de distance peuvent être échantillonnée sur une grille de pas Δ, proportionnel à s

$$d_{min} = k_{min}\Delta \tag{8.29}$$
$$d_{max} = k_{max}\Delta \tag{8.30}$$
$$\Delta \propto s. \tag{8.31}$$

8.2.4 Relations multiples dans un k-uplet d'extrema

S'il fallait considérer toutes les questions portant sur des k-uplets d'extrema, on aurait à construire un gigantesque dictionnaire de questions. De plus le coût algorithmique nécessaire pour poser chaque question à un signal donné serait lui aussi très grand. Pour éviter ces problèmes, en vue de l'utilisation de ces questions pour construire un arbre de décision, on construit les questions sur des k-uplets par *raffinement* de questions sur des $(k-1)$-uplets.

On s'inspire en cela de la construction de dictionnaire proposée par Amit *et al.* pour la reconnaissance de caractères [AGW97, AG97] ou de chiffres prononcés [AM99]. Les questions de ce dictionnaire sont construites à partir de la *conjonction* d'un certain nombre de relations simples (entre paires d'extrema).

Raffinement d'une question

Considérons q une question, associée à une relation $\mathcal{R}_q(e_1, \ldots, e_k)$, portant sur k extrema. Un raffinement de q est une question qui précise la structure de $\mathcal{E}(x)$ déjà mise en lumière par q. Il y a deux façons de le faire en n'utilisant que des conjonctions de relations entre paires d'extrema :

- Raffinement interne : on précise la structure des k extrema $(e_1, \ldots, e_k)$, à l'aide d'une relation $\mathcal{R}'(e_i, e_j)$ entre une certaine paire (e_i, e_j) d'extrema
$$\mathcal{R}(e_1, \ldots, e_k) \triangleq \mathcal{R}_q(e_1, \ldots, e_k) \wedge \mathcal{R}'(e_i, e_j). \tag{8.32}$$
- Raffinement externe : on met en jeu un $k+1$-ème extremum e_{k+1}, en lui imposant une relation $\mathcal{R}'(e_i, e_{k+1})$ avec un certain extremum e_i parmi les k extrema de $(e_1, \ldots, e_k)$
$$\mathcal{R}(e_1, \ldots, e_k, e_{k+1}) \triangleq \mathcal{R}_q(e_1, \ldots, e_k) \wedge \mathcal{R}'(e_j, e_{k+1}). \tag{8.33}$$

On notera $q \wedge \mathcal{R}'$ la question obtenue par raffinement de q à l'aide de la relation $\mathcal{R}'$, et $\mathcal{R}_q \wedge \mathcal{R}'$ la relation associée.

Exemples

En pratique, on va employer trois types de raffinement :

1. Propagation inter-échelle d'un extremum : c'est le premier type de raffinement, à l'aide duquel on peut caractériser l'exposant de Hölder des singularités déjà détectées. Il s'agit d'un raffinement *externe*. Soit par exemple q une question de distance portant sur une paire d'extrema (e_1, e_2). Un raffinement q_1 "de propagation" vérifie s'il existe *trois* extrema (e_1, e_2, e_3) dans $\mathcal{E}(x)$ tels que
   - $\mathcal{R}_q(e_1, e_2) = 1$ (représenté schématiquement sur la figure 8.1 par des traits pointillés fins "reliant" e_1 et e_2),
   - l'extremum e_3 est la propagation à l'échelle inférieure de e_1, selon une relation $\mathcal{R}_1$ de propagation inter-échelle du type (8.26) (représentée par un "cône" en traits gras sur la figure 8.1).

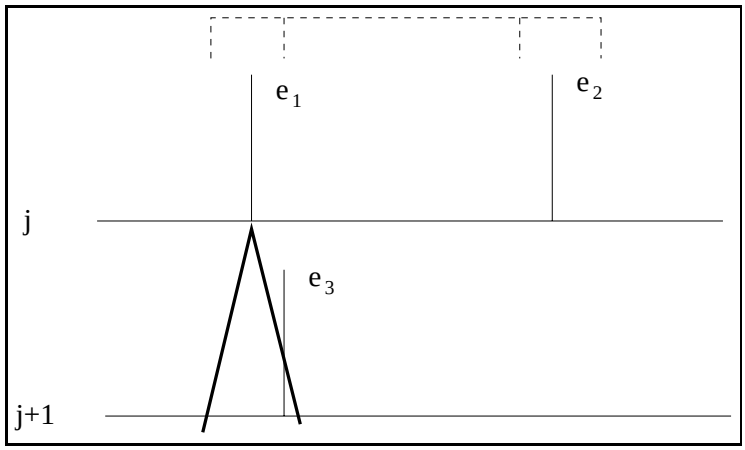

FIG. 8.1 – Exemple de raffinement : propagation inter-échelle.

2. Relation de distance avec un nouvel extremum : c'est maintenant un raffinement *externe*. Si q_2 est un tel raffinement de q_1, alors $q_2(x) = 1$ si, et seulement si, il existe *quatre* extrema (e_1, e_2, e_3, e_4) dans $\mathcal{E}(x)$ tels que
   - $\mathcal{R}_{q_1}(e_1, e_2, e_3) = 1$ (en pointillés fins sur la figure 8.2),
   - les extrema e_1 et e_4 vérifient une certaine relation $\mathcal{R}_2$ de distance du type (8.23) (en trait gras).
3. Relation de distance supplémentaire : il s'agit d'un raffinement *interne*. Si q_3 est un tel raffinement de q_2, alors $q_1(x) = 1$ si, et seulement si, il existe quatre extrema (e_1, e_2, e_3, e_4) dans $\mathcal{E}(x)$ tels que
   - $\mathcal{R}_q(e_1, e_2, e_3) = 1$ (en traits pointillés sur la figure 8.3),
   - les extrema e_2 ete_4 vérifient une certaine relation $\mathcal{R}_3$ de distance du type (8.23) (en traits gras).

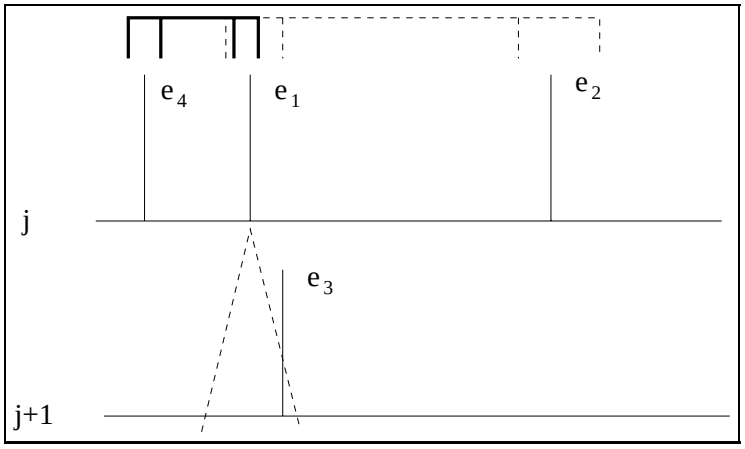

FIG. 8.2 – Exemple de raffinement : relation de distance avec un nouvel extremum.

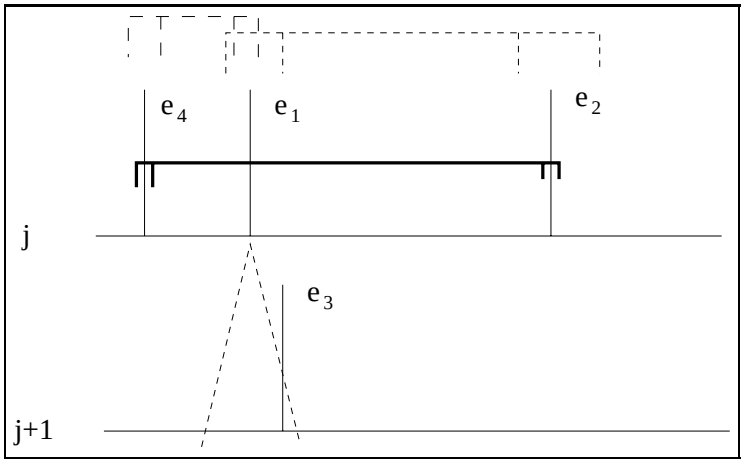

FIG. 8.3 – Exemple de raffinement : relation de distance supplémentaire entre extrema.

8.2.5 Définition du dictionnaire par raffinements successifs

Pour une question q donnée, on définit l'ensemble de ses raffinement à l'*ordre* 1

$$\mathrm{Raff}_1(q) \triangleq \{q \wedge \mathcal{R}, \mathcal{R} \text{ du type (8.23) ou (8.26)}\} \tag{8.34}$$

puis itérativement les raffinements à l'ordre k

$$\mathrm{Raff}_{k+1}(q) \triangleq \cup_{q' \in \mathrm{Raff}_k(q)} \mathrm{Raff}_1(q') = \cup_{q' \in \mathrm{Raff}_1(q)} \mathrm{Raff}_k(q'), \tag{8.35}$$

et enfin l'ensemble de tous les raffinements de q

$$\mathrm{Raff}(q) \triangleq \cup_{k=1}^{\infty} \mathrm{Raff}_k(q). \tag{8.36}$$

Le dictionnaire $\mathcal{Q}$ de questions est alors simplement l'ensemble des raffinements de l'ensemble $\mathcal{Q}_{elem}$ des questions élémentaires.

$$\mathcal{Q} \triangleq \mathcal{Q}_{elem} \cup \cup_{q \in \mathcal{Q}_{elem}} \mathrm{Raff}(q). \tag{8.37}$$

Ordre d'une question

L'ordre d'une question est le nombre de relations entre paires d'extrema nécessaires pour la déterminer.

$$\forall q \in \mathcal{Q}, \mathrm{ordre}(q) \triangleq 1 + \min\{k, q \in \mathrm{Raff}_k(\mathcal{Q}_{elem})\}. \tag{8.38}$$

Les questions élémentaires sont donc toutes d'ordre 1, tandis que leurs raffinements à l'ordre k sont *au plus* d'ordre $k+1$.

8.3 Construction gloutonne d'arbres de décision binaires

Pour classifier des signaux en fonction des singularités qu'ils contiennent, nous avons choisi d'utiliser des arbres binaires de décision, en raison de l'adaptativité qu'ils permettent. Rappelons brièvement le vocabulaire des arbres de décision, et le principe selon lequel est effectuée la classification à l'aide de tels arbres.

8.3.1 Notations et vocabulaire

Un arbre est constitué de *noeuds* et de *branches*. Chaque noeud t de l'arbre T peut être représenté par une suite

$$t = \epsilon_1 \ldots \epsilon_D \tag{8.39}$$

de 0 et de 1 qui déterminent le parcours (0 pour la branche de gauche, 1 pour la branche de droite) qu'il a fallu effectuer dans l'arbre pour y parvenir à partir de la *racine*. Le nombre de signes D est la *profondeur* du noeud t. La racine $\emptyset$ est de profondeur 0.

On munit l'ensemble des noeuds d'un ordre partiel : $t \succeq t'$ si, et seulement si, t' est un descendant de t. A chaque noeud interne t est associée une question binaire q_t. A chaque feuille t est associé une classe y_t[2]. L'arbre T associe une feuille $T(x)$ à un signal x. Elle est déterminée par le *parcours* de x dans l'arbre : depuis la racine de l'arbre jusque vers les feuilles en fonction des réponses $q_t(x)$ aux questions qu'il rencontre aux différents noeuds.

On dira que x passe par le noeud t si, et seulement si, $t \succeq T(x)$. La classe associée par l'arbre au signal est alors

$$y = y_{T(x)}.$$

8.3.2 Principe de la construction gloutonne

Pour construire un arbre de décision pour un problème donné, nous utilisons la construction *gloutonne* de Breiman *et al.* [BFOS84], en utilisant comme critère d'impureté l'entropie. Nous rappelons ici son principe.

On effectue une récursion pour faire *croître* l'arbre : à l'aide d'un échantillon $\mathcal{L}$ de L signaux d'apprentissage dont on connaît les classes

$$\mathcal{L} \stackrel{\Delta}{=} \{(x_1, y_1), \ldots, (x_L, y_L)\} \tag{8.40}$$

et à partir d'un arbre T_0 initialement réduit à sa racine, on construit successivement des arbres T_m de plus en plus étoffés. Pour passer de T_m à T_{m+1}, on commence par choisir une feuille t de T_m. A l'aide de l'ensemble $\mathcal{L}_t$ des échantillons

$$\mathcal{L}_t \stackrel{\Delta}{=} \{(x_l, y_l),\ T(x_l) \preceq t\} \tag{8.41}$$

dont le parcours dans T_m aboutit en t, on sélectionne une "bonne" question q_t pour ce noeud, et l'on transforme la feuille t en lui ajoutant deux fils $t0$ et $t1$.

Critère de choix d'une question

On choisit la question q_t de façon à faire décroître le plus possible l'*impureté* lorsque l'on passe de t à ses deux fils. On mesure cette impureté en un noeud t avec l'entropie $H_t(Y) = H(Y|t \succeq T(X))$ de la variable aléatoire de classe Y, conditionnée par le passage de X par le noeud t. La question est donc

[2] on peut également associer une classe à chaque noeud interne.

choisie de façon à optimiser l'information mutuelle

$$\begin{aligned} I_t(Y;q(X)) &= H_t(Y) \quad -H_t(Y|q(X)) && (8.42)\\ &= H_t(Y) \quad -\mathcal{P}_t\left(q(X)=0\right) H_{t0}(Y) \\ &\qquad\quad -\mathcal{P}_t\left(q(X)=1\right) H_{t1}(Y) && (8.43) \end{aligned}$$

entre sa réponse $q(X)$ et la classe Y, conditionnellement au passage de x par le noeud t. On estime celle-ci à partir de l'échantillon $\mathcal{L}_t$.

Critère d'arrêt

Lorsque l'échantillon $\mathcal{L}_t$ devient trop petit[3], les estimations d'entropie sont si biaisées qu'il vaut mieux cesser de faire croître la branche de l'arbre issue de t. On dit alors que l'on utilise le critère d'arrêt

$$\#\mathcal{L}_t < N_{min}. \tag{8.44}$$

L'usage de ce critère d'arrêt n'est pas anodin. En effet, pour sélectionner l'arbre T "idéal", on voudrait minimiser le critère entropique

$$\arg\min_T H(Y|T(X)),$$

ce qui revient bien sûr à maximiser l'information mutuelle

$$\arg\max_T I(Y;T(X)).$$

Comme l'optimisation globale est *a priori* difficile[4], on emploie la méthode gloutonne, qui est analogue à une descente de gradient du critère entropique. Comme Breiman *et al.* l'ont fait remarquer, l'emploi du critère d'arrêt

$$\sup_q I_t(Y;q(X)) < \varepsilon,$$

fait donc courir le risque de se retrouver piégé dans un minimum local de $T \mapsto H(Y|T(X))$. C'est pourquoi ils prônent l'emploi d'un critère d'arrêt ne faisant pas directement intervenir la grandeur à optimiser. Lorsque l'on ne peut plus gagner suffisamment d'information, il vaut mieux continuer tout de même la construction, car il est possible que la question *suivante* apporte enfin de l'information. Dans le pire des cas, on pourra en effet procéder après coup à un *élagage* [EMS97] de l'arbre (trop grand) que l'on aura construit.

[3] Le cas extrême est lorsqu'il ne reste plus qu'*un* signal dans l'échantillon.

[4] Elle est cependant envisageable avec des méthodes d'optimisation stochastique globale, telles que la construction d'arbres de décision avec l'algorithme de Metropolis proposée par Blanchard [Bla98]

Classe associée à une feuille

La classe y_t associée à une feuille t est celle que l'arbre T fera correspondre à chaque signal x tel que $T(x) = t$. Afin de minimiser le taux d'erreur de classification, on associe donc à chaque feuille sa classe majoritaire

$$y_t \stackrel{\Delta}{=} \arg\max_y \#\left\{l \mid (x_l, y) \in \mathcal{L}_t\right\}. \tag{8.45}$$

8.3.3 Élagage et sélection d'arbres

L'élagage (ou *pruning*) consiste à construire à partir d'un arbre T_{max} une famille de sous-arbres et à sélectionner un "meilleur" sous-arbre, en utilisant un échantillon auxiliaire

$$\mathcal{T} \stackrel{\Delta}{=} \{(x_{L+1}, y_{L+1}), \ldots (x_{L+P}, y_{L+P})\}$$

de P signaux et de leurs classes. Les méthodes d'élagage de ce type ont été étudiées et comparées en détails par Esposito *et al.* [EMS97].

On généralise ce principe en sélectionnant le meilleur arbre $T_{\widehat{\eta}}$ parmi une famille paramétrique T_η d'arbres, qui ne sont pas forcément des sous-arbres d'un même arbre. C'est ce que nous serons amenés à faire un peu plus loin, pour choisir un arbre adapté à la classification et à la détection de signaux bruités.

A partir d'un même échantillon d'apprentissage $\mathcal{L}$, on construit une famille T_η d'arbres avec la méthode gloutonne. Ce qui différencie ces arbres, c'est que le dictionnaire $\mathcal{Q}_\eta$ de questions employé pour construire T_η est l'ensemble des questions dont le paramètre θ de seuil vérifie $\theta \geq \eta$. Le paramètre η est donc un paramètre de seuillage des extrema locaux de x, et permet d'effectuer un débruitage. Plus il est élevé, plus le débruitage est sévère.

Pour chacun de ces arbres, on estime le taux d'erreur de classification sur l'échantillon auxiliaire $\mathcal{T}$, et l'on sélectionne le seuil optimal

$$\widehat{\eta} \stackrel{\Delta}{=} \arg\max_\eta \mathcal{P}\left(Y \neq y_{T_\eta(X)}\right). \tag{8.46}$$

L'arbre $T_{\widehat{\eta}}$ ainsi sélectionné est associé à un débruitage optimal en termes d'erreur de classification.

8.4 Dictionnaires adaptés de questions

La question sélectionnée au noeud t de l'arbre est choisie dans un dictionnaire $\mathcal{Q}_t \subset \mathcal{Q}$. Dans la version standard de la construction gloutonne d'arbres, ce dictionnaire est fixé une fois pour toutes, *i.e.* $\forall t,\ \mathcal{Q}_t = \mathcal{Q}_\emptyset$. Cependant, pour des raisons de complexité, on a intérêt à employer à chaque noeud un dictionnaire aussi petit que possible, mais suffisamment grand pour

contenir des questions pertinentes. Il est donc utile d'adapter le dictionnaire $\mathcal{Q}_t$ au fur et à mesure de la construction de l'arbre. Nous détaillons ci-après comment procéder.

8.4.1 Élimination de questions inutiles

Lors de la construction de l'arbre, la partie utile du dictionnaire s'appauvrit définitivement des questions qui ont déjà été posées, puisque

$$\forall t' \succeq t,\ I_t(Y; q_{t'}) = 0. \tag{8.47}$$

Au noeud t, on peut donc se contenter de chercher les questions dans l'ensemble

$$\mathcal{Q}_t \diagdown \left\{ q_{t'},\ t' \preceq t \right\}. \tag{8.48}$$

Ce phénomène de *masquage* rend définitivement inutiles un certain nombre de questions.

Définition d'un ordre partiel sur les questions

Pour tenir compte de manière efficace de ce phénomène de masquage, on utilise l'existence d'un ordre partiel sur les questions

$$q \leq q' \iff \forall x,\ \big(q(x) = 1 \Rightarrow q'(x) = 1\big) \tag{8.49}$$

$$\iff \forall x,\ \big(q'(x) = 0 \Rightarrow q(x) = 0\big) \tag{8.50}$$

$$\iff \forall x,\ q(x) \leq q'(x). \tag{8.51}$$

auquel est naturellement associé un ordre partiel sur les relations entre extrema : $q \leq q' \Leftrightarrow \mathcal{R}_q \leq \mathcal{R}_{q'}$. Avec cet ordre, $q \leq q'$ correspond simplement au fait que q est plus *fine*, *i.e.* plus *sélective* que q'. Pour tous les signaux passant par des noeuds descendants de $t1$, on sait donc non seulement que $q_t(x) \equiv 1$, mais aussi que $q(x) \equiv 1$ dès que $q_t \leq q$. On dispose d'une propriété analogue dans la branche issue de $t0$. On en déduit la propriété suivante :

Proposition 2 *Pour tout noeud t muni de la question q_t*

$$\forall t' \preceq t1,\quad \forall q \geq q_t,\quad I_{t'}(Y; q(X)) = 0 \tag{8.52}$$

$$\forall t' \preceq t0,\quad \forall q \leq q_t,\quad I_{t'}(Y; q(X)) = 0 \tag{8.53}$$

Par ailleurs, quand on parcours le dictionnaire $\mathcal{Q}_t$ pour déterminer la meilleure question q_t en mesurant $I_t(Y; q(X))$, on peut en profiter pour repérer l'ensemble $\mathcal{Q}_t^0$ (respectivement $\mathcal{Q}_t^1$) des questions q_i telles que $q_i(X) \equiv 0$ (respectivement $q_i(X) \equiv 1$) pour tous les signaux passant par le noeud t

$$\mathcal{Q}_t^\epsilon \triangleq \{q \in \mathcal{Q}_t,\ \forall l,\ T(x_l) \preceq t \Rightarrow q(x) = \epsilon\}. \tag{8.54}$$

On peut donc également masquer ces questions et celles qui sont plus fines (respectivement plus grossières) dans les noeuds issus de t.

Proposition 3 *Pour tout noeud t, $\forall t' \preceq t$,*

$$\forall q \geq q', \quad q' \in \mathcal{Q}_t^1 \quad I_{t'}(Y; q(X)) = 0 \tag{8.55}$$
$$\forall q \leq q', \quad q' \in \mathcal{Q}_t^0 \quad I_{t'}(Y; q(X)) = 0 \tag{8.56}$$

Grâce à cet ordre partiel sur les questions, on est donc capable de masquer bien plus de questions que les seules questions qui ont déjà été posées sur le chemin menant au noeud t. A condition que l'ordre partiel soit lisible immédiatement sur les paramètres qui définissent les questions, on peut donc ôter au fur et à mesure du dictionnaire l'ensemble des questions masquées.

Ordre partiel dans le dictionnaire de raffinements

On peut facilement caractériser *en partie* cet ordre partiel dans le dictionnaire $\mathcal{Q}$. En effet, on peut comparer deux questions élémentaires en comparant leurs paramètres. A défaut de dresser la caractérisation exhaustive de l'ordre partiel dans le domaine des paramètres (ce qui n'est pas difficile, mais présente peu d'intérêt ici), contentons nous de la faire observer sur un exemple :

$$\theta \leq \theta' \Longrightarrow \mathcal{R}[s, d_{min}, d_{max}, \epsilon, \theta] \geq \mathcal{R}[s, d_{min}, d_{max}, \epsilon, \theta'] \tag{8.57}$$

En outre, on peut comparer deux raffinements $q \wedge \mathcal{R}_i, i = 1, 2$ d'une même question dès que l'on peut comparer $\mathcal{R}_1$ et $\mathcal{R}_2$:

$$\mathcal{R}_1 \leq \mathcal{R}_2 \Longrightarrow q \wedge \mathcal{R}_1 \leq q \wedge \mathcal{R}_2 \tag{8.58}$$

Enfin, soit $q' \in \text{Raff}(q)$ un raffinement d'une question q. Comme q' est défini à l'aide de *conjonctions* de la relation $\mathcal{R}_q$ et d'autres relations $\mathcal{R}_i$ entre extrema, on a

$$\forall q' \in \text{Raff}(q), q' \leq q \tag{8.59}$$

8.4.2 Extension adaptée du dictionnaire

Le masquage appauvrit le dictionnaire utile au fur et à mesure que l'on descend dans l'arbre. Par ailleurs, les caractéristiques de l'échantillon $\mathcal{L}_t$ sont de plus en plus homogènes lorsque le noeud t est plus profond dans l'arbre. En effet, tous les signaux de cet échantillon sont regroupés au noeud t parce qu'ils ont présenté les mêmes réponses aux questions posées sur la branche menant à ce noeud. S'il existe dans $\mathcal{L}_t$ des signaux de différentes classes, il faudra donc des questions assez subtiles (d'ordre élevé) pour repérer les structures qui permettront de les distinguer.

Si l'on part d'un dictionnaire initial $\mathcal{Q}_\emptyset$ très riche, contenant déjà toutes les questions subtiles potentiellement nécessaires, un problème de complexité algorithmique se pose manifestement.

Nous optons pour une autre stratégie, inspirée de la technique employée par Amit, Geman et Wilder [AGW97] pour classifier des caractères. On commence avec un dictionnaire $\mathcal{Q}_\emptyset$ assez frustre, composé de questions d'ordre 1, qu'on *étend* ensuite, au fur et à mesure des connaissances acquises sur les signaux, en lui ajoutant de façon judicieuse des questions d'ordre plus grand. On parlera d'extension adaptée du dictionnaire.

Lorsque l'on sait que le signal x aboutit au noeud $t = \epsilon_1 \dots \epsilon_D$, on a acquis sur lui une certaine quantité d'information. Une partie de cette information est lisible dans la branche menant à t. Comme

$$(x_l, y_l) \in \mathcal{L}_{\epsilon_1 \dots \epsilon_D} \Leftrightarrow \forall d \in [\![1, D-1]\!],\ q_{\epsilon_1 \dots \epsilon_d}(x_l) = \epsilon_{d+1} \tag{8.60}$$

elle est contenue dans

- les paramètres des questions $q_{\epsilon_1 \dots \epsilon_d}, d \in [\![1, D-1]\!]$ qui lui ont été posées jusque là ;
- les réponses $q_{\epsilon_1 \dots \epsilon_d}(x) = \epsilon_{d+1}$ à ces questions ;

Mais cette information n'est pas spécifique à x, car elle est commune à tout l'échantillon $\mathcal{L}_t$. On a cependant glané au passage de l'information supplémentaire, spécifique à chaque signal x_l tel que $(x_l, y_l) \in \mathcal{L}_t$, et qui peut nous servir pour le distinguer des autres signaux de l'échantillon. On a en effet pu repérer les *occurrences*[5] des questions pour lesquelles $\epsilon_{d+1} = 1$. On peut construire les raffinements de ces questions : ils sont susceptibles de repérer dans ces occurrences les structures qui permettront de distinguer les signaux de $\mathcal{L}_t$ qui doivent l'être. De plus, comme les occurrences ont déjà été repérées, la réponse à ces raffinements peut être calculée plus rapidement qu'en parcourant tous les k-uplets d'extrema possibles.

C'est seulement lorsque l'on a une nouvelle question dont la réponse est positive qu'il nous faut ajouter ses raffinements au dictionnaire. Afin de ne pas augmenter outre mesure la complexité, on se contentera d'ajouter ses raffinements à l'ordre 1. On va donc étendre *itérativement* le dictionnaire, en lui ajoutant lorsque nécessaire les raffinements de la dernière question posée :

$$\mathcal{Q}_t \mapsto \mathcal{Q}_{t1} \subset \mathcal{Q}_t \cup \mathrm{Raff}_1(q_t) \tag{8.61}$$

De plus il n'est pas nécessaire d'introduire tous les raffinements possibles de q_t, car un sous-ensemble discret des paramètres $(s, [h_{min}, h_{max}], C_\psi)$ d'une part, $(s, [d_{min}, d_{max}])$ est suffisant. Pour ce qui est des raffinements de distance, on procède comme pour la définition du dictionnaire de questions élémentaires (8.27). Quand aux raffinements mesurant la propagation inter-échelle, on se fixe une fois pour toutes la taille relative C_ψ du cône d'influence, et une famille dichotomique d'intervalles utiles

$$[h_{min}, h_{max}] = [k/2^n, (k+1)/2^n]. \tag{8.62}$$

[5] Une *occurrence* est un k-uplet d'extrema de $\mathcal{E}(x)$ vérifiant la relation associée à une question.

8.4.3 Discrétisation du seuil adaptée aux données

L'échantillon $\mathcal{L}_t$ des signaux arrivant au noeud t est *fini*. Cela va nous permettre de discrétiser le seuil θ (utilisé dans les questions de distance) de façon adaptée aux données $(x_l, y_l) \in \mathcal{L}_t$. L'utilisation d'un nombre fini de seuils, aussi petit que possible, permet d'éviter une trop grande complexité.

Soient en effet $s, [d_{min}, d_{max}], \epsilon$ des valeurs fixées, et q_θ la question associée à $\mathcal{R}[s, d_{min}, d_{max}, \epsilon, \theta]$. Grâce à l'ordre partiel (8.57) existant entre les relations de distance, on sait que la fonction

$$\theta \mapsto N(\theta) \stackrel{\Delta}{=} \#\left\{(x_l, y_l) \in \mathcal{L}_t \mid q_\theta(x_l) = 1\right\} \tag{8.63}$$

est décroissante. Comme elle est à valeurs dans $\mathbb{N}$, elle est constante par morceaux, avec un nombre fini de discontinuités

$$0 < \theta_0 < \ldots < \theta_i < \ldots < \theta_I. \tag{8.64}$$

De plus, comme le seuil est défini en (8.22) avec une inégalité large, $N(\theta)$ est continue à gauche. En vertu de l'ordre partiel, si $\theta_i < \theta \leq \theta' \leq \theta_{i+1}$, on a donc

$$\forall (x_l, y_l) \in \mathcal{L}_t,\ q_\theta(x_l) = q_{\theta'}(x_l) \tag{8.65}$$

c'est-à-dire que les questions q_θ et $q_{\theta'}$ sont *indiscernables* sur l'échantillon $\mathcal{L}_t$. Il suffira donc d'*un* représentant de chaque ensemble $\{q_\theta, \theta \in]\theta_i, \theta_{i+1}]\}$ pour être aussi expressif que si l'on disposait de tous les seuils. On pourra par exemple d'utiliser

$$q_{\theta_0}, \cdots, q_{\theta_I}.$$

Il est en effet inutile d'utiliser un représentant de $\theta \in]\theta_I, +\infty[$, car les questions associés à ces paramètres sont indiscernables de la question $q(x) \equiv 0$.

Pour chaque valeur de $s, [d_{min}, d_{max}], \epsilon$, on utilise donc un nombre fini de seuils θ_i.

8.4.4 Algorithme glouton de construction d'arbres avec des dictionnaire adaptés

En utilisant les mécanismes que l'on vient de voir, on peut définir itérativement des dictionnaires adaptés $\mathcal{Q}_t$ de taille finie. L'algorithme glouton de construction d'arbres avec un dictionnaire adaptatif de questions prend donc la forme suivante

1. Le dictionnaire initial $\mathcal{Q}_\emptyset$ est constitué de l'ensemble de questions d'ordre 1, où le seuil est discrétisé selon le mécanisme précédemment décrit

$$\mathcal{Q}_\emptyset \stackrel{\Delta}{=} \mathcal{Q}_{elem} \tag{8.66}$$

2. Traitement du noeud t :
   (a) Parcours des questions $q \in \mathcal{Q}_t$ et détermination de q_t, $\mathcal{Q}_t^0$ et $\mathcal{Q}_t^1$, à l'aide de $\mathcal{L}_t$. On cesse éventuellement la construction si le critère d'arrêt est atteint.
   (b) Partage de $\mathcal{L}_t$ en $\mathcal{L}_{t0}$ et $\mathcal{L}_{t1}$.
   (c) Masquage des questions qui sont inutiles dans les deux fils de t

$$\mathcal{Q}_t^* = \mathcal{Q}_t \diagdown \left\{ q \in \mathcal{Q}_t \mid q \geq q', q' \in \mathcal{Q}_t^1 \text{ ou } q \leq q', q' \in \mathcal{Q}_t^0 \right\} \quad (8.67)$$

   (d) Construction de $\mathcal{Q}_{t1}$ à partir de $\mathcal{Q}_t^*$
   – masquage des questions inutiles

$$\mathcal{Q}_{t1}^{**} \stackrel{\Delta}{=} \mathcal{Q}_t^* \diagdown \{q \geq q_t\} \quad (8.68)$$

   – ajout des raffinements à l'ordre 1 de q_t :

$$\mathcal{Q}_{t1} = \mathcal{Q}_{t1}^{**} \cup \mathrm{Raff}_1(q_t) \quad (8.69)$$

   (e) Construction de $\mathcal{Q}_{t0}$ à partir de $\mathcal{Q}_t$

$$\mathcal{Q}_{t0}^{**} = \mathcal{Q}_t^* \diagdown \{q \leq q_t\} \quad (8.70)$$

8.4.5 Nécessité d'une classe de rejet

Le mécanisme d'adaptation du dictionnaire que nous venons de décrire appauvrit systématiquement le dictionnaire $\mathcal{Q}_t$ le long de la branche "000..." de l'arbre. Les signaux arrivant au noeud "00...0" ont en effet répondu négativement à toutes les questions qui leur ont été posées, qui n'ont donc aucune occurrence, et ne peuvent être raffinées. Il devient donc difficile de classer les signaux de cette branche, faute de question intéressante. Ce phénomène peut se faire sentir dès le premier noeud 0, où le dictionnaire est plus pauvre qu'en 1. Il est gênant car il compromet les possibilités de classification. En outre, il introduit une asymétrie artificielle entre les classes dont les signaux partent majoritairement à gauche, et celles partant majoritairement à droite.

En utilisant l'idée selon laquelle pour classifier des données, il faut non seulement trouver ce qui les *distingue*, mais aussi ce qui les *regroupe*, nous proposons pour remédier à ce problème une solution qui a l'avantage de permettre de simultanément *classifier* et *détecter* les signaux qui nous intéressent.

On introduit une classe supplémentaire, dite de *rejet*, constituée de signaux n'appartenant à aucune des classes de signaux de l'échantillon d'apprentissage $\mathcal{L}$. On peut par exemple utiliser comme classe de rejet une classe de bruits blancs gaussiens, mais il peut suffire de considérer la classe des signaux *nuls*. L'essentiel est que le squelette des signaux de cette classe ait

suffisamment peu de structure pour toujours répondre négativement aux questions qu'on lui pose.

On ajoute alors à l'échantillon d'apprentissage $\mathcal{L}$ un échantillon de signaux de rejet, et l'on construit l'arbre selon la méthode que l'on vient de décrire, mais à partir de cette base de données agrandie. Les signaux qui parcourent la branche "000..." sont alors les signaux de rejet, et tous les signaux intéressants sont à un moment ou à un autre *regroupés* par une question qui les distingue de ces signaux sans structure.

L'arbre ainsi construit permet alors d'effectuer simultanément une classification et une détection. Lorsqu'un signal aboutit à une feuille étiquetée par la classe de rejet, c'est qu'il n'est pas détecté comme appartenant à l'une des classes intéressantes.

8.5 Classification de singularités glissantes

Les techniques d'analyse de signaux à l'aide de dictionnaires redondants ont montré, aux chapitres précédents, une remarquable capacité à caractériser simultanément les phénomènes transitoires et les parties oscillantes. Elles permettent également de traiter séparément ces différents phénomènes. C'est donc, idéalement, à partir de telles représentations que l'on aimerait développer des outils de reconnaissance automatique de transitoires, de classification d'attaques, *etc.* Cependant la taille de ces dictionnaires peut poser des problèmes de complexité. Nous avons préféré dans un premier temps explorer la classification active à partir du "petit" dictionnaire des ondelettes. Nous nous sommes donc "restreints" pour cette exploration à la classification de singularités. Ce cadre est déjà suffisamment riche pour aborder un certain nombre de problèmes réels de reconnaissance automatique. Ainsi on peut observer sur la figure 8.4-(a) le "profil" d'un avion, c'est-à-dire le tracé de son contour en coordonnées polaires $r(2\pi t)$. On peut constater que le nez $(t = 0)$, les bouts des ailes $(t = 0.4$ et $t = 0.6)$ et les divers objets placés sous les ailes $(t = 0.1$ à $t = 0.3)$ donnent lieu à des singularités de ce contour qui sont bien visibles sur la transformée en ondelettes 8.4-(b) et son squelette 8.4-(c). En pratique, l'observation d'un même avion peut donner autant de profils $r(2\pi(t-t_0))$ que d'orientations possibles $\theta = 2\pi t_0$ de l'avion. Le même problème est susceptible de surgir pour la classification d'attaques d'instruments, car l'instant d'arrivée de la note associée n'est pas forcément connu. La classification doit donc être invariante par translation.

Nous avons donc choisi d'évaluer notre stratégie de reconnaissance active sur un exemple où les classes, invariantes par translation, sont constituées de signaux présentant des singularités plus ou moins bruitées. Nous nous sommes attachés à évaluer les performances de notre méthode, son comportement en présence de bruit, ses limites. Nous avons ensuite procédé à des comparaisons avec des outils de classification plus classiques.

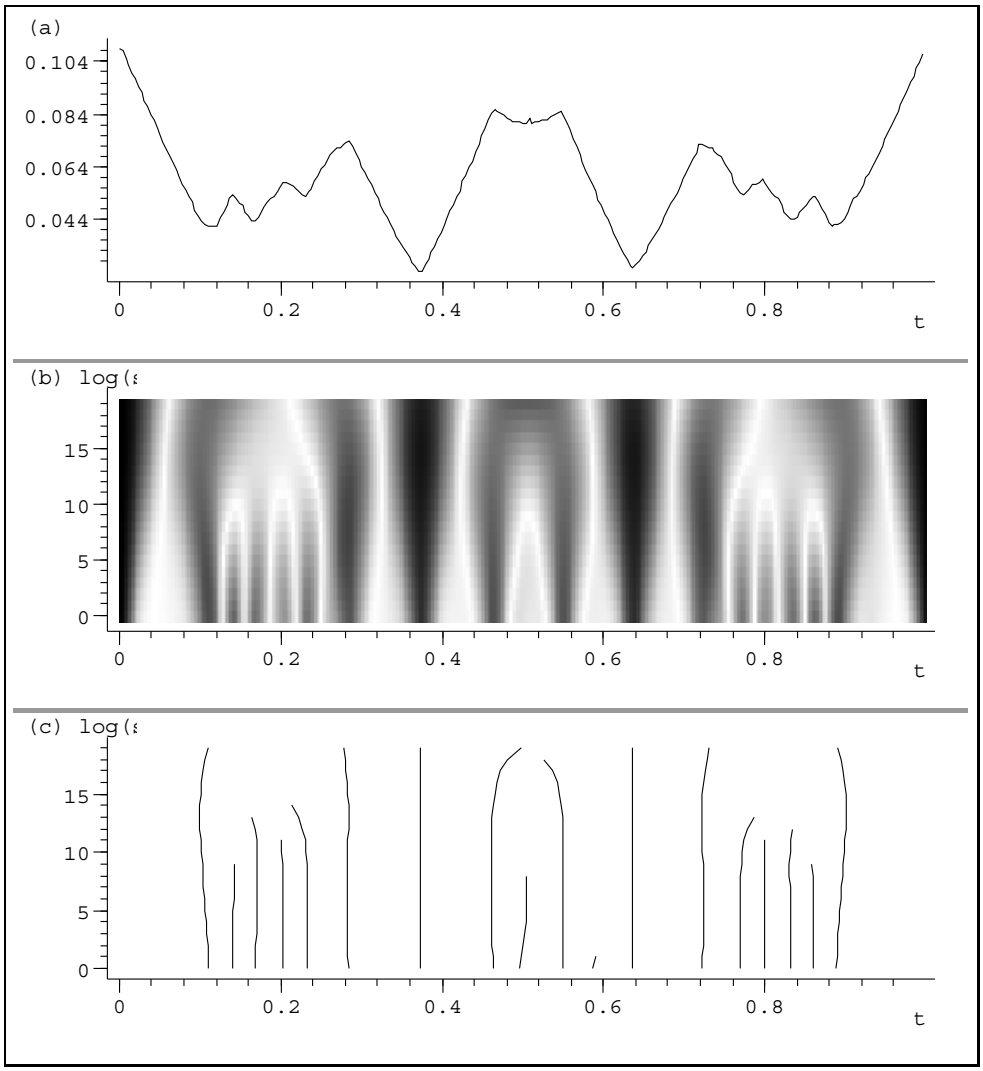

FIG. 8.4 – Le contour d'un avion, en coordonnées polaires (a), sa transformée en ondelettes continue (10 voies par octave) avec une ondelette dérivée seconde de gaussienne (b), et le squelette de celle-ci (c).

8.5.1 Signaux et classes

On considère des signaux constitués d'une *paire de singularités*, de la forme

$$x(t) \stackrel{\Delta}{=} A(t-a)^{\alpha} + B(t-b)^{\beta} + \sigma dW_t \tag{8.71}$$

où dW_t est un brownien. La figure 8.5 représente un signal de ce type (non bruité), ainsi que sa transformée en ondelettes continue (avec 10 voies par octave) et son squelette. La transformée en ondelettes *dyadique* utilisée pour la classification est moins redondante, avec 1 voie par octave. La figure 8.6 représente, elle, l'évolution de l'amplitude de la transformée en ondelettes le long de la ligne d'extrema associée à la singularité de gauche, repérée en gras sur la figure 8.5. On y mesure une pente $\widehat{\alpha} = 0.199$ qui correspond à l'exposant $\alpha = 0.2$ choisi sur cet exemple.

On considère huit classes de tels signaux, caractérisées par les lois des paramètres aléatoires A, B, a, b, α et β. Les forces α et β des deux singularités sont tirées aléatoirement dans deux intervalles

$$\alpha \in [\alpha_{min}, \alpha_{max}] \tag{8.72}$$
$$\beta \in [\beta_{min}, \beta_{max}] \tag{8.73}$$

selon des lois uniformes, et leurs amplitudes A et B sont réparties uniformément entre 1 et 5. Leurs emplacements a et b vérifient

$$b - a \in [d_{min}, d_{max}] \tag{8.74}$$

et sont tirés au hasard de façon à ce que chaque classe soit "invariante par translation"[6]. L'intensité σ^2 du bruit est la même pour toutes les classes : sur les signaux discrets

$$x_d[n] = x(n/N) \tag{8.75}$$

utilisés en pratique, σ^2 est la variance par échantillon. La table 8.1 résume les intervalles nécessaires pour définir les huit classes. La figure 8.7 donne un aperçu de ces classes, pour un niveau de bruit de $\sigma = 0.04$. Elle représente, pour chacune d'elles, une réalisation, sa transformée en ondelettes dyadique et son squelette.

La distribution *a priori* des huit classes est uniforme

$$\mathcal{P}(Y = i) = cste. \tag{8.76}$$

On désire classifier automatiquement les signaux issus de ces classes.

[6] pour éviter les effets de bord dans la transformée en ondelettes, on se limite aux translation gardant les singularités à distance des bords.

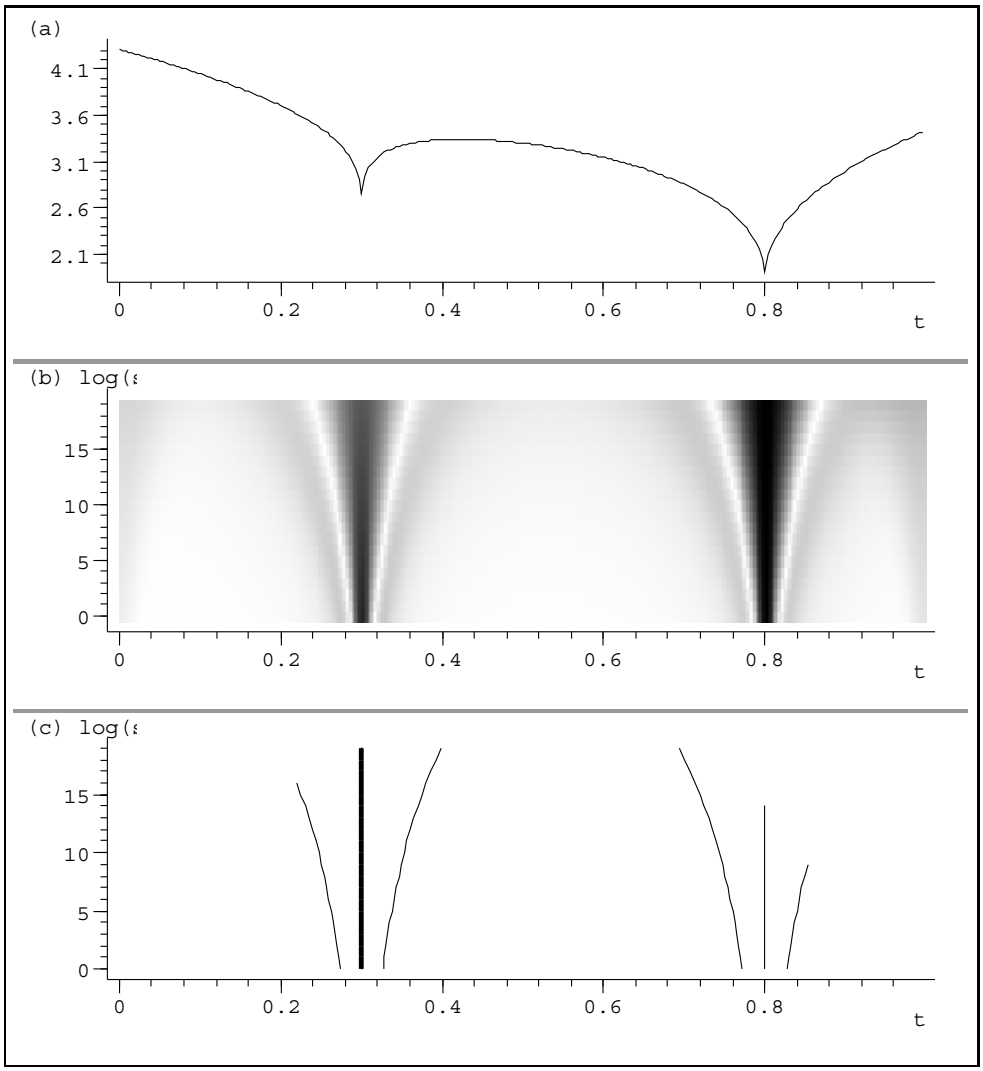

FIG. 8.5 – Un exemple de paire de singularités glissantes (a), sa transformée en ondelettes (10 voies par octave) avec une ondelette dérivée seconde de gaussienne (b) et le squelette de celle-ci (c). La figure 8.6 représente le logarithme de l'amplitude de la transformée en ondelettes le long de la ligne d'extrema représentée en gras.

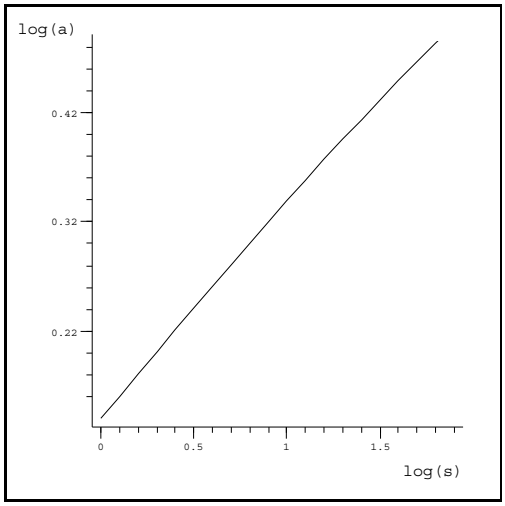

FIG. 8.6 – Évolution du logarithme de la transformée en ondelettes le long de la ligne d'extrema repérée en gras sur la figure 8.5.

Classe	$[d_{min}, d_{max}]$	$[\alpha_{min}, \alpha_{max}]$	$[\beta_{min}, \beta_{max}]$
0	$[0.31, 0.39]$	$[0.35, 0.45]$	$[0.35, 0.45]$
1	$[0.31, 0.39]$	$[0.35, 0.45]$	$[0.55, 0.65]$
2	$[0.31, 0.39]$	$[0.55, 0.65]$	$[0.35, 0.45]$
3	$[0.31, 0.39]$	$[0.55, 0.65]$	$[0.55, 0.65]$
4	$[0.53, 0.61]$	$[0.35, 0.45]$	$[0.35, 0.45]$
5	$[0.53, 0.61]$	$[0.35, 0.45]$	$[0.55, 0.65]$
6	$[0.53, 0.61]$	$[0.55, 0.65]$	$[0.35, 0.45]$
7	$[0.53, 0.61]$	$[0.55, 0.65]$	$[0.55, 0.65]$

TAB. 8.1 – Valeurs possibles des forces et des distances entre les deux singularités, selon la classe.

8.5.2 Arbres de décision avec des extrema

Les extrema de la transformée en ondelettes ne sont pas tous des reflets de singularités des signaux, car du bruit est présent. Beaucoup de ces extrema sont en effet seulement des pics de bruit, comme on peut l'observer sur les squelettes représentés sur la figure 8.7. Pour les éliminer on peut effectuer un débruitage par seuillage des extrema. Le seuillage des coefficients dans une base orthonormale d'ondelettes a été étudié par Donoho et Johnstone [DJ94]. Lorsque N est grand, le meilleur seuil est asymptotiquement

$$\eta_{opt} \approx \sigma\sqrt{2\log N}. \tag{8.77}$$

Cependant le squelette est issu d'une représentation redondante, qui n'est pas une base orthonormale d'ondelettes. Par ailleurs le niveau de bruit σ n'est

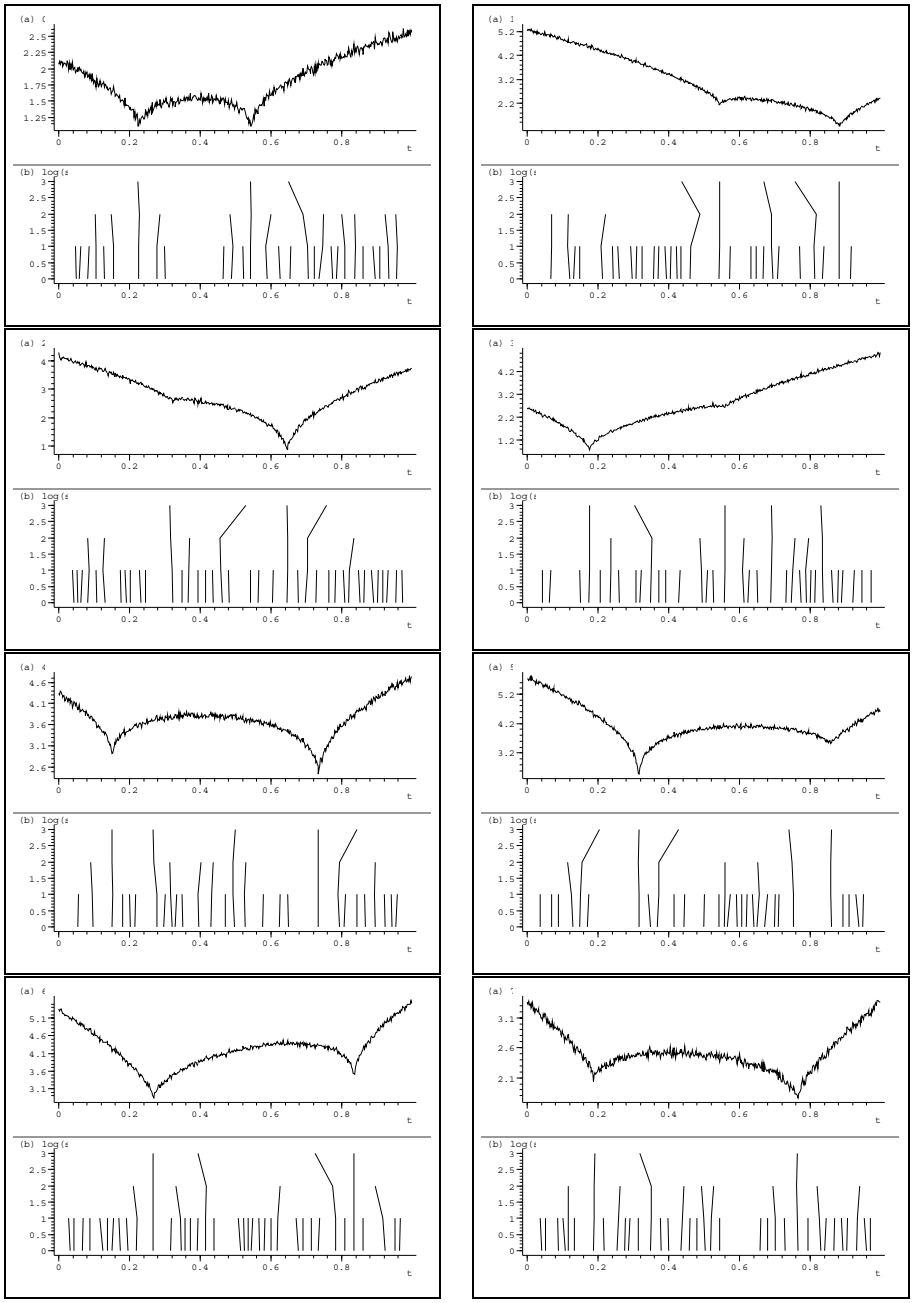

FIG. 8.7 – Exemples de signaux des huit classes considérées. Le niveau de bruit est caractérisé par l'écart type $\sigma = 0.04$ pour chaque échantillon.

a priori pas disponible dans une situation réelle. En effet, pour classifier des attaques d'instruments par exemple, il est difficile de distinguer le bruit porteur d'information du bruit perturbateur. On n'est donc pas dans une situation où l'on peut utiliser sans scrupules le seuil (8.77).

Pour chaque valeur de η, on peut construire un arbre T_η comme on l'a décrit à la section 8.4.4, à partir des squelettes débruités $\mathcal{E}_\eta(x_l)$. Cela revient à peu près[7] à se limiter au dictionnaire $\mathcal{Q}_\eta$ des questions dont le seuil θ vérifie $\theta \geq \eta$. Le seuillage préalable limite le nombre d'extrema présents dans un squelette, et donc la complexité de recherche de la meilleure question.

On sélectionne ensuite le meilleur des arbres de la famille T_η à l'aide d'un ensemble d'échantillons d'élagage, comme expliqué à la section 8.3.3. Cela correspond à déterminer *expérimentalement* le meilleur seuil. Au lieu d'optimiser le seuil au sens du critère énergétique mesurant la dégradation du signal, on l'optimise ici au sens de l'information contenue dans le signal débruité. Si un bruit non gaussien est porteur d'information dans le signal, le seuil choisi sera sans doute plus faible que celui défini par Donoho et Johnstone.

8.5.3 Taux de reconnaissance avant sélection du meilleur seuil

Nous avons appliqué notre méthode pour différents niveaux de bruit. Les arbres sont construits à partir d'une petite base de données contenant 20 échantillons de chaque classe. On utilise pour la classe de rejet des signaux nuls. Le tableau 8.2 montre les taux d'erreur mesurés sur une base de donnée d'élagage de 100 échantillons par classe. Ils correspondent aux arbres T_η construits avec un niveau de bruit $\sigma = 0.005$ fixé. Les erreurs sont réparties en faux négatifs (FN) d'une part et erreurs de classification (FC) d'autre part.

η	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
%FN	4	5.1	3.4	3.4	6.4	9.4	16	16	21	27	44
%FC	27	18	2.3	2.3	2.9	2.5	1.6	1.6	1.4	3.1	3.0
%FN+%FC	31	23.1	5.7	5.7	9.3	11.9	17.6	17.6	22.4	30.1	47

TAB. 8.2 – Taux d'erreur pour la classification et la détection de singularités glissantes. Le niveau de bruit $\sigma = 0.005$ est fixé. On utilise des arbres de décision T_η avec un seuil η variable.

L'arbre sélectionné en minimisant le taux total d'erreur %FN+%FC correspond ici au seuil $\eta_{opt} = 0.10$. Ses taux de faux négatifs et d'erreur de

[7] Comme les questions mesurant la propagation inter-échelle ne sont associées à aucun seuil, le seuillage préalable n'est pas rigoureusement équivalent à l'utilisation du dictionnaire $\mathcal{Q}_\eta$

classification, estimés sur une base de données de test de 100 échantillons de chaque classe, sont

$$\begin{aligned} \%FN &= 3.1 \\ \%FC &= 3.6 \end{aligned}$$

ce qui fait un total de 6.7% d'erreurs.

8.5.4 Performances en fonction du niveau de bruit

Le tableau 8.3 récapitule les seuils optimaux et les taux d'erreur des arbres associés, pour des niveaux de bruit σ allant de 0 à 0.045. Le taux

$\sigma \times 10^{-3}$	0	5	10	15	20	25	30	35	40	45
η_{opt}	0.1	0.1	0.15	0.15	0.1	0.15	0.2	0.2	0.2	0.2
%FN	2.5	3.1	2.9	2.4	1	3.2	3.9	3.4	0.25	1.5
%FC	2.3	3.6	7.6	6.5	13	16	16	16	23	29
%FN+%FC	4.8	6.7	10.5	8.9	14	19.2	19.9	19.4	23.25	30.5

TAB. 8.3 – Évolution du seuil optimal η_{opt} et des taux d'erreur associés, pour différents niveaux de bruit.

d'erreur de classification augmente lorsque le niveau de bruit croît. A partir d'un certain niveau de bruit, les performances se dégradent : il devient en effet difficile de mesurer les exposants de Hölder des singularités, car le niveau σ du bruit est de l'ordre de l'amplitude $As^{\alpha+1/2}$ des extrema à petite échelle qui caractérise ces exposants. La distance entre les singularités permet de distinguer les classes $\{0, 1, 2, 3\}$ des classes $\{4, 5, 6, 7\}$. Par contre seules les forces de leurs singularités distinguent les signaux à l'intérieur de chacun de ces groupes de quatre classes. Ils deviennent donc très difficiles à classifier. Toutefois, si la classification ne tenait compte que de la distance entre les singularités, la classe serait, au mieux, tirée au hasard parmi les quatre classes possibles. Cela mènerait à un taux d'erreur de 75%. Les taux d'erreur beaucoup plus faibles observés avec les niveaux de bruit considérés montrent que la classification tient compte de la force des singularités.

8.5.5 Comparaison avec l'Analyse Discriminante Linéaire

Pour comparer ces résultats avec les performances de méthodes plus classiques de classification, on décompose notre méthode en deux étapes :

1. projection adaptative sur l'espace engendré par les extrema locaux de la transformée en ondelettes ;
2. classification "active" à partir de cette projection, avec des arbres de décision.

Les seuils optimaux η_{opt} déterminés lors de la sélection d'un arbre sont tels que le squelette seuillé $\mathcal{E}_{\eta_{opt}}$ ne contient plus qu'une trentaine d'extrema, alors que la dimension totale du signal est de 512. On compare donc cette méthode avec une Analyse Discriminante Linéaire effectuée sur les 30 composantes principales les plus énergétiques.

Analyse en Composantes Principales

Les composantes principales sont déterminées à partir d'un échantillon de 100 signaux par classe. On estime pour cela la covariance globale K de l'échantillon de 800 signaux. Les 30 composantes principales sélectionnées sont les vecteurs propres de K associés aux 30 plus grandes valeurs propres.

Analyse Discriminante Linéaire

On estime ensuite la covariance K_i et le centre μ_i de chaque classe, à l'aide des mêmes 100 signaux par classe. La classification d'un signal x est effectuée à l'aide de sa distance de Mahalanobis $\left\langle x-\mu_i, K_i^{-1}(x-\mu_i)\right\rangle$ à chacune des classes. On assigne au signal x la classe qui lui est le plus proche au sens de cette distance.

8.5.6 Effet de l'invariance par translation

A cause de l'invariance par translation des classes, les composantes principales sont extraites de la base de Fourier. On peut lire sur les composantes de Fourier la régularité de Lipschitz *uniforme* d'un signal : si x est uniformément Lipschitz-α, alors

$$\widehat{x}(\omega) = \mathcal{O}(1/\omega^{\alpha}). \tag{8.78}$$

Cependant les signaux à classifier sont C^{∞} partout sauf en leurs deux singularités. Comme c'est précisément la force de leurs singularités qui les distingue, leur régularité uniforme n'est pas suffisante pour les caractériser.

La classification par Analyse Discriminante Linéaire sur les 30 premières composantes principales mène à des taux d'erreur de classification de l'ordre de 60%, quel que soit le niveau de bruit. Elle fait donc un peu mieux que les 75% d'erreur d'un classifieur caractérisant uniquement la distance entre singularités. Le tableau 8.4 résume ces résultats.

$\sigma \times 10^{-3}$	0	5	10	15	20	25	30	35	40	45
%FC	59	57	60	53	59	61	60	60	57	62

TAB. 8.4 – Taux d'erreur pour la classification de singularités glissantes par Analyse Discriminante Linéaire sur les 30 premières composantes principales, en fonction du niveau de bruit.

8.5.7 Intérêt de l'adaptativité

Les extrema de la transformée en ondelettes ont comme avantage, par rapport aux composantes principales, de s'adapter automatiquement à la translation aléatoire imposée aux singularités. Pour mesurer les effets de cet avantage, nous avons mené une autre expérience, où les composantes principales sont calculées à partir de signaux "calés" temporellement. En pratique, on change légèrement la définition des classes, en fixant l'instant $a = 0.2$ de la première singularité. Seul l'instant b de la seconde singularité se change aléatoirement de réalisation en réalisation. Les 15 composantes principales les plus énergétiques, calculées à partir de 100 réalisations de chaque classe, sont représentées sur la figure 8.8. La classification est effectuée par Analyse

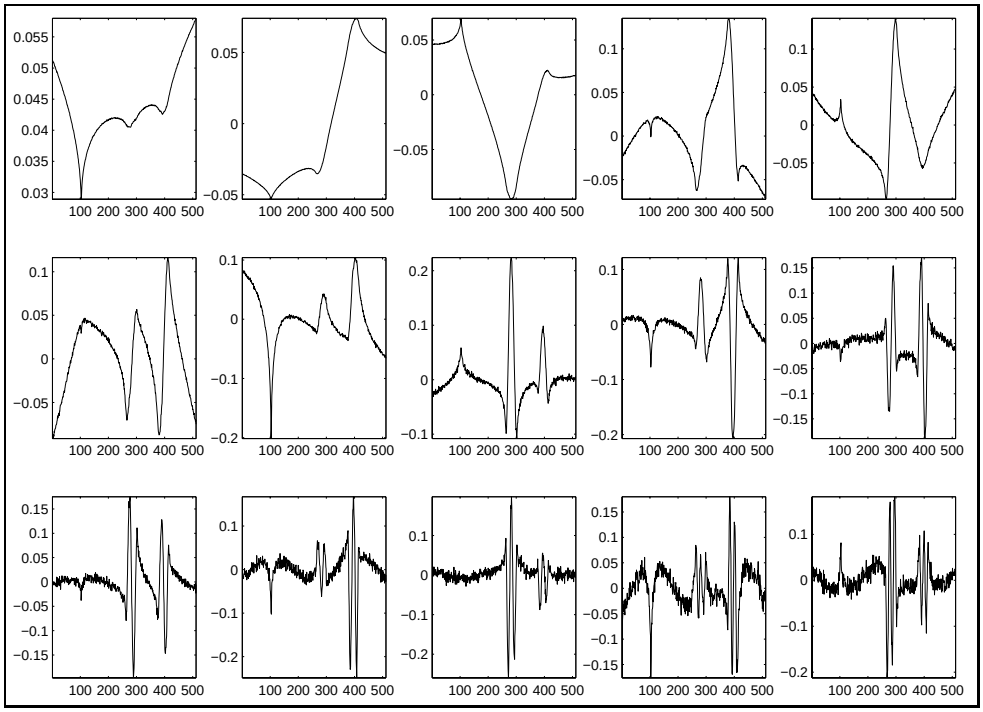

FIG. 8.8 – Les 15 premières composantes principales lorsque la première singularité est calée en $a = 0.2$ (100-ème échantillon).

Discriminante Linéaire sur les 30 premières composantes principales. Les résultats sont rassemblés dans le tableau 8.5 et comparés à ceux obtenus par notre méthode.

L'amélioration considérable des performances de l'Analyse Discriminante Linéaire dans cette seconde expérience montre que l'adaptativité de la représentation des signaux est primordiale pour une bonne classification. Les bons taux de classification obtenus s'expliquent par le fait que la première composante principale (voir figure 8.8) peut s'adapter à la forme en $|t-a|^{\alpha}$

$\sigma \times 10^{-3}$	0	5	10	15	20	25	30	35	40	45
%FC ACP+ADL	16	19	20	18	20	23	24	27	25	31
%FC MMTO+CART	4.8	6.7	10.5	8.9	14	19.2	19.9	19.4	23.25	30.5

TAB. 8.5 – Taux d'erreur en fonction du niveau de bruit : (ACP+ADL) classification de singularités *non* glissantes par Analyse Discriminante Linéaire sur les 30 composantes principales les plus énergétiques ; (MMTO+CART) classification de singularités glissantes avec les extrema de leur transformée en ondelettes et des arbres de décision.

d'une singularité.

Le tableau 8.5 montre que la classification de singularités *glissantes* avec notre technique donne de meilleurs résultats que l'Analyse Discriminante Linéaire sur des singularités *calées temporellement*. En effet la seconde singularité est, elle aussi, susceptible de subir des translations aléatoires. Dans notre technique, les arbres de décision permettent de s'adapter aux variations de la position relative et de la force de cette seconde singularité. En complément de la représentation adaptative de bas niveau constituée du squelette de la transformée en ondelettes, le choix des arbres de décision comme classificateur est donc un facteur déterminant dans les performances de notre méthode.

Lorsque le niveau de bruit augmente, les performances de l'Analyse Discriminante Linéaire sur les signaux calés se rapprochent de celles de notre méthode. En effet les forces des singularités deviennent alors difficiles à mesurer. L'avantage de notre technique, sa capacité à mesurer une grandeur locale fine (l'exposant de Hölder local), s'amenuise en effet lorsque le niveau de bruit ne permet plus de mesurer fiablement ces grandeurs locales.

En définitive, notre méthode de classification de transitoires s'avère systématiquement meilleure que l'Analyse Discriminante Linéaire. L'un de ses points forts est sa robustesse vis-à-vis de l'invariance par translation. Sa capacité à mesurer des caractéristiques fines des signaux, telles que l'exposant de Hölder local, a également été démontrée.

Chapitre 9

Conclusion et perspectives de recherche

Nous avons introduit dans cette thèse des représentations de signaux qui permettent d'en extraire des structures caractéristiques. La poursuite harmonique décompose les signaux musicaux en structures harmoniques qui caractérisent la fréquence fondamentale et la durée des notes. Les variations de fréquence instantanée sont finement analysées par le Matching Pursuit "Chirpé". L'étude des *ridges* du dictionnaire de Gabor que nous avons effectuée pour développer cette technique rapide d'analyse nous a permis de mieux comprendre l'information apportée par l'*échelle* des atomes temps-fréquence sélectionnés dans une poursuite. Le Matching Pursuit Haute Résolution préserve les transitoires des signaux analysés, en imposant des contraintes de résolution temporelle. Il est capable de discriminer des structures temporelles proches que les techniques linéaires usuelles ne peuvent distinguer.

Par ailleurs nous avons pu montrer l'intérêt de l'"Analyse Discriminante Non-linéaire" pour la classification de signaux. L'Analyse Discriminante Linéaire projette le signal qu'elle doit classifier sur un sous-espace déterminé à l'avance, indépendamment de ce signal. L'analyse d'un exemple académique, l'identification de la couleur d'un bruit gaussien, nous a permis de montrer qu'il est payant de choisir la projection de façon adaptative. Nous avons alors exploré une technique de classification de transitoires utilisant les extrema de la transformée en ondelettes et des arbres de décision.

Plusieurs voies de recherche sont naturellement ouvertes par ces travaux. En ce qui concerne la représentation de signaux, les poursuites harmonique et "chirpée" appellent à être fusionnées en une technique unique de décomposition en structures harmoniques dont la fréquence fondamentale peut varier. Le spectre d'une telle structure est constitué de pics dont l'étalement fréquentiel croît avec le rang. A partir d'un certain rang, l'étalement est tel que les pics voisins se recouvrent [MA89]. L'analyse de Fourier à fenêtre ne peut donc détecter correctement les pics en haute fréquence. La poursuite

"harmonique chirpée" permettra sans doute d'y remédier et d'améliorer les outils de compression de la parole [Oud98] qui, jusqu'ici, codent les hautes fréquences comme du bruit.

Par ailleurs, la poursuite harmonique présente le même type d'artefacts que la poursuite usuelle lors des transitoires. Nous avons commencé à travailler sur une adaptation aux molécules harmoniques du critère haute-résolution. Une approche complémentaire sera d'employer des enveloppes asymétriques [Goo97] pour les atomes du dictionnaire.

Il est envisageable de construire des dictionnaires de molécules de dimension variable. Il faudra alors pénaliser les molécules de grande dimension. En effet entre deux molécules emboîtées, la poursuite moléculaire choisira naturellement celle de plus grande dimension, car elle ôtera plus d'énergie au signal analysé. Ainsi, si on a le choix entre deux molécules harmoniques de fondamentales ξ_1 et $\xi_1/2$ et de dimensions K et $2K$, la seconde molécule sera assurément choisie. C'est ainsi que des erreurs d'octave [Dov94] peuvent se produire. Des outils tels que le principe de "Minimum Description Length" [Ris83] pourront s'avérer utiles pour effectuer cette pénalisation.

Le Matching Pursuit Chirpé nous a permis de détecter *localement* la fréquence instantanée et ses variations. Il faut à partir de cette information locale reconstituer le trajet *global* de la fréquence instantanée, par exemple pour caractériser un vibrato. Cela nécessite de *chaîner* les ridges ainsi repérées. Les techniques de recuit simulé [CHT95a],[CHT95b], ou les Modèles de Markov Cachés [GDR93] sont très efficaces pour y parvenir, mais n'utilisent que l'information de fréquence instantanée. L'estimation du chirp à l'aide de notre technique peut sûrement renforcer leur robustesse, en particulier face aux croisements de fréquences instantanées. Par ailleurs, nous avons entamé une étude du *suivi actif de partiels* utilisant les idées développées par Geman et Jedynak [GJ96] pour le suivi de routes dans les images satellitaires. Cette technique présente un intérêt manifeste du point de vue de la complexité algorithmique. Il est en effet inutile de calculer le spectrogramme tout entier pour détecter les ridges, car seules les régions potentiellement intéressantes sont explorées.

Notre méthode de classification de singularités ouvre la porte à la classification de transitoires à l'aide de représentations redondantes adaptatives. L'extension de nos résultats à des dictionnaires temps-fréquence multi-échelle, très redondants, rendra nécessaire la définition de stratégies nouvelles de réduction de la complexité. On pourra par exemple construire les arbres de décision avec une stratégie gloutonne *stochastique*, en utilisant à chaque nœud un sous-dictionnaire *aléatoire* de questions, de façon analogue à notre technique de poursuite rapide dans des sous-dictionnaires de maxima locaux. Les travaux de Amit *et al.* [AG97] [AGW97] [AM99] montrent que cette construction stochastique améliore la qualité de classification en réduisant l'adaptation aux données. On effectuera ainsi quelques pas de plus vers la modélisation et la compréhension des transitoires.

Troisième partie

Annexes

Annexe A

Calcul rapide de produits scalaires d'atomes temps-fréquence gaussiens

Le but de cette annexe est d'établir les formules qui vont nous permettre de calculer *rapidement* les produits scalaires entre des atomes temps-fréquence gaussiens, afin de pouvoir efficacement réaliser le Matching Pursuit Rapide détaillé au chapitre 4. Pour cela on commence par énoncer une formule analytique valable pour des atomes gaussiens à temps continu, éventuellement *chirpés*. Si cette formule est bien connue [Pap86][Pap87][MA89], on en rappelle néanmoins rapidement la démonstration, et surtout on la met sous une forme qui va nous être utile pour calculer le produit scalaire d'atomes gaussiens à temps discret. On rappelle ensuite en effet comment le produit scalaire d'atomes à temps discret se déduit de celui des atomes à temps continu, et l'on en déduit une formule approchée de calcul du produit scalaire d'atomes gaussiens discrets, avec une précision arbitraire.

A.1 Expression analytique du produit scalaire d'atomes gaussiens à temps continu

On va montrer ici que le produit scalaire de deux atomes temps-fréquence Gaussiens "chirpés" :

$$g_{\gamma_j}(t) = \frac{1}{\pi^{1/4}\sqrt{s_j}} e^{-\frac{\left(t-u_j\right)^2}{2s_j^2} + i\xi(t-u_j) + i\frac{c_j}{2}(t-u_j)^2}, j = 1, 2 \qquad \text{(A.1)}$$

est donné par

$$\langle g_{\gamma_1}, g_{\gamma_2} \rangle = \frac{1}{\sqrt{\pi s_1 s_2}} \sqrt{\rho} e^{i\frac{\theta}{2}} e^{R+iI} \qquad \text{(A.2)}$$

avec

$$R = -\frac{\Delta u^2}{2(s_1^2+s_2^2)} - \frac{1}{4}\frac{\overline{s}^2}{\mu}\left(\Delta\xi - \overline{\xi}_R\right)^2 \quad (A.3)$$

$$I = -\frac{(\xi_1+\xi_2)\Delta u}{2} + \Delta c\frac{\Delta u^2}{8} + \frac{1}{8\mu}\left(\frac{s_2^2-s_1^2}{s_1^2+s_2^2}\right)^2 \Delta c\Delta u^2 + \frac{1}{2\mu}\frac{s_2^2-s_1^2}{s_1^2+s_2^2}\Delta u\left(\Delta\xi - \overline{\xi}_I\right) - \frac{1}{8\mu}\overline{s}^4\Delta c\left(\Delta\xi - \overline{\xi}_I\right)^2 \quad (A.4)$$

$$\rho = \frac{\pi\overline{s}^2}{\sqrt{\mu}} \quad (A.5)$$

$$\theta = \arctan\left(\frac{\overline{s}^2}{2}\Delta c\right) \quad (A.6)$$

Pour définir ces différents termes, on a utilisé les abréviations suivantes :

$$\begin{array}{llllllll} \Delta u & \triangleq & u_1-u_2 & \overline{\xi}_R & \triangleq & \frac{c_1 s_1^2+c_2 s_2^2}{s_1^2+s_2^2}\Delta u & \overline{s}^2 & \triangleq & \frac{2s_1^2 s_2^2}{s_1^2+s_2^2} \\ \Delta c & \triangleq & c_1-c_2 & \overline{\xi}_I & \triangleq & \frac{c_1+c_2}{2}\Delta u & \mu & \triangleq & 1+\frac{1}{4}(\overline{s}^2\Delta c)^2 \\ \Delta\xi & \triangleq & \xi_1-\xi_2 & & & & & & \end{array}$$

On peut remarquer que $\overline{s}^2$ est la *moyenne harmonique* des carrés des échelles des deux atomes.

Démonstration

Pour faciliter le travail, on introduit temporairement la notation $\alpha_j \triangleq \frac{1}{2}(1/s_j^2 - ic_j)$. Le produit scalaire entre les deux atomes s'écrit donc, après un changement de variable évident :

$$\begin{aligned} \langle g_{\gamma_1}, g_{\gamma_2}\rangle &= \frac{1}{\sqrt{\pi s_1 s_2}}\int_{-\infty}^{+\infty}\exp\left\{-\alpha_1 t^2 - \overline{\alpha_2}\left(t+\Delta u\right)^2\right\}\exp\left\{i\xi_1 t - i\xi_2\left(t+\Delta u\right)\right\}dt \\ &= \frac{1}{\sqrt{\pi s_1 s_2}}\exp\left\{-\overline{\alpha_2}\Delta u^2 - i\xi_2\Delta u\right\}\int_{-\infty}^{+\infty}\exp\left\{-\left(\alpha_1+\overline{\alpha_2}\right)t^2 + \left(i\Delta\xi - 2\overline{\alpha_2}\Delta u\right)t\right\}dt \end{aligned}$$

En introduisant les raccourcis $\alpha = \alpha_1 + \overline{\alpha_2}$, $\beta = i\Delta\xi - 2\overline{\alpha_2}\Delta u$ et $\lambda = -\overline{\alpha_2}\Delta u^2 - i\xi_2\Delta u$, ce produit scalaire a donc pour expression

$$\begin{aligned} \langle g_{\gamma_1}, g_{\gamma_2}\rangle &= \frac{1}{\sqrt{\pi s_1 s_2}}\exp\{\lambda\}\int_{-\infty}^{+\infty}\exp\left\{-\alpha t^2 + \beta t\right\}dt \\ \langle g_{\gamma_1}, g_{\gamma_2}\rangle &\overset{(a)}{=} \frac{1}{\sqrt{\pi s_1 s_2}}\exp\left\{\lambda + \frac{\beta^2}{4\alpha}\right\}\int_{-\infty}^{+\infty} e^{-\alpha t'^2}dt' \end{aligned}$$

On a obtenu la relation (a) en effectuant le changement de variable $t' = t - \frac{\beta}{2\alpha}$. On aboutit à l'expression

$$\langle g_{\gamma_1}, g_{\gamma_2} \rangle = \frac{1}{\sqrt{\pi s_1 s_2}} \exp\left\{\lambda + \frac{\beta^2}{4\alpha}\right\} \sqrt{\frac{\pi}{\alpha}} \tag{A.7}$$

qui donne bien la formule (A.2) une fois développés les différents termes et facteurs. Ce qui suit est le script Maple qui établit l'égalité en question.

Fin de la preuve

```
#
# Ce fichier est la verification sous Maple de l'exactitude des formules
# de mise a jour pour les atomes gaussiens
#

# Les echelles et les pentes sont reelles
assume(s1,real);
assume(s2,real);
assume(c1,real);
assume(c2,real);

a1 := (1/s1^2-I*c1)/2;
a2 := (1/s2^2-I*c2)/2;
dc := c1-c2;

# De meme que les frequences
assume(xi1,real);
assume(xi2,real);
dXi := xi1-xi2;

# Et meme chose pour les temps
assume(du,real);

a := a1+conjugate(a2);
b := I*dXi-2*conjugate(a2)*du;
l := -conjugate(a2)*du^2-I*xi2*du;

#
# L'expression de depart
#
expr  := l+b^2/(4*a);
Rexpr := normal(evalc(Re(expr)));
Iexpr := normal(evalc(Im(expr)));
```

```
# Les raccourcis pour l'expression a l'arrivee
XiMeanR       := du*(c1*s1^2+c2*s2^2)/(s1^2+s2^2);
XiMeanI       := du*(c1+c2)/2;
sHarm2        := 2*s1^2*s2^2/(s1^2+s2^2);
mu            := 1+(sHarm2*dc/2)^2;
sDiffQuotient := (s2^2-s1^2)/(s1^2+s2^2);

# Partie reelle
R             := -du^2/(2*(s1^2+s2^2))-sHarm2*(dXi-XiMeanR)^2/(4*mu);

# Partie imaginaire
I1            := -(xi1+xi2)*du/2 +dc*du^2/8 +dc*du^2*sDiffQuotient^2/(8*mu)\
+du*sDiffQuotient*(dXi-XiMeanI)/(2*mu)-dc*sHarm2^2*(dXi-XiMeanI)^2/(8*mu);

# Ces resultats doivent valoir zero
print(‘Ces resultats doivent etre nuls‘);
normal(Rexpr-R);
normal(Iexpr-I1);
```

A.2 Effet de la discrétisation sur le produit scalaire

Nous prenons ici les conventions du libre *A Wavelet Tour of Signal Processing* [Mal98] pour la transformée de Fourier. Soit $g_j^d[n]$ une version échantillonnée de l'atome temps-fréquence $g_j(t)$. On considérera que la fréquence d'échantillonnage est de 1, et que l'on n'a pas vérifié le critère de Shannon-Nyquist, *i.e.* bien que g_j ne soit pas à bande limitée on à défini g_j^d par $g_j^d[n] = g_j(n)$, *i.e.* $g_j^d = \underline{\text{|||}}_1 g_j$, où $\underline{\text{|||}}_T$ est le peigne de Dirac de période T.

Cette section à pour objet d'établir les relations entre le produit scalaire $\sum_n g_1^d[n] g_2^d[n]$ des atomes *discrets* g_1^d, g_2^d et des grandeurs analogues pour leurs homologues *continus* g_1, g_2, afin d'établir une formule approchée de calcul du produit scalaire des atomes discrets. On commence par utiliser la formule de Poisson qui nous donne la transformée de Fourier des atomes discrets :

$$\widehat{g_j^d}(\omega) = \underline{\text{|||}}_{2\pi} \star \widehat{g_j}(\omega) = \sum_{k=-\infty}^{+\infty} \widehat{g_j}(\omega - 2k\pi)$$

puis à l'aide de l'identité de Parseval on obtient

$$\begin{aligned}
\left\langle g_1^d, g_2^d \right\rangle &= \sum_{n=-\infty}^{+\infty} g_1^d[n]\overline{g_2^d}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{g_1^d}(\omega)\overline{\widehat{g_2^d}}(\omega)d\omega \\
&= \frac{1}{2\pi}\sum_{k_1,k_2}\int_{-\pi}^{\pi} \widehat{g_1}(\omega-2k_1\pi)\overline{\widehat{g_2}}(\omega-2k_2\pi)d\omega \\
&= \frac{1}{2\pi}\sum_{k,q}\int_{-\pi}^{\pi} \widehat{g_1}(\omega-2k\pi)\overline{\widehat{g_2}}(\omega-2(k+q)\pi)d\omega \\
&= \frac{1}{2\pi}\sum_{k,q}\int_{-\pi+2k\pi}^{\pi+2k\pi} \widehat{g_1}(\omega)\overline{\widehat{g_2}}(\omega-2q\pi)d\omega \\
&= \frac{1}{2\pi}\sum_{q}\int_{-\infty}^{+\infty} \widehat{g_1}(\omega)\overline{\widehat{g_2}}(\omega-2q\pi)d\omega \\
&= \frac{1}{2\pi}\sum_{q}\int_{-\infty}^{+\infty} \widehat{g_1}(\omega+q\pi)\overline{\widehat{g_2}}(\omega-q\pi)d\omega \\
&= \frac{1}{2\pi}\sum_{q} 2\pi\int_{-\infty}^{+\infty} g_1(t)e^{-iq\pi t}\overline{g_2(t)e^{+iq\pi t}}dt \\
&= \sum_{q}\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle
\end{aligned}$$

En définitive, et ce *quelle que soit l'enveloppe* des atomes continus, la formule de discrétisation

$$\left\langle g_1^d, g_2^d \right\rangle = \sum_{q}\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle \tag{A.8}$$

permet d'exprimer le produit scalaire d'atomes discrets en fonction de leur version à temps continu.

A.3 Formule approchée pour les atomes gaussiens discrets

A partir de la formule analytique (A.2) on peut calculer chacun des termes de la série (A.8) de façon *rapide*, avec un nombre d'opérations $\mathcal{O}(1)$ indépendant de la dimension N des signaux considérés. En effectuant la somme (A.8) sur un petit nombre d'indices $q \in [\![q_1, q_2]\!]$, on calcule une valeur approchée du produit scalaire entre deux atomes discrets gaussiens chirpés. Pour contrôler l'erreur de troncature

$$\varepsilon(q_1,q_2) \leq \sum_{q=-\infty}^{q_1-1}\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right| + \sum_{q=q_2+1}^{+\infty}\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right| \tag{A.9}$$

on majore les termes $\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right|$ de manière à contrôler la somme de la série majorante. D'après l'expression analytique (A.2), on a

$$\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right| = Ae^{-\frac{\lambda}{2}(2\pi q - M)^2} \tag{A.10}$$

où $\lambda \triangleq \frac{\overline{s}^2}{2\mu}$ et $M \triangleq \Delta\xi - \overline{\xi}_R$ ne dépendent pas de q. Distinguons alors deux cas, selon la position de q par rapport à $M/2\pi$:

– si $2\pi(q-1) > M$, alors

$$\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right| \leq A\int_{q-1}^{q} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx \tag{A.11}$$

– si $2\pi(q+1) < M$, alors

$$\left|\left\langle g_1(t)e^{-iq\pi t}, g_2(t)e^{+iq\pi t}\right\rangle\right| \leq A\int_{q}^{q+1} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx \tag{A.12}$$

.

Dès que $2\pi q_1 < M < 2\pi q_2$ on a donc la majoration

$$\begin{aligned}
\varepsilon(q_1, q_2) &\leq A\sum_{q=-\infty}^{q_1-1}\int_{q}^{q+1} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx + A\sum_{q=q_2+1}^{+\infty}\int_{q-1}^{q} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx \\
&\leq A\int_{-\infty}^{q_1} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx + A\int_{q_2}^{+\infty} e^{-\frac{\lambda}{2}(2\pi x - M)^2}dx
\end{aligned}$$

Soit δ un paramètre et

$$q_1 \triangleq \lfloor M/2\pi - \delta \rfloor \tag{A.13}$$

et

$$q_2 \triangleq \lceil M/2\pi + \delta \rceil . \tag{A.14}$$

Déterminons une condition suffisante sur δ pour se garantir que l'erreur de troncature est petite. Par changement de variable $t = \sqrt{\lambda}(2\pi x - M)$

$$\begin{aligned}
\varepsilon(\delta) &\leq \frac{A}{2\pi\sqrt{\lambda}}\left(\int_{-\infty}^{\sqrt{\lambda}(2\pi q_1 - M)} e^{-t^2/2}dt + \int_{\sqrt{\lambda}(2\pi q_2 - M)}^{+\infty} e^{-t^2/2}dt\right) && \text{(A.15)} \\
&\leq \frac{A}{2\pi\sqrt{\lambda}}\left(\int_{-\infty}^{-2\pi\delta\sqrt{\lambda}} e^{-t^2/2}dt + \int_{2\pi\delta\sqrt{\lambda}}^{+\infty} e^{-t^2/2}dt\right) && \text{(A.16)} \\
&\leq \frac{2A}{2\pi\sqrt{\lambda}}\int_{2\pi\delta\sqrt{\lambda}}^{+\infty} e^{-t^2/2}dt. && \text{(A.17)}
\end{aligned}$$

Comme pour tout x

$$\int_{x}^{\infty} e^{-t^2/2}dt \leq e^{-x^2/2}/x$$

on obtient la majoration

$$\varepsilon(\delta) \leq \frac{A}{2\pi^2\lambda\delta} e^{-2\pi^2\lambda\delta^2} \tag{A.18}$$

Par conséquent pour une précision $\varepsilon \ll 1$ visée, il suffit d'utiliser

$$\delta \sim \sqrt{\frac{1}{2\pi^2\lambda} \log \frac{1}{\varepsilon}}. \tag{A.19}$$

En général, si λ n'est pas trop petit, $\delta = 1$ suffit.

Annexe B

Démonstration des théorèmes de ridges

L'objet de cette annexe est la démonstration des divers théorèmes de ridges de dictionnaires multi-échelle de Gabor gaussiens, chirpés ou non, que nous avons énoncés au cours de cette thèse.

B.1 Démonstration des théorèmes d'approximation 4 et 5

On commence par démontrer le théorème 4, puis le théorème 5. Ces théorèmes d'approximation consistent à montrer qu'un signal analytique "régulier" ressemble localement, du point de vue du dictionnaire gaussien $\mathcal{D}$ ou $\mathcal{D}^+$ d'analyse, à un atome gaussien chirpé.

B.1.1 Démonstration du théorème 4

- Soit $x(t) = a(t)e^{i\phi(t)}$ un signal analytique. Les développements de Taylor à l'ordre 1 de a et à l'ordre 3 de ϕ, au voisinage de l'instant u, s'écrivent

$$\begin{aligned} a(u+t) &= a(u) + a'(\theta_1)t \\ \text{et} & \\ \phi(u+t) &= \phi(u) + \phi'(u)t + \frac{\phi''(u)}{2}t^2 + \frac{\phi'''(\theta_2)}{6}t^3 \end{aligned}$$

  où θ_1 et θ_2 sont compris entre u et $u+t$.
- Le produit scalaire entre x et $g_{(s,u,\xi,0)}$ s'écrit donc, en faisant apparaître

les termes qui nous intéressent par un changement de variable,

$$\begin{aligned}\left\langle x, g_{(s,u,\xi,0)}\right\rangle &= a(u)e^{i\phi(u)}\int_{-\infty}^{+\infty}\overline{g_{(s,0,\xi,0)}(t)}e^{i\phi'(u)t}e^{i\frac{\phi''(u)}{2}t^2}e^{i\frac{\phi'''(\theta_2)}{6}t^3}dt\\ &\quad+\int_{-\infty}^{+\infty}a'(\theta_1)t\overline{g_{(s,0,\xi,0)}(t)}e^{i\phi(u+t)}dt\\ &= a(u)e^{i\phi(u)}\int_{-\infty}^{+\infty}g_{(s,0,0,\phi''(u))(t)}e^{-i(\xi-\phi'(u))t}dt\\ &\quad+a(u)e^{i\phi(u)}\int_{-\infty}^{+\infty}g_{(s,0,0,\phi''(u))}(t)e^{-i(\xi-\phi'(u))t}\left(e^{i\frac{\phi'''(\theta_2)}{6}t^3}-1\right)dt\\ &\quad+\int_{-\infty}^{+\infty}a'(\theta_1)t\overline{g_{(s,0,\xi,0)}(t)}e^{i\phi(u+t)}dt\end{aligned}$$

- En mettant en valeur le premier terme du développement, on obtient

$$\left\langle x, g_{(s,u,\xi,0)}\right\rangle = a(u)e^{i\phi(u)}\left(\widehat{g}_{(s,0,0,\phi''(u))}(\xi-\phi'(u))+\epsilon_1(s,u,\xi)\right) \quad \text{(B.1)}$$

où le terme d'erreur $|\epsilon_1(s,u,\xi)|$ est manifestement majoré par

$$\int_{-\infty}^{+\infty}g_s(t)\left|e^{i\frac{\phi'''(\theta_2)}{6}t^3}-1\right|dt+\frac{\|a'(u)\|_\infty}{a(u)}\int_{-\infty}^{+\infty}|t|\,g_s(t)dt \quad \text{(B.2)}$$

avec $g_s(t)=1/\sqrt{s}g(t/s)$.

- Majorons maintenant le premier terme de (B.2), en découpant l'intégrale en deux morceaux, avec un paramètre η

$$\int_{-\infty}^{+\infty}g_s(t)\left|e^{i\frac{\phi'''(\theta_2)}{6}t^3}-1\right|dt=\int_{|t|>\eta s}+\int_{|t|\leq\eta s}.$$

Pour ce qui est du premier morceau

$$\int_{|t|>\eta s}\leq 2\int_{|t|>\eta s}g_s(t)dt=2\sqrt{s}\int_{|t|>\eta}g(t)dt\leq 2\sqrt{s}\frac{e^{-\frac{\eta^2}{2}}}{\eta} \quad \text{(B.3)}$$

quand au second, comme pour tout $z\in\mathbb{C}$, $|e^z-1|\leq e^{|z|}-1\leq|z|\,e^{|z|}$, il est majoré par

$$\begin{aligned}\int_{|t|\leq\eta s} &\leq \int_{|t|\leq\eta s}g_s(t)\frac{|\phi'''(\theta_2)|}{6}|t|^3\,e^{\frac{|\phi'''(\theta_2)|}{6}|t|^3}dt\\ &\leq \frac{\|\phi'''\|_\infty}{6}e^{\frac{\|\phi'''\|_\infty}{6}\eta^3s^3}\int_{|t|\leq\eta s}|t|^3\,g_s(t)dt\\ &\leq \frac{\|\phi'''\|_\infty}{6}e^{\frac{\|\phi'''\|_\infty}{6}\eta^3s^3}\int_{-\infty}^{+\infty}|t|^3\,g_s(t)dt \quad \text{(B.4)}\end{aligned}$$

- En notant $\sigma_k^k \stackrel{\Delta}{=} \int |t|^k g(t)dt$, et en réunissant (B.2), (B.3) et (B.4), on obtient

$$|\epsilon_1(s,u,\xi)| \leq \sqrt{s}\left(s\sigma_1\frac{\|a'(u)\|_\infty}{a(u)} + \frac{s^3\sigma_3^3}{6}\left\|\phi'''\right\|_\infty e^{\frac{\|\phi'''\|_\infty}{6}\eta^3 s^3} + 2\frac{e^{-\frac{\eta^2}{2}}}{\eta}\right) \tag{B.5}$$

- Il reste à bien choisir la valeur de η pour que (B.5) soit aussi serrée que possible. Idéalement, on dispose de la majoration

$$|\epsilon_1(s,u,\xi)| \leq \sqrt{s}\inf_\eta\left\{\frac{s^3\sigma_3^3}{6}\left\|\phi'''\right\|_\infty e^{\frac{\|\phi'''\|_\infty}{6}\eta^3 s^3} + 2\frac{e^{-\frac{\eta^2}{2}}}{\eta}\right\} \tag{B.6}$$

En prenant $\eta^3 = 1/(s^3\left\|\phi'''\right\|_\infty)$, on obtient la majoration (5.30).
□.

B.1.2 Démonstration du théorème 5

- Soit $x(t) = a(t)e^{i\phi(t)}$ un signal analytique. On peut exprimer l'amplitude et la phase au voisinage de u par les développements de Taylor suivants

$$\begin{aligned} a(u+t) &= a(u)e^{-\alpha'(u)t-\frac{\alpha''(u)}{2}t^2-\frac{\alpha'''(\theta_1)}{6}t^3} \\ \text{et} & \\ \phi(u+t) &= \phi(u)+\phi'(u)t+\frac{\phi''(u)}{2}t^2+\frac{\phi'''(\theta_2)}{6}t^3 \end{aligned}$$

où $\alpha(t) = -\log a(t)$ et θ_1, θ_2 sont compris entre u et $u+t$.
- Le produit scalaire entre x et $g_{(s,u,\xi,c)}$ s'exprime alors, avec un changement de variable :

$$\begin{aligned} \left\langle x, g_{(s,u,\xi,c)}\right\rangle &= a(u)e^{i\phi(u)}\int_{-\infty}^{+\infty} e^{-\alpha'(u)t-\frac{\alpha''(u)}{2}t^2}e^{i\phi'(u)t+i\frac{\phi''(u)}{2}t^2} \\ &\quad \frac{1}{\sqrt{s}}e^{-\frac{1}{2s^2}t^2}e^{-i\xi t-i\frac{c}{2}t^2} \\ &\quad e^{\frac{t^3}{6}(-\alpha'''(\theta_1)+i\phi'''(\theta_2))}dt \\ &= \frac{a(u)e^{i\phi(u)}}{(\alpha''(u))^{1/4}}e^{\frac{(\alpha')^2}{2\alpha''}}\int_{-\infty}^{+\infty} g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',0,0)}(t)\overline{g_{(s,u,\xi-\phi',c-\phi'')}(t)} \\ &\quad e^{\frac{t^3}{6}(-\alpha'''(\theta_1)+i\phi'''(\theta_2))}dt \end{aligned}$$

- On obtient donc, en mettant en valeur le premier terme du développement :

$$\left\langle x, g_{(s,u,\xi,c)}\right\rangle = \frac{a(u)e^{i\phi(u)}}{(\alpha'')^{1/4}} e^{\frac{(\alpha')^2}{2\alpha''}} \left(\left\langle g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',0,0)}, g_{(s,u,\xi-\phi',c-\phi'')} \right\rangle + \epsilon(s,u,\xi,c) \right) \tag{B.7}$$

où le terme d'erreur est

$$\epsilon(s,u,\xi,c) = \int_{-\infty}^{+\infty} g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',0,0)}(t)\overline{g_{(s,u,\xi-\phi',c-\phi'')}(t)} \left(e^{\frac{t^3}{6}(-\alpha'''(\theta_1)+i\phi'''(\theta_2))} - 1 \right) dt \tag{B.8}$$

- On s'attache maintenant à établir une majoration de ce terme. On a d'abord immédiatement :

$$\begin{aligned} |\epsilon(s,u,\xi,c)| &\leq (\alpha''/s^2)^{1/4} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2s^2}-\frac{\alpha''}{2}(t+\alpha'/\alpha'')^2} \left| e^{\frac{t^3}{6}(-\alpha'''(\theta_1)+i\phi'''(\theta_2))} - 1 \right| dt \\ &\leq (\alpha''/s^2)^{1/4} \left(\int_{|t|>\eta s} + \int_{|t|\leq \eta s} \right) \end{aligned}$$

On traite séparément les deux morceaux d'intégrale écrits ci-dessus. Pour ce qui est du premier terme :

$$\begin{aligned} \int_{|t|>\eta s} &= \int_{|t|>\eta s} e^{-\frac{t^2}{2s^2}} \left| e^{-\alpha' t - \frac{\alpha''}{2}t^2 - \frac{t^3}{6}\alpha'''(\theta_1) - \frac{(\alpha')^2}{2\alpha''}} e^{i\frac{t^3}{6}\phi'''(\theta_2)} - e^{-\frac{\alpha''}{2}(t+\alpha'/\alpha'')^2} \right| dt \\ &= \int_{|t|>\eta s} e^{-\frac{t^2}{2s^2}} \left| \frac{a(u+t)}{a(u)} e^{-\frac{(\alpha')^2}{2\alpha''}} e^{i\frac{t^3}{6}\phi'''(\theta_2)} - e^{-\frac{\alpha''}{2}(t+\alpha'/\alpha'')^2} \right| dt \\ &\leq \int_{|t|>\eta s} e^{-\frac{t^2}{2s^2}} \left(\frac{\|a\|_\infty}{|a(u)|} + 1 \right) dt \leq \frac{2\|a\|_\infty}{|a(u)|} \int_{|t|>\eta s} e^{-\frac{t^2}{2s^2}} dt \\ &\leq \frac{2\|a\|_\infty}{|a(u)|} s \frac{2}{\eta} e^{-\eta^2/2} \leq s \frac{\|a\|_\infty}{|a(u)|} \frac{4}{\eta} e^{-\eta^2/2} \end{aligned} \tag{B.9}$$

quand au second, comme pour tout $z \in \mathbb{C}$, $|e^z - 1| \leq e^{|z|} - 1 \leq |z|\, e^{|z|}$,

$$\begin{aligned} \int_{|t|\leq\eta s} &\leq \int_{|t|\leq\eta s} e^{-\frac{t^2}{2s^2}} \frac{|t|^3}{6} \left(\left|\alpha'''(\theta_1)\right| + \left|\phi'''(\theta_2)\right| \right) e^{\frac{|t|^3}{6}(|\alpha'''(\theta_1)|+|\phi'''(\theta_2)|)} dt \\ &\leq \frac{\|\alpha'''\|_\infty + \|\phi'''\|_\infty}{6} e^{\frac{\eta^3 s^3}{6}(\|\alpha'''\|_\infty + \|\phi'''\|_\infty)} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2s^2}} |t|^3 \, dt \\ &\leq s \frac{\|\alpha'''\|_\infty + \|\phi'''\|_\infty}{6} e^{\frac{\eta^3 s^3}{6}(\|\alpha'''\|_\infty + \|\phi'''\|_\infty)} s^3 \sigma_3^3 \end{aligned} \tag{B.10}$$

- Notons alors $K = \|\alpha'''\|_\infty + \|\phi'''\|_\infty$. En rassemblant les majorations (B.9) et (B.10), on a donc montré que pour tout η

$$|\epsilon(s,u,\xi,c)| \leq (\alpha'' s^2)^{1/4} \left(\frac{\|a\|_\infty}{a(u)} \frac{4e^{-\frac{\eta^2}{2}}}{\eta} + \frac{Ks^3\sigma_3^3}{6} e^{\frac{K s^3}{6}\eta^3} \right) \tag{B.11}$$

Pour terminer la démonstration il faut choisir η de manière à obtenir dans (B.11) la meilleure majoration possible, puisque

$$|\epsilon(s,u,\xi,c)| \leq (\alpha'' s^2)^{1/4} \inf_{\eta} \left(\frac{\|a\|_\infty}{a(u)} \frac{4e^{-\frac{\eta^2}{2}}}{\eta} + \frac{K s^3 \sigma_3^3}{6} e^{\frac{K s^3}{6}\eta^3} \right) \tag{B.12}$$

On obtient avec $\eta^3 = K^{-1} s^{-3}$ la majoration (5.60).

- Pour conclure, on met en forme le premier terme du développement (B.7)

$$\begin{aligned} \left\langle g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',0,0)}, g_{(s,u,\xi-\phi',c-\phi'')} \right\rangle &= e^{-i\phi'\alpha'/\alpha'' + i\frac{\phi''}{2}(\alpha'/\alpha'')^2} \\ &\quad \left\langle g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',\phi'-\phi''\alpha'/\alpha'',\phi'')}, g_{(s,u,\xi,c)} \right\rangle \end{aligned}$$

B.2 Démonstration des corollaires 1 et 2

A partir de ces théorèmes d'approximation, on peut montrer que les maxima $\xi(s,u)$ de $\xi \mapsto \langle x, g_{(s,u,\xi,0)} \rangle$ caractérisent la position de la fréquence instantanée. C'est l'objet du corollaire 1. Par ailleurs les maxima $s(u)$ de $s \mapsto \langle x, g_{(s,u,\xi(s,u),0)} \rangle$ caractérisent le chirp instantané, comme l'exprime le corollaire 2. On démontre ici ces deux corollaires.

B.2.1 Démonstration du corollaire 1

- Le premier terme du développement (5.29) de $\langle x, g_{(s,u,\xi,0)} \rangle$ établi au théorème 4 ne dépend de ξ que par la fonction $\widehat{g}_{(s,0,0,\phi''(u))}(\xi - \phi'(u))$, dont le module

$$\left| \widehat{g}_{(s,0,0,\phi''(u))}(\xi - \phi'(u)) \right| = \left(\frac{4\pi s^2}{1 + s^4 |\phi''(u)|^2} \right)^{1/4} e^{-\frac{s^2}{2\left(1+s^4|\phi''(u)|^2\right)}(\xi-\phi'(u))^2} \tag{B.13}$$

atteint son maximum en $\xi = \phi'(u)$.

- D'après le théorème 4, la valeur de $\left| \langle x, g_{(s,u,\xi,0)} \rangle \right|$ en $\xi = \phi'(u)$ est au moins

$$\left| \langle x, g_{(s,u,\phi'(u),0)} \rangle \right| \geq a(u) \left((4\pi\lambda)^{1/4} - \sqrt{s}\epsilon_{max} \right) \tag{B.14}$$

où l'on a noté $\lambda \stackrel{\Delta}{=} s^2/(1 + s^4 |\phi''(u)|^2)$ pour abréger les notations, et où ϵ_{max} est le majorant du terme d'erreur $\epsilon(s,u,\xi)$ établi en (5.30).
On veut montrer que, lorsque ξ est loin de $\phi'(u)$, le produit scalaire est plus petit que cette valeur, de sorte que le maximum sera atteint

au voisinage de $\phi'(u)$. Pour cela considérons un réel $\delta\xi > 0$: lorsque $|\xi - \phi'(u)| > \delta\xi$, on a, toujours d'après le théorème 4

$$\left|\left\langle x, g_{(s,u,\xi,0)}\right\rangle\right| < a(u)\left((4\pi\lambda))^{1/4} e^{-\frac{\lambda}{2}(\delta\xi)^2} + \sqrt{s}\epsilon_{max}\right). \tag{B.15}$$

- A condition que

$$(4\pi\lambda)^{1/4} - \sqrt{s}\epsilon_{max} \geq (4\pi\lambda)^{1/4} e^{-\frac{\lambda(\delta\xi)^2}{2}} + \sqrt{s}\epsilon_{max} \tag{B.16}$$

on a

$$\begin{aligned} \sup_{\xi\in[\phi'(u)-\delta\xi,\phi'(u)+\delta\xi]} \left|\left\langle x, g_{(s,u,\xi,0)}\right\rangle\right| &\geq (4\pi\lambda)^{1/4} - \sqrt{s}\epsilon_{max} \\ &\geq (4\pi\lambda)^{1/4} e^{-\frac{\lambda(\delta\xi)^2}{2}} + \sqrt{s}\epsilon_{max} \\ &> \sup_{\xi\notin[\phi'(u)-\delta\xi,\phi'(u)+\delta\xi]} \left|\left\langle x, g_{(s,u,\xi,0)}\right\rangle\right|. \end{aligned} \tag{B.17}$$

donc $\left|\left\langle x, g_{(s,u,\xi,0)}\right\rangle\right|$ atteint son maximum absolu pour $|\xi - \phi'(u)| \leq \delta\xi$, ce qui est le résultat cherché.
- Pour prouver que le maximum absolu $\xi(u,s)$ vérifie

$$\left|\xi(u,s) - \phi'(u)\right| \leq \delta\xi(u,s) \tag{B.18}$$

il suffit donc que $e^{-\frac{\lambda(\delta\xi)^2}{2}} \leq 1 - 2\epsilon_{max}\sqrt{s}/(4\pi\lambda)^{1/4}$. Cela n'est possible que si

$$\epsilon_{max} \leq \frac{(4\pi\lambda)^{1/4}}{2\sqrt{s}} = \left(\frac{\pi}{4\left(1 + s^4\left|\phi''(u)\right|^2\right)}\right)^{1/4} \tag{B.19}$$

La plus petite valeur $\delta\xi(u,s)$ qui convienne est alors

$$\delta\xi(u,s) = \sqrt{\frac{2\left(1 + s^4\left|\phi''(u)\right|^2\right)}{s^2}\log\left(1 - \epsilon_{max}\left(\frac{4\left(1 + s^4\left|\phi''(u)\right|^2\right)}{\pi}\right)^{1/4}\right)^{-1}} \tag{B.20}$$

- D'après (B.17) et le théorème 4, la valeur atteinte au maximum est dans l'intervalle $\left[a(u)\left((4\pi\lambda)^{1/4} - \sqrt{s}\epsilon_{max}\right), a(u)\left((4\pi\lambda)^{1/4} + \sqrt{s}\epsilon_{max}\right)\right]$ ce qui conclut la démonstration.
□.

B.2.2 Démonstration du corollaire 2

- D'après le corollaire 1, en notant $\beta \stackrel{\Delta}{=} \log s^2 \left|\phi''(u)\right|$, on a

$$\left|\left\langle x, g_{(s,u,\xi(s,u),0)}\right\rangle\right| = a(u)\left(\frac{2\pi}{\left|\phi''(u)\right|}\right)^{1/4}\left(\left(\frac{1}{\cosh\beta}\right)^{1/4} + \left(\frac{s^2\left|\phi''(u)\right|}{2\pi}\right)^{1/4}\epsilon(s,u)\right). \tag{B.21}$$

  où $\epsilon(s,u)$ est majoré par $\epsilon_{max}(s,u)$ défini en (5.30). Le premier terme de cette expression atteint son maximum en $\beta = 0$, *i.e.* lorsque s vaut

$$s_0 = 1/\sqrt{\left|\phi''(u)\right|}. \tag{B.22}$$

- Comme $\epsilon_{max}(s,u)$ est fonction croissante de s, le terme d'erreur dans (B.21) est majoré sur $]0,s]$ par

$$\begin{aligned}\eta_{max}(s,u) &\stackrel{\Delta}{=} \left(\frac{s^2\left|\phi''(u)\right|}{2\pi}\right)^{1/4}\epsilon_{max}(s,u)\\ &= \left(\frac{s}{s_0}\right)^{1/2}(2\pi)^{-1/4}\epsilon_{max}(s,u)\end{aligned} \tag{B.23}$$

  et donc sur $[0,\lambda s_0]$ par

$$\eta_{max}(\lambda s_0,u) = \sqrt{\lambda}(2\pi)^{-1/4}\epsilon_{max}(\lambda s_0,u) \tag{B.24}$$

- La valeur maximale de $\left|\left\langle x, g_{(s,u,\xi(s,u),0)}\right\rangle\right|$ est au moins la valeur prise en $s = s_0$, que l'on peut minorer par

$$\left|\left\langle x, g_{(s_0,u,\phi'(u),0)}\right\rangle\right| \geq a(u)\left(\frac{2\pi}{\left|\phi''(u)\right|}\right)^{1/4}(1-\eta_{max}(\lambda s_0,u)) \tag{B.25}$$

  On veut montrer que, lorsque $s \in]0,\lambda s_0]$ est loin de s_0, *i.e.* lorsque β est loin de zéro, le produit scalaire est nécessairement plus petit que le membre de droite de (B.25), de sorte que le maximum sur $]0,\lambda s_0]$ sera atteint au voisinage de s_0. A cet effet, on utilisera la technique employée pour la démonstration du lemme 1.
- Pour tout $s \in]0,\lambda s_0]$ tel que $|\beta| > \beta_1 > 0$, (*i.e.* $s > s_0 e^{\beta_1/2}$ ou $s < s_0 e^{-\beta_1/2}$), on a

$$\left|\left\langle x, g_{(s,u,\xi(s,u),0)}\right\rangle\right| < a(u)\left(\frac{2\pi}{\left|\phi''(u)\right|}\right)^{1/4}\left(\left(\frac{1}{\cosh\beta_1}\right)^{1/4} + \eta_{max}(\lambda s_0,u)\right) \tag{B.26}$$

  En raisonnant comme dans la démonstration du corollaire 1, si

$$\left(\frac{1}{\cosh\beta_1}\right)^{1/4} + \eta_{max}(\lambda s_0,u) \leq 1 - \eta_{max}(\lambda s_0,u) \tag{B.27}$$

alors

$$\sup_{s\in[s_0e^{-\beta_1/2},s_0e^{\beta_1/2}]}\left|\left\langle x,g_{(s,u,\xi(s,u),0)}\right\rangle\right| > \sup_{s\in]0,\lambda s_0]\setminus[s_0e^{-\beta_1/2},s_0e^{\beta_1/2}]}\left|\left\langle x,g_{(s,u,\xi(s,u),0)}\right\rangle\right| \tag{B.28}$$

La condition (B.27) ne peut être remplie que si $2\eta_{max}(\lambda s_0,u) < 1$, c'est-à-dire si

$$\epsilon_{max}(\lambda s_0,u) \leq \frac{(2\pi)^{1/4}}{2\sqrt{\lambda}}. \tag{B.29}$$

et la plus petite valeur $\underline{\beta}(\lambda s_0,u)$ de β_1 qui convienne est alors

$$\underline{\beta}(\lambda s_0,u) = \arg\cosh\left(1-2\eta_{max}(\lambda s_0,u)\right)^{-4} \tag{B.30}$$

La relation (B.29) est bien vérifiée en vertu de l'hypothèse (5.39).

- Montrons maintenant que

$$s_0e^{\underline{\beta}(\lambda s_0,u)/2} < \lambda s_0, \tag{B.31}$$

Comme pour tout x, $e^x < 2\cosh x$, on a

$$e^{\underline{\beta}(\lambda s_0,u)/2} < \sqrt{2\cosh\underline{\beta}(\lambda s_0,u)} = \sqrt{2\left(1-2\eta_{max}(\lambda s_0,u)\right)^{-4}}. \tag{B.32}$$

Il suffit donc de montrer que $\sqrt{2\left(1-2\eta_{max}(\lambda s_0,u)\right)^{-4}} \leq \lambda$, *i.e.*

$$\eta_{max}(\lambda s_0,u) \leq (1-2^{1/4}/\sqrt{\lambda})/2 \tag{B.33}$$

soit encore

$$\epsilon_{max}(\lambda s_0,u) \leq \frac{(2\pi)^{1/4}}{2\sqrt{\lambda}}(1-2^{1/4}/\sqrt{\lambda}) \tag{B.34}$$

ce qui est vrai d'après l'hypothèse (5.39).

- On peut maintenant conclure la démonstration, car en vertu de (B.31) on a

$$[s_0e^{-\beta(\lambda s_0,u)/2}, s_0e^{\beta(\lambda s_0,u)/2}] \subsetneq]0,\lambda s_0] \tag{B.35}$$

et la relation (B.28) montre alors que le maximum absolu $s(u)$ sur $]0,\lambda s_0]$ vérifie

$$s(u)/s_0 \in \left[e^{-\underline{\beta}(\lambda s_0,u)/2}, e^{+\underline{\beta}(\lambda s_0,u)/2}\right]. \tag{B.36}$$

De plus ce maximum absolu est un *maximum local*, car il est intérieur à $]0,\lambda s_0]$.

- D'après (B.21) la valeur $\left|\left\langle x,g_{(s(u),u,\xi(s(u),u),0)}\right\rangle\right|$ atteinte en ce maximum vérifie l'encadrement exprimé en (5.41)

□.

B.3 Démonstration de la proposition 1

Le produit scalaire $\langle x(t), g_{(s,u,\xi,0)} \rangle = Ae^{i\Phi}$ est proportionnel à

$$\left\langle g_{(1/\sqrt{\alpha''},u-\alpha'/\alpha'',0,0)}, g_{(s,u,\xi-\phi',-\phi'')} \right\rangle \tag{B.37}$$

D'après l'expression analytique (A.2), utilisée avec $\overline{s}^2 = \frac{2(1/\alpha'')s^2}{(1/\alpha'')+s^2}$, $\Delta c = \phi''$ et $\mu = 1 + \frac{\overline{s}^4\phi''^2}{4}$, $\Phi(\xi)$ et $\log A(\xi)$ sont donc des polynômes d'ordre 2 en ξ, et

$$\Phi''(\xi) = -\frac{\overline{s}^4\phi''}{4\mu} \tag{B.38}$$

et

$$(\log A)''(\xi) = -\frac{\overline{s}^2}{2\mu} \tag{B.39}$$

Soit

$$z = (\log A)''(\xi) + i\Phi''(\xi) = -\frac{\overline{s}^2}{2\mu}\left(1 + i\frac{\overline{s}^2\phi''}{2}\right) \tag{B.40}$$

Comme

$$|z|^2 = (\log A''(\xi))^2 + (\Phi''(\xi))^2 = \frac{\overline{s}^4}{4\mu^2}\left(1 + \frac{1}{4}\overline{s}^4\phi''^2\right) \tag{B.41}$$

$$= \frac{\overline{s}^4}{4\mu} = -\frac{\overline{s}^2}{2}\log A''(\xi) = -\frac{\Phi''(\xi)}{\phi''} \tag{B.42}$$

on a

$$\overline{s}^2 = -2\frac{(\log A'')^2 + (\Phi'')^2}{\log A''} \tag{B.43}$$

et

$$\phi'' = -\frac{\Phi''}{(\log A'')^2 + (\Phi'')^2} \tag{B.44}$$

Pour finir il suffit de se rappeler de la définition de $\overline{s}^2$ pour établir

$$\frac{1}{(1/\alpha'')} + \frac{1}{s^2} = \frac{2}{\overline{s}^2} = -\frac{\log A''}{(\log A'')^2 + (\Phi'')^2}$$

ce qui conduit au résultat cherché

□.

B.3.1 "Corollaire de la démonstration" de la proposition 1

Corollaire 3 *Avec les hypothèses de la proposition 1 on a*

$$\begin{aligned} \left|\Phi''(\xi)\right| &\leq \frac{s^2}{2} \\ 0 < -(\log A)''(\xi) &\leq s^2 \\ \left|\Phi''(u)\right| &\leq \frac{1}{2s^2} \end{aligned}$$

Démonstration

Les formules (B.38) et (B.39) permettent immédiatement d'établir que

$$\begin{aligned} \left|\Phi''(\xi)\right| &= \frac{\left|\phi''\right| \overline{s}^2/2}{1+(\left|\phi''\right| \overline{s}^2/2)^2} \frac{\overline{s}^2}{2} \leq \frac{\overline{s}^2}{4} \\ 0 > \log A''(\xi) &\geq -\frac{\overline{s}^2}{2} \end{aligned}$$

Or $\overline{s}^2 = 2s^2 \dfrac{1}{1+s^2/(1/\alpha'')} \leq 2s^2$, d'où les deux premières inégalités. De plus, avec la même démarche que précédemment, on peut obtenir l'expression de la dérivée seconde de la phase par rapport au temps :

$$\Phi''(u) = \frac{\phi''}{\mu} \left(\frac{(1/\alpha'')}{s^2 + (1/\alpha'')} \right)^2 \tag{B.45}$$

d'où

$$\begin{aligned} \left|\Phi''(u)\right| &= \frac{\left|\phi''\right|}{1+(\left|\phi''\right| \overline{s}^2/2)^2} \frac{\overline{s}^4}{4s^4} \\ &= \frac{\left|\phi''\right| \overline{s}^2/2}{1+(\left|\phi''\right| \overline{s}^2/2)^2} \frac{\overline{s}^2}{2s^4} \\ &\leq \frac{\overline{s}^2}{4s^4} \leq \frac{2s^2}{4s^4} = \frac{1}{2s^2} \end{aligned}$$

Annexe C

Mélange de gaussiennes et information mutuelle

Ce chapitre est consacré à la démonstration des propriétés, lemmes et théorèmes utilisés et énoncés au chapitre 7.

C.1 Rappels : lois conditionnelles de bruits gaussiens

Soit w un bruit gaussien centré de matrice de covariance K. La variable aléatoire $Q_g = \langle w, g \rangle$, conditionnée par rapport à n'importe quelle famille finie de variables aléatoires $Q_{g_k} = \langle w, g_k \rangle, 1 \leq k \leq m$, a une loi gaussienne d'espérance μ_m linéaire en $\{Q_{g_k}, 1 \leq k \leq m\}$. Il existe donc des constantes $\lambda_1, \ldots, \lambda_m$ et σ_m^2 telles que :

$$\mathcal{P}(Q_g = q | Q_{g_k} = q_k, 1 \leq k \leq m) \sim \mathcal{N}\left(\sum_{k=1}^{m} \lambda_k q_k, \sigma_m^2\right) \tag{C.1}$$

Les lemmes suivant expriment les valeurs de μ_m et de σ_m^2.

Lemme 4 *(Espérance conditionnelle) L' espérance de Q_g conditionnellement à $B_m = \{Q_{g_k} = q_k, 1 \leq k \leq m\}$ est*

$$\mu_m = \mathbb{E}_m \{\langle w, g \rangle\} = \langle w, P_{m,K} g \rangle = \left\langle P_{m,K}^* w, g \right\rangle \tag{C.2}$$

où

$$\boldsymbol{V}_m = Vect\{g_k, 1 \leq k \leq m\}$$

et

- *$P_{m,K}$ est le projecteur orthogonal sur $\boldsymbol{V}_m$ relativement au produit scalaire $\langle ., . \rangle_K \stackrel{\Delta}{=} \langle ., K. \rangle$,* i.e. *le projecteur sur $\boldsymbol{V}_m$ parallèlement à*

$$\boldsymbol{V}_m^{\perp_K} = (K\boldsymbol{V}_m)^{\perp} = K^{-1}\boldsymbol{V}_m^{\perp},$$

- $P^*_{m,K}$ *est l'adjoint de* $P_{m,K}$ *pour le produit scalaire usuel,* i.e. *le projecteur sur* $K\boldsymbol{V}_m$ *parallèlement à* $\boldsymbol{V}_m^{\perp}$.

Démonstration du Lemme 4

L'espérance conditionnelle μ_m est de la forme

$$\mu_m = \mathbb{E}_m \left\{ \langle w, g \rangle \right\} = \sum_{k=1}^{m} \lambda_k \langle w, g_k \rangle , \tag{C.3}$$

où les λ_k sont caractérisés par la dé-corrélation entre $\langle w, g \rangle - \sum_{k=1}^{m} \lambda_k \langle w, g_k \rangle$ et tous les $\langle w, g_l \rangle , 1 \leq l \leq m$. La dé-corrélation

$$\forall l,\ \mathbb{E} \left\{ \left(\langle w, g \rangle - \sum_{k=1}^{m} \lambda_k \langle w, g_k \rangle \right) \langle w, g_l \rangle \right\} = 0 \tag{C.4}$$

s'écrit comme une orthogonalité au sens du produit scalaire $\langle ., K. \rangle$, *i.e.* pour tout $l, 1 \leq l \leq m$, :

$$\langle g, g_l \rangle_K = \langle g, K g_l \rangle = \left\langle \sum_{k=1}^{m} \lambda_k g_k, K g_l \right\rangle = \left\langle \sum_{k=1}^{m} \lambda_k g_k, g_l \right\rangle_K \tag{C.5}$$

c'est-à-dire que $\sum_{k=1}^{m} \lambda_k g_k$ est le projeté orthogonal, au sens de $\langle ., . \rangle_K$, de g sur $\boldsymbol{V}_m$. Donc

$$\mu_m = \left\langle w, \sum_{k=1}^{m} \lambda_k g_k \right\rangle = \langle w, P_{m,K} g \rangle = \left\langle P^*_{m,K} w, g \right\rangle \tag{C.6}$$

□.

Lemme 5 *(Variance conditionnelle) La variance conditionnelle de* Q_g *est*

$$\sigma_m^2 = \mathbb{E}_m \left\{ \left(\langle w, g \rangle - \langle w, P_{m,K} g \rangle \right)^2 \right\} = \| (Id - P_{m,K}) g \|_K^2 = \| R_{m,K} g \|_K^2 \tag{C.7}$$

où $R_{m,K} = Id - P_{m,K}$ *est le projecteur sur* $(K\boldsymbol{V}_m)^{\perp}$ *parallèlement à* $\boldsymbol{V}_m$.

Démonstration du Lemme 5

La variance conditionnelle de Q_g est par définition

$$\sigma_m^2 = \mathbb{E}_m \left\{ \left(\langle w, g \rangle - \langle w, P_{m,K} g \rangle \right)^2 \right\}. \tag{C.8}$$

Comme $\langle w, g \rangle - \langle w, P_{m,K} g \rangle$ est indépendant de $\langle w, g_k \rangle$ pour tout $1 \leq k \leq m$, on a en définitive :

$$\begin{aligned} \sigma_m^2 &= \mathbb{E} \left\{ \langle w, (Id - P_{m,K}) g \rangle^2 \right\} \\ &= \langle (Id - P_{m,K}) g, K (Id - P_{m,K}) g \rangle \\ &= \| (Id - P_{m,K}) g \|_K^2 \end{aligned}$$

Remarque

Cette loi est éventuellement dégénérée en dirac si $g \in \boldsymbol{V}_m = \text{Vect}\,(g_k)_{1\leq k\leq m}$, puisque dans ce cas Q_g est *fonction* (linéaire) de $\{Q_{g_k}, 1 \leq k \leq m\}$.

Corollaire 4 *Soit $x = f_i + w_i$ un signal, où f_i est un vecteur fixé et w_i un bruit gaussien d'opérateur de covariance K_i. L'espérance conditionnelle de $Q_g = \langle x, g\rangle$ est*

$$\begin{array}{rcccl} \mu_{m,i}[g] & \triangleq & \mathbb{E}_m\left\{\langle f_i + w_i, g\rangle\right\} & = & \langle f_i, g\rangle + \langle w_i, P_{m,K_i} g\rangle \\ & = & \langle f_i, g\rangle + \langle x - f_i, P_{m,K_i} g\rangle & = & \langle f_i, (Id - P_{m,K_i})g\rangle + \langle x, P_{m,K_i} g\rangle\,. \end{array}$$

d'où

$$\mu_{m,i}[g] = \underbrace{\langle f_i, R_{m,K_i} g\rangle}_{appris} + \underbrace{\langle x, P_{m,K_i} g\rangle}_{observable}. \tag{C.9}$$

C.2 Expression de l'information mutuelle conditionnelle

Si le processus X, conditionnellement à la classe $Y = y$, est gaussien, alors son conditionnement supplémentaire par rapport à l'événement $B_m(x)$

$$B_m(x) = \left\{Q_{g_k(x)}(X) = Q_{g_k(x)}(x), 1 \leq k \leq m\right\} \tag{C.10}$$

est également un processus gaussien. Pour tout vecteur g, la caractéristique $Q_g(X) = \langle X, g\rangle$ conditionnée par rapport à $B_m(x)$ et $Y = y$ est donc une variable aléatoire gaussienne de loi

$$\mathcal{P}_{m,y}[Q_g] = \mathcal{N}(\mu_{m,y}[g], \sigma_{m,y}[g]^2). \tag{C.11}$$

où les paramètres $\mu_{m,y}[g]$ et $\sigma^2_{m,y}[g]$ peuvent dépendre de x, par l'intermédiaire du conditionnement, comme on l'a vu à la section précédente.

Comme le mélange (7.42) ne dépend que des espérances conditionnelles $\mu_{m,y}[g]$, des variances conditionnelles $\sigma^2_{m,y}[g]$, et de la densité de mélange $p_{m,y}$, l'information mutuelle conditionnelle s'écrit

$$I(Q_g(X); Y | B_m(x)) = I\left((\mu_{m,y}[g], \sigma^2_{m,y}[g], p_{m,y})_{y\in\mathcal{Y}}\right). \tag{C.12}$$

La meilleure caractéristique est associée à l'atome $g_{m+1}(x)$ qui maximise cette expression. Pour déterminer celui-ci, on étudie le comportement de (C.12) en fonction des paramètres $\mu_{m,y}, \sigma^2_{m,y}, p_{m,y}$. On commence par en étudier les *invariances*, puis on s'intéressera a son *sens de variation*. On en déduira, lorsque c'est possible, une formulation explicite du critère à maximiser.

Afin d'alléger les notations, on omettra l'indice m de conditionnement. Dans le même but on omettra la dépendance en g des espérances μ_y et des variances σ^2_y, lorsque cela ne portera pas à confusion.

Invariances

L'entropie différentielle vérifie les propriétés suivantes [CT91]

$$\begin{aligned} H(X+\mu) &= H(X) \\ H(\alpha X) &= H(X) + \log|\alpha| \end{aligned}$$

On en déduit l'invariance de l'information mutuelle vis-à-vis des translations et dilatations : pour tout μ et tout $\alpha \neq 0$

$$I\left((\mu_y+\mu, \sigma_y^2, p_y)_{y\in\mathcal{Y}}\right) = I\left((\mu_y, \sigma_y^2, p_y)_{y\in\mathcal{Y}}\right) \quad \text{(C.13)}$$

$$I\left((\alpha\mu_y, \alpha^2\sigma_y^2, p_y)_{y\in\mathcal{Y}}\right) = I\left((\mu_y, \sigma_y^2, p_y)_{y\in\mathcal{Y}}\right). \quad \text{(C.14)}$$

L'étude de (C.12) se ramène donc à celle de

$$I\left(\left(\frac{\mu_y-\mu_{y_0}}{\sigma_{y_0}}, \frac{\sigma_y^2}{\sigma_{y_0}^2}\right)_{y\in\mathcal{Y}\setminus\{y_0\}}, (p_y)_{y\in\mathcal{Y}}\right). \quad \text{(C.15)}$$

Cas particulier : mélange de *deux* gaussiennes

Lorsqu'on a seulement *deux* classes y_0 et y_1, grâce à l'invariance

$$I\left(0, 1, p_0, \frac{\mu_1-\mu_0}{\sigma_0}, \frac{\sigma_1^2}{\sigma_0^2}, p_1\right) = I\left(0, 1, p_0, \frac{\mu_1-\mu_0}{-\sigma_0}, \frac{\sigma_1^2}{\sigma_0^2}, p_1\right) \quad \text{(C.16)}$$

et au fait que $p_1 = 1-p_0$, on peut se ramener à l'étude de

$$I\left(\left|\frac{\mu_1-\mu_0}{\sigma_0}\right|, \frac{\sigma_1^2}{\sigma_0^2}, p_0\right). \quad \text{(C.17)}$$

Expression "analytique"

On s'est ramené à l'étude de l'information mutuelle $I(\mu, \sigma, p)$ dans un mélange de deux gaussiennes $\mathcal{N}(\mu, 1)$ et $\mathcal{N}(0, \sigma^2)$ en proportions $(p, 1-p)$. La densité de probabilité associée à la loi normale $\mathcal{N}(\mu, \sigma^2)$ est $1/\sigma g(t/\sigma)$, où

$$g(t) = 1/\sqrt{2\pi}e^{-\frac{t^2}{2}}.$$

Son entropie [CT91] est indépendante de l'espérance μ et vaut

$$H(\mathcal{N}) = \frac{1}{2}\log 2\pi e\sigma^2.$$

Lorsque le mélange est en proportion $(p, 1-p)$, sa loi a pour densité

$$h(y) = pg(y-\mu) + (1-p)1/\sigma g(y/\sigma).$$

En notant $\psi(x) = x \log x$, l'information mutuelle [CT91] vaut donc

$$I(\mu, \sigma, p) = -\int \psi\left[h(y)\right] dy - \frac{p}{2} \log 2\pi e - \frac{1-p}{2} \log 2\pi e \sigma^2.$$

Afin de simplifier les calculs par la suite, on utilisera le paramètre $\alpha = \sigma^{-1}$. On a donc

$$h(y) = pg(y-\mu) + (1-p)\alpha g(\alpha y) \tag{C.18}$$

et

$$I(\mu, \alpha^{-1}, p) = -\int \psi\left[h(y)\right] dy - \frac{1}{2} \log 2\pi e + (1-p) \log |\alpha| \tag{C.19}$$

C.3 Variations de l'information mutuelle

L'écriture (C.19) nous permet de déterminer *explicitement* l'atome optimum $g_{m+1}(x)$, dans le cas du mélange de *deux* gaussiennes. Les lemmes suivants nous éclairent sur le sens de variations de l'information mutuelle dans un mélange de *deux* processus gaussiens.

Lemme 6 *Lorsque* $\dfrac{\sigma_1^2}{\sigma_0^2} = 1$*, l'information mutuelle est fonction* croissante *de* $\left|\dfrac{\mu_1 - \mu_0}{\sigma_0}\right|$.

Lemme 7 *Lorsque* $\mu_1 = \mu_0$*, l'information mutuelle est fonction*

- décroissante *de* $\dfrac{\sigma_1^2}{\sigma_0^2}$ *sur la partie* $\dfrac{\sigma_1^2}{\sigma_0^2} \leq 1$
- croissante *de* $\dfrac{\sigma_1^2}{\sigma_0^2}$ *sur la partie* $\dfrac{\sigma_1^2}{\sigma_0^2} \geq 1$

Remarques

- L'intuition laisse penser que le premier résultat doit se généraliser au cas où σ_1^2/σ_0^2 est constant, mais différent de 1. Cependant on n'en a pas établi de démonstration à ce jour.
- Le cas de plus de deux classes est problématique pour un traitement analytique complet : même dans le cas où $\sigma_y^2/\sigma_{y_0}^2$ est fixé (c'est-à-dire dans les conditions du lemme 6), I est fonction d'au moins deux variables $(\mu_y - \mu_0)/\sigma_0, i = 1, 2$, ce qui rend le critère de maximisation *a priori* malaisé.

Démontrons maintenant ces lemmes, à l'aide des méthodes usuelles d'étude des variations des fonctions.

C.3.1 Démonstration du lemme 6 : variations à α fixé

Lorsque $\alpha = \sigma_0/\sigma_1$ est fixé, on étudie le signe de la dérivée partielle $\partial_\mu I(\mu, \alpha^{-1}, p)$ pour établir le sens de variation de $I(\mu, \alpha^{-1}, p)$ en fonction de $\mu = |\mu_1 - \mu_0| / \sigma_0$.

$$\begin{aligned}
\partial_\mu I(\mu, \alpha^{-1}, p) &= -\int \partial_\mu h(y) \psi' [h(y)]\, dy = +\int p \overbrace{g'(y-\mu)}^{\partial_y g(y-\mu)} [1 + \log h(y)]\, dy \\
&\overset{(a)}{=} p \left\{ \underbrace{[g(y-\mu)\,(1+\log h(y))]_{-\infty}^{+\infty}}_{=0} - \int g(y-\mu) \frac{h'(y)}{h(y)} dy \right\} \\
&= +p \int g(y-\mu) \frac{(y-\mu) p g(y-\mu) + \alpha^2 y (1-p) \alpha g(\alpha y)}{h(y)} dy \\
&\overset{(b)}{=} p \int g(y-\mu) \frac{(y-\mu) h(y) + \left(\alpha^2 y - (y-\mu)\right)(1-p) \alpha g(\alpha y)}{h(y)} dy \\
&\overset{(c)}{=} p(1-p) \int \frac{\left((\alpha^2-1) y + \mu\right) \alpha g(\alpha y) g(y-\mu)}{h(y)} dy
\end{aligned}$$

On a utilisé en (a) une intégration par parties, en (b) on a fait apparaître $h(y)$ au numérateur, et en (c) on a utilisé la nullité de l'intégrale de la fonction impaire $(y-\mu)g(y-\mu)$.

Comme I est une fonction paire de μ, il suffit pour conclure de déterminer le signe de $\partial_\mu I$ pour $\mu > 0$. Lorsque $\alpha = 1$, on peut facilement conclure, car $\partial_\mu I(\mu, 1, p) = C(\mu, 1, p)\mu$, où

$$C(\mu, 1, p) = p(1-p) \int \frac{g(y) g(y-\mu)}{h(y)} dy \geq 0.$$

C.3.2 Démonstration du lemme 7 : variations pour $\mu = 0$

Lorsque $\mu = 0$ l'expression (C.18) de h se simplifie, et sa dérivation mène a

$$\begin{aligned}
h(y) &= p g(y) + (1-p) \alpha g(\alpha y) \\
h'(y) &= -p y g(y) - (1-p) \alpha^3 y g(\alpha y) \\
\partial_\alpha h(y) &= (1-p) g(\alpha y) [1 - (\alpha y)^2]
\end{aligned}$$

La dérivée partielle de I dans la direction de α s'écrit donc, en notant pour simplifier $I(\alpha^{-1},p)$ au lieu de $I(0,\alpha^{-1},p)$,

$$\begin{aligned}
\partial_\alpha I(\alpha^{-1},p) &= -\int \partial_\alpha h(y)\psi'[h(y)]dy + \frac{(1-p)}{\alpha} \\
&= \frac{(1-p)}{\alpha} - \int \frac{(1-p)}{\alpha}\alpha g(\alpha y)[1-(\alpha y)^2][1+\log h(y)]dy \\
&\overset{(a)}{=} \frac{(1-p)}{\alpha}\left\{1 - \int \overbrace{g(u)(1-u^2)}^{-g''(u)}\,(1+\log h(u/\alpha))\,du\right\} \\
&= \frac{(1-p)}{\alpha}\left\{1 - \underbrace{\left[-g'(u)\,(1+\log h(u/\alpha))\right]_{-\infty}^{+\infty}}_{=0} + \int -g'(u)\frac{\frac{1}{\alpha}h'(u/\alpha)}{h(u/\alpha)}du\right\} \\
&= \frac{(1-p)}{\alpha}\left\{1 + \int ug(u)\frac{1}{\alpha}\frac{-p\frac{u}{\alpha}g(u/\alpha) - (1-p)\alpha^3\frac{u}{\alpha}g(u)}{h(u/\alpha)}du\right\} \\
&= \frac{(1-p)}{\alpha}\left\{1 - \int u^2 g(u)\frac{p\frac{1}{\alpha^2}g(u/\alpha) + (1-p)\alpha g(u)}{h(u/\alpha)}du\right\}.
\end{aligned}$$

On a effectué en (a) le changement de variable $u=\alpha y$. Comme $\int u^2 g(u)du = 1$ ($g(u)$ est une densité gaussienne normalisée, donc de variance 1) on peut continuer les calculs comme suit

$$\begin{aligned}
\partial_\alpha I(\alpha^{-1},p) &= \frac{(1-p)}{\alpha}\int u^2 g(u)\frac{h(u/\alpha) - p\frac{1}{\alpha^2}g(u/\alpha) - (1-p)\alpha g(u)}{h(u/\alpha)}du \\
&= \frac{(1-p)}{\alpha}\int u^2 g(u)\frac{p\left(1-\frac{1}{\alpha^2}\right)g(u/\alpha)}{h(u/\alpha)}du \\
&= p(1-p)\left(\alpha - \frac{1}{\alpha}\right)\underbrace{\int \frac{y^2 g(y)g(\alpha y)}{h(y)}\alpha dy}_{>0}
\end{aligned}$$

On en déduit que $\partial_\alpha I(\alpha^{-1},p)$ est du signe de $\alpha - 1/\alpha$, c'est-à-dire que I est fonction croissante de α lorsque celui-ci est supérieur à 1, et décroissante sinon. En fonction de $\sigma = \alpha^{-1}$, le comportement est le même.

C.4 Démonstration du théorème 7

D'après le lemme 5, le rapport des variances conditionnelles

$$\sigma^2_{m,1}/\sigma^2_{m,0} = 1$$

est indépendant de g. L'information mutuelle (C.17) ne dépend donc de g que par l'intermédiaire de la grandeur

$$\eta_m[g] = \left| \frac{\mu_{m,1}[g] - \mu_{m,0}[g]}{\sigma_{m,0}[g]} \right|. \tag{C.20}$$

De plus, d'après le lemme 6, l'information est fonction *croissante* de cette grandeur, si bien que

$$g_{m+1}(x) = \arg\max_g I_m(Q_g(X); Y) = \arg\max_g \eta_m[g]. \tag{C.21}$$

Exprimons maintenant $\eta_m[g]$ de façon plus simple, à l'aide des lemmes 4 et 5 et de leur corollaire 4. Comme les deux classes ont même opérateur de covariance K, en utilisant (C.9) on a

$$\mu_{m,1}[g] - \mu_{m,0}[g] = \langle f_1 - f_0, R_m g \rangle = \left\langle K^{-1}(f_1 - f_0), R_m g \right\rangle_K \tag{C.22}$$

où R^m est le projecteur orthogonal, au sens du produit scalaire[1]

$$\langle ., . \rangle_K = \langle ., K. \rangle$$

parallèlement au sous-espace vectoriel

$$V_m(x) = Vect\left\{ g_1(x), \ldots, g_m(x) \right\}.$$

Par ailleurs comme $\sigma^2_{m,0}[g] = \|R_m g\|^2_K$, on a

$$\eta_m[g] = \left| \left\langle K^{-1}(f_1 - f_0), \frac{R^m g}{\|R^m g\|_K} \right\rangle_K \right| \tag{C.23}$$

La famille de vecteurs unitaires (au sens de la norme euclidienne $\|x\|_K = \langle x, Kx \rangle$ associée au produit scalaire $\langle ., . \rangle_K$)

$$u_m(x) = \frac{R^{m-1} g_m(x)}{\|R^{m-1} g_m(x)\|_K} \tag{C.24}$$

est l'orthonormalisée de Gram-Schmidt de la famille $g_m(x)$, et forme donc une base orthonormale de $\boldsymbol{V}_m(x)$, si bien que

$$\left\| P_{\boldsymbol{V}_m(x)} K^{-1}(f_1 - f_0) \right\|^2_K = \sum_{k=1}^{m} \left| \left\langle K^{-1}(f_1 - f_0), u_k(x) \right\rangle \right|^2_K \tag{C.25}$$

Le choix (C.21) se résume alors à

$$\begin{aligned} g_{m+1}(x) &= \arg\max_{g \in \mathcal{D}} \left| \left\langle K^{-1}(f_1 - f_0), \frac{R^m g}{\|R^m g\|_K} \right\rangle \right|^2_K \\ &= \arg\max_{g \in \mathcal{D}} \left| \left\langle R^m K^{-1}(f_1 - f_0), \frac{R^m g}{\|R^m g\|_K} \right\rangle \right|^2_K . \end{aligned}$$

[1] R^m est le projecteur parallèlement à V_m, sur le sous-espace $V_m^{\perp_K} = (KV_m)^\perp$.

et maximise

$$\left\| P_{\boldsymbol{V}_{m+1}(x)} K^{-1}(f_1 - f_0) \right\|_K^2 = \left\| P_{\boldsymbol{V}_{m}(x)} K^{-1}(f_1 - f_0) \right\|_K^2 + \left| \left\langle R^m K^{-1}(f_1 - f_0), \frac{R^m g}{\|R^m g\|_K} \right\rangle \right|_K^2 .$$

Les atomes $g_m(x)$ sont donc bien obtenus par un Matching Pursuit Orthogonal [Zha93] [Dav94] [PRK93] sur le signal $K^{-1}(f_0 - f_1)$.

□.

C.5 Classification active de bruits gaussiens

On démontre maintenant les trois lemmes permettant de prouver que la sélection d'atomes, pour la classification de bruits gaussiens, est véritablement active.

C.5.1 Démonstration du lemme 1

Soit (u_k) une base de diagonalisation de $K_0^{-1} K_1$

$$K_0^{-1} K_1 u_k = \lambda_k^2 u_k, \ \lambda_1^2 \leq \ldots \leq \lambda_N^2. \tag{C.26}$$

On va montrer par récurrence que pour tout m, il existe un indice $k_m(x)$ tel que

$$g_m(x) = \arg\max_g I\left(\langle X, g\rangle ; Y | B_{m-1}(x)\right) = u_{k_m(x)} \tag{C.27}$$

$\underline{m = 0}$ Comme les classes sont centrées, $\mu_{m,1}[g] = \mu_{m,0}[g] = 0$. D'après le lemme 7, le meilleur atome

$$g_1(x) = \arg\max_g I\left(0, \frac{\sigma_1^2[g]}{\sigma_0^2[g]}, p_0, p_1\right) \tag{C.28}$$

est un extremum de

$$\frac{\sigma_1^2[g]}{\sigma_0^2[g]} = \frac{\langle g, K_1 g\rangle}{\langle g, K_0 g\rangle} \tag{C.29}$$

On peut montrer à l'aide de multiplicateurs de Lagrange que les valeurs extrémales de cette expression sont atteintes lorsque g est vecteur propre de $K_0^{-1} K_1$, associé à une valeur propre extrémale. Le meilleur premier atome $g_1(x)$ est donc un vecteur propre

$$g_1(x) = u_{k_1(x)}. \tag{C.30}$$

$\underline{m \to m+1}$ D'après l'hypothèse de récurrence, comme pour $1 \leq l \leq m$ $g_l(x) = u_{k_l(x)}$ est vecteur propre de $K_0^{-1}K_1$, $\boldsymbol{V}_m(x)$ est stable par $K_0^{-1}K_1$, et l'on a donc

$$K_0 \boldsymbol{V}_m(x) = K_1 \boldsymbol{V}_m(x). \tag{C.31}$$

Les projecteurs P_{m,K_i} sur $\boldsymbol{V}_m(x)$ parallèlement à

$$(K_i \boldsymbol{V}_m(x))^\perp = K_i^{-1}(\boldsymbol{V}_m(x))^\perp$$

coïncident donc. D'après le lemme 4 on a donc

$$\mu_{m,1}[g] - \mu_{m,0}[g] = \langle x, (P_{m,K_1} - P_{m,K_0})g \rangle = 0.$$

En raisonnant comme précédemment, le meilleur atome

$$g_{m+1}(x) = \arg\max_g I\left(0, \frac{\sigma^2_{m,1}[g]}{\sigma^2_{m,0}[g]}, p_{m,0}, p_{m,1}\right) \tag{C.32}$$

est un extremum[2] de

$$\frac{\sigma^2_{m,1}[g]}{\sigma^2_{m,0}[g]} = \frac{\langle R^m g, K_1 R^m g \rangle}{\langle R^m g, K_0 R^m g \rangle}. \tag{C.33}$$

A l'aide des multiplicateurs de Lagrange, on montre que les extrema g de cette expression vérifient l'égalité

$$(R^m)^* K_1 R^m g = \lambda^2 (R^m)^* K_0 R^m g. \tag{C.34}$$

Montrons qu'alors g est vecteur propre de $K_0^{-1}K_1$. L'égalité (C.34) est vérifiée si, et seulement si, g remplit la suite de conditions suivantes

$$\begin{aligned}
\langle x, (R^m)^* K_1 R^m g \rangle &= \lambda^2 \langle x, (R^m)^* K_0 R^m g \rangle, \ \forall x & \text{(C.35)} \\
\langle R^m x, K_1 R^m g \rangle &= \lambda^2 \langle R^m x, K_0 R^m g \rangle, \ \forall x & \text{(C.36)} \\
\langle y, K_1 R^m g \rangle &= \lambda^2 \langle y, K_0 R^m g \rangle, \ \forall y \in \mathrm{Im} R^m. & \text{(C.37)}
\end{aligned}$$

Comme K_0 est auto-adjoint et

$$\mathrm{Im} R^m = K_0^{-1} \boldsymbol{V}_m^\perp(x) \tag{C.38}$$

[2] On a fait appel au lemme 5 pour écrire

$$\sigma^2_{m,i}[g] = \langle R^m g, K_i R^m g \rangle$$

où $R^m = Id - P_{m,K_i}$ est le projecteur sur $(K_i \boldsymbol{V}_m)^\perp$ parallèlement à $\boldsymbol{V}_m$.

cela équivaut à

$$\begin{aligned}
\left\langle K_0 y, K_0^{-1} K_1 R^m g \right\rangle &= \lambda^2 \left\langle K_0 y, R^m g \right\rangle \ \forall y \in K_0^{-1} \boldsymbol{V}_m^{\perp}(x) && \text{(C.39)}\\
\left\langle z, K_0^{-1} K_1 R^m g \right\rangle &= \lambda^2 \left\langle z, R^m g \right\rangle \ \forall z \in \boldsymbol{V}_m^{\perp}(x) && \text{(C.40)}\\
\left\langle z, (K_0^{-1} K_1 - \lambda^2 Id) R^m g \right\rangle &= 0 \ \forall z \in \boldsymbol{V}_m^{\perp}(x) && \text{(C.41)}\\
(K_0^{-1} K_1 - \lambda^2 Id) R^m g &\in \boldsymbol{V}_m(x) && \text{(C.42)}
\end{aligned}$$

De plus le projecteur R^m commute avec $K_0^{-1}K_1$, car son image et son noyau sont stables par $K_0^{-1}K_1$. La relation (C.42) est donc vérifiée si, et seulement si,

$$R^m (K_0^{-1} K_1 - \lambda^2 Id) g \in \boldsymbol{V}_m(x). \tag{C.43}$$

Comme $\boldsymbol{V}_m(x)$ est le noyau de R^m, cela équivaut en définitive à

$$K_0^{-1} K_1 g = \lambda^2 g \tag{C.44}$$

c'est-à-dire que g est vecteur propre de $K_0^{-1}K_1$, associé à la valeur propre λ^2. On sait donc désormais que le meilleur $m+1$-ème vecteur est

$$g_{m+1}(x) = u_{k_{m+1}(x)}. \tag{C.45}$$

Comme les m vecteurs propres $u_{k_l}(x)$, $1 \leq l \leq m$ de $K_0^{-1}K_1$ sont dans le noyau $\boldsymbol{V}_m(x)$ de R^m et n'apportent plus aucune information, $g_{m+1}(x)$ est choisi dans les vecteurs propres restants

$$k_{m+1}(x) \notin \{k_l(x), 1 \leq l \leq m\}. \tag{C.46}$$

□.

C.5.2 Démonstration du lemme 2

D'après le lemme 1, l'indice $k_{m+1}(x)$ est choisi à chaque étape selon le critère

$$k_{m+1}(x) = \arg \max_{k \notin \{k_l(x), 1 \leq l \leq m\}} I\left(0, \frac{\sigma^2_{m,1}[u_k]}{\sigma^2_{m,0}[u_k]}, p_{m,0}, p_{m,1}\right). \tag{C.47}$$

Comme l'expression (C.33) prend en $u_k \notin \boldsymbol{V}_m$ la valeur

$$\frac{\sigma^2_{m,1}[u_k]}{\sigma^2_{m,0}[u_k]} = \lambda_k^2 \tag{C.48}$$

le meilleur indice est

$$k_{m+1}(x) = \arg \max_{k \notin \{k_l(x), 1 \leq l \leq m\}} I\left(0, \lambda_k^2, p_{m,0}, p_{m,1}\right) \tag{C.49}$$

En faisant appel au lemme 7, on est amenés à distinguer trois cas :

- Si $\underline{\lambda}_m^2(x) \geq 1$, alors

$$\left\{k \notin \{k_l(x), 1 \leq l \leq m\} \,|\, \lambda_k^2 \leq 1\right\} = \emptyset \tag{C.50}$$

donc

$$\max_{k \notin \{k_l(x), 1 \leq l \leq m\}} I\left(0, \lambda_k^2, p_{m,0}, p_{m,1}\right) = I\left(0, \overline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right) \tag{C.51}$$

et $k_{m+1}(x)$ est déterminé par

$$\lambda_{k_{m+1}(x)}^2 = \overline{\lambda}_m^2(x) \tag{C.52}$$

- Si $\overline{\lambda}_m^2(x) \leq 1$, alors

$$\left\{k \notin \{k_l(x), 1 \leq l \leq m\} \,|\, \lambda_k^2 \geq 1\right\} = \emptyset \tag{C.53}$$

donc

$$\max_{k \notin \{k_l(x), 1 \leq l \leq m\}} I\left(0, \lambda_k^2, p_{m,0}, p_{m,1}\right) = I\left(0, \underline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right) \tag{C.54}$$

si bien que $k_{m+1}(x)$ est déterminé par

$$\lambda_{k_{m+1}(x)}^2 = \underline{\lambda}_m^2(x) \tag{C.55}$$

- Si $\overline{\lambda}_m^2(x) \geq 1 \geq \underline{\lambda}_m^2(x)$, alors il faut choisir entre ces deux valeurs propres extrémales, car

$$\max_{k \notin \{k_l(x), 1 \leq l \leq m\}} I\left(0, \lambda_k^2, p_{m,0}, p_{m,1}\right) = \max\left\{I\left(0, \overline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right), I\left(0, \underline{\lambda}_m^2(x), p_{m,0}, p_{m,1}\right)\right\}. \tag{C.56}$$

□.

C.5.3 Démonstration du lemme 3

On commence par établir un lemme technique, dont le lemme 3 sera un corollaire.

Lemme 8 *Si $\alpha^2 > 2/3$ alors au voisinage de $p \approx 1$, l'information mutuelle se développe comme suit*

$$I(\alpha^{-1}, p) = (1-p)\left\{3/2 + \log\alpha + \frac{1}{2\alpha^2}\right\} + \mathcal{O}((1-p)^2) \tag{C.57}$$

Démonstration

On utilise l'expression (C.19). En notant

$$t(y) = \frac{1-p}{p}\frac{\alpha g(\alpha y)}{g(y)}$$

on peut écrire

$$\begin{aligned}
\int \psi[h] &= \int h \log h = \int h \log pg + \int h \log[1+t] && \text{(C.58)}\\
&= \int h(y)[\log p + \log \frac{1}{\sqrt{2\pi}} - \frac{y^2}{2}]dy + \int h \log[1+t] && \text{(C.59)}\\
&= \log p + \log \frac{1}{\sqrt{2\pi}} - \int h(y)\frac{y^2}{2}dy + \int h \log[1+t] && \text{(C.60)}
\end{aligned}$$

car $\int h = 1$. On calcule alors les intégrales qui nous intéressent. D'abord, en utilisant les variances 1 et $1/\alpha^2$ des distributions $g(y)$ et $\alpha g(\alpha y)$, on obtient

$$\begin{aligned}
\int h(y)y^2 &= p\int g(y)y^2 + (1-p)\int \alpha g(\alpha y)y^2 && \text{(C.61)}\\
&= p + \frac{(1-p)}{\alpha^2}. && \text{(C.62)}
\end{aligned}$$

Ensuite comme pour tout $t \geq 0$ on dispose de l'encadrement

$$-t^2/2 \leq \log(1+t) - t \leq 0$$

on a

$$\begin{aligned}
\int h \log[1+t] &= \int ht - \eta(p) && \text{(C.63)}\\
&\text{où}\\
0 &\leq \eta(p) \leq \frac{1}{2}\int ht^2. && \text{(C.64)}
\end{aligned}$$

On doit donc maintenant calculer

$$\begin{aligned}
\int ht &= p\int gt + (1-p)\int \alpha g(\alpha y)t(y)dy && \text{(C.65)}\\
&\text{et}\\
\int ht^2 &= p\int gt^2 + (1-p)\int \alpha g(\alpha y)t^2(y)dy. && \text{(C.66)}
\end{aligned}$$

Pour cela on calcule, si $\alpha^2 > 1 - 1/k$,

$$\begin{aligned}
\int gt^k &= \alpha^k \left(\frac{1-p}{p}\right)^k \frac{1}{\sqrt{k\alpha^2 - (k-1)}} && \text{(C.67)}\\
&\text{et}\\
(1-p)\int \alpha g(\alpha y)t^k(y)dy &= p\int gt^{k+1}. && \text{(C.68)}
\end{aligned}$$

On est dans les cas $k = 2$ et $k = 3$, donc il suffit que $\alpha^2 > 2/3$ pour que

$$\int ht = (1-p) + \mathcal{O}((1-p)^2) \tag{C.69}$$

$$\int ht^2 = \mathcal{O}((1-p)^2). \tag{C.70}$$

En définitive

$$\int h \log[1+t] = (1-p) + \mathcal{O}((1-p)^2) \tag{C.71}$$

Comme par ailleurs

$$\log p = \log(1-(1-p)) = -(1-p) + \mathcal{O}((1-p)^2) \tag{C.72}$$

on a

$$\int \psi[h] = \log \frac{1}{\sqrt{2\pi}} - (1-p) - \frac{p}{2} - \frac{1-p}{2\alpha^2} - (1-p) + \mathcal{O}((1-p)^2) \tag{C.73}$$

et

$$I(\alpha^{-1}, p) = -\log \frac{1}{\sqrt{2\pi}} + 2(1-p) + \frac{p}{2} + \frac{1-p}{2\alpha^2} - \log\sqrt{2\pi e} \tag{C.74}$$

$$+(1-p)\log\alpha + \mathcal{O}((1-p)^2)$$

$$= (1-p)\left\{2 - 1/2 + \frac{1}{2\alpha^2} + \log\alpha\right\} + \mathcal{O}((1-p)^2) \tag{C.75}$$

d'où le résultat.

□.

Démontrons maintenant le lemme 3, sous la forme du corollaire suivant

Corollaire 5 *Si* $\alpha^2 \in]2/3, 1[$, *alors au voisinage de* $p = 1$

$$I(\alpha^{-1}, p) > I(1/\alpha^{-1}, p) \tag{C.76}$$

et

$$I(\alpha^{-1}, 1-p) < I(1/\alpha^{-1}, 1-p) \tag{C.77}$$

Démonstration

Comme on peut montrer par un changement de variable que

$$I(1/\alpha^{-1}, 1-p) = I(\alpha^{-1}, p),$$

si p vérifie (C.76), alors (C.77) sera également vérifiée. Maintenant comme $\alpha^2 > 2/3$ et $1/\alpha^2 > 2/3$, on utilise le lemme précédent pour établir que

$$\lim_{p\to 1} \frac{I(\alpha^{-1}, p) - I(1/\alpha^{-1}, p)}{1-p} = 2\log\alpha + \frac{1}{2\alpha^2} - \frac{\alpha^2}{2} \tag{C.78}$$

Une brève étude de fonction montre que

$$2\log\alpha + \frac{1}{2\alpha^2} - \frac{\alpha^2}{2} > 0 \Longleftrightarrow \alpha < 1 \tag{C.79}$$

Pour $\alpha^2 \in]2/3, 1[$, si p est suffisamment proche de 1, on a donc bien (C.76), d'où le résultat.

□.

Bibliographie

[AG97] Y. Amit et D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9 :1545–1588, 1997. 113, 119, 136, 160

[AGW97] Y. Amit, D. Geman et K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(11) :1300–1305, novembre 1997. 113, 136, 145, 160

[AM99] Y. Amit et A. Murua. Speech recognition using randomized relational decision trees. Rapport technique 487, Department of Statistics, University of Chicago, avril 1999. Http ://galton.uchicago.edu/ amit/Papers/sound.ps.gz. 136, 160

[AO95] A. Antoniadis et G. Oppenheim, rédacteurs. *Wavelets and Statistics*, chapitre R. Carmona. Extrema reconstruction and spline smooting : variations on an algorithm of Mallat and Zhong, pages 96–108. Springer-Verlag, Berlin, 1995. 132

[BCG94] J. Berger, R. Coifman et M. Goldberg. A method of denoising and reconstructing audio signals. Dans *Proc. Int. Computer Music Conf. (ICMC'94)*, pages 344–347. septembre 1994. 106

[BCR91] G. Beylkin, R. Coifman et V. Rokhlin. Fast wavelet transforms and numerical algorithms. *Commun. on Pure and Appl. Math.*, 44 :141–183, 1991. 27

[Ber95] F. Bergeaud. *Représentations adaptatives d'images numériques,* Matching Pursuit. Thèse de doctorat, Ecole Centrale Paris, 1995. 18, 37, 41, 63, 66, 77

[BFOS84] L. Breiman, J. Friedman, R. Olshen et C. Stone. *Classification And Regression Trees.* Chapman & Hall, 1984. 20, 114, 121, 140

[BJ93a] R. G. Baraniuk et D. L. Jones. Shear madness : New orthonormal bases and frames using chirp functions. *IEEE Trans. Signal Process.Special Issue on Wavelets in Signal Procesing*, 41(12) :3543–3548, décembre 1993. 28

[BJ93b] R. G. Baraniuk et D. L. Jones. Signal-dependent time-frequency analysis using a radially gaussian kernel. *Signal Process.*, 32(3) :263–284, juin 1993. 95

[BJ93c] R. G. Baraniuk et D. L. Jones. A signal-dependent time-frequency representation : Optimal kernel design. *IEEE Trans. Signal Process.*, 41(4) :1589–1602, avril 1993. 95

[Bla98] G. Blanchard. The "progressive mixture" estimator for regression trees. Rapport technique, Ecole Normale Supérieure, Département de Mathématiques et Informatique, mai 1998. Available at http ://www.dmi.ens.fr/preprints. 141

[BM96] F. Bergeaud et S. Mallat. Matching pursuit : adaptive representations of images and sounds. *Journal of Computational and Applied Mathematics*, 1996. 63

[BM97] L. Birgé et P. Massart. *Festschrift for Lucien Le Cam*, chapitre From model selection to adaptive estimation, pages 55–87. Springer, New York, 1997. 114

[Bou91] P. Boulez. “Pierre Boulez”. Erato Disques, 1991. 56, 57

[BS95] A. J. Bell et T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7 :1129–1159, 1995. 118

[Bul95] A. Bultan. *A Four-Parameter atomic decomposition and the related time-frequency distribution.* Thèse de doctorat, Middle East Technical University, décembre 1995. 18, 71, 71, 76, 76, 76

[Bul99] A. Bultan. A four-parameter atomic decomposition of chirplets. *IEEE Trans. Signal Process.*, 47(3) :731–745, mars 1999. 18, 71, 71, 76, 76, 76

[CBB+97] G. Ciuperca, L. Bellanger, M. Bobbia, D. Dacunha-Castelle, P. Jackubowicz, G. Oppenheim et R. Tomassone. Prévision de l’ozone en région parisienne. Dans *XXIXème Journées de Statistique*, pages 265–266. Carcassonne, mai 1997. 114

[CD95] S. Chen et D. Donoho. Atomic decomposition by basis pursuit. Rapport technique, Statistics Department, Stanford University, 1995. 17, 19, 33, 63, 106

[CHT95a] R. A. Carmona, W. L. Hwang et B. Torresani. Characterization of signals by the ridges of their wavelet transforms. Rapport technique, Centre de Physique Théorique - CNRS - Luminy, 1995. 93, 160

[CHT95b] R. A. Carmona, W. L. Hwang et B. Torresani. Identification of chirps with continuous wavelet transform. Rapport technique, Centre de Physique Théorique - CNRS - Luminy, 1995. 93, 160

[CM91] R. R. Coifman et Y. Meyer. Remarques sur l'analyse de fourier à fenêtre. *Comptes Rendus Acad. Sci. Paris(A)*, 312 :259–261, 1991. 17, 27, 28, 29, 31

[CM97] M. Clerc et S. Mallat. Identification de processus localement dilatés. Dans *XXIXème Journées de Statistique*, pages 267–270. Carcassonne, mai 1997. 17

[Com94] P. Comon. Independent component analysis, a new concept ? *Signal Process.*, 36 :287–314, 1994. 19, 116

[CT91] T. M. Cover et J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley-Interscience, 1991. 116, 117, 184, 184, 185

[CW92] R. Coifman et M. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2) :713–718, mars 1992. 17, 29, 29, 118

[Dal93] M.-A. Dalbavie. "Marc-André Dalbavie". Dans *compositeurs d'aujourd'hui.* IRCAM, 19991-93. 99

[Dau88] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. on Pure and Appl. Math.*, 41 :909–996, novembre 1988. 28

[Dau92] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 1992. 90

[Dav94] G. Davis. *Adaptive Nonlinear Approximations.* Thèse de doctorat, New York University, septembre 1994. 33, 35, 45, 63, 189

[DEG+92] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian et B. Torrésani. Asymptotic wavelet and gabor analysis : Extraction of instantaneous frequency. *IEEE Trans. Inform. Theory*, 38(2) :644–664, mars 1992. 30, 78

[Del92] N. Delprat. *Analyse Temps-Fréquence de Sons Musicaux : Exploration d'une Nouvelle Méthode d'Extraction de Données Pertinentes pour un Modèle de Synthèse.* Thèse de doctorat, Univ. d'Aix-Marseille II, Institut de Mécanique des Fluides, avril 1992. 30, 78

[DeV98] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998. 17, 28, 35

[DJ94] D. L. Donoho et I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus Acad. Sci. Paris Série I*, 319 :1317–1322, 1994. 17, 152

[DMA] G. Davis, S. Mallat et M. Avellaneda. Adaptive nonlinear approximations. Rapport technique, NY University, Courant Institute. 45

[DMA97] G. Davis, S. Mallat et M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13 :57–98, 1997. 35

[DMvS97] D. L. Donoho, S. Mallat et R. von Sachs. Estimating covariances of locally stationnary processes : Rate of convergence of best basis methods. Rapport technique, Dept of Statistics, Stanford University, 1997. 17

[DO96] G. Deco et D. Obradovic. *An Information-Theoretic Approach to Neural Computing*. Perspectives in Neural Computing. Springer, 1996. 118

[Dov94] B. Doval. *Estimation de la fréquence fondamentale des signaux sonores*. Thèse de doctorat, Université de Paris VI, 1994. 48, 48, 49, 160

[DT96] S. Dubnov et N. Tishby. Influence of frequency modulating jitter on higher order moments of sound residual with applications to synthesis and classification. Dans *Proc. Int. Computer Music Conf. (ICMC'96)*, page 378 385. Hong-Kong, 1996. 19

[EMS97] F. Esposito, D. Malerba et G. Semeraro. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5) :476–491, may 1997. 122, 141, 142

[Fla93] P. Flandrin. *Temps-Fréquence*. Hermes, Paris, France, 1993. 31, 46, 72, 95

[Fle62] H. Fletcher. Normal vibration frequencies of a stiff piano string. *J.A.S.A.*, 36(1) :203–209, 1962. 48

[Fuk72] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Electrical Science. Academic Press, 1972. 20, 116, 119

[GBM+96] R. Gribonval, E. Bacry, S. Mallat, P. Depalle et X. Rodet. Analysis of sound signals with high resolution matching pursuit. Dans *Proc. IEEE Conf. Time-Freq. and Time-Scale Anal. (TFTS'96)*, pages 125–128. Paris, juin 1996. 19, 106

[GDR93] G. García, P. Depalle et X. Rodet. Tracking of partial for additive sound synthesis using hidden markov models. Dans *Proc. Int. Computer Music Conf. (ICMC'93)*, pages 94–97. 1993. 160

[GDR+96] R. Gribonval, P. Depalle, X. Rodet, E. Bacry et S. Mallat. Sound signals decomposition using a high resolution matching pursuit. Dans *Proc. Int. Computer Music Conf. (ICMC'96)*, pages 293–296. août 1996. 19, 106

[GJ96] D. Geman et B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(1) :1–14, janvier 1996. 119, 122, 160

[GKM96] P. Guillemain et R. Kronland-Martinet. Characterization of acoustic signals through continuous linear time-frequency representations. *Proceedings of the IEEE*, 84(4) :561–585, avril 1996. Special issue on time-frequency and time-scale analysis. 30

[GMSV98] I. Guyon, J. Makhoul, R. Schwartz et V. Vapnik. What size test set gives good error rate estimates ? *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1) :52–64, janvier 1998. 122

[Goo97] M. Goodwin. Matching pursuit with damped sinusoids. Dans *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'97)*. 1997. 37, 41, 42, 160

[Gre75] J. M. Grey. *An exploration of musical timbre using computer-based techniques for analysis, synthesis and perceptual scaling*. Thèse de doctorat, Stanford University, 1975. 129

[Gri95] R. Gribonval. *Application de l'algorithme de Matching Pursuit Haute Résolution à l'analyse temps-fréquence des signaux sonores*. Mémoire de DEA, Université Paris VI, juillet 1995. 63, 68, 77, 106

[Gri96] R. Gribonval. *Approximations adaptatives de signaux sonores*. Mémoire de magistère, Université Paris VI, novembre 1996. 63, 68, 77

[HM89] B. Hummel et R. Moniot. Reconstruction from zero-crossings in scale-space. *IEEE Trans. Acoust. Speech Signal Process.*, 37(12), décembre 1989. 132

[HT91] M. Holschneider et P. Tchamitchian. Pointwise analysis of riemman's "non differentiable" function. *Inventiones Mathematicae*, (105) :157–176, 1991. 131

[Hub85] P. J. Huber. Projection pursuit. *The annals of statistics*, 13(2) :435–475, 1985. 35, 37, 39

[Jaf91] S. Jaffard. Pointwise smoothness, two microlocalization and wavelet coefficients. *Publucacions Matemàtiques*, 35 :155–168, 1991. 131

[JB95] D. L. Jones et R. G. Baraniuk. An adaptive optimal-kernel time-frequency representation. *IEEE Trans. Signal Process.*, 43(11) :2361–2371, octobre 1995. Http ://www.ece.rice.edu/ baraniuk/publications/pub/runrgk3.ps.Z. 95

[JCMW95] S. Jaggi, W. Carl, S. Mallat et A. Willsky. A fine scale version of the matching pursuit algorithm. Rapport technique, MIT, novembre 1995. 106

[JN84] N. J. Jayant et P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, Englewods-Cliffs, NJ, 1984. 17

[Jon87] L. Jones. On a conjecture of Huber concerning the convergence of PP-regression. *The Annals of Statistics*, 15 :880–882, 1987. 34, 39

[Kal99] J. Kalifa. *Minimax restoration and deconvolution in mirror wavelet bases.* Thèse de doctorat, Ecole Polytechnique, France, mai 1999. 17

[KJ96] H. K. Kwok et D. L. Jones. Improved FM demodulation in a fading environment. Dans *Proc. IEEE Conf. Time-Freq. and Time-Scale Anal. (TFTS'96)*, pages 9–12. Paris, juin 1996. 90

[KMG96] R. Kronland-Martinet et P. Guillemain. Ridges associated to continuous linear time-frequency representations of asymptotic and transients signals. Dans *Proc. IEEE Conf. Time-Freq. and Time-Scale Anal. (TFTS'96)*, pages 177–180. Paris, juin 1996. 30

[Knu98] D. E. Knuth. *The Art of Computer Programming : Sorting and Searching*, tome 3 de *Art of Computer Programming.* Addison-Wesley Pub Co, deuxième édition, juin 1998. 65

[LL99] B. Liu et S.-F. Ling. On the selection of informative wavelets for machine diagnosis. *Journal of Mechanical Systems and Signal Processing*, 13(1) :145–162, 1999. (ID mssp.1998.0177). 19, 20, 118, 123, 123

[MA86] J. S. Marques et L. B. Almeida. A background for sinusoid based representation of voiced speech. Dans *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'86)*, pages 1233–1236. Tokyo, 1986. 87

[MA89] J. S. Marques et L. B. Almeida. Frequency-varying sinusoidal modeling of speech. *IEEE Trans. Speech and Audio Process.*, 37(5) :763–765, mai 1989. 87, 159, 163

[Mal89] S. Mallat. A theory for multiresolution signal decomposition : the wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 674–693, juillet 1989. 27

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998. 29, 54, 72, 79, 132, 166

[MB96] P. Masri et A. Bateman. Improved modelling of attack transients in music analysis-resynthesis. Dans *Proc. Int. Computer Music Conf. (ICMC'96)*, pages 100–103. Hong-Kong, 1996. 106

[MC97] M. McClure et L. Carin. Matching pursuits with a wave-based dictionary. *IEEE Trans. Signal Process.*, 45(12) :2912–2927, décembre 1997. 31, 106

[MD92] C. McIntyre et D. Dermott. A new fine-frequency estimation algorithm based on parabolic regression. Dans *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'92)*, pages 541–544. 1992. 88

[Men91] J. M. Mendel. Tutorial on higher-order statistics (spectra) in signal processing and system theory. *Proceedings of the IEEE*, 79(3), juillet 1991. 19

[Mey94] Y. Meyer. *Les Ondelettes : algorithmes et applications.* Acquis Avancés de l'Informatique. Armand Colin, deuxième édition, 1994. 132

[MH91] S. Mallat et W. Hwang. Singularity detection and procesing with wavelets. Rapport technique, Courant Institute of Mathematical Science, New York University, New York, NY 10012, mars 1991. 132

[MH95] S. Mann et S. Haykin. The chirplet transform : Physical considerations. *IEEE Trans. Signal Process.*, 43(11) :2745–2761, novembre 1995. Http ://www.wearcam.org/chirplet/chirplet.html. 71

[Moo78] J. Moorer. The use of the phase vocoder in computer music applications. *Journal of the AES*, (26) :42–45, 1978. 106

[MZ92] S. Mallat et S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 40 :2464–2482, juillet 1992. 132

[MZ93] S. Mallat et Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12) :3397–3415, décembre 1993. 17, 18, 31, 33, 34, 39, 40, 46, 118

[NP94] J.-P. Nadal et N. Parga. Nonlinear neurons in the low-noise limit : a factorial code maximizes information transfer. *Network : Computation in Neural Systems*, 5(4) :565–581, novembre 1994. 118

[Oud98] M. Oudot. *Analyse/synthese des signaux de parole a partir d'un modele de sinusoides et de bruit. Application au codage bas debit et aux transformations prosodiques.* Thèse de doctorat, ENST-Paris, 1998. 160

[Pap86] A. Papoulis. *Signal analysis.* Mcgraw-Hill Book Co., 1986. 123, 163

[Pap87] A. Papoulis. *The fourier integral and its applications.* Mcgraw-Hill Publisher Co., 1987. 80, 163

[Pea91] E. Pearson. *The Multiresolution Fourier Transform and its application to Polyphonic Audio Analysis.* Thèse de doctorat, University of Warwick, septembre 1991. 30

[PRK93] Y. Pati, R. Rezaiifar et P. Krishnaprasad. Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition. Dans *Proceedings of the* 27^{th} *Annual Asilomar Conf. on Signals, Systems and Computers.* novembre 1993. 35, 189

[QC94] S. Qian et D. Chen. Signal representation using adaptive normalized gaussian functions. *Signal Process.*, 36(1) :1–11, 1994. 31, 71, 73

[Ris83] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2) :416–431, 1983. 114, 160

[Rod80] X. Rodet. Time-domain formant-wave functions synthesis. Dans J. Simon, rédacteur, *Spoken Language Generation and Understanding*, C : Mathematical and Physical Sciences, chapitre 4-Speech Synthesis, pages 429–441. D. Reidel Publishing Company, 1980. 95

[RW92] L. Rejtö et G. Walter. Remarks on projection pursuit regression and density estimation. *Stochastic Analysis and Applications*, 10(2) :213–222, 1992. 39

[RY96] J. Rissanen et B. Yu. *Learning and Geometry : Computational Approaches*, chapitre MDL Learning, pages 3–19. Birkhäuser, 1996. 114

[Sai94] N. Saito. *Local Feature Extraction and Its Application Using a Library of Bases.* Thèse de doctorat, Yale University, décembre 1994. 19, 118

[Sai98] N. Saito. Least statistically-dependent basis and its application to image modeling. Dans A. Laine, M. Unser et A. Aldroubi, rédacteurs, *Wavelet Applications in Signal and Image Processing*, tome 3458 de *Proc. SPIE.* San Diego CA., juillet 1998. 19, 20, 118, 124

[SC94] N. Saito et R. Coifman. Local discriminant bases. Dans A. Laine et M. Unser, rédacteurs, *Mathematical Imaging : Wavelet Applications in Signal and Image Processing*, tome 2303 de *Proc. SPIE.* 1994. 19, 20, 118

[Sco92] D. W. Scott. *Multivariate Density Estimation : Theory, Practice and Visualization.* John Wiley & Sons, New York, 1992. 114, 122

[Shr98] A. Shrijver. *Theory of Linear and Integer Programming.* John Wiley, 1998. 33

[Tem98] V. Temlyakov. The best m-term approximation and greedy algorithms. *Advances in Comp. Math.*, (8) :249–265, 1998. 35

[Tem99a] V. Temlyakov. Universal bases and greedy algorithms. Rapport technique 9908, Dept of Mathematics, University of South Carolina, Columbia, SC 29208, 1999. Http ://www.math.sc.edu/ imip/99papers/9908.ps. 35

[Tem99b] V. Temlyakov. Weak greedy algorithms. Rapport technique 9903, Dept of Mathematics, University of South Carolina, Columbia, SC 29208, 1999. Http ://www.math.sc.edu/ imip/99papers/9903.ps. 35, 39

[Vap95] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, septembre 1995. 114

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, septembre 1998. 114

[Vio95] P. A. Viola. *Alignment by Maximization of Mutual Information*. Thèse de doctorat, Massachusetts Institute of Technology, 1995. 122

[VK95] M. Vetterli et J. Kovacevic. *Wavelets and Subband Coding*. Prentice-Hall, Englewods-Cliffs, NJ, 1995. 17

[VSS95] P. A. Viola, N. N. Schraudolph et T. J. Sejnowski. Empirical entropy manipulation for real-world problems. Dans M. M. David S. Touretzky et M. Perrone, rédacteurs, *Advances in Neural Information Processing*, tome 8. MIT Press, Cambridge, Denver 1995, 1995. 122

[WG98] G. Watson et K. Gilholm. Signal and image feature extraction from local maxima of generalized correlation. *Pattern Recognition*, 31(11) :1733–1745, 1998. 32, 73, 76, 77, 93

[Wic91] M. V. Wickerhauser. Fast approximate factor analysis. Dans *Curves and Surfaces in Computer Vision and Graphics*, tome 1610 de *Proc. SPIE*, pages 23–32. octobre 1991. 118

[YP86] A. Yuille et T. Poggio. Scaling theorems for zero crossings. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8, janvier 1986. 132

[Zha93] Z. Zhang. *Matching Pursuit*. Thèse de doctorat, New York University, juillet 1993. 35, 189