THÈSE

présentée pour obtenir le titre de

Docteur de l'École Nationale Supérieure des Télécommunications
Spécialité: Signal et son

Thierry BLU

Bancs de filtres itérés en fraction d'octave
Application au codage de son

soutenue le 1er Avril 1996, devant le jury composé de

Martin VETTERLI	*Rapporteur et Président*
Stéphane MALLAT	*Rapporteur*
Alain LE GUYADER	*Examinateur*
Michel LEVER	*Examinateur*
Pierre DUHAMEL	*Directeur de thèse*

Table des matières

Introduction

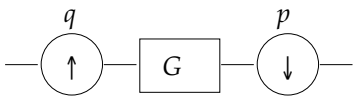
Cette thèse s'intéresse à un outil particulier de traitement du signal: le banc de filtres à taux d'échantillonnage fractionnaire, que l'on appellera plus brièvement, “banc de filtres rationnel”. Cet outil est une transformation qui peut s'appliquer à tout signal à temps discret, c'est -à dire dans la pratique à tout signal à temps continu échantillonné.

Ce que l'on entend par banc de filtres rationnels est en fait une simple extension des bancs de filtres que l’on trouve classiquement dans la littérature de traitement de signal et que l'on appellera désormais “banc de filtres dyadiques” ou “banc de filtres entiers”. On définit deux classes de bancs de filtres: les bancs de filtres d’analyse (qui associent plusieurs sorties pour une seule entrée) et les bancs de filtres de synthèse (qui associent plusieurs entrées pour une sortie). Tout banc de filtres est constitué de “branches” dont les opérations élémentaires sont (pour les notations graphiques utilisées, voir la fin de ce chapitre)

- le sur-échantillonnage 

- le filtrage 

- le sous-échantillonnage 

pour donner la branche 

Les bancs de filtres sont bien connus dans le traitement du signal depuis le début des années 1970 qui a marqué le passage des signaux analogiques aux signaux numériques, avec entre autres la banalisation et la montée en puissance des processeurs de traitement de signal. L’intérêt des bancs de filtres était évident dans la mesure où les signaux véhiculent en général l’information pertinente en fréquence et qu’il est donc utile de bien pouvoir décomposer leur spectre, et d’autre part car ils ne modifient pas le débit numérique total —du moins pour les bancs de filtres à échantillonnage critique—.

Une première “espèce” de bancs de filtres a vu le jour au milieu des années 1970 [CEG]: il s’agissait des “filtres miroirs en quadrature” (QMF), bancs de deux bandes sous-échantillonnées chacune par 2 dont la particularité était d’annuler complètement le repliement de spectre quand on utilisait les mêmes filtres lors de la reconstruction. La composition d’un banc de filtres QMF d’analyse avec son miroir était donc un simple filtrage dont on pouvait à loisir rendre la réponse fréquentielle aussi plate que possible [Jo].

En parallèle dans les années 1980 on a eu de moins en moins tendance à se contenter de la transformée de Fourier glissante pour analyser les signaux non stationnaires, pour s’intéresser aux transformées “à Q constant” c’est-à dire à résolution fréquentielle dépendant linéairement de la valeur locale de la fréquence. Il s’agit là en particulier de l’introduction de la transformée en ondelettes continue [GGM].

Le camp de la transformée en ondelettes et celui du banc de filtres se réunirent finalement quand il fut mis en évidence par Y. Meyer et S. Mallat qu’un banc de filtres itérés en octave était simplement la version discrétisée d’une transformée en ondelettes continue [Ma1]. La théorie unificatrice était appelée “analyse multirésolution” et proposait une vision très simple du banc de filtres itérés en octave [Ma1,Mey1]. À ce stade les résultats accumulés dans les

deux domaines —discret et continu— ont été mis en commun, permettant dans un cas de construire des bases d'ondelettes biorthogonales —ou tout simplement orthogonales— et d'inverser sous cette forme une transformée en ondelettes discrète [CDF], de produire des algorithmes rapides de calcul de transformées en ondelettes discrètes [HKMT,RD1,She], et dans le camp banc de filtres, de faire intervenir une nouvelle propriété du banc de filtres, la régularité, dont l'estimation a fait couler beaucoup d'encre [Dau1,DauL1,DauL2,Ri1,Ri2].

En fait les deux camps se séparaient sur le problème de l'utilisation de la transformation: les premiers voyaient dans la transformée en ondelettes un outil précieux d'analyse [KMG], alors que les seconds étaient plus intéressés par le banc de filtres dans un but de codage et de compression. On comprend que dans ce cas, la régularité n'ait pas la même signification pour tout le monde…

Le gros problème de l'unification de ces deux domaines est qu'elle n'est pas complète. En effet, celle-ci se fait seulement pour des facteurs d'échelle entiers —en particulier 2—, ce qui oblige, si l'on souhaite implémenter une transformation dont le rapport $\Delta f/f$ soit plus proche de un, à décomposer l'octave en voix [RD1] c'est-à dire à décomposer la branche passe-haut (non-itérée) en autant de filtres.

C'est à cette solution peu naturelle qu'est censée répondre l'utilisation de bancs de filtres rationnels. Bien sûr, rien ne s'oppose à étendre la notion d'analyse multirésolution au cas de facteurs non entiers, du moins tant que ces facteurs sont fractionnaires [Au]. Il a cependant pu être montré [CD,Ko,KV1,KV3] que dans le cas non entier, les ondelettes ont alors nécessairement un support non borné, les filtres correspondants étant à réponse impulsionnelle infinie. Afin d'obtenir des transformations de rapport fractionnaire, Kovačević et Vetterli se sont intéressés à l'itération d'un banc de filtres rationnel [KV3] et ont conclu qu'il n'existait pas de fonction limite associée à ces schémas. Cette affirmation fut le départ de cette thèse car il devait s'avérer que la situation était en fait plus complexe, qu'il devenait nécessaire de perdre la propriété —implicite— d'invariance par translation ce qui mettait en évidence non plus une fonction limite, mais une infinité dénombrable [Blu1].

Pour mener à bien ce travail, il était d'abord nécessaire de prendre connaissance des recherches déjà existantes sur le sujet. Elles se sont avérées très succinctes [Bi,Hsi] montrant la possibilité de transformer le banc de filtres rationnel en un banc de filtres classique, insistant sur l'intérêt de l'utilisation de telles transformations pour l'interpolation fractionnaire [CR,Vai2], ou proposant un algorithme de conception de filtres [NBS1,NBS2]. Ce sont en fait essentiellement les articles de Kovačević et Vetterli [Ko,KV1,KV2,KV3] qui m'ont été les plus utiles, puisqu'ils s'attaquaient effectivement à bras le corps au problème très spécifique que constituent les bancs de filtres rationnels et leur conception.

Les bancs de filtres entiers et rationnels partagent un certain nombre de propriétés communes que je vais rappeler

- ce sont des transformations linéaires
- ce sont des transformations qui associent plusieurs sorties pour une seule entrée (bancs de filtres d'analyse), ou bien une seule sortie pour plusieurs entrées (bancs de filtres de synthèse)

- à la différence des simples opérations de convolution (filtrage), ces transformations ne sont pas à invariance temporelle, mais vérifient tout de même une sorte de cyclostationnarité que l'on précisera dans le chapitre I. On verra également dans le chapitre II cette fois que si l'on met les échantillons sous forme vectorielle le banc de filtres rationnel se réduit à une filtrage matriciel

- pourvu que le banc de filtres vérifie une condition dite d'échantillonnage critique, ou surcritique, il pourra en général être inversé. Cependant, dans le cas général, l'inverse du banc de filtres d'analyse ne prend pas la forme d'un banc de filtres de synthèse, sauf dans certains cas (deux bandes, itération de deux bandes, etc…)

Cette dernière restriction nous incitera à regarder plus précisément la forme de l'inversion. Comme cela avait été indiqué par Hoang et Vaidyanathan [HV], l'inverse d'un banc de filtres non uniforme ne prend pas nécessairement la forme miroir de celle du banc de filtres d'analyse. C'est évidemment encore plus fréquemment le cas dans le cas rationnel. Une façon de résoudre le problème qui surgit alors —quelle est donc la forme de l'inverse?— est d'introduire les bancs de filtres généralisés (chapitre I): on constatera ainsi que les bancs de filtres rationnels ne sont pas la forme la plus générale des transformations linéaires à invariance cyclique. On précisera comment construire toutes ces transformations à l'aide de bancs de filtres généralisés, eux-même construits à partir de "branches généralisées". À l'aide de ce type de transformation, on pourra alors résoudre notre problème d'inversion. Cependant, dans le reste de ce document on en restera au cas des bancs de filtres non-généralisés et l'on se restreindra même à l'étude des bancs de filtres itérés.

Les bancs de filtres itérés sont en effet l'objet principal de ce travail de thèse, et l'on s'attachera à mettre constamment en parallèle les propriétés communes avec les bancs de filtres itérés en octave. On pourra ainsi mettre en évidence l'existence de fonctions limites comme dans le cas dyadique. L'originalité du cas rationnel est que ces fonctions ont perdu la propriété d'invariance par translation comme on l'a dit plus haut, ce qui a pour principale conséquence de compliquer les calculs. C'est cette propriété perdue que l'on désigne sous le vocable "amnésie" —ou "shift error" en version anglaise— des fonctions limites.

On montrera pourtant que, en gros, la plupart des résultats du cas dyadique s'étendent au cas rationnel de façon assez naturelle en particulier en ce qui concerne la régularité au sens de Hölder [BR,RB] qui se confirmera être mieux adaptée que la régularité au sens de Sobolev pour l'étude de ces fonctions. En particulier, le banc de filtres itéré est le pendant discret d'une transformation à temps continu que l'on précisera et dont on verra qu'elle "ressemble" à une transformée en ondelettes —on parlera de pseudo-ondelette— quand l'amnésie est suffisamment petite. On indiquera comment calculer cette amnésie et à quelles propriétés une faible amnésie est reliée. On verra également qu'à ces propriétés qui concernent les ondelettes limites, c'est-à dire des fonctions continues, on peut relier des propriétés qui concernent cette fois directement les bancs de filtres.

Enfin on s'intéressera à la conception de bancs de filtres d'analyse-synthèse dans le but de réaliser une transformation crédible en fractions d'octave pour certains signaux qui nécessitent une telle résolution. C'est d'ailleurs l'une des motivations originelles des études sur les bancs

de filtres itérés en fractions d'octave. Rappelons que les bancs de filtres dyadiques ne présentent pas une finesse suffisante pour analyser efficacement —au sens psychoacoustique— des signaux sonores destinés à être captés par l'oreille humaine [ZF]. Les recherches en psychoacoustique ont en effet mis en évidence une bande de cohérence associée à chaque fréquence, la bande de Bark, dont la largeur est approximativement d'un tiers d'octave pour les fréquences supérieures à 500 Hz [ZF]. Ainsi l'oreille humaine pourrait être efficacement modélisée, en ce qui concerne les effets linéaires et pour les fréquences supérieures à 500 Hz, par un banc de filtres itérés en tiers d'octave. Le dernier chapitre s'étendra un peu sur les résultats classiques de psychoacoustique qui permettent de comprendre comment il faut coder les sorties de notre transformation en tiers d'octave afin de minimiser la perte d'information subjective dans le processus de compression. À cette occasion, on proposera un nouveau modèle de masquage psychoacoustique à base de transformation en ondelettes qui permet de mieux prendre en compte les qualités dynamiques de la perception auditive.

Notations—Définitions

La fonction "partie entière" qui retourne le plus grand entier inférieur ou égal à un réel donné x sera notée $\mathrm{E}(x)$, tandis que la partie fractionnaire résiduelle sera notée $[x]$. Quand on aura besoin d'utiliser des nombres complexes, on désignera le complexe conjugué d'un nombre z par $\bar{z}$.

On note $\binom{x}{p}$ le nombre

$$\binom{x}{p} = \begin{cases} 1 \text{ si } p = 0 \\ \dfrac{x(x-1)\ldots(x-p+1)}{p!} \text{ sinon} \end{cases}$$

défini pour tout entier positif ou nul p, et tout complexe x.

À toute suite de réels x_n on associera sa transformée en z

$$X(z) = \sum_n x_n z^n$$

Cependant, quand le nombre d'indices et d'exposants sera devenu important, ou quand l'indice courant prendra une expression trop longue, on notera $x_n = x[n]$ de façon à préserver un peu de l'intelligibilité des équations. D'autre part les lettres minuscules sont consacrées aux valeurs "échantillon", tandis que les majuscules sont réservées aux transformées en z associées à ces valeurs "échantillon".

On va maintenant définir les opérateurs qui permettent de construire les bancs de filtres.

Interpolateur

Il s'agit de l'opérateur de sur-échantillonnage entier pour les signaux à temps discret. Soit N un nombre entier positif, l'interpolateur de facteur N associe à la suite x_n, la suite y_n ainsi définie

$$\begin{aligned} y_{nN} &= x_n \\ y_{nN+n_0} &= 0 \quad \text{pour } n_0 = 1..N-1 \end{aligned}$$

En terme de transformée en z, cette définition devient

$$Y(z) = X(z^N)$$

Graphiquement, cet opérateur se représente par 

.

Décimateur

Il s'agit de l'opérateur de sous-échantillonnage entier pour les signaux à temps discret. Soit N un nombre entier positif, cet opérateur transforme la suite x_n en $y_n = x_{nN}$ ce qui donne pour la transformée en z

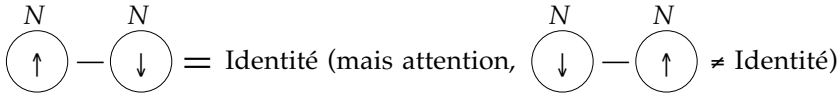
$$Y(z^N) = \frac{1}{N}\sum_{k=0}^{N-1} X(ze^{-2i\pi\frac{k}{N}})$$

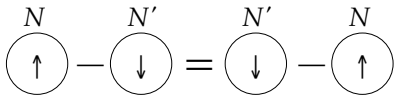
Graphiquement, cet opérateur se représente par 

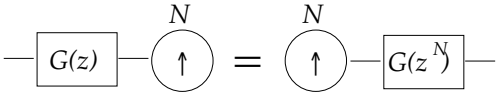
.

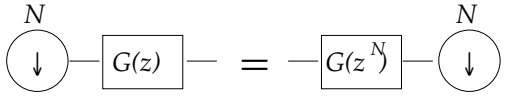
Lois de composition

Ces opérateurs, combinés avec le filtrage présentent certaines propriétés que l'on appelle lois de composition. Elles sont rappelées ci-dessous [Vet,KV3]

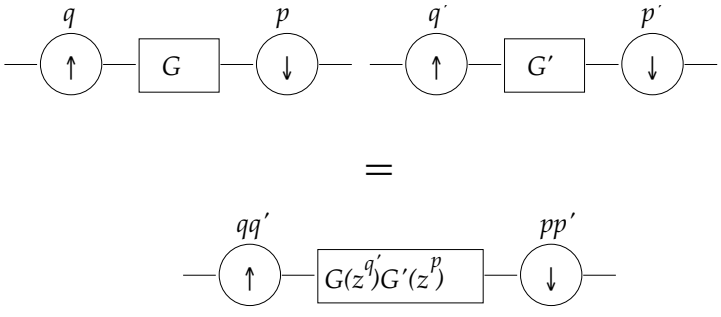
- 

- 

- 

- 

De ces relations, on peut déduire une cinquième qui nous sera fort utile par la suite: c'est la loi de composition des branches rationnelles



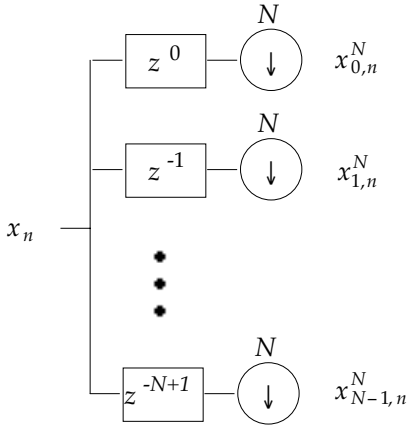
pourvu que pgcd(p,q')=1, c'est-à dire: la combinaison de deux branches donne encore une branche.

Transformations polyphases

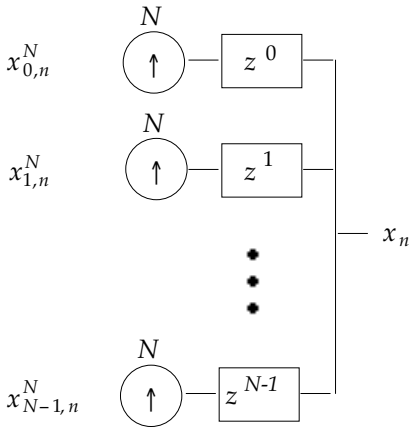
Ce sont les exemples les plus simples de bancs de filtres entiers. La transformation polyphase d'ordre N associe à un signal x_n, N signaux $x_j^N[n]$ où j=0…N-1 définis de la façon suivante

$$x_j^N[n] = x_{j+nN}$$

Cette notation ainsi que son équivalent en transformée en z (c'est-à dire $X_j^N(z)$) sera fréquemment utilisée dans le reste du document. On représente l'équation ci-dessus de manière graphique par



Cette transformation est bien évidemment inversible, et son inverse se représente schématiquement par la transformation polyphase inverse

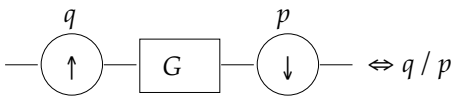


ce qui illustre la relation de reconstruction

$$X(z) = \sum_{j=0}^{N-1} z^j X_j^N(z^N)$$

L'utilisation des transformations polyphases permettra de réduire les bancs de filtres à de simples produits matriciels (chapitre II).

On simplifiera parfois la notation d'une branche quand il ne sera pas nécessaire de préciser le filtre, de la façon suivante



afin de pouvoir écrire un banc de filtres d'analyse sous la forme $(q_0 / p_0, q_1 / p_1, \ldots q_{N-1} / p_{N-1})$ et un banc de filtres de synthèse sous la forme $(q_0 / p_0, q_1 / p_1, \ldots q_{N-1} / p_{N-1})^T$.

Fonctions/Distributions

La théorie définit les distributions comme des fonctions généralisées caractérisées par le résultat obtenu lors de leur produit scalaire avec les fonctions indéfiniment dérivables à support compact, dénommées pour cette raison "fonctions test" [GM]. On notera cet espace fonctionnel C_0^∞.

Pour le produit scalaire défini par les distributions, on utilise les trois formes suivantes comme étant équivalentes, suivant la propriété que l'on cherche à mettre en évidence

$$\langle \varphi, f \rangle = \int_D \varphi f = \int_D \varphi(t) f(t) dt$$

qui désignent donc ici le résultat de l'application de la distribution φ à la fonction C_0^∞ f (la première forme sera utilisée quand seules les propriétés de linéarité de la distribution seront nécessaires, la deuxième quand le domaine d'intégration n'est plus l'espace des réels et la troisième quand on a besoin de faire apparaître en outre la variable d'intégration, pour bénéficier par exemple de toutes les ressources du calcul intégral).

On désignera également la dérivée d'une distribution φ soit sous la forme $\varphi', \varphi^{(N)}$ si les indices et exposants ne compliquent pas trop l'écriture de la distribution, soit sous la forme $\partial\varphi, \partial^N\varphi$ dans le cas contraire.

I. Interpolation

Par leur nature même, les signaux considérés sont au départ des signaux à temps continu: champs électriques, champs de pression, etc… Après leur mesure, ils sont échantillonnés à intervalles réguliers et l'on fait l'hypothèse que cet échantillonnage préserve la totalité de l'information du signal. Cette supposition est en général basée sur l'hypothèse d'un signal à bande limitée à la moitié de la fréquence d'échantillonnage; la formule d'interpolation exacte de Nyquist

$$f(t) = \sum_n f(nT)(-1)^n \frac{\sin\left(\pi \frac{t}{T}\right)}{\pi\left(\frac{t}{T} - n\right)} \qquad \text{(I.1)}$$

donne alors la valeur $f(t)$ que l'on aurait pu mesurer en un temps différent d'un des échantillons nT choisis. On a là un exemple de passage du continu au discret et vice-versa. Le seul inconvénient de cette formule est que les fonctions d'interpolation les "sinus cardinaux" sont à support infini, et de surcroît, lentement décroissants, ce qui nécessite dans la réalité d'autres fonctions qui approximent l'interpolation. On va cependant se restreindre dans un premier temps, à la formule de Nyquist pour montrer comment il est possible de passer simplement du continu au discret et vice-versa dans diverses transformations.

A. Passage temps continu–temps discret

Le traitement numérique du signal s'intéresse justement aux signaux échantillonnés ou à temps discret, et non pas à temps continu. Il s'agit de faire subir à la version échantillonnée plutôt qu'à la version temps continu, un certain nombre de transformations destinées à révéler les paramètres propres du signal en vue de la compression ou de l'analyse du signal. Ces transformations auraient été difficiles, fastidieuses et génératrices de bruit parasite sur le signal continu lui-même, alors que l'intervention de l'informatique pour le traitement apporte désormais une souplesse incomparable à leur implantation.

Un premier exemple de transformation continue chère aux physiciens est la transformée de Fourier. La formule d'interpolation de Nyquist mène directement à la formule de sa version discrétisée, c'est-à dire de la DFT (Transformée de Fourier Discrète — *Discrete Fourier Transform*)

$$\hat{f}(\nu) = \int f(t) e^{-2i\pi\nu t} dt \quad \leftrightarrow \quad DFT_f(\nu) = T \sum_n f(nT) e^{-2i\pi\nu nT}$$

La formule discrète laisse apparaître une périodicité de $1/T$ pour la fréquence, et si le signal comporte un nombre fini N d'échantillons —bien qu'en contradiction avec l'hypothèse de support fréquentiel borné, cette hypothèse engendre une erreur qui reste négligeable tant que le signal est prépondérant dans la bande de Nyquist $[-1/2T, 1/2T]$— les fréquences peuvent être

échantillonnées suivant $\nu=n/NT$, menant à une transformation bijective entre les échantillons du signal, et les échantillons de sa DFT.

1. Transformations “locales” uniformes

L’opération la plus simple et qui réunit les propriétés qui sont considérées comme les plus importantes pour le traitement du signal, à savoir l’invariance temporelle et la linéarité, est le filtrage. On voit tout de suite que l’équivalent discret de la convolution est ce que l’on appelle aussi la convolution discrète

$$y(t) = (h * x)(t) \Leftrightarrow y_n = \sum_k h'_{n-k} x_k$$

où h' est un filtre discret qui s’obtient en convoluant h avec la fonction d’interpolation de Nyquist et en l’échantillonnant.

C'est lorsque l'on entreprend de qualifier des signaux localement, et non plus globalement dans le but d'en évaluer les caractéristiques non stationnaires que l'on fait apparaître le premier exemple de bancs de filtres. En effet la transformée de Fourier à court terme (*Short-Time Fourier Transform*) d'un signal $f(t)$ dont on ne suppose plus le support temporel borné

$$STFT_f(t,\nu) = \int f(\tau) w(\tau - t) e^{-2i\pi\nu\tau} dt$$

n'est rien d'autre, à fréquence ν constante, que le filtrage par la fenêtre modulée $w(-t)e^{2i\pi\nu t}$ du signal d'entrée (et multiplié par un terme de démodulation qui ne présente pas d'intérêt puisque c'est généralement le module de la transformation qui est exploité). Si maintenant l'on échantillonne les fréquences (qui sont dans l'intervalle compact $[-1/2T, 1/2T]$) sur N valeurs, on obtient un banc de N filtres que l'on peut également échantillonner en temps et qui fournit des convolutions discrètes. On se rend alors compte qu'il n'est pas nécessaire d'échantillonner les sorties des filtres à la fréquence de Nyquist T du signal, mais qu'il est suffisant d'échantillonner à NT si l'on souhaite reconstruire le signal —il faut également ajouter quelques restrictions suivant la fenêtre w—. Cette transformation est représentée graphiquement dans la figure 1 à droite et sa reconstruction à gauche.

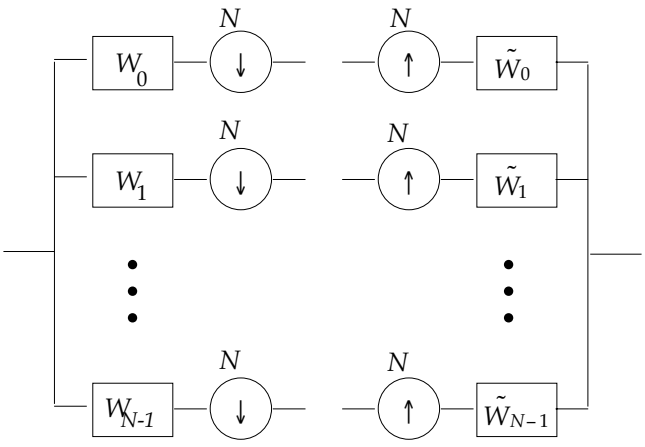


figure 1

où les filtres W_k se déduisent de la fenêtre $w(t)$. Cette relation n'est d'ailleurs pas très différente d'une modulation simple —i.-e $w_k[n] = h[n]\cos\left(\frac{\pi}{N}(n+\alpha)(k+\beta)\right)$ $w_k[n] = w_0[n]e^{2i\pi\frac{kn}{N}}$— sous certaines conditions sur w (par exemple que le support fréquentiel de la fenêtre soit contenu dans $[1\text{-}1/N,1]$).

En général, cependant, les signaux considérés sont réels ce qui impose une certaine symétrie de la transformée de Fourier, divisant par deux l'information pertinente fournie par la transformée. Cette constatation a conduit à l'élaboration de la transformée en cosinus basée sur la discrétisation de la partie réelle de la transformée de Fourier à court terme. La symétrie entre les fréquences négatives et positives incite à ne conserver que les fréquences contenus dans $[0,1/2T]$ ce qui, si l'on garde le même nombre de bandes, double la résolution de la transformation par rapport à la transformée complexe. Là encore, sous des contraintes de support fréquentiel de la fenêtre, les filtres W_k se déduisent les uns des autres par une modulation de la forme $w_k[n] = h[n]\cos\left(\frac{\pi}{N}(n+\alpha)(k+\beta)\right)$. Quand α et β sont demi-entiers, on sait construire des bancs de filtres à reconstruction parfaite FIR basés sur cette relation: on obtient alors ce qui est connu sous le nom de ELT —*extended lapped transform*— ou banc de filtres modulés [Malv1,Malv2]. Un algorithme rapide est d'ailleurs disponible pour en calculer les sorties [DMP]. Cette transformation est uniforme car les taux d'échantillonnage de chaque branche sont identiques.

2. Transformations non uniformes

Une autre interprétation du banc de filtres de la figure 1 présente cette transformation comme la décomposition d'un signal échantillonné sur N bandes de fréquences, plutôt que comme la valeur d'une transformée de Fourier en N points particuliers de la bande de Nyquist. Dans ces conditions, les filtres impliqués dans chaque branche n'ont plus de raison d'être engendrés par un seul filtre, comme c'est le cas dans la STFT discrète. On généralise donc aux bancs de filtres uniformes.

Une autre extension naturelle est de supposer que les taux d'échantillonnage de chaque branche ne sont plus identiques. Et finalement, comme il peut apparaître arbitraire de n'avoir que des termes de sous-échantillonnage à l'analyse, et des termes de sur-échantillonnage à la synthèse, on peut à nouveau généraliser les bancs de filtres aux bancs de filtres rationnels schématisés en figure 3 (voir chapitre II): chaque branche est alors constituée d'un sur-échantillonneur, d'un filtre et d'un sous-échantillonneur, permettant d'atteindre un échantillonnage fractionnaire.

Ces extensions ne sont pas dues au seul plaisir de la généralisation mathématique. On s'est en effet rendu compte qu'il est utile d'analyser certains signaux sur des bandes de fréquence variables en fonction de la fréquence centrale: c'est le cas des bruits en $1/\nu$, mais aussi des sons destinés à être perçus par l'oreille humaine. Or les largeurs de bande de fréquence sont directement liées aux taux d'échantillonnage de chaque branche, d'où l'intérêt de faire varier ces facteurs. Néanmoins, jusqu'à l'arrivée des ondelettes comme outil classique de traitement du signal, les difficultés liées à la conception des bancs de filtres à échantillonnage non uniforme ont empêché le large développement de ces transformations. On verra en effet qu'une manière efficace de concevoir un banc de filtres non uniforme est d'itérer un banc de deux filtres. Le banc de filtres équivalent peut alors s'interpréter comme une transformation en ondelettes [Ma1,Mey1] dont la régularité est une clef pour caractériser la sélectivité de l'ensemble des filtres du banc (voir chapitre V).

3. Ondelettes

C'est au début des années 1980 que la transformation en ondelettes (Wavelet Transform) est apparue sous la forme classique

$$WT_f(t,a) = \int f(\tau)\psi\left(\frac{\tau - t}{a}\right)\frac{d\tau}{\sqrt{a}}$$

où le paramètre d'échelle a joue le rôle de l'inverse d'une fréquence. Certaines conditions sont imposées à l'ondelette ψ que nous ne détaillerons pas. La "philosophie" sous-jacente à cette transformation est très différente de celle de la transformée de Fourier.

Dans l'analyse de Fourier on tente de déceler les périodicités propres à un signal et pour ce faire la sélectivité fréquentielle des filtres mis en jeu est le paramètre essentiel, bien plus que leur support. Le but est en effet d'avoir la meilleure résolution fréquentielle possible ce qui conduit à l'uniformité de cette résolution pour tout le spectre. Ce choix est particulièrement adapté aux signaux stationnaires ou quasi-stationnaires, mais peut facilement devenir un handicap dans les autres cas.

En revanche dans l'analyse en ondelettes, la dimension temporelle du signal n'est jamais éclipsée par ses caractéristiques fréquentielles —on remplace d'ailleurs le terme de "fréquence" par "échelle"—. Même si l'on continue à imposer à l'ondelette d'analyse une bonne résolution fréquentielle, cela ne signifiera pas que l'analyse sera elle-même uniformément fine: en fait, les fréquences élevées seront moins bien résolues que les fréquences plus basses. En échange, leur localisation temporelle sera plus fine. Cette interdépendance fréquence-temps est indispensable à l'analyse des signaux non-stationnaires.

Là encore, pour chaque a fixé, la tranformée est le résultat d'un simple filtrage: la transformée est alors redondante. Afin d'éliminer cette redondance, il convient d'échantillonner chaque sortie à la fréquence qui permet de rendre suffisamment faible le repliement de spectre induit. Cette fréquence n'est plus une constante, mais dépend linéairement de $1/a$, et donc l'échantillonnage correspondant dépend linéairement de a. Ainsi, si l'on suppose que l'on échantillonne à la période αTa et que la fonction f est connue par ses échantillons f_k à la fréquence de Nyquist $1/T$, on aura

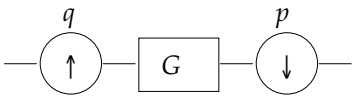
$$WT_f[n] = \sum_k g(n\alpha a - k)f_k$$

où l'on a posé

$$g(s) = \frac{T}{\sqrt{a}}\int \chi(x)\psi\left(T\,\frac{x-s}{a}\right)dx$$

Cette formule pose évidemment des problèmes dès que αa n'est pas entier puisque l'on ne sait pas réellement sous-échantillonner d'un rapport non entier.

Cependant si a est un nombre fractionnaire —ou rationnel, terme que l'on utilisera fréquemment dans cette thèse— alors, on peut définir —voir plus bas— un échantillonneur parfait de rapport $a=p/q$. Cet échantillonneur prend la forme schématique suivante



qui est la brique de base des bancs de filtres rationnels. Ainsi, la transformation discrète associée à une transformée en ondelettes sera un banc de filtres rationnel, c'est-à dire basé sur des échantillonneurs de rapport fractionnaire.

4. Échantillonnage fractionnaire idéal

Toujours en utilisant la formule de Nyquist (I.1) pour un signal $x(t)$ à bande limitée voyons maintenant à quoi correspond un changement d'échelle temporelle du signal continu reconstruit à partir de ses échantillons.

Si a est le facteur de changement d'échelle, on veut ainsi trouver l'équivalent discret de l'opération $x(t) \rightarrow y(t) = x(at)$. Bien sûr, si a est supérieur à 1, il n'y a aucune raison pour que le signal $y(t)$ échantillonné à la même fréquence que $x(t)$ puisse être reconstruit à l'aide de la formule de Nyquist puisque sa bande ne sera pas nécessairement limitée à la moitié de la fréquence d'échantillonnage.

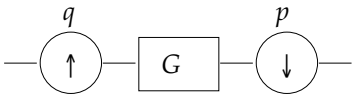
À l'aide de (I.1) on obtient directement (légèrement adapté de [SR])

$$y_n = \sum_k x_k (-1)^k \frac{\sin(\pi n a)}{\pi(na - k)}$$

Dans le cas particulier où a est un nombre fractionnaire de la forme $a=p/q$, le changement d'échelle s'écrit

$$y_n = \sum_k g_{np-kq} x_k$$
$$\text{avec} \quad g_n = \frac{\sin(\pi n / q)}{\pi n / q}$$

Cette équation peut se représenter graphiquement à l'aide des opérateurs de sur-échantillonnage, de filtrage et de sous-échantillonnage sous la forme



Malheureusement, le filtre G est ici un filtre parfait à support temporel infini qui a la mauvaise propriété de ne pas être réalisable. On va donc, dans la réalité, devoir remplacer ce filtre par un autre dont la transformée en z puisse se représenter sous la forme d'une fraction rationnelle tout en restant le plus proche possible de ce filtre idéal.

B. Échantillonnage et interpolation

On a jusqu'à présent parlé de manière intuitive de l'échantillonnage des signaux à temps réels. Cette notion doit cependant être généralisée afin de tenir compte, en particulier de la

façon effective selon laquelle un échantillonnage est réalisé. Cette généralisation va nous permettre d'introduire de façon naturelle les espaces multirésolution [Ma1,Mey1] qui sont un des outils mathématiques majeurs de l'étude des signaux non stationnaires dits "à Q-constant". L'originalité de ce qui est présenté ici ne réside pas dans la formulation de l'interpolation et de l'échantillonnage vus comme, respectivement, la projection sur un espace V_0 et les coefficients de cette projection dans une base particulière de cet espace. Ce point de vue a été par exemple exposé dans [UA] et utilisé pour caractériser l'interpolation issue d'échantillonneurs non-idéaux.

Ici on se limite à proposer une extension naturelle de ce formalisme pour le cas où l'opérateur d'échantillonnage dépendrait du temps. Le but n'est pas de proposer une nouvelle théorie complète adaptée à ce type d'échantillonnage —à la différence de [UA] elle buterait très vite sur le fait qu'il n'est plus du tout naturel de passer dans l'espace de Fourier, à cause justement de cette perte d'invariance temporelle—, mais plutôt de poser les bases d'une analyse multirésolution généralisée et d'en étudier quelques caractéristiques. Cette généralisation est en effet indispensable pour conserver une interprétation simple des bancs de filtres itérés en fraction d'octave et que nous étudierons au chapitre IV.

1. Échantillonnage

Il s'agit dans le principe, d'un opérateur linéaire $\mathfrak{E}$ qui transforme un signal à temps continu, en un signal à temps discret et dont l'expression la plus courante est

$$\{x(t)\}_{t\in\mathbf{R}} \xrightarrow{\mathfrak{E}} \{x(n\tau)\}_{n\in\mathbf{Z}} \tag{I.2}$$

où τ est l'intervalle d'échantillonnage.

Cependant, on ne peut généralement pas avoir accès par la mesure, à des valeurs instantanées comme c'est sous-entendu par (I.2). Même si c'était le cas, les signaux réels étant par nature à support fréquentiel infini ou du moins très grand, il est habituel de préfiltrer ces signaux afin de les rendre à bande limitée. Il est bien clair que dans ce dernier cas, aucune technique d'interpolation ne pourra permettre de récupérer l'information perdue par cette mise à zéro d'une portion du spectre du signal. Dans les deux cas, on sera conduit à écrire à la place de (I.2)

$$\{x(t)\}_{t\in\mathbf{R}} \xrightarrow{\mathfrak{E}} \left\{\int \phi(u-n)x(\tau u)du\right\}_{n\in\mathbf{Z}} \tag{I.3}$$

où ϕ est une fonction centrée sur 0 qui décrit l'inertie de l'appareil de mesure. Dans le cas idéal de valeurs instantanées, on a bien sûr $\phi(u) = \delta(u)$.

Et finalement, si notre appareil de mesure a le mauvais goût d'être variable avec le temps, on pourra remplacer cette expression par

$$\{x(t)\}_{t\in\mathbf{R}} \xrightarrow{\mathfrak{E}} \left\{\int \phi_n(u)x(\tau u)du\right\}_{n\in\mathbf{Z}} \tag{I.4}$$

où les fonctions ϕ_n sont supposées être centrées au voisinage de n. Cette expression est la plus générale dès que l'on impose la linéarité de l'opérateur et sa continuité, ce qui semble être un minimum. Il est clair que dans ce cas, les formules de reconstruction doivent être revues. On va

donc maintenant s'intéresser en détail au problème de l'interpolation qui est l'opérateur inverse de l'échantillonnage.

2. Interpolation

Dans la mesure où l'échantillonnage est un procédé qui perd de l'information, on est obligé d'imposer certaines propriétés au signal que l'on cherche à retrouver afin de rendre l'opération inversible. Dans le cas de l'interpolation de Nyquist, on impose ainsi que le signal soit à bande limitée.

Dans la pratique, les signaux réels, qui sont en particulier à durée finie ne peuvent mathématiquement être à support fréquentiel borné. Qui plus est, les fonctions d'interpolation de Nyquist sont à support temporel infini et correspondent à des filtres qui sont irréalisables dans la pratique.

On est donc conduit à définir de nouvelles propriétés que devront vérifier les signaux que nous souhaitons interpoler, et qui nous permettront de nous rapprocher de situations plus réalistes.

a. Linéarité

La première de ces propriétés découle directement de la linéarité du processus d'échantillonnage. Il faut donc que son inverse soit également linéaire, c'est-à dire que l'espace d'interpolation V_0 soit un espace vectoriel.

Désignons par $\Im$ l'opérateur linéaire d'interpolation. On écrira

$$x(t) = \Im\left(\left\{x_n\right\}_{n\in\mathbf{Z}}\right)$$

et si l'on définit

$$\varphi_k(t) = \Im\left(\left\{\delta_{n-k}\right\}_{n\in\mathbf{Z}}\right)$$

alors les fonctions φ_k constituent une base de l'espace V_0. On obtient alors la formule d'interpolation suivante

$$x(t) = \sum_n x_n \varphi_n(t)$$

Afin que cette formule corresponde à une véritable interpolation, il est bien sûr nécessaire que

- $\varphi_k(n) = \delta_{n-k}$ si l'échantillonnage est défini par (I.2)
- $\int \phi(u-n)\varphi_k(u)du = \delta_{n-k}$ si l'échantillonnage est défini par (I.3)
- $\int \phi_n(u)\varphi_k(u)du = \delta_{n-k}$ si l'échantillonnage est défini par (I.4)

ceci pour tout k,n entier.

b. Invariances

Revenons quelques instants sur l'interpolation de Nyquist. Dans ce cas particulier, l'espace V_0 est l'espace des fonctions à support fréquentiel limité à [-1/2,1/2], ce qui conduit à la définition des fonctions d'interpolation

$$\begin{cases} \varphi_n(t) = \varphi(t-n) \\ \varphi(t) = \dfrac{\sin \pi t}{\pi t} \end{cases}$$

On peut remarquer d'autre part que l'on a au moins deux possibilités pour la fonction d'échantillonnage ϕ_n puisque $\phi_n(u) = \delta(u-n)$ *ou* $\phi_n(u) = \varphi(u-n)$ conviennent. Cette non unicité de l'interpolation est une propriété générale qui sera illustrée dans les bancs de filtres rationnels quand on sera amené à calculer le filtre d'interpolation associé au filtre d'analyse: pour un opérateur d'échantillonnage donné il existe une infinité d'opérateurs d'interpolation. Les deux propriétés les plus remarquables de l'espace d'interpolation V_0 sont

- *l'invariance temporelle*

$$\forall k \in \mathbf{Z}\ \forall x(t) \in V_0 \quad \mathfrak{E}\big(x(t-k)\big)_n = \mathfrak{E}\big(x(t)\big)_{n-k} \tag{I.5}$$

- *l'invariance d'échelle*

$$\forall a \geq 1 \quad x(t) \in V_0 \quad \Rightarrow \quad x(t/a) \in V_0 \tag{I.6}$$

Sur ces propriétés on peut faire les commentaires suivants

i. Invariance temporelle

La définition utilisée, bien que masquée par le formalisme mathématique est en fait très naturelle: on demande qu'un retard d'un nombre entier d'échantillons se traduise par un décalage identique du signal échantillonné. On vérifie aisément que cela impose que l'espace d'interpolation soit engendré par une seule fonction $\phi(t)$ et ses translatées $\phi(t-\tau) = \phi_\tau(t)$. Par contre, cela n'impose pas ce type de contrainte sur les fonctions d'échantillonnage.

ii. Invariance d'échelle

Cette invariance est la clef principale des espaces multirésolutions. Elle concrétise simplement l'idée qu'un niveau de résolution donné doit contenir les niveaux de résolution plus grossiers (a>1). Comme on est amené à discrétiser l'ensemble des résolutions admissibles, la progression géométrique s'impose comme étant la plus naturelle, de la même façon que la progression arithmétique était adaptée à l'échelle temporelle. On doit alors choisir une échelle minimale a>1, permettant de définir l'invariance d'échelle sous la forme restreinte

$$x(t) \in V_0 \quad \Rightarrow \quad x(t/a) \in V_0$$

ce qui équivaut bien sûr à $x(t/a^n) \in V_0$ pour tout n entier positif.

c. Support

À l'inverse, la fonction de Nyquist souffre du fait qu'elle est à support infini, ou plutôt qu'elle décroît trop lentement à l'infini. Intuitivement, il n'est en effet pas naturel d'admettre que des échantillons séparés d'une distance aussi grande que l'on souhaite, doivent être pris en compte pour reconstituer le signal continu en un point donné. Bien sûr cela est dû au fait que l'espace d'interpolation V_0 n'est constitué que de telles fonctions: il n'est pas acceptable en pratique.

On est donc amené à imposer la contrainte supplémentaire de la compacité du support des fonctions de reconstruction, ainsi que celles d'analyse. Surgit alors un autre problème: les deux contraintes d'invariance temporelle et d'échelle sont incompatibles avec cette hypothèse de support fini dès que le paramètre d'échelle a n'est pas entier [CD,KV1]. Il faut donc choisir. Dans cette thèse on abandonnera délibérément l'invariance temporelle, sachant que celle-ci peut être récupérée approximativement quand on choisit de manière adéquate les paramètres de changement d'échelle —en fait, les filtres d'un banc de filtres rationnel— dans le cas où a est fractionnaire.

Notons également que la non-invariance par translation implique de facto que les fonctions d'échantillonnage ne sont pas non plus engendrées par une unique fonction translatée. Le degré de non invariance par translation sera mesuré par une quantité que l'on appellera amnésie, pour bien signifier le phénomène d'oubli qui caractérise alors l'appareil de mesure.

3. Analyse multirésolution

Comme on décide de conserver la propriété d'invariance d'échelle, si f appartient à notre espace d'interpolation V_0 alors sa version plus grossière $f(t/a)$ appartient également à V_0. De la même manière, on peut s'intéresser à des résolutions plus fines que V_0, ce qui définit une suite d'espaces

$$V_N = \underset{n\in\mathbf{Z}}{\mathrm{Vect}}\left\{\phi_n\left(\frac{p^N}{q^N}t\right)\right\} \cap L^2$$

qui sont imbriqués grâce à la propriété d'invariance d'échelle

$$\ldots \subset V_N \subset V_{N-1} \subset \ldots V_0 \subset V_1 \subset \ldots$$

Il faut noter qu'en général on impose à V_0 d'être contenu dans L^2, afin de munir les espaces d'un produit scalaire qui permettra, entre autres de calculer des compléments orthogonaux. On impose enfin

$$\bigcap_N V_N = \{0\} \quad \text{et} \quad \overline{\bigcup_N V_N} = L^2$$

C'est cette suite d'ensembles que l'on appellera analyse multirésolution. Cette définition est à peine différente de celle qui a cours dans la littérature de traitement de signal [Mey1,Dau2], la différence avec notre définition venant du fait que les fonctions d'interpolation ϕ_n sont engendrées par une seule fonction translatée et que le facteur d'échelle utilisé est un nombre entier. Cette propriété supplémentaire nous est, comme on l'a vu plus haut, interdite puisque nous utilisons des facteurs d'échelle fractionnaires et des fonctions à support borné. Il

faut cependant noter que rien ne s'oppose à ce que l'on définisse une analyse multirésolution à facteur d'échelle non entier: les fonctions engendrant l'espace V_0 seront simplement à support infini. Des exemples de construction d'analyse multiresolution avec des facteurs d'échelle fractionnaires sont donnés par exemple dans [Au].

À partir de ces espaces, on peut définir des espaces complémentaires qui contiennent l'information perdue quand on passe d'une résolution à une résolution plus basse. Ces espaces W_N seront définis par la relation

$$V_N = V_{N-1} \oplus W_{N-1}$$

Il y a donc plusieurs choix. L'un de ceux-ci consiste à prendre le complément orthogonal de V_N-1 dans V_N. Dans tous les cas, il existera une suite de fonctions ψ_n telle que

$$W_N = \underset{n\in\mathbf{Z}}{\mathrm{Vect}}\left\{\psi_n\left(\frac{p^N}{q^N}t\right)\right\}$$

On a alors la propriété de disjonction

$$W_N \cap W_{N'} = \{0\}$$

si $N \neq N'$. En tout état de cause, on aura la relation de décomposition d'un espace multirésolution $V_{N'}$ en espaces de plus basse résolution

$$V_{N'} = \bigoplus_{n=N}^{N'-1} W_n \oplus V_N$$

ce qui conduit à la décomposition de L^2 en espaces disjoints

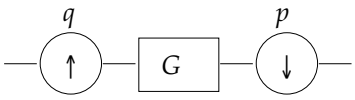
$$L^2 = \bigoplus_N W_N$$

Le problème —l'analyse— consistant à décomposer toute fonction de L^2 en éléments de ces espaces W_N à résolution fixée sera résolu dans le cas des bancs de filtres rationnels, et mettra en évidence une suite de fonctions d'échantillonnage φ_n et ψ_n, permettant de définir une analyse multirésolution duale. On aura alors

$$f = \sum_{N,n} \psi_n(a^N t) \int a^N f(t) \psi_n(a^N t) dt$$

C. Sous- et sur-échantillonnage

Revenant aux méthodes de passage du temps continu au temps discret, on se souvient que l'on a pu définir un opérateur d'échantillonnage fractionnaire qui s'appliquait à un signal à temps discret pour donner un autre signal à temps discret, en utilisant la formule d'interpolation de Nyquist. On obtenait alors le schéma suivant



pour un filtre bien précis qui prenait en compte l'interpolation de Nyquist. On décide d'étendre ce résultat, sans plus se préoccuper de cohérence avec le cas continu à tous les filtres et l'on va étudier les caractéristiques de cette transformation, que l'on appellera par analogie "échantillonnage fractionnaire".

1. L'opérateur de base

Mathématiquement, il se définit par les formulations "échantillon" ou "transformée en z"

$$\begin{aligned} y_n &= \sum_k g_{np-kq} x_k \\ Y(z^p) &= \frac{1}{p} \sum_{k=0}^{p-1} G\left(z e^{2i\pi \frac{k}{p}}\right) X\left(z^q e^{2i\pi k \frac{q}{p}}\right) \end{aligned} \tag{I.7}$$

si x est l'entrée et y la sortie. On peut faire les observations suivantes

- les entiers p et q sont nécessairement premiers entre eux. Si ce n'était pas le cas, il serait très simple de vérifier que l'on peut réécrire l'opérateur rationnel comme une branche où les opérateurs de sur- et sous-échantillonnage seraient divisés par leur pgcd, et où le filtre serait un élément polyphasé du filtre initial

- le débit de sortie (c'est-à dire le nombre d'échantillons par unité de temps) est exactement de q/p le débit d'entrée, ce qui justifie une fois de plus l'adjectif "rationnel" associé à ce type d'opérateur

- bien que non temporellement invariant puisque ce n'est pas un filtrage, cet opérateur vérifie la propriété d'invariance suivante: un retard de p échantillons sur x correspond à un retard de q échantillons sur y, et inversement

- le filtre G n'agit pas directement sur x, mais sur sa version sur-échantillonnée de q, c'est-à dire que si ce filtre est censé sélectionner un domaine fréquentiel de x, il devra être q fois plus sélectif qu'un filtre en bande de base. Cette observation est évidemment primordiale pour la conception de filtres

- le filtre G sera parfait en particulier s'il permet d'éviter le repliement de spectre après le sous-échantillonnage par p. Par exemple, si l'on considère le filtre comme étant passe-bas et à support fréquentiel connexe (donc du type intervalle), alors ce support fréquentiel sera entièrement contenu dans $[-1/2p, 1/2p]$. Dans ces conditions, la branche rationnelle ne laissera passer du signal d'entrée que la bande de fréquences contenue dans $[-q/2p, q/2p]$. Cependant, pour obtenir un résultat semblable, on aurait pu préférer que G soit passe-bande et filtre sélectivement les répliques passe-bas du signal. C'est ainsi la particularité des bancs de filtres rationnels de ne pas avoir un unique gabarit utilisable

dans la conception de filtres. On reviendra sur ce sujet de manière plus précise dans le chapitre II

- à la différence d'un filtre simple, ou d'une branche à échantillonnage entier, cet opérateur ne transforme pas une sinusoïde en une autre sinusoïde, du moins en général (cela peut cependant arriver quand le filtre est parfait, c'est-à dire en particulier de longueur infinie). On peut en fait vérifier qu'il transforme une sinusoïde en q sinusoïdes: supposons en effet que $x_n = e^{2i\pi n\nu}$ alors d'après la formule (I.7) on trouve

$$y_n = \frac{1}{q}\sum_{s=0}^{q-1} G\left(e^{-2i\pi(\nu+s)/q}\right)e^{2i\pi n\frac{p}{q}(\nu+s)}$$

qui est bien la somme de q sinusoïdes dont par ailleurs, les fréquences sont $\left[p(\nu+s)/q\right]$ pour s=0…q-1

- si L est la longueur du filtre G, la complexité de ce type d'opérateur est approximativement de L/p multiplications et additions par échantillon de sortie

a. Un filtrage matriciel

Les équations (I.7) cachent une simple formulation de filtrage matriciel que je vais maintenant indiquer. Réécrivons l'équation "échantillon" sous la forme

$$y^q_{n_0}[n] = \sum_{k_0=0}^{p-1}\sum_k g^{pq}_{n_0p-k_0q}[n-k]x^p_{k_0}[k]$$

c'est-à dire qu'on a introduit une transformée polyphase d'ordre p à l'entrée et une transformée polyphase inverse d'ordre q à la sortie de façon à avoir un opérateur associant q sorties à p entrées. Cet opérateur réalise donc clairement un filtrage du p-vecteur d'entrées $\xi_n = (x^p_0[n], x^p_1[n], \ldots x^p_{p-1}[n])^T$ par le filtre matriciel $\mathbf{\Gamma}$ défini par

$$\gamma_{k,l}[n] = g^{pq}_{kp-lq}[n]$$

pour donner le q-vecteur de sortie $\eta_n = (y^q_0[n], y^q_1[n], \ldots y^q_{q-1}[n])^T$. En termes de transformées en z on a

$$\mathrm{H}(z) = \Gamma(z)\Xi(z)$$

On verra (chapitre II) que cette relation reste vraie si l'on considère un banc de filtres rationnels comportant donc une entrée et plusieurs sorties, ou plusieurs entrées et une sortie.

2. Généralisation de l'opérateur de base

Ayant délibérément oublié le lien entre la transformation continue et la transformation discrète, on doit repenser la notion d'échantillonnage dans le champ discret afin de donner une

signification à l'échantillonnage fractionnaire. Ce que l'on peut entendre par échantillonnage ne recouvre pas nécessairement la notion d'interpolation, c'est-à dire que le premier but de l'échantillonnage est la réduction ou l'augmentation du débit d'échantillons. C'est seulement après que l'on s'intéresse au fait que les échantillons de sortie sont bien des versions du signal d'entrée à des échelles supérieure ou inférieure à 1. Or si l'on se limite à la modification du débit, tout en conservant la linéarité, l'ensemble des transformations linéaires est considérablement plus étendu. Celles-ci s'écrivent sous la forme

$$y_n = \sum_k r_{n,k} x_k$$

et dans le cas de l'opérateur rationnel on a $r_{n,k} = g_{np-kq}$, c'est-à dire qu'il vérifie $r_{n+q,k+p} = r_{n,k}$. On doit se poser la question suivante: comment obtenir une transformée qui modifie le débit? En fait le problème ainsi posé est trop vaste, puisque le débit est une moyenne sur un temps à priori infini. On va donc se resteindre aux opérateurs qui préservent une certaine cyclostationnarité de l'opérateur.

a. Invariance cyclique

On va ainsi supposer qu'un décalage de a échantillons en entrée entraîne un décalage de b échantillons en sortie. Il est alors facile de vérifier que le débit de la sortie est dans un rapport b/a avec le débit de l'entrée et si l'on souhaite avoir une variation de débit de rapport q/p, il faut donc que

$$b = \lambda q$$
$$a = \lambda p$$

où λ est un entier positif quelconque. D'autre part, l'invariance cyclique va impliquer la relation

$$r_{n+\lambda q,k+\lambda p} = r_{n,k}$$

pour tout couple d'entiers n,k. Dès que $\lambda \neq 0$, cette équation ne peut pas se réduire à une branche rationnelle simple qui vérifie, ainsi qu'on l'a vu plus haut, $r_{n+q,k+p} = r_{n,k}$. On va donc appeler "branche rationnelle généralisée" ce nouvel opérateur dont on démontrera qu'il est la base de tout opérateur de synthèse dans la reconstruction d'un banc de filtres rationnel "simple" (chapitre II). En effet, on se rendra compte qu'un banc de filtres d'analyse simple n'admet pas, en général (sauf dans le cas deux bandes), de banc de filtres de synthèse simple également, comme opérateur de reconstruction (voir à ce sujet l'exemple de Hoang et Vaidyanathan sur le cas (1/2,1/3,1/6) [HV]). Par contre, si l'on étend la définition des bancs de filtres aux bancs de filtres généralisés, c'est-à dire utilisant la branche rationnelle généralisée pour brique de base, alors l'inverse est également un banc de filtres généralisé.

b. Représentation graphique

On s'intéresse maintenant à la représentation graphique d'une branche généralisée afin de bien mettre en évidence la différence avec une branche rationnelle "normale". On peut obtenir plusieurs résultats différents, mais on va se contenter d'un seul. Posons donc

$$\begin{array}{l} \lambda q = q_1 q_2 \\ \lambda p = p_1 p_2 \end{array} \quad \text{où} \quad \begin{array}{l} \text{pgcd}(q_1, \lambda p) = \text{pgcd}(q_1, q_2) = 1 \\ \text{pgcd}(p_1, \lambda q) = \text{pgcd}(p_1, p_2) = 1 \end{array}$$

où q_1 et p_1 sont les plus grands entiers vérifiant ces conditions de primalité. On peut montrer qu'alors

$$\begin{array}{ll} p_1 = \dfrac{p}{\text{pgcd}(p, \lambda)} & q_1 = \dfrac{q}{\text{pgcd}(q, \lambda)} \\ p_2 = \lambda \text{pgcd}(p, \lambda) & q_2 = \lambda \text{pgcd}(q, \lambda) \end{array} \quad \text{et} \qquad \text{(I.8)}$$

Preuve

Comme λ divise $q_1 q_2$ et que q_1 est premier avec λp, donc avec λ, il reste que λ divise q_2: $q_2 = \lambda q_2'$.

De même on trouve $p_2 = \lambda p_2'$. On a alors

$$\begin{aligned} q &= q_1 q_2' \\ p &= p_1 p_2' \end{aligned}$$

on est donc maintenant assuré que q_1 est premier avec λp et que p_1 est premier avec λq. Il suffit donc de trouver le plus grand q_1 qui soit premier avec $q_2 = \lambda q'_2$ et le plus grand p_1 premier avec $p_2 = \lambda p'_2$.

Comme q_1 est premier avec λ, q_1 divise $q / \text{pgcd}(q, \lambda)$ qui est premier avec $\text{pgcd}(q,\lambda)$: le couple $q_1 = q / \text{pgcd}(q, \lambda)$ et $q_2 = \text{pgcd}(q,\lambda)$ est donc notre solution. Même chose pour p_1 et p_2, ce qui démontre (I.8).

Considérons l'opérateur défini par le noyau $r_{n_0}[n,k] = r_{n_0 + nq_2, k}$ où $n_0 = 0..q_2\text{-}1$. On a clairement

$$y_{n_0}^{q_2}[n] = \sum_k r_{n_0}[n,k] x[k]$$

Le théorème suivant va montrer que l'opérateur défini par $r_{n_0}[n,k]$ est une branche rationnelle simple avec pour taux d'échantillonnage $q_1 / p_1 p_2$. On observe en effet que $r_{n_0}[n + q_1, k + p_1 p_2] = r_{n_0}[n,k]$ où q_1 et $p_1 p_2$ sont premiers entre eux.

Théorème I.1 *Soit $s_{n,k}$ le noyau d'un opérateur linéaire vérifiant la relation $s_{n+q,k+p} = s_{n,k}$ où p et q sont premiers entre eux, alors il existe un filtre $G(z)$ tel que $s_{n,k} = g_{np-kq}$. En d'autres termes, cet opérateur représente une branche rationnelle simple.*

Preuve

Puisque p et q sont premiers entre eux, il existe p' et q' tels que $pp'\text{-}qq' = 1$ (relation de Bezout), et donc

$$\begin{aligned} \varphi_n(t) &= \sum_k r_a[n,k] \varphi_k(at) \\ s_{n,k} &= s[n(pp' - qq'), k(pp' - qq')] \\ &= s[p'(np - kq), q'(np - kq)) \end{aligned}$$

c'est-à dire qu'en posant $g_n = s_{p'n,q'n}$ on obtient le résultat annoncé.

On voit donc que l'opérateur rationnel généralisé peut se représenter comme un banc de filtres rationnel simple dont chacune des q_2 bandes est de taux d'échantillonnage q_1/p_1p_2 suivi d'une transformation polyphase inverse d'ordre q_2, comme c'est indiqué en figure 2

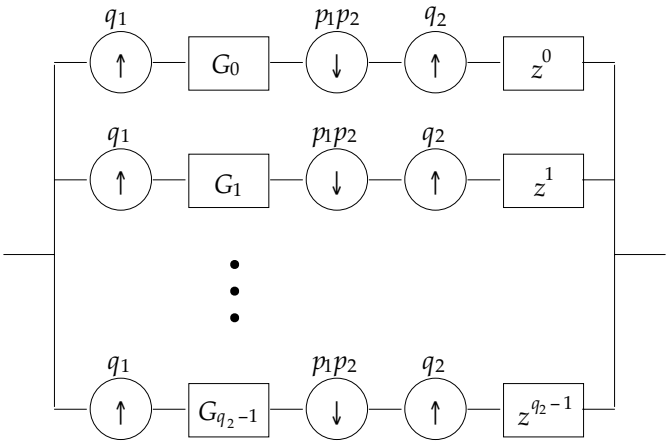


figure 2

Ce qui est différent de tout ce que nous avons vu jusqu'à présent, c'est la présence de l'opérateur de sous-échantillonnage par p_2 immédiatement suivi par l'opérateur de sur-échantillonnage par q_2 où p_2 et q_2 ont des facteurs communs, et qui ne peuvent donc être interchangés. Constatons également que les filtres G_k sont uniquement définis: en d'autres termes si $r_{n,k}$=0 pour tout n,k alors chaque branche est nulle et par voie de conséquence, chaque filtre G_k est nul.

D. Résumé du chapitre

Nous nous sommes intéressé ici essentiellement aux liens entre les signaux à temps continu et ceux à temps discret qui conduisent aux notions d'échantillonnage et d'interpolation. Il n'y a rien d'original là-dedans, le seul but étant de montrer comment les bancs de filtres rationnels peuvent être vus comme l'une des manières les plus générales d'implémenter en temps discret les transformations en ondelettes continues, et de façon à préparer l'analogie entre bancs de filtres itérés et ondelettes. On s'est ainsi intéressé à une version plus générale d'analyse multirésolution ne vérifiant plus la propriété d'invariance par translation: ceci est en effet nécessaire si l'on veut conserver les avantages de l'analyse multirésolution, la compacité du support des fonctions génératrices et l'utilisation de facteurs d'échelle non entiers. Enfin, dans le but de montrer que les bancs de filtres rationnels ne sont pas la forme la plus générale de transformation linéaire continue *à temps discret*, nous avons introduit une généralisation de l'opérateur d'échantillonnage fractionnaire de base en décidant de modifier le notion d'échantillonnage en temps discret. Ce nouvel opérateur, loin d'être une simple pathologie, sera utilisé dans le prochain chapitre pour inverser tout banc de filtres non uniforme. On n'insistera cependant pas sur le sujet après le chapitre II.

II. Cas discret

Après avoir étudié les liens —interpolation et échantillonnage— entre transformées continues et transformées discrètes, nous allons maintenant nous mettre résolument du côté discret. À ce sujet nous nous intéresserons maintenant aux différentes propriétés des bancs de filtres rationnels. Le cas de bancs de deux bandes nous retiendra particulièrement car c'est à partir de tels bancs de filtres que nous allons en construire, par itérations, de plus grands [KV1]. On verra en particulier dans le chapitre IV comment récupérer une transformation continue à partir de ces bancs de filtres itérés [Blu1].

A. Bancs de filtres rationnels

Un banc de filtres rationnel d'analyse est défini graphiquement de la façon suivante

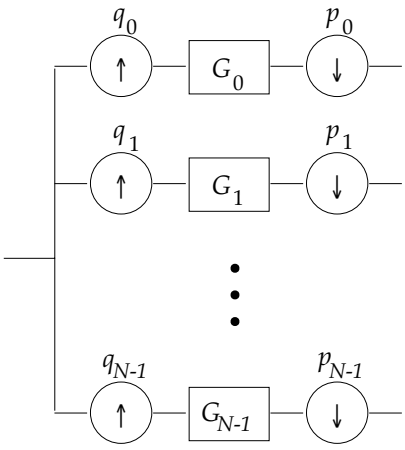


figure 3

Les questions que l'on peut se poser concernant une telle transformation sont

- à quelles conditions ce banc de filtres est-il inversible?
- dans ce cas quelle est la forme de l'inverse?
- quelle est la complexité de cette transformation?
- quelle est la signification des filtres $G_j(z)$?

1. Inversibilité

Dans un but de conception mais aussi d'inversion il avait été présenté dans [Hsi,KV1,KV3] une méthode composée de deux transformations qui permettaient de transformer un banc de filtres rationnel en un banc de filtres uniforme. Cette technique permettait alors de construire des bancs de filtres à taux d'échantillonnage fractionnaire et à reconstruction parfaite FIR, un problème alors ouvert.

Nous faisons ici la même chose, mais en condensant ces deux opérations et en présentant le résultat sous forme mathématique plutôt que graphique. On va d'abord montrer de la même manière que dans le chapitre I, qu'un tel banc de filtres se réduit à un filtrage matriciel: il s'agit donc là d'une extension de la formule qui concernait seulement une branche rationnelle.

a. Un filtrage matriciel pour le banc de filtres d'analyse

L'équation qui donne la j-ème sortie $y_{j,n}$ en fonction de l'entrée x_n est

$$y_{j,n} = \sum_k g_j[np_j - kq_j]x_k$$

ce que l'on peut réécrire, en notant P un multiple de p_j et en posant $Q_j = Pq_j/p_j$

$$y_{j,n_0}^{Q_j}[n] = \sum_{k_0=0}^{P-1}\sum_k g_{j,n_0p_j-k_0q_j}^{Pq_j}[n-k]x_{k_0}^P[k] \qquad \text{(II.1)}$$

Si maintenant on choisit P comme le plus petit multiple commun à tous les p_j, alors on obtiendra l'équation de filtrage matriciel suivante

$$\begin{pmatrix} \left[Y_{0,k}^{Q_0}(z)\right]_{0\le k\le Q_0-1} \\ \left[Y_{1,k}^{Q_1}(z)\right]_{0\le k\le Q_1-1} \\ \vdots \\ \left[Y_{N-1,k}^{Q_{N-1}}(z)\right]_{0\le k\le Q_{N-1}-1} \end{pmatrix} = \begin{pmatrix} \left[G_{0,kp_0-lq_0}^{q_0P}(z)\right]_{\substack{0\le k\le Q_0-1\\ 0\le l\le P-1}} \\ \left[G_{0,kp_1-lq_1}^{q_1P}(z)\right]_{\substack{0\le k\le Q_1-1\\ 0\le l\le P-1}} \\ \vdots \\ \left[G_{N-1,kp_{N-1}-lq_{N-1}}^{q_{N-1}P}(z)\right]_{\substack{0\le k\le Q_{N-1}-1\\ 0\le l\le P-1}} \end{pmatrix} \left(\left[X_k^P(z)\right]_{0\le k\le P-1}\right) \qquad \text{(II.2)}$$

ce que l'on écrit de façon compacte $\mathbf{Y}(z) = \mathbf{\Gamma}(z)\mathbf{X}(z)$ où $\mathbf{\Gamma}$ est une matrice de taille

$$\left(P\sum_{j=0}^{N-1}\frac{q_j}{p_j}\right) \times P$$

b. Échantillonnage critique

On se trouve donc en présence de trois possibilités

- $\sum_{j=0}^{N-1}\frac{q_j}{p_j} < 1$: le banc de filtres n'est pas inversible
- $\sum_{j=0}^{N-1}\frac{q_j}{p_j} = 1$: le banc de filtres est inversible si et seulement si $\det(\mathbf{\Gamma}) \neq 0$
- $\sum_{j=0}^{N-1}\frac{q_j}{p_j} > 1$: le banc de filtres est inversible si et seulement si $\mathbf{\Gamma}$ est de rang P

Dans le premier cas on dit que l'échantillonnage est sous-critique, dans le second cas, qu'il est critique, et dans le troisième qu'il est sur-critique. Dans une optique de codage de signal, il est bien clair que l'on souhaite être proche de, ou à l'échantillonnage critique, afin de ne pas avoir une redondance artificielle d'information (sauf cas où l'on souhaite rendre moins sensible au bruit l'inversion du banc de filtres).

c. Un filtrage matriciel pour le banc de filtres de synthèse

La formule d'inversion s'écrit alors $\mathbf{X}(z) = \Gamma^{-1}(z)\mathbf{Y}(z)$. Cependant cette formule ne détermine pas nécessairement un banc de filtres de synthèse. En effet, en utilisant la même technique que pour l'analyse, on peut vérifier qu'après avoir inséré des transformées polyphases en entrée et en sortie, le banc de filtres de synthèse se met sous une forme matricielle. On a alors

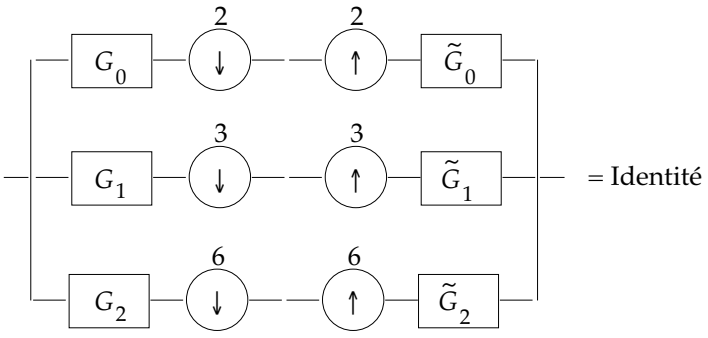
$$\mathbf{X}(z) = \tilde{\Gamma}(z)\mathbf{Y}(z)$$

$$\text{où} \quad \tilde{\Gamma}(z) = \begin{pmatrix} \left[\tilde{G}^{q_0 P}_{0,lq_0-kp_0}(z)\right]_{\substack{0\le k\le Q_0-1\\ 0\le l\le P-1}} \\ \left[\tilde{G}^{q_1 P}_{1,lq_1-kp_1}(z)\right]_{\substack{0\le k\le Q_1-1\\ 0\le l\le P-1}} \\ \vdots \\ \left[\tilde{G}^{q_{N-1} P}_{N-1,lq_{N-1}-kp_{N-1}}(z)\right]_{\substack{0\le k\le Q_{N-1}-1\\ 0\le l\le P-1}} \end{pmatrix}^T \qquad \text{(II.3)}$$

Il n'y a évidemment aucune raison pour que la forme très particulière de cette matrice puisse coïncider en général avec l'inverse d'une matrice de la forme de $\mathbf{\Gamma}$ sauf pour des valeurs particulières de p_i et q_i [KV3].

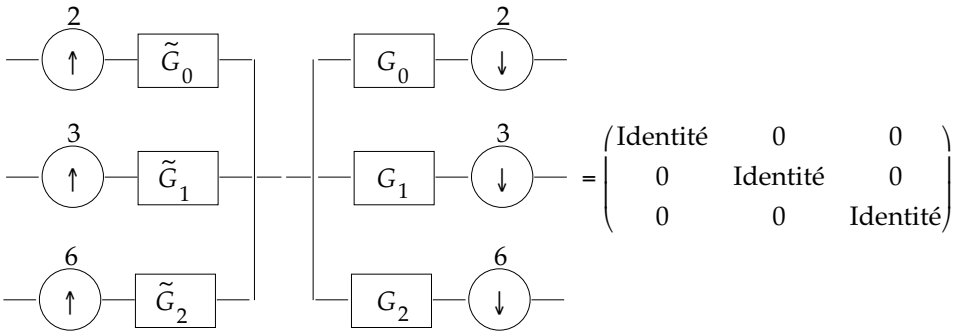
d. Contre-exemple à la proposition $\Gamma^{-1} = \tilde{\Gamma}$

Il n'est en fait même pas nécessaire que le banc de filtres soit rationnel pour cela: c'est le cas avec la structure 1/2 1/3 1/6 dont on vérifie qu'elle est bien à échantillonnage critique. Cet exemple est dû à Hoang-Vaidyanathan [HV]. Supposons que l'on puisse avoir schématiquement l'égalité suivante

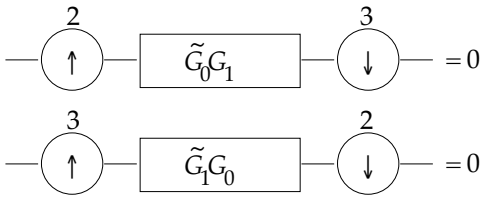


(II.4)

c'est-à dire reconstruction parfaite avec un schéma miroir de celui d'analyse alors on aura aussi



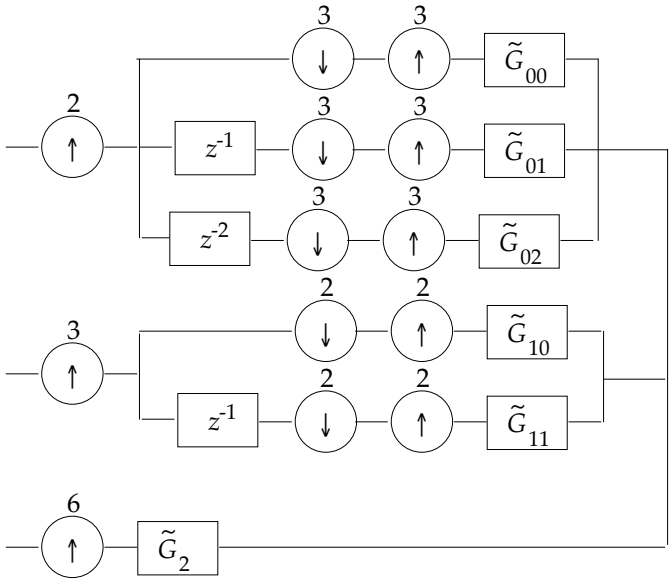
en particulier il faut



c'est-à dire $\tilde{G}_0(z)G_1(z) = 0$ et $\tilde{G}_1(z)G_0(z) = 0$. Ces conditions impliquent que les filtres de synthèse ou d'analyse ne sont pas réalisables (on rappelle qu'un filtre est réalisable s'il peut s'écrire sous forme de fraction rationnelle). A fortiori, l'analyse-synthèse avec des filtres à réponse impulsionnelle finie est impossible… si l'on impose comme ci-dessus la forme de la synthèse.

e. Bancs de filtres généralisés

En fait, on pourrait voir que la transformation inverse associée à l'analyse (1/2,1/3,1/6) ci-dessus est de la forme



qui ne peut pas se ramener sous la forme plus simple d'un banc de filtres classique de synthèse. Elle se décrit donc à l'aide d'opérateurs rationnels généralisés —en fait dans ce cas précis, les taux d'échantillonnage sont entiers—. Ces observations conduisent aux résultats suivants

Théorème *L'inverse d'un banc de filtres rationnel simple ou généralisé est un banc de filtres généralisé*

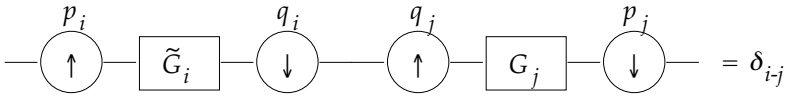
Preuve

C'est en fait assez rapide car l'inverse d'un banc de filtres rationnels ou généralisé s'écrit à l'aide d'un filtrage matriciel, que l'on peut découper en N branches, la $j^{\text{ème}}$ admettant en entrée $y_j[n]$. Ces branches sont à la fois des opérateurs linéaires et à invariance cyclique. En conséquence, d'après le théorème I.1, chaque branche peut s'écrire comme un opérateur d'échantillonnage rationnel généralisé, ce qui achève la démonstration.

Théorème II.1 *Pour que le schéma de reconstruction soit miroir du schéma d'analyse avec des filtres d'analyse réalisables (FIR ou fraction rationnelles) il est nécessaire que pour chaque couple de branches (i,j) on ait* $\text{pgcd}(p_i,p_j)\neq 1$.

Preuve

Reprenant le même esprit que pour la démonstration concernant le cas (1/2,1/3,1/6) on observe que



ce qui se traduit par

$$\sum_k g_j[n'p_j - kq_j]\tilde{g}_i[kq_i - np_i] = \delta_{i-j}\delta_{n'-n}$$

pour tout entier n, n' et i,j=0…N-1. En posant $n = n_0 + n_1 q_i$ et $n' = n'_0 + n'_1 q_j$ on obtient

$$\sum_k g^{q_j}_{j,n'_0 p_j}[n'_1 p_j - n_1 p_i - k]\tilde{g}^{q_i}_{i,-n_0 p_i}[k] = \delta_{i-j}\delta_{n-n'}$$

c'est-à dire en particulier

$$\left(G^{q_j}_{j,n'_0 p_j}(z)\tilde{G}^{q_i}_{i,-n_0 p_i}(z)\right)_0^{\text{pgcd}(p_i,p_j)} = 0$$

pour tout $n_0 = 0\ldots q_i - 1,\ n_0' = 0\ldots q_j - 1$ et $i \neq j$. Ce qui signifie que si pgcd(p_i,p_j)=1, alors nécessairement

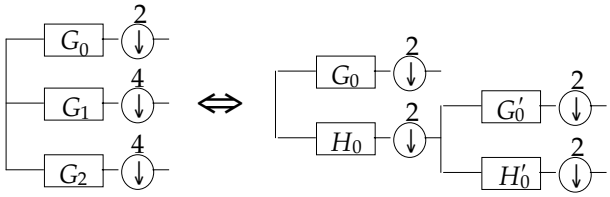
$$G_j(z^{q_i})\tilde{G}_i(z^{q_j}) = 0$$

ce que l'on ne peut obtenir avec des filtres réalisables.

Ceci dit, une condition nécessaire et suffisante pour avoir la même structure à l'analyse et à la synthèse, avec des filtres réalisables reste à trouver… Une condition suffisante serait par exemple que le banc de filtres d'analyse puisse s'écrire sous la forme itérée de bancs de deux bandes. En effet lorsque le banc de filtres se réduit à deux bandes (à échantillonnage critique bien sûr), alors le schéma de synthèse est toujours le miroir du schéma d'analyse, car à ce moment, tous les éléments de la matrice $\mathbf{\Gamma}$ sont indépendants. En d'autre termes, toute matrice polynômiale de dimension pxp peut être associée à un banc de filtres à deux bandes $(q/p,(p\text{-}q)/p)$ et inversement bien sûr. Comme exemple d'un tel banc de filtres itérés on pourrait avoir (1/15,2/15,1/5,3/5) qui peut se voir comme l'itération suivante

$$(1/5,4/5)\begin{cases}(1/3,2/3)\\(1/4,3/4)\end{cases}$$

Toutes ces considérations ne prennent cependant en compte que les facteurs d'échantillonnage. Il est clair que les filtres également doivent vérifier certaines propriétés afin que $\tilde{\mathbf{\Gamma}}^{-1}$ soit du type de $\mathbf{\Gamma}$. Ainsi, on sait construire un banc de filtres (1/2,1/4,1/4) par itération d'un banc de filtres dyadique, mais les filtres G_0, G_1 et G_2 associés ne sont pas quelconques selon cette construction. En fait, on peut montrer que tout banc de filtres (1/2,1/4,1/4) à inverse de type miroir est nécessairement le produit de l'itération de deux bancs de deux filtres dyadiques



Une démonstration simple consiste à écrire les équations de synthèse-analyse suivantes (en notant de façon évidente les filtres de reconstruction par des tildes "~")

$$\begin{aligned}\tilde{G}_0(z)G_0(z) + \tilde{G}_0(-z)G_0(-z) &= 2\\ \tilde{G}_0(z)G_1(z) + \tilde{G}_0(-z)G_1(-z) &= 0\\ \tilde{G}_0(z)G_2(z) + \tilde{G}_0(-z)G_2(-z) &= 0\end{aligned}$$

qui nous montrent que, d'une part, $\tilde{G}_0(z)$ est premier avec $\tilde{G}_0(-z)$, et d'autre part que G_1 et G_2 sont multiples de $\tilde{G}_0(-z)$. Plus précisément, on vérifie que

$$G_1(z) = zG_0'(z^2)\tilde{G}_0(-z)$$
$$G_2(z) = zH_0'(z^2)\tilde{G}_0(-z)$$

d'où l'assertion.

2. Complexité

Ainsi que nous l'avons vu, une branche peut se mettre sous la forme d'un filtrage matriciel après avoir fait une transformée polyphase de la sortie et une transformée polyphase inverse de l'entrée. On rappelle l'équation (II.1) où l'on fait $P=p=p_j$, $Q_j=q=q_j$ et où l'on élimine l'indice j

$$y_{n_0}^q[n] = \sum_{k_0=0}^{p-1} \sum_k g_{n_0 p-k_0 q}^{pq}[n-k] x_{k_0}^p[k]$$

La complexité d'une transformation polyphase étant nulle (il s'agit d'un échantillonnage avec différentes phases) ainsi que d'une transformation polyphase inverse, on doit donc estimer la complexité de q résultats d'une somme de p filtrages. Si Mult(L,M) (respectivement Add(L,M)) donne le nombre de multiplications (respectivement additions) nécessaires au calcul du produit de deux polynômes, de degré L et M, alors le nombre maximum de multiplications sera

$$pq\text{Mult}\left(\frac{L}{pq}, \frac{M}{p}\right)$$

et le nombre maximum d'additions sera

$$pq\text{Add}\left(\frac{L}{pq}, \frac{M}{p}\right) + (p-1)q\left(\frac{L}{pq} + \frac{M}{p} + 1\right)$$

Sans utiliser d'algorithme de convolution spécialement rapide, on a donc approximativement LM/p multiplications et autant d'additions. On peut donc dire qu'un banc de filtres rationnel a une complexité d'environ

$$M\sum_{i=0}^{N-1} L_i \,/\, p_i$$

additions et autant de multiplications si L_i est le degré de G_i. On peut voir cependant que si G_i est un filtre désiré sélectif, alors la largeur de bande de transition $\delta\omega_i$ de la branche rationnelle associée diminuera approximativement comme q_i/L_i (et non comme $1/L_i$) et donc que l'on

pourra écrire que la complexité du banc de filtres rationnel est approximativement M/ω en supposant tous les ω_i égaux.

Anticipons légèrement sur les bancs de filtres itérés en fractions d'octave et calculons la complexité de ce type de bancs de filtres. Dans ce cas on tire parti du calcul par itération qui réduit la complexité. Supposons qu'après la j$^{\text{ème}}$ itérations le signal à transformer comporte N_j échantillons. La complexité par échantillon du banc de filtres itéré jusqu'à l'infini sera alors de

$$\frac{L+l}{pN_0}\sum_j N_j$$

et comme $N_j = \frac{q^j}{p^j}N_0$ on en déduit que la complexité maximale du banc de filtres itéré sera $\frac{L+l}{p-q}$ multiplications et additions par échantillon.

3. Délai–Effets de bord

Les deux notions de délai et d'effet de bord sont étroitement liées. La première indique le retard en échantillons que l'on peut attendre d'une transformée d'analyse-synthèse qui serait causale. La seconde est une mesure de l'augmentation du nombre d'éléments du signal d'entrée (supposé de longueur finie) après analyse-synthèse.

a. Délai

Dans le cas d'un banc de filtres rationnel, on va ainsi s'intéresser au délai induit par la composition d'une branche d'analyse avec une branche de synthèse. En laissant tomber les indices de branche et en notant x'_n le résultat on a ainsi $x'_n = \sum_{k,k'} g[nq-k'p]g[k'p-kq]x_k$ avec des notations évidentes. Posons également

$$G(z) = \sum_l^L g_n z^n$$
$$\tilde{G}(z) = \sum_f^{\tilde{L}} \tilde{g}_n z^n$$

On a alors

$$k_{\max} \le E\left(\frac{k'_{\max}p-l}{q}\right) \le E\left(\frac{E\left(\frac{nq-\tilde{f}}{p}\right)p-l}{q}\right)$$

cette inégalité étant atteinte. Le délai de la transformation est alors défini par la quantité $d = \max_n(k_{\max}(n)-n)$. On obtient donc

$$d = \mathrm{E}\left(-\frac{l+\tilde{f}}{q}\right) \qquad \text{(II.5)}$$

Le délai du banc de filtres d'analyse-synthèse de la figure 3 sera alors la valeur maximale des délais de chaque branche, c'est-à dire avec des notations évidentes

$$d = \max_{0 \le j \le N-1} \mathrm{E}\left(-\frac{l_j + f_j}{q_j} \right) \tag{II.6}$$

Dans ce cas, une transformation simple comme

$$x_n \rightarrow \begin{cases} x_{2n'} & \text{où} \quad n' = E\left(\frac{n}{2}\right) \\ x_{2n''+1} & \text{où} \quad n'' = E\left(\frac{n-1}{2}\right) \end{cases} \rightarrow x_n$$

est à délai nul (transformation polyphase analyse-synthèse).

b. Effet de bord

Cependant, en général dans un banc de filtres on groupe les données d'entrée par bloc, ce qui induit automatiquement un délai de la taille de ce bloc que l'on appellera B. Le signal d'entrée pourra donc s'écrire sous la forme $x = \sum_b x_b$ où x_b est un signal de support temporel [bB,bB+B-1] et égal à x sur ce support. Le support du signal résultat de l'application d'une branche d'analyse-synthèse à x_b sera alors

$$\left[\mathrm{E}\left(\frac{f + q - 1 + p\mathrm{E}\left(\frac{l+p-1+bBq}{p}\right)}{q} \right), \mathrm{E}\left(\frac{L + p\mathrm{E}\left(\frac{L+bBq+q(B-1)}{p}\right)}{q} \right) \right]$$

ce qui est un résultat exact un peu trop compliqué. On remarque que cet intervalle est contenu dans

$$\left[-\mathrm{E}\left(-\frac{l+f}{q}\right) + Bb, \mathrm{E}\left(\frac{L+f}{q}\right) + Bb + B - 1 \right]$$

Il y a donc des effets de bords antérieurs de longueur Eb_a inférieure ou égale à $\mathrm{E}\left(-\frac{l+f}{q}\right)$ et des effets de bord postérieurs de taille Eb_p inférieure ou égale à $\mathrm{E}\left(\frac{L+f}{q}\right)$. Il faut donc noter que le délai de la transformation s'identifie à peu près aux effets de bords antérieurs. Pour la transformée totale, les effets de bord antérieur et postérieur sont majorés de la façon suivante

$$\begin{aligned} Eb_a &\le \max_{0 \le j \le N-1} \mathrm{E}\left(-\frac{l_j + f_j}{q_j} \right) \\ Eb_p &\le \max_{0 \le j \le N-1} \mathrm{E}\left(\frac{L_j + f_j}{q_j} \right) \end{aligned} \tag{II.7}$$

C'est bien sûr l'effet de bloc antérieur qui va imposer un délai pour la reconstruction du signal. En général, on considérera $B \geq Eb_a$ de telle sorte que le retard n'impose pas de garder plus d'un bloc en mémoire. Le délai associé à une telle transformation sera ainsi de 2 blocs, c'est-à dire $2B$.

c. Banc de filtres itéré

Dans le cas où l'on itère N fois l'une des branches d'un banc de deux filtres on obtient un banc de N-1 filtres grâce aux lois de composition des branches rationnelles. On note G le filtre passe-bas destiné à être itéré, H le filtre passe-haut et $\tilde{G}$, $\tilde{H}$ les filtres de reconstruction correspondant. Les entiers l, m, $\tilde{l}$, $\tilde{m}$ seront les plus petites puissances de z du développement de ces filtres.

Alors le délai minimum correspondant à la branche passe-bas itérée N fois et celui d'un passe-bas itéré j fois suivi d'un passe-haut seront respectivement

$$\mathrm{E}\left(-(l+\tilde{l})\frac{p^N-q^N}{q^N(p-q)}\right)$$
$$\mathrm{E}\left(-(l+\tilde{l})\frac{p^j-q^j}{q^j(p-q)}-(m+\tilde{m})\frac{p^j}{q^j(p-q)}\right)$$

et le délai total sera alors

$$d=\max\left[\mathrm{E}\left(-(l+\tilde{l})\frac{p^N-q^N}{q^N(p-q)}\right),\mathrm{E}\left(-(l+\tilde{l})\frac{p^{N-1}-q^{N-1}}{q^{N-1}(p-q)}-(m+\tilde{m})\frac{p^{N-1}}{q^{N-1}(p-q)}\right)\right]$$

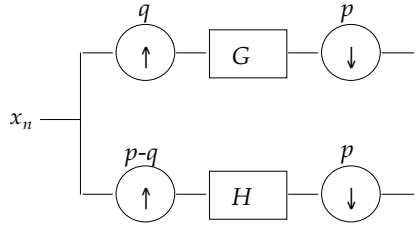
qui croît comme $\left(\frac{p}{q}\right)^N$ pour N grand. Dans le cas d'un banc de filtres orthonormés, on peut exprimer ce délai en fonction du seul degré L du polynôme passe-bas ce qui donne

$$\text{délai}=\mathrm{E}\left(L\frac{p^N-q^N}{q^N(p-q)}\right)$$

B. Cas de deux bandes

À l'instar du cas dyadique, on va construire des bancs de filtres rationnels par l'itération de la branche passe-bas d'un banc de deux filtres. On verra que cette itération engendre des fonctions limites dont les propriétés de régularité (entre autres) permettent de décrire de façon précise le banc de filtres final, qui définira donc une transformation à $\Delta f/f$ constant.

On s'intéresse maintenant au banc de filtres suivant



Selon la position des transformations polyphases, on aboutit à trois types de relations d'analyse–synthèse différentes

- la relation polyphase–polyphase
- la relation modulation–polyphase
- la relation modulation–modulation

1. Les relations de base

a. La relation polyphase–polyphase

Dans ce cas, comme on l'a vu plus haut, on insère une transformation polyphase et son inverse à l'entrée du banc de filtres mais aussi à la sortie. Pour une branche, on obtient ainsi les relations de filtrage matriciel démontrées plus haut

$$y_{n_0}^q[n] = \sum_{k_0=0}^{q-1} \sum_k g_{n_0 p - k_0 q}^{pq}[n-k] x_{k_0}^p[k] \quad \text{pour } n_0 = 0..q-1$$

$$Y_{n_0}^q(z) = \sum_{k_0=0}^{q-1} G_{n_0 p - k_0 q}^{pq}(z) X_{k_0}^p(z) \quad \text{pour } n_0 = 0..q-1$$

Dans ces conditions, les équations d'analyse ou de synthèse s'écrivent matriciellement sous la forme

$$\left(Y_k^q\right)_{0\le k\le q-1} = \underbrace{\left(G_{kp-lq}^{pq}\right)_{\substack{0\le k\le q-1\\ 0\le l\le p-1}}}_{\text{matrice polyphase-polyphase}} \left(X_k^p\right)_{0\le k\le p-1} \tag{II.8}$$

pour une branche, ici. Si l'on y ajoute la partie due au filtre passe-haut, on obtient donc la formulation polyphase-polyphase du problème.

La matrice du banc de filtres de synthèse s'obtient en échangeant formellement p et q, ce qui donne

$$\left(X_k^p\right)_{0\le k\le p-1} = \left(\tilde{G}_{kq-lp}^{pq}\right)_{\substack{0\le k\le p-1\\ 0\le l\le q-1}} \left(Y_k^q\right)_{0\le k\le q-1}$$

b. La relation modulation–polyphase

On insère cette fois une transformée polyphase et son inverse en sortie du banc de filtres seulement ce qui donne, pour une branche

$$y_{n_0}^q[n] = \sum_k g_{n_0 p}^q[np-k] x_k$$

$$Y_{n_0}^q(z^p) = \frac{1}{p} \sum_{k=0}^{p-1} G_{n_0 p}^q\left(z e^{2i\pi k/p}\right) X\left(z e^{2i\pi k/p}\right)$$

relation qui fait apparaître des versions polyphasées et modulées du filtre G. Cette méthode permet de transformer un banc de filtres à taux d'échantillonnage fractionnaire, en un banc de filtres uniforme, cas que l'on sait mieux traiter classiquement. Il s'agit en fait de la même opération qui est proposée par [KV3].

On aurait pu récupérer cette équation matricielle en utilisant la formule

$$\left(X\left(ze^{2i\pi k/p}\right)\right)_{0\le k\le p-1} = \mathbf{F}_p \,\mathrm{diag}(1, z, \dots z^{p-1})\left(X_k^p(z^p)\right)_{0\le k\le p-1}$$

où l'on a posé $\mathbf{F}_p = \left(e^{2i\pi kl/p}\right)_{0\le k,l\le p-1}$ la matrice de Fourier discrète. On obtient alors après quelques calculs

$$\left(Y_k^q(z^p)\right)_{0\le k\le q-1} = \underbrace{\frac{1}{p}\left(G_{kp}^q(ze^{2i\pi l/p})\right)_{\substack{0\le k\le q-1\\0\le l\le p-1}}}_{\text{matrice modulation/polyphase}}\left(X\left(ze^{2i\pi k/p}\right)\right)_{0\le k\le p-1}$$

On a ainsi un hybride entre modulation et transformation polyphase. En effectuant des calculs semblables, on obtient pour la branche passe-bas de synthèse

$$\left(X\left(ze^{2i\pi k/p}\right)\right)_{0\le k\le p-1} = \frac{1}{p}\left(\tilde{G}_{kq}^p(ze^{2i\pi l/q})\right)_{\substack{0\le k\le p-1\\0\le l\le q-1}}\left(Y_k^q(z^p)\right)_{0\le k\le q-1}$$

c. La relation modulation–modulation

Cette fois on fait apparaître des termes de modulation aussi bien pour x que pour y. Définissant p' et q' les entiers associés à p et q dans la relation de Bezout $pp' - qq' = 1$, on obtient alors

$$Y\left(z^p e^{2i\pi k/q}\right)_{0\le k\le q-1} = \underbrace{\frac{1}{p}\left(G(ze^{-2i\pi(k\frac{p'}{q}+l\frac{q'}{p})})\right)_{\substack{0\le k\le q-1\\0\le l\le p-1}}}_{\text{matrice modulation-modulation}} X\left(z^q e^{2i\pi k/p}\right)_{0\le k\le p-1}$$

et en intervertissant formellement p et q, on obtient la forme de l'opérateur correspondant de synthèse

$$X\left(z^q e^{2i\pi k/p}\right)_{0\le k\le p-1} = \frac{1}{q}\left(\tilde{G}(ze^{-2i\pi(k\frac{p'}{q}+l\frac{q'}{p})})\right)^T_{\substack{0\le k\le q-1\\0\le l\le p-1}} Y\left(z^p e^{2i\pi k/q}\right)_{0\le k\le q-1}$$

On constate ici que, même si sa forme est plus complexe que dans les deux précédentes descriptions, la matrice de synthèse s'obtient formellement par simple transposition (et bien sûr remplacement des filtres d'analyse par ceux de synthèse). Si l'on dénote $\mathbf{\Gamma}_{mm}$ la matrice modulation-modulation totale (c'est-à dire incluant également le filtre passe-haut) alors on a la relation de reconstruction parfaite $\Gamma_{mm}(z)\tilde{\Gamma}_{mm}(z) = \tilde{\Gamma}_{mm}(z)\Gamma_{mm}(z) = \mathbf{I}$. De la première on tire les relations suivantes

$$\begin{aligned}
&\sum_{k=0}^{p-1} G(ze^{2i\pi\frac{k}{p}})\overline{G}(ze^{2i\pi(\frac{k}{p}+\frac{s}{q})}) = pq\delta_s && \text{pour } s = 0..q-1\\
&\sum_{k=0}^{p-1} H(ze^{2i\pi\frac{k}{p}})\overline{H}(ze^{2i\pi(\frac{k}{p}+\frac{s}{p-q})}) = p(p-q)\delta_s && \text{pour } s = 0..p-q-1\\
&\sum_{k=0}^{p-1} G(z^{p-q}e^{2i\pi\frac{k}{p}})\overline{H}(z^q e^{2i\pi(-\frac{k}{p}+\frac{s}{p-q})}) = 0 && \text{pour } s = 0..p-q-1\\
&\sum_{k=0}^{p-1} \overline{G}(z^{p-q}e^{2i\pi\frac{k}{p}})H(z^q e^{2i\pi(-\frac{k}{p}+\frac{s}{p-q})}) = 0 && \text{pour } s = 0..p-q-1
\end{aligned} \qquad \text{(II.9)}$$

et de la deuxième on tire

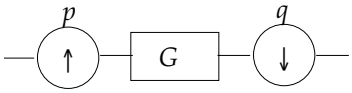
$$\frac{1}{pq}\sum_{l=0}^{q-1} \overline{G}(z^{p-q}e^{-2i\pi\frac{l}{q}})G(z^{p-q}e^{-2i\pi(\frac{k}{p}+\frac{l}{q})}) + \frac{1}{p(p-q)}\sum_{l=0}^{p-q-1} \overline{H}(z^q e^{-2i\pi\frac{l}{p-q}})H(z^q e^{-2i\pi(\frac{k}{p}+\frac{l}{p-q})}) = \delta_k \qquad \text{(II.10)}$$

pour $0 \le k,k' \le p-1$. Bien sûr, si l'on se limite à $p=2$ et $q=1$, on retrouve les équations simples du cas dyadique, et en particulier $G(z)\overline{G}(z) + G(-z)\overline{G}(-z) = 2$.

Ces équations nous seront très utiles par la suite, en particulier lorsque les filtres considérés seront orthonormaux.

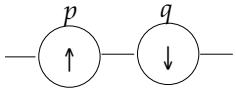
2. Conditions d'inversion

On peut remarquer, de manière préliminaire que comme dans le cas dyadique, il y a dualité entre analyse et synthèse: les bancs de filtres d'analyse et de synthèse peuvent être échangé tout en constituant à nouveau un système à reconstruction parfaite. Cela est dû au fait que l'échantillonnage est critique qui rend les matrices polyphases carrées. Se poser la question de l'inversibilité d'un banc de filtres d'*analyse* dont le passe-bas G est connu, revient donc à trouver un inverse de l'opérateur de *synthèse* (ou d'interpolation)



a. Condition par les déterminants (p-q=1)

Il faut tout d'abord remarquer que, contrairemement au cas où $q=1$, on peut trouver des opérateurs de sur-échantillonnage fractionnaire qui perdent strictement de l'information. Dès que $q>1$, il est en effet facile de construire de tels opérateurs. Par exemple, l'opérateur



perd toute l'information des échantillons du signal d'entrée d'indice $nq+s$ pour s=1…q-1. Il n'y a donc aucun espoir de reconstruction parfaite même avec un filtre de longueur infinie…

D'autre part, si l'on souhaite une inversion FIR, le cas dyadique apparaît comme extrêmement simple par rapport au cas rationnel. En effet, dans le cas dyadique, l'inversibilité FIR se résume à la condition que G ne soit pas divisible par un polynôme de la forme $z^2 - a$

Dans le cas rationnel au contraire, pour p-q=1, on va voir que la condition est plus complexe, impliquant des déterminants et irréductible à la simple divisibilité du polynôme G par un facteur de la forme $z^p - a$. On peut citer à ce sujet le polynôme suivant

$$G(z) = \frac{1}{9}\left(10 + 2z + 4z^2 + 5z^3 + 4z^4 + 2z^5\right)$$

dont la branche rationnelle (avec p/q=3/2) n'est pas inversible à l'aide de filtres FIR, bien que ses racines ne présentent pas de caractéristique particulière.

Théorème II.2 *Soit* $\mathbf{G}$ *la matrice polyphase de taille qxp issue du filtre* G

$$\mathbf{G} = \left[G^{pq}_{kp-lq} \right]_{\substack{0 \le k \le q-1 \\ 0 \le l \le p-1}}$$

et où p-q=1. Alors, l'opérateur de sur-échantillonnage fractionnaire $p\uparrow G\downarrow q$ *sera inversible FIR si et seulement si les mineurs d'ordre* q *de* $\mathbf{G}$ *sont globalement premiers entre eux.*

Preuve

Il s'agit de montrer qu'il existe un filtre H FIR tel que le banc de deux filtres composé de G et H soit inversible FIR, c'est-à dire vérifie $\det(\Gamma) = \text{Cte} \times z^n$ où $\mathbf{\Gamma}$ est la matrice polyphase associée au banc de filtres. Le développement du déterminant de $\mathbf{\Gamma}$ s'écrit comme une relation de Bezout entre tous les mineurs d'ordre q de la sous matrice $\mathbf{G}$ de $\mathbf{\Gamma}$ ce qui montre que ces mineurs sont premiers entre eux.

Supposons à l'inverse que les mineurs soient premiers entre eux. On peut donc écrire une relation de Bezout les reliant. Ces coefficients sont alors directement les éléments polyphasés du filtre passe-bas d'où le résultat.

On a dû se restreindre au cas p-q=1 car c'était le plus simple. Si p-q>1, alors on peut simplement affirmer que la condition sur les déterminants est nécessaire.

b. Obtention du passe-haut

Étant donné les filtres passe-bas, on peut en déduire des filtres passe-haut associés. C'est seulement dans le cas où p-q=1 que les filtres passe-haut sont uniques (à une constante multiplicative et un délai près). Alors il existe une formule utilisant les déterminants permettant de les donner exactement comme on va le voir maintenant.

Rappelons que l'équation d'analyse-synthèse du système deux bandes s'écrit sous la forme

$$\mathbf{I} = \begin{pmatrix} \left[\tilde{G}^{qp}_{lq-kp}(z) \right]_{\substack{0\le k\le q-1 \\ 0\le l\le p-1}} \\ \left[\tilde{H}^{(p-q)p}_{l(p-q)-kp}(z) \right]_{\substack{0\le k\le p-q-1 \\ 0\le l\le p-1}} \end{pmatrix}^T \begin{pmatrix} \left[G^{qp}_{kp-lq}(z) \right]_{\substack{0\le k\le q-1 \\ 0\le l\le p-1}} \\ \left[H^{(p-q)p}_{kp-l(p-q)}(z) \right]_{\substack{0\le k\le p-q-1 \\ 0\le l\le p-1}} \end{pmatrix}$$

D'autre part, on sait écrire l'inverse d'une matrice $\mathbf{\Gamma}$ donnée. Si l'on note $|\Gamma|^{i,j}$ le mineur d'ordre *i,j* de la matrice $\mathbf{\Gamma}$ —le déterminant de la matrice $\mathbf{\Gamma}$ à laquelle on a retiré la ligne *i* et la colonne *j*— on a en effet

$$\Gamma_{i,j} = (-1)^{i+j} \frac{|\tilde{\Gamma}|^{j,i}}{\det(\tilde{\Gamma})} \tag{II.11}$$

Bien sûr dans notre cas, le déterminant de $\tilde{\Gamma}$ est de la forme $\text{Constante} \times z^n$. Or dans le cas où p-q=1, la partie de chaque matrice liée au filtre passe-haut est un vecteur —ligne pour la matrice d'analyse et colonne pour la matrice de reconstruction—, ce qui implique que la formule (II.11) ne dépend, pour le calcul de *H*, que du filtre $\tilde{G}$. De même pour le calcul du filtre $\tilde{H}$.

Une observation simple concerne la longueur du filtre passe-haut: supposons que *G* et $\tilde{G}$ soient de longueur *L*. Alors on peut alors voir que la longueur de *H* est également de l'ordre de *L* par la formule (II.11).

Logiquement pourtant, dans la mesure où *H* agit sur le domaine même du signal d'entrée, alors que *G* agit sur le domaine *suréchantillonné de q*, le filtre *H* devrait être plutôt de taille L/q. On peut donc penser que dans le cas de la conception de filtres sélectifs, la formule (II.11) donne des filtres passe-haut dont la plupart des coefficients sont nuls. On verra que c'est effectivement le cas, ce qui nous incitera à considérer une autre méthode pour récupérer les filtres passe-haut, basée sur la décomposition de la matrice polyphase $\mathbf{\Gamma}$ en facteurs élémentaires (voir chapitre III)).

3. Propriétés statistiques

On ne s'est pour l'instant intéressé qu'aux propriétés mécaniques d'un banc de filtres, en particulier tout ce qui concerne l'inversion. Ceci ne dépend aucunement du signal analysé et en particulier, on a pour l'instant inclus dans les transformées admissibles toutes celles qui n'ont aucun pouvoir de "séparation fréquentielle" et celles qui ont la mauvaise propriété d'amplifier les erreurs apportées par des calculs numériques, ou par des opérateurs de quantification.

Il est donc important de considérer de nouvelles contraintes sur la transformée, ce qui nécessitera certaines hypothèses statistiques sur les signaux étudiés. En général, on considérera des signaux stationnaires jusqu'aux moments d'ordre deux à densité spectrale de puissance finie. Si x_n est notre signal d'entrée, on posera ainsi $\rho_{n-n'} = \langle x_n \bar{x}_{n'} \rangle$ et $R(z) = \sum_n \rho_n z^n$. On rappelle tout d'abord que l'opérateur branche n'est en général pas stationnaire. En particulier, si y_n est la sortie d'une branche donné par (I.7) y_n n'est pas stationnaire dès que *q* est différent de un. Cependant, il est cyclostationnaire si, outre une moyenne statistique, on effectue une moyenne temporelle sur un nombre fini d'éléments. On peut en effet démontrer la relation

$$\sum_{k=0}^{q-1}\langle|y_{n+k}|^2\rangle=\int_0^1\left|G(e^{-2i\pi\nu})\right|^2 R(e^{-2i\pi q\nu})d\nu \qquad \text{(II.12)}$$

qui ne dépend plus de n. On reconnaît ici la densité spectrale de puissance $R(e^{-2i\pi\nu})$ de x_n, qui a la propriété d'être réelle et positive. Cette relation est formellement très proche de celle que l'on obtient dans le cas d'un filtrage simple. Elle présente bien sûr des analogies avec la formule en z de (I.7).

Dans le but de préciser un peu plus la non-stationnarité de l'opérateur, on peut affiner la démonstration qui conduit à (II.12) et l'on obtient la majoration suivante

$$\left|\langle|y_n|^2\rangle-\tfrac{1}{q}\sum_{k=0}^{q-1}\langle|y_{n+k}|^2\rangle\right|\le\rho_0\,\tfrac{q-1}{q}\max_{1\le k\le q-1}\sup_{\nu}\left|G(e^{-2i\pi\nu})G(e^{-2i\pi(\nu+\frac{k}{q})})\right|$$

Le terme de droite est en fait petit par rapport à ρ_0 dès que le filtre G est suffisamment passe-bas, en particulier si sa bande atténuée contient l'intervalle $[1/2q,1\text{-}1/2q]$, c'est-à dire que si le G laisse passer l'intégralité du signal x_n, alors l'erreur de non-stationnarité sera très faible en valeur relative. Par contre, si le signal est filtré de manière importante, le bruit de non stationnarité pourra concurrencer le résidu basse fréquence de x_n dans y_n. On verra (chapitre IV) que le terme qui régit ce bruit est étroitement lié au comportement non invariant des fonctions limites, l'amnésie. Ceci nous incite à considérer la forme des filtres idéaux, dans une optique de conception de filtres.

a. Forme des filtres idéaux

Dans notre cas, les filtres idéaux seront des filtres à coefficients réels qui sépareront de manière parfaite le signal d'entrée en ses composantes passe-bas et passe-haut, tout en ne perdant pas d'information bien sûr, afin de pouvoir reconstruire. Dans ce cas, la répartition des bandes de fréquence entre le passe-bas et le passe-haut sera donnée par les intervalles $[0,q/2p]$ et $[q/2p,1/2]$.

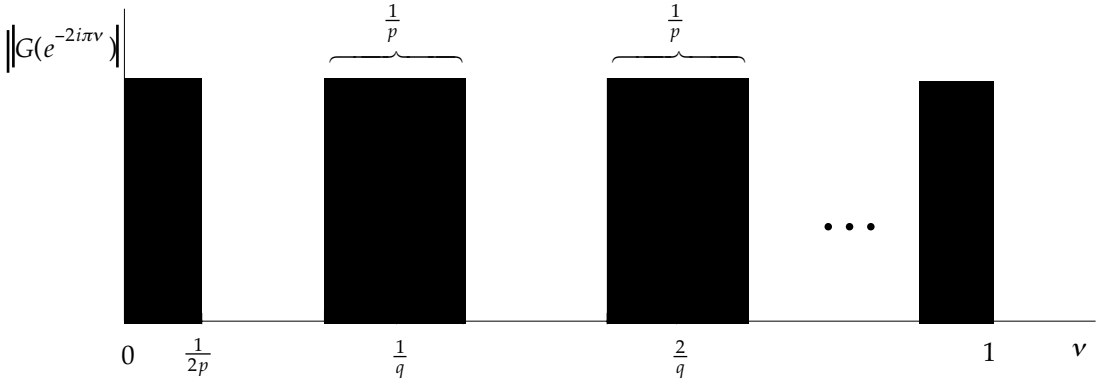
Puisque la branche passe-bas ne laisse passer que les fréquences de l'intervalle $[0,q/2p]$ on déduit de (II.12) que

$$G(e^{-2i\pi\frac{\nu+k}{q}})=0 \qquad \begin{matrix}\forall k=0\ldots q-1\\ \forall\nu\in\left[\frac{q}{2p},\frac{1}{2}\right]\end{matrix}$$

Le support de $G(e^{-2i\pi\nu})$ est donc contenu dans la réunion d'intervalles

$$\bigcup_{k=0}^{q-1}\left[\frac{k}{q}-\frac{1}{2p},\frac{k}{q}+\frac{1}{2p}\right]$$

c'est-à dire une suite d'intervalles centrés en les fréquences $1/q$ et de largeur $1/p$



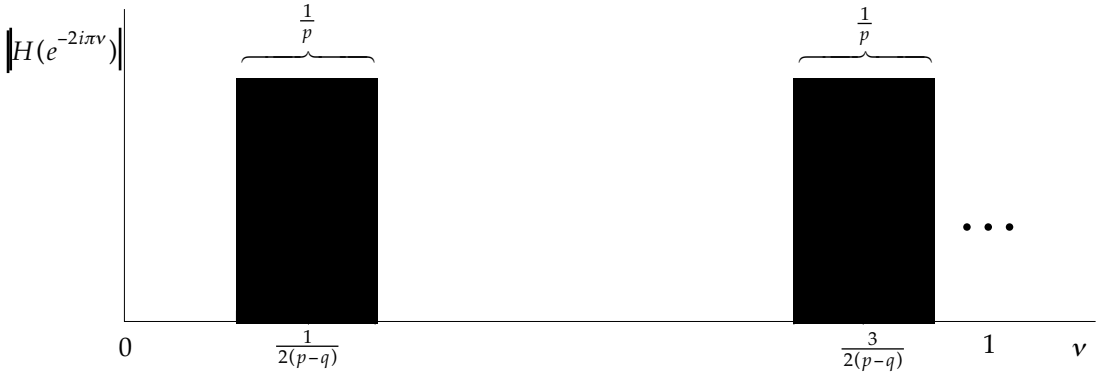
Cela diffère du cas dyadique où le support de $G(e^{-2i\pi\nu})$, que l'on suppose de taille $1/p$ (donc $1/2$ cas dyadique) pour assurer l'inversibilité du banc de filtres, est connexe par nature. Il suffira cependant de considérer le filtre dont le support est [-1/2*p*,1/2*p*]: en effet les intervalles [-1/2*p*+*k*/*q*,1/2*p*+*k*/*q*] n'apportent aucune information supplémentaire sur le signal puisqu'ils correspondent à des répliques de la partie qui est contenue dans [-1/2*p*,1/2*p*] et d'autre part cette bande de fréquence n'est pas dégradée par le repliement de spectre.

Examinons maintenant le support de $H(e^{-2i\pi\nu})$. De même que pour *G*, on a

$$H(e^{-2i\pi\frac{\nu+k}{p-q}}) = 0 \quad \begin{matrix} \forall k = 0 \ldots p-q-1 \\ \forall \nu \in \left[0, \frac{q}{2p}\right] \end{matrix}$$

et donc le support de $H(e^{-2i\pi\nu})$ est contenu dans

$$\bigcup_{k=0}^{p-q-1} \left[\frac{k+1/2}{p-q} - \frac{1}{2p}, \frac{k+1/2}{p-q} + \frac{1}{2p}\right]$$



Bien sûr, on souhaite que le repliement de spectre dû au sous échantillonnage par *p* soit nul, ce qui implique que le support fréquentiel de *H* soit de taille $1/p$. Si *H* était un filtre complexe, il suffirait donc de prendre l'un quelconque des intervalles $\left[\frac{k+1/2}{p-q} - \frac{1}{2p}, \frac{k+1/2}{p-q} + \frac{1}{2p}\right]$ pour que le support de *H* soit connexe, puisque chacun vérifie la propriété de non repliement (leur taille est exactement $1/p$).

Les choses deviennent plus compliquées quand H est réel, car alors son support est symétrique par rapport à $1/2$. Il est cependant encore possible de lui choisir un support connexe quand p-q est un nombre impair. En effet, dans ce cas on prend $k=(p-q-1)/2$ ce qui donne l'intervalle

$$\left[\frac{1}{2}-\frac{1}{2p},\frac{1}{2}+\frac{1}{2p}\right]$$

qui est bien symétrique autour de $1/2$ (et n'est pas repliable).

Par contre, si p-q est pair le filtre H ne sera pas connexe sur [0,1] puisque $1/2$ n'appartient pas à la réunion d'intervalles donnée plus haut. Cependant, on peut encore le choisir connexe sur $[0,1/2]$: cet intervalle sera sur $[0,1/2]$ de la forme $\left[\frac{n}{2p},\frac{n+1}{2p}\right]$ où n est un entier, ce qui nous assure que, réuni avec son symétrique $\left[-\frac{n+1}{2p},-\frac{n}{2p}\right]$, le repliement après échantillonnage par p sera nul. Il faut (et il est suffisant) alors qu'il existe k,k' tels que

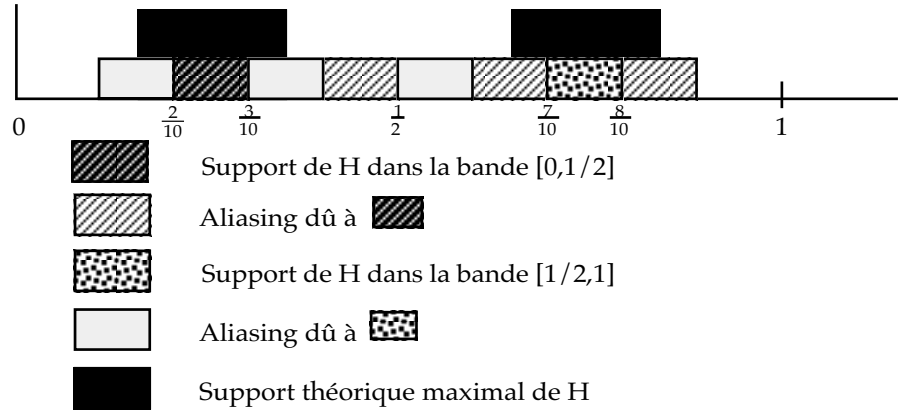
$$\frac{k+1/2}{p-q}-\frac{1}{2p}\le\frac{n}{2p} \qquad \frac{n+1}{2p}\le\frac{k+1/2}{p-q}+\frac{1}{2p} \qquad \text{et} \qquad \frac{k'+1/2}{p-q}-\frac{1}{2p}\le-\frac{n+1}{2p} \qquad -\frac{n}{2p}\le\frac{k'+1/2}{p-q}+\frac{1}{2p}$$

Sous ces conditions, le support $\left[\frac{n}{2p},\frac{n+1}{2p}\right]\cup\left[-\frac{n+1}{2p},-\frac{n}{2p}\right]$ sera bien admissible puisqu'il sera contenu dans $\bigcup_{k=0}^{p-q-1}\left[\frac{k+1/2}{p-q}-\frac{1}{2p},\frac{k+1/2}{p-q}+\frac{1}{2p}\right]$ et ne sera pas sujet au repliement de spectre. L'étude de ces inégalités conduit rapidement à $k+k'+1=0$ et réduit à deux le nombre d'inégalités. Il est alors facile de voir que si l'on se donne k, alors n existe toujours et est donné de manière unique (car $p-q\equiv 1 \bmod(2)$) par

$$n=E\left(\frac{(2k+1)p}{p-q}\right)$$

Un choix qui permet de conserver à H sa notion de "passe-haut" (disons plutôt "passe-bande haut"…) consiste à prendre $k=E\left(\frac{p-q}{2}\right)$. Il faut noter que si on fait p-q impair, ce choix permet de retrouver le support donné plus haut $\left[\frac{1}{2}-\frac{1}{2p},\frac{1}{2}+\frac{1}{2p}\right]$ correspondant effectivement à p-q impair.

Ci dessous est donné un exemple dans le cas $p/q=5/3$



En général, le problème de la conception de filtre est rendu assez difficile du fait des nombreux paramètres qui entrent en ligne de compte, de la nonlinéarité du problème, et aussi du fait que les contraintes imposées aux filtres sont toujours contradictoires (sélectivité du filtre, mais également longueur finie), ou discutables (doit-on minimiser au sens de la norme L^∞ ou bien au sens L^2?).

b. Filtres orthonormaux

On s'est cependant rendu compte qu'il était possible d'imposer une contrainte relativement simple qui, assez miraculeusement, ramène le problème de la conception d'un banc de deux filtres à celui du seul filtre passe-bas d'analyse. La conception de ce filtre est même simplifiée par le fait que le problème d'optimisation équivaut à la seule minimisation de la bande atténuée de ce filtre. Cette contrainte est la paraunitarité, ou orthonormalité. Dans la littérature, un filtre la vérifiant est fréquemment dit "sans perte" ("lossless" en anglais). En outre, ce type de banc de filtres jouit d'un certain nombre de propriétés statistiques remarquables.

Il s'agit de contraindre la reconstruction à être semblable à l'analyse. Plus précisément, on impose $\tilde{G}(z) = G(z^{-1})$ et $\tilde{H}(z) = H(z^{-1})$. Il est facile de voir d'après la forme des matrices que cette contrainte est équivalente à imposer $\tilde{\mathbf{\Gamma}}(z) = \mathbf{\Gamma}(z^{-1})^T$ et donc en ne conservant qu'une seule matrice

$$\mathbf{\Gamma}(z^{-1})^T \mathbf{\Gamma}(z) = \mathbf{I}$$

Ce type de matrice a été étudié assez intensément dans la littérature, en particulier par Vaidyanathan [Vai2,VH,VNDS], et l'on peut donner une façon de construire toutes les matrices vérifiant cette équation matricielle. Cette méthode (en fait une factorisation de toutes les matrices paraunitaires) sera exposée dans le chapitre suivant, ainsi que sa généralisation au cas quelconque, c'est-à dire biorthogonal.

Certaines propriétés rendent ce type de transformation particulièrement attrayant, en particulier pour la conception de filtres. On montre ainsi à l'aide des relations (II.9) que

$$\sum_{k=0}^{p-1} \left| G(e^{-2i\pi\frac{\nu+k}{p}}) \right|^2 = pq$$
$$\sum_{k=0}^{p-1} \left| H(e^{-2i\pi\frac{\nu+k}{p}}) \right|^2 = p(p-q)$$

Comme le filtre passe-bas idéal a un support contenu dans $[-1/2p_j, 1/2p_j]$ sur lequel il est constant en module, il suffira de minimiser la norme du filtre sur la bande atténuée pour s'assurer de sa constance sur la bande passante.

La situation est plus complexe pour H si $p - q \not\equiv 1 (\mathrm{mod}\, 2)$ puisque le filtre idéal, passe-bande, n'est plus connexe. En tous cas, (II.10) conduit à

$$\frac{1}{pq} \sum_{l=0}^{q-1} \left| G(e^{-2i\pi\frac{\nu+l}{q}}) \right|^2 + \frac{1}{p(p-q)} \sum_{l=0}^{p-q-1} \left| H(e^{-2i\pi\frac{\nu+l}{p-q}}) \right|^2 = 1 \qquad \text{(II.13)}$$

ce qui montre que la sélectivité de la branche passe-bas impose automatiquement la sélectivité de la branche passe-haut (en utilisant la relation (II.15) ci-après).

c. Dynamique des coefficients de la transformée

Ce que l'on entend par dynamique d'un signal est en général la différence entre la plus grande et la plus petite valeur absolue de ses échantillons

$$D = \max_n |x_n| - \min_n |x_n|$$

Pour un signal aléatoire, on modifie la définition dans un sens plus faible: on fixera une probabilité arbitrairement faible ε dans le but de définir le maximum M et le minimum m du signal dans un sens probabiliste sous la forme $P(|x_n| \geq M) = \varepsilon$ et $P(|x_n| \leq m) = \varepsilon$. On pose alors

$$D = M - m$$

En général nous aurons des signaux dont les valeurs proches de zéro seront fréquentes d'où m=0. On peut d'autre part relier assez facilement M à l'écart-type du signal (à l'aide du théorème de Bienaymé-Tchebitcheff) $M \leq \varepsilon^{-1/2} \langle |x_n|^2 \rangle^{1/2}$. On pourra ainsi dire que

$$D = C \sqrt{\langle |x_n|^2 \rangle} \tag{II.14}$$

où C est une constante dépendant de ε. Par exemple, si la loi de probabilité de x_n est équirépartie, $C = \sqrt{3}$ permet d'avoir ε=0. Bien évidemment, si la loi de probabilité change, cette constante peut prendre des valeurs très différentes, dans les limites du théorème de Bienaymé-Tchebitcheff. Cependant, comme on n'aura pas réellement besoin de la valeur *exacte* de la dynamique D, mais une valeur la majorant, on considèrera en général que l'on peut identifier écart moyen et dynamique à l'aide de (II.14) pour une constante C fixée (par exemple 2), indépendante de la loi de probabilité des échantillons.

En utilisant (II.12), on peut estimer la dynamique des coefficients de sortie de la transformée. On a en effet

$$\langle |x_n|^2 \rangle \inf_\nu A(e^{-2i\pi\nu}) \leq \frac{1}{p} \sum_{k=0}^{q-1} \langle |y_{n+k}|^2 \rangle + \frac{1}{p} \sum_{k=0}^{p-q-1} \langle |z_{n+k}|^2 \rangle \leq \langle |x_n|^2 \rangle \sup_\nu A(e^{-2i\pi\nu}) \tag{II.15}$$

si l'on pose

$$A(e^{-2i\pi\nu}) = \frac{1}{pq} \sum_{l=0}^{q-1} \left| G(e^{-2i\pi \frac{\nu+l}{q}}) \right|^2 + \frac{1}{p(p-q)} \sum_{l=0}^{p-q-1} \left| H(e^{-2i\pi \frac{\nu+l}{p-q}}) \right|^2$$

En particulier, si le système est orthonormal, A=1 et cette double inégalité se transforme en égalité. Si l'erreur due à la non invariance temporelle du système est faible, on aura ainsi

$$\langle |y_n|^2 \rangle \leq \frac{p}{q} \langle |x_n|^2 \rangle$$
$$\langle |z_n|^2 \rangle \leq \frac{p}{p-q} \langle |x_n|^2 \rangle$$

d. Gain de codage

On va ici considérer un banc de filtres général de la forme indiquée en figure 3. Afin de pouvoir transmettre, ou stocker numériquement les coefficients de la transformée, il est indispensable de les quantifier, une fois leur dynamique évaluée. On va alors pouvoir compter, en bits, la quantité d'information retenue I. On souhaite cependant que l'erreur sur le signal reconstruit soit inférieure à un certain seuil (de perception par exemple). On est ainsi conduit à minimiser I sous contrainte que cette erreur soit bornée par une constante fixée à l'avance.

Supposons que nous n'utilisions que de la quantification linéaire, alors posons $\sigma_j = \sqrt{\langle |x_j[n]|^2 \rangle}$ et soit C la constante invariable permettant de relier dynamique et écart-type par (II.14). Soit également η_j le pas de quantification utilisé pour la sortie j. On peut donc écrire la quantité d'information par échantillon (ou débit) sous la forme

$$R = \sum_{j=0}^{N-1} \frac{q_j}{p_j} \log_2\left(\frac{C\sigma_j}{\eta_j}\right) \qquad \text{(II.16)}$$

Supposons que l'opération de quantification soit équivalente à l'addition d'une variable aléatoire $e_j[n]$ à chaque sortie $x_j[n]$, ayant les caractéristiques d'un bruit blanc et statistiquement indépendante d'une sortie à l'autre. Alors, η_j pourra être identifiée à l'écart-type de la variable aléatoire e_j, c'est-à dire que l'on aura $\langle |e_j[n]|^2 \rangle = \eta_j^2$. Définissons enfin, comme au début de ce chapitre, P comme le plus petit commun multiple des p_j. Les variables aléatoires e_j induiront sur le signal reconstruit une erreur $e[n]$ dont l'expression est

$$e[n] = \sum_{j=0}^{N-1} \sum_k g_j[nq_j - kp_j] e_j[k]$$

qui n'est pas un bruit blanc et n'est même pas stationnaire. L'écart moyen $\eta^2[n]$ de l'erreur entre le signal reconstruit après quantification avec le signal d'entrée sera alors donnée par

$$\eta^2[n] = \sum_{j=0}^{N-1} \eta_j^2 \sum_k |g_j[nq_j - kp_j]|^2$$

qui est périodique de période P et l'on aura

$$\eta^2 = \frac{1}{P} \sum_{n=0}^{P-1} \eta^2[n] = \sum_{j=0}^{N-1} \eta_j^2 \frac{1}{p_j} \sum_k |g_j[k]|^2$$

On peut par ailleurs démontrer à l'aide d'inégalités triangulaires et de Cauchy-Schwartz que

$$\left|\sum_k |g[nq-kp]|^2 - \tfrac{1}{p}\sum_k |g[k]|^2\right| \le C_0\left(\int_{\frac{1}{2p}}^{1-\frac{1}{2p}} \left|G(e^{-2i\pi\nu})\right|^2 d\nu\right)^{1/2}$$

où $C_0 = \frac{1}{p}\left[2\left(p\sum_k |g[k]|^2\right)^{1/2} + 1\right]$ ce qui signifie que si les filtres G_j sont suffisamment proches des filtres idéaux on aura $\eta^2[n] \approx \eta^2$ (noter que l'on peut considérer la constante C_0 comme relativement indépendante de G puisque pour des filtres orthonormaux on a $C \le 2$). En notant $a_j = \frac{1}{p_j}\sum_k \left|g_j[k]\right|^2$ —remarquons que si les filtres sont orthogonaux $G_j(z) = G_j(z^{-1})$ alors $a_j = \frac{q_j}{p_j}$— le lien entre pas de quantification sur chaque branche et écart moyen de l'erreur de reconstruction s'écrira

$$\sum_k a_j \eta_j^2 = \eta^2 \tag{II.17}$$

Le problème de minimisation se traduira mathématiquement par la recherche des paramètres de quantification η_j, minimisant (II.16) sous la contrainte (II.17). La solution est alors

$$\eta_j = \eta\sqrt{\frac{q_j}{a_j p_j}} \tag{II.18}$$

Le gain de quantification s'exprimera alors comme l'exponentielle de la différence du débit qui aurait été nécessaire en quantifiant directement le signal d'entrée avec un pas de quantification égal à η avec celui obtenu après optimisation, d'où

$$G = \frac{\sigma}{\prod_{j=0}^{N-1}\sigma_j^{q_j/p_j}}\sqrt{\prod_{j=0}^{N-1}\left(\frac{q_j}{a_j p_j}\right)^{q_j/p_j}} \tag{II.19}$$

C. Résumé du chapitre

On s'est ici préoccupé essentiellement des propriétés de reconstruction des bancs de filtres rationnels: on a ainsi montré comment calculer la matrice polyphase du système, sans cependant utiliser la technique à deux transformées de [KV1]: les résultats sont bien sûr équivalents.

On s'est également penché sur les effets de bords dus au banc de filtres rationnel ainsi que le délai induit par une transformation complète et son inverse, dont on a donné les formules en évaluant l'amplitude.

Comme nous serons préoccupés essentiellement par les bancs de filtres itérés, il était important de donner les formules précises concernant le cas deux bandes, en particulier pour ce qui concerne l'inversion du banc de filtres et la forme des filtres idéaux. On a également anticipé

sur la transformation issue de l'itération afin d'en déduire certaines caractéristiques statistiques qui en faciliteront l'interprétation.

III. Factorisation

L'utilisation des matrices polynômiales apparaît indispensable à l'étude des bancs de filtres [Vai2,Vet]. Il n'y a là, en fait, qu'une simple extension des polynômes ($\mathbf{R}[X]$ à coefficients réels ou $\mathbf{C}[X]$ à coefficients complexes) qui sont utilisés couramment pour décrire les opérations de filtrage. Simplement les coefficients du développement polynômial sont des matrices, au lieu d'être des scalaires (réels ou complexes).

On va se restreindre ici aux matrices carrées qui correspondent, ainsi qu'on l'a vu au chapitre II aux bancs de filtres à échantillonnage critique. Il peut cependant être parfois nécessaire de recourir à des matrices non carrées, par exemple si l'on considère des bancs de filtres à échantillonnage non critique, ou si l'on doit prendre en compte une sous-matrice d'une matrice carrée (par exemple celle concernant la branche passe-bas d'un banc de filtres). On indiquera alors des résultats plus généraux qui s'appliquent à de telles matrices rectangulaires.

Dans le cas des matrices carrées précisément, on bénéficie d'un grand nombre de propriétés qui apparentent les polynômes aux entiers naturels, et les fractions rationnelles aux nombres rationnels. En particulier on a

- une structure d'anneau commutatif muni des lois de composition "+" et "×" pour les polynômes
- une structure de corps commutatif muni des lois de composition "+" et "×" pour les fractions rationnelles
- l'existence d'éléments premiers (de degré deux dans le cas de $\mathbf{R}[X]$, et de degré 1 dans le cas de $\mathbf{C}[X]$) permettant la factorisation de tous les polynômes
- la division euclidienne (dont le degré du reste est inférieur à celui du diviseur), pgcd et ppcm (plus grand diviseur et multiple commun), la relation de Bezout

Il se trouve que dès que l'on passe aux matrices polynômiales carrées d'ordre D, on perd un grand nombre de ces propriétés: la plus évidente (mais pas la plus importante) est la commutativité de la multiplication. En fait on se rend également compte qu'il existe des "diviseurs de zéro", ce qui empêche d'étendre l'anneau (non commutatif) des matrices polynômiales à un corps, puisque tout élément n'aurait pas d'inverse. Bien sûr, si l'on se restreint aux seules matrices polynômiales non diviseurs de zéro, on perd la propriété de groupe additif…

Le plus gros problème soulevé par les matrices polynômiales carrées est le fait que leur inversibilité FIR nécessite la condition simple mais très non-linéaire que son déterminant soit un délai pur. Une façon d'éliminer ce problème dans le cas de matrices paraunitaires (voir la définition plus loin), il a été mis en œuvre la factorisation de ces matrices en éléments paraunitaires de degré un [Vai1,Va2,VH,VNDS]. Ce résultat important sera rappelé dans la suite de ce chapitre: quelques améliorations y seront d'ailleurs apportées, en particulier pour ce qui est de la factorisation des matrices rectangulaires (un sujet également traité dans [CV]), et sur la minimalité de cette factorisation. L'intérêt d'une telle approche est que la condition de reconstruction parfaite FIR est *structurellement* vérifiée, et donc que pour la conception de filtres par exemple il n'y a plus qu'à minimiser sur un ensemble de paramètres libres —ou presque—, en tous cas sans la contrainte non-linéaire de reconstruction parfaite.

Cependant malgré l'ancienneté de techniques de factorisations, une factorisation générale des matrices biorthogonales, par opposition à orthogonales ou paraunitaires, n'a pas encore été publiée dans la littérature de traitement de signal. Diverses factorisations sont bien

connues et permettent de décomposer en facteurs simples des matrices de filtres à phase linéaire [NV,SVN] ou mettent en évidence des facteurs intéressants [VG]. Le problème avait d'ailleurs été décrit, à tort, par Vaidyanathan comme ouvert [Vai2]. Très récemment [VC1,VC2], ce même auteur a proposé une classe importante de facteurs simples qui permettent de construire des matrices unimodulaires: on verra en particulier la similitude entre la forme des éléments de base proposés et celle que je propose. Dans [VC2] il est ainsi montré que l'on peut factoriser toute BOLT —*Biorthogonal Lapped Transform* à qui correspond une matrice polyphase de degré matriciel un— sous forme de facteurs $\mathbf{I} + (z-1)uv^T$ où l'on impose $v^T u = 1$ et qui sont des facteurs FIR inversibles. Il reste que la factorisation générale des matrices reste jusqu'à maintenant un problème dont la solution semble peu connue dans la communauté de traitement de signal: c'est cela que je me propose de traiter dans ce chapitre.

A. Degrés d'une matrice polynômiale

Pour une matrice polynômiale de taille $D \times D$, on peut bien évidemment définir au moins D^2 degrés différents correspondant à chaque composante de la matrice. Cependant, par la suite on utilisera essentiellement trois types de degrés, sur lesquels on pourra effectuer des raisonnements par récurrence. Il s'agit des

- degré matriciel, abrégé en "degM"
- degré vectoriel, abrégé en "degV"
- degré déterminant, abrégé en "degD"

que l'on va maintenant définir.

1. Degré matriciel

Par définition une matrice polynômiale (non nécessairement carrée) $\mathbf{M}(z)$ s'écrit sous la forme

$$\mathbf{M}(z) = \sum_{n=M}^{N} \mathbf{M}_n z^n$$

et on appellera alors degré matriciel (degM) de la matrice $\mathbf{M}$, l'entier N. C'est d'ailleurs une définition que l'on appliquera indifféremment à des matrices carrées ou rectangulaires.

Il s'agit là, comme on peut le constater immédiatement, d'un invariant par changement de base. Ce degré permet d'avoir accès directement au nombre d'éléments non nuls dans l'expansion de la matrice, par $\mathrm{degM}(\mathbf{M}(z)+\mathrm{degM}(\mathbf{M}(z^{-1}))+1$. On a bien sûr la relation

$$\mathrm{degM}(\mathbf{M}_1\mathbf{M}_2) \le \mathrm{degM}(\mathbf{M}_1) + \mathrm{degM}(\mathbf{M}_2)$$

2. Degré vectoriel

À la différence du degré matriciel il s'agit là d'une grandeur non invariante par changement de base. Ainsi, si l'on pose $\mathbf{M}(z) = \left[C_{k,l}(z)\right]_{1\le k,l\le D}$, on définit le degré vectoriel (degV) de la matrice $\mathbf{M}(z)$ par

$$\mathrm{degV}(\mathbf{M}) = \sum_{k=1}^{D} \max_{1\le l\le D} \deg\left(C_{k,l}\right)$$

c'est-à dire par la somme des degrés matriciels des vecteurs-ligne de la matrice $\mathbf{M}(z)$. Cette définition nous sera très utile lors de la récurrence destinée à factoriser la matrices unimodulaires. Comme le degré matriciel, cette définition s'étend sans difficulté au cas des matrices rectangulaires.

3. Degré déterminant

On définit enfin le degré déterminant (degD) par l'expression

$$\mathrm{degD}(\mathbf{M}) = \deg \det(\mathbf{M})$$

qui est donc un invariant par changement de base. Ce degré est utile lors de la factorisation des matrices paraunitaires. On a alors exactement $\mathrm{degD}(\mathbf{M}_1\mathbf{M}_2) = \mathrm{degD}(\mathbf{M}_1) + \mathrm{degD}(\mathbf{M}_2)$. À la différence des deux autres degrés, il ne présente pas de sens de l'étendre aux matrices non carrées.

B. Matrices polynômiales FIR–inversibles

Jouant le même rôle que les monômes z^n pour les polynômes, les matrices polynomiales de taille $D\times D$ FIR-inversibles divisent toutes les matrices polynômiales de même taille. On note donc G l'ensemble des matrices de taille $D\times D$ et défini par la propriété

$$\forall \mathbf{M}\in G \quad \exists \tilde{\mathbf{M}}\in G \quad \text{telle que} \quad \mathbf{M}(z)\tilde{\mathbf{M}}(z)^T = \mathbf{I} \qquad \text{(III.1)}$$

L'opérateur tilde "˜" ainsi défini équivaut donc à la transposition et à l'inversion d'une matrice. On constate immédiatement les propriétés suivantes

- $\mathbf{M}\in G$ si et seulement si on peut trouver un entier n et une constante C non nulle tels que $\det(\mathbf{M}(z)) = Cz^n$. On se ramène ainsi au cas des polynômes; malheureusement cette relation, sous sa simplicité formelle est trop complexe pour être exploitée, par exemple dans la conception de filtres

- G est un groupe multiplicatif —mais malheureusement pas additif!—

- si $\mathbf{M} \in G$ alors $\mathbf{M}^T \in G$ également. De même si $a \neq 0$, $\mathbf{M}(az^n) \in G$

Les matrices de G apparaissent naturellement dans les bancs de filtres [Vet] ainsi qu'on l'a vu dans le chapitre précédent, dès que l'on désire avoir des filtres d'analyse et de synthèse à réponse impulsionnelle finie. Le rôle du présent chapitre est donc de mieux décrire ces matrices, en particulier, on va voir qu'il est possible de les factoriser en éléments simples qui sont de deux types, paraunitaires et unimodulaires, que je présente maintenant.

1. Matrices paraunitaires

Ces matrices polynômiales constituent un sous-ensemble P de G. Une matrice $\mathbf{M}$ appartient à P si et seulement si elle vérifie la relation [Vai2]

$$\mathbf{M}(z)\mathbf{M}\left(\frac{1}{z}\right)^T = \mathbf{Id} \qquad \text{(III.2)}$$

Cette équation implique bien évidemment (III.1), et on a alors l'égalité $\tilde{\mathbf{M}}(z) = \mathbf{M}\left(z^{-1}\right)^T$. Bien que l'on définisse fréquemment l'opérateur tilde sous cette forme dans la littérature des bancs de filtres "orthogonaux", ou "lossless", ou encore "paraunitaires", il ne faut pas perdre de vue que dans le corps de cette thèse où l'on s'intéresse également aux bancs de filtres "biorthogonaux" l'acception de l'opérateur tilde est plus large. Un exemple simple de matrice paraunitaire est $\mathbf{M}(z) = \mathbf{Id} + (z-1)uu^T$ où le vecteur u a été choisi unitaire: $u^T u = 1$.

Les matrices $\mathbf{M}$ de P vérifient les propriétés suivantes

- $\det(\mathbf{M}(z)) = \pm z^n$ où n est un entier. Donc, la restriction de P aux matrices constantes est le groupe orthogonal des matrices de déterminant égal à ± 1
- P est un sous groupe multiplicatif de G

2. Matrices unimodulaires

À l'inverse des matrices paraunitaires, ces matrices ont un succès bien moindre en traitement du signal, probablement à cause de leurs propriétés un peu contre nature quand on se réfère aux polynômes: elles présentent un inverse anticausal dont l'intérêt a seulement récemment été mis en évidence [VC1]. Il faut dire que, employées seules, elles présentent de piètres performances fréquentielles. D'autre part leur maniement est plus difficile que celui des matrices paraunitaires, puisqu'il n'y a pas de relation simple entre une matrice et son inverse comme c'est le cas des matrices paraunitaires. D'un point de vue théorique, et en tous cas du côté des systèmes ces matrices sont mieux connues [Kai] et participent même à la décomposition des matrices polynômiales quelconques (non nécessairement FIR inversibles) à travers la forme de Smith-McMillan que l'on rappellera plus loin.

Notons U l'ensemble des matrices unimodulaires définies par

$$\mathbf{M} \in U \quad \text{ssi} \quad \exists l \in \mathbf{Z} \quad \text{tel que} \quad \begin{cases} z^{-l}\mathbf{M}(z) = \sum_{0 \le n \text{ fini}} \mathbf{M}_n z^n \\ \det\left(z^{-l}\mathbf{M}(z)\right) = \text{Constante} \neq 0 \end{cases}$$

Remarquons que cela implique automatiquement que $\mathbf{M}_0$ n'est pas nulle et est inversible.

On voit tout de suite ce qui paraît contre nature dans cette définition: la matrice polynômiale $z^l\mathbf{M}$ ne comporte que des puissances positives de z ce qui en général implique que le déterminant comporte au moins une puissance de z non nulle… Un exemple simple de matrice unimodulaire est $\mathbf{M}(z) = \mathbf{Id} + zuv^T$ où les vecteurs u et v sont choisis orthogonaux.

On a alors les propriétés suivantes

- U est un sous-groupe multiplicatif de G. En effet, il est facile de voir que le produit de deux matrices unimodulaires l'est également. Pour montrer que l'inverse d'une matrice unimodulaire l'est aussi, il faut recourir à la formule d'inversion des matrices (qui fait intervenir les déterminants mineurs de la matrice): on se rend alors immédiatement compte que $\left(z^l\mathbf{M}(z)\right)^{-1}$ ne comporte dans son développement polynômial que des puissances positives de z, ce qui permet de conclure puisque son déterminant est une constante.

- les matrices communes à P et U sont les matrices orthogonales multipliées par un monôme z^n. En effet, si $\mathbf{M}$ est une telle matrice alors il existe l tel que l'on ait conjointement

$$z^l\mathbf{M}(z) = \sum_{n \ge 0} \mathbf{M}_n z^n$$
$$z^{-l}\mathbf{M}^T\left(z^{-1}\right) = \sum_{n \ge 0} \mathbf{M}'_n z^n$$

  ce qui implique nécessairement que $z^l\mathbf{M}(z) = \mathbf{C}$. Enfin, à cause de la propriété de paraunitarité, on doit en outre avoir $\mathbf{C}^T\mathbf{C} = \mathbf{Id}$ d'où le résultat.

C. Résultats de factorisations

On va maintenant voir que ces deux sous-groupes multiplicatifs peuvent être engendrés par des produits d'éléments simples. Il s'agira, dans le cas des matrices paraunitaires d'éléments de la forme [VNDS,DVN]

$$\begin{aligned} &\mathbf{P}(z) = \mathbf{I} + (z-1)uu^T \\ &\text{avec } u^T u = 1 \end{aligned} \tag{III.3}$$

et pour les matrices unimodulaires, d'éléments de la forme

$$\begin{aligned}&\mathbf{U}(z) = \mathbf{I} + z^n uv^T \\ &\text{avec } u^T v = 0\end{aligned} \qquad \text{(III.4)}$$

Il est aisé de vérifier que ces matrices appartiennent à G. On vérifie en effet immédiatement que

$$\begin{aligned}\mathbf{P}^{-1}(z) &= \mathbf{I} + (z^{-1} - 1)uu^T = \mathbf{P}^T(z) \\ \mathbf{U}^{-1}(z) &= \mathbf{I} - z^n uv^T\end{aligned}$$

Enfin on a les propriétés suivantes

$$\begin{aligned}\det \mathbf{P}(z) &= z \\ \det \mathbf{U}(z) &= 1\end{aligned}$$

qui peuvent se démontrer, par exemple en faisant un changement de base ramenant les vecteurs u et v sur la base canonique de R^D.

1. Matrices paraunitaires

Il s'agit d'un résultat que l'on trouve abondamment chez Vaidyanathan [Vai1,Vai2,VH,VNDS]. Cependant j'y ai ajouté quelques précisions qui ne sont pas indiquées par cet auteur, en particulier concernant les matrices non carrées (voir cependant [CV] pour des résultats semblables). La factorisation des matrices paraunitaires en éléments simples de la forme (III.3) est issue des deux lemmes suivants

Lemme III.1 *Soit* $\mathbf{M}(z)$ *une matrice paraunitaire dont le développement polynômial s'écrit*

$$\mathbf{M}(z) = \sum_{n=M}^{N} \mathbf{M}_n z^n$$

avec $M \neq N$*, alors on a nécessairement*

$$\mathbf{M}_N \mathbf{M}_M^T = \mathbf{0}$$

Preuve

Il suffit de développer le produit $\mathbf{M}(z)\mathbf{M}^T(z^{-1})$ dont le terme de plus haut degré $(N\text{-}M)$ doit être nul.

Lemme III.2 *Soit* $\mathbf{M}(z)$ *une matrice paraunitaire non constante, alors il existe un vecteur u tel que, si l'on forme la matrice* $\mathbf{M'}$

$$\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I} + (z^{-1} - 1)uu^T\right)$$

alors cette matrice vérifie les relations suivantes sur ses degrés matriciels et déterminants

i. degM($\mathbf{M'}(z)$)≤degM($\mathbf{M}(z)$)
ii. degM($\mathbf{M'}(z^{-1})$)≤degM($\mathbf{M}(z^{-1})$)
iii. degD($\mathbf{M'}(z)$)=degD($\mathbf{M}(z)$)-1

Preuve
Il suffit de prendre n'importe quelle ligne non nulle de $\mathbf{M}_N^T$ pour le vecteur u. Dans ces conditions, grâce au lemme III.1 on a $\mathbf{M}_0 u=0$. Puis on prouve

i. le terme de plus haut degré de $\mathbf{M'}$ est évidemment inférieur ou égal à celui de $\mathbf{M}$
ii. le terme de plus bas degré de $\mathbf{M'}$ est de degré M; En effet, en faisant directement le produit, on voit que le terme de degré M-1 s'annule à cause du choix de u puisque $\mathbf{M}_M u u^T = 0$
iii. vient de la relation $\det \mathbf{M}(z) = z^{-1} \det \mathbf{M}(z)$ puisque les éléments simples paraunitaires (III.3) sont de déterminant égal à z

Grâce à ces deux lemmes on peut maintenant énoncer le théorème de factorisation des matrices paraunitaires.

Théorème III.3 *Soit* $\mathbf{M}(z)$ *une matrice paraunitaire. Il existe un entier* M*, une suite de vecteurs unitaires* $u_1, u_2, \ldots u_K$ *et une matrice orthogonale* $\mathbf{R}$ *tels que*

$$\mathbf{M}(z) = z^M \mathbf{R}\left(\mathbf{I} + (z-1)u_1 u_1^T\right)\left(\mathbf{I} + (z-1)u_2 u_2^T\right)\ldots\left(\mathbf{I} + (z-1)u_K u_K^T\right) \quad \text{(III.5)}$$

Preuve
On choisit pour M le plus bas degré du développement polynômial de $\mathbf{M}$, donc on peut écrire

$$z^{-M}\mathbf{M}(z) = \sum_{n=0}^{N} \mathbf{M}_n z^n$$

La démonstration se fait alors par récurrence sur le degré déterminant de la matrice $\mathbf{M}(z)$. Le lemme III.2 a en effet montré que l'on peut trouver un élément paraunitaire simple tel que la multiplication de la matrice par l'inverse de cet élément

- ne modifie pas le degré le plus bas du développement polynômial de la matrice
- fait décroître le degré déterminant de la matrice d'exactement une unité

Il est alors facile de conclure, puisqu'une matrice dont le plus bas degré est 0, ne peut avoir un degré déterminant inférieur à 0.

Cette preuve indique d'autre part que si $\mathbf{M}(z)$ est de degré déterminant égal à P, alors il y a exactement P facteurs paraunitaires simples dans (III.5), c'est-à dire K=degD($\mathbf{M}(z)$). On a alors

$$K \le (D-1)\mathrm{degM}\left(\mathbf{M}(z)\right)$$

En effet, on pourrait voir que, parallèlement au degré déterminant la dimension du noyau de $\mathbf{M}_N$, le coefficient matriciel de plus haut degré, décroît de 1 à chaque multiplication par un élément paraunitaire simple (voir démonstration du théorème III.12). Comme cette dimension est constamment supérieure ou égale à 1, il suffit de D-1 multiplications pour réduire le degré matriciel de $\mathbf{M}(z)$ d'une unité.

En poussant plus loin les investigations, on peut se rendre compte qu'il est possible de factoriser toute matrice paraunitaire de degré matriciel N comme un produit de N matrices paraunitaires de degré matriciel 1. C'est l'objet du théorème suivant qui n'est à ma connaissance pas énoncé dans la littérature sur le sujet

Théorème III.4 *Soit* $\mathbf{M}(z)$ *une matrice paraunitaire de degré matriciel N. Alors il existe une suite de N opérateurs de projection orthogonale* $\mathbf{P}_1, \mathbf{P}_2, \ldots \mathbf{P}_N$ *(c'est-à dire vérifiant conjointement* $\mathbf{P}_k^{\,2} = \mathbf{P}_k$ *et* $\mathbf{P}_k = \mathbf{P}_k^T$*) , une matrice orthogonale constante* $\mathbf{R}$ *et un entier M tels que*

$$\mathbf{M}(z) = z^M \mathbf{R} \prod_{k=1}^{N} \left(\mathbf{I} + (z-1)\mathbf{P}_k\right)$$

Preuve

On anticipe un peu en reprenant, plutôt que la preuve de III.3, celle plus générale de III.12. On pose donc

$$\mathbf{M}(z) = \sum_{k=0}^{N} \mathbf{M}_k z^k$$

après multiplication par z^{-M}. On verra alors qu'il existe u tel que $\mathbf{M}_0 u$=0 et tel que u soit orthogonal à Ker($\mathbf{M}_N$). Après multiplication par l'élément paraunitaire $\mathbf{I} + (z^{-1}-1)uu^T$, la matrice coefficient du terme de degré le plus élevé est $\mathbf{M}'_N = \mathbf{M}_N\left(\mathbf{I} - uu^T\right)$ c'est-à dire que $\mathrm{Ker}\left(\mathbf{M}'_N\right) = \mathrm{Ker}\left(\mathbf{M}_N\right) \oplus \mathrm{Vect}(u)$. On utilise ce fait dans la preuve de III.12 pour montrer que la dimension du noyau de la matrice de plus haut degré décroît strictement. On peut en fait être encore plus précis puisque l'on voit maintenant aisément que, tant que le degré matriciel de la matrice $\mathbf{M}'$ est N, les vecteurs u que l'on trouvera successivement seront tous orthogonaux entre eux, jusqu'à remplir complètement le noyau de $\mathbf{M}'_N$.

Soit donc $\mathbf{P}$ la projection orthogonale sur l'orthogonal du noyau de $\mathbf{M}_N$ (noté $\mathrm{Ker}(\mathbf{M}_N)^\perp$): on a alors $\mathbf{M}_N\left(\mathbf{I} - \mathbf{P}\right) = \mathbf{0}$. D'autre part, on a la relation bien connue $\mathrm{Ker}\left(\mathbf{M}_N\right)^\perp = \Im m\left(\mathbf{M}_N^T\right)$ qui peut se démontrer de la façon suivante: $\Im m\left(\mathbf{M}_N^T\right)^\perp$ est

composé des vecteurs v orthogonaux aux vecteurs de la forme $\mathbf{M}_N^T a$ (où a est quelconque), c'est-à dire tels que $v^T \mathbf{M}_N^T a = 0$ pour tout a. C'est équivalent à dire que $\mathbf{M}_N v = 0$, donc que v décrit le noyau de $\mathbf{M}_N$ ce qui prouve que $\mathrm{Ker}\left(\mathbf{M}_N\right) = \Im m\left(\mathbf{M}_N^T\right)^{\perp}$. Cette relation entraîne que $\mathbf{M}_0\mathbf{P} = \mathbf{0}$, et en mettant bout à bout nos résultats, on constate que la matrice **M'** définie par

$$\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I} + (z^{-1} - 1)\mathbf{P}\right)$$

est de degré matriciel strictement inférieur à celui de $\mathbf{M}(z)$, est paraunitaire, et ne comporte dans son développement polynômial que des puissances positives de z. Ce qui démontre, par récurrence, le théorème.

Il faut cependant noter que la factorisation n'est, là encore pas unique. Elle se rapproche cependant un peu plus des factorisations de polynômes que nous connaissons habituellement dans l'anneau des polynômes.

2. Matrices unimodulaires

Le résultat que je vais énoncer n'est pas connu sous cette forme dans la littérature. Il est cependant équivalent à un autre résultat, cité en exercice chez Kailath [Kai], c'est-à dire que toute matrice unimodulaire peut se décomposer en un produit d'opérations simples

- ajout du multiple d'une colonne (ou d'une ligne) à une autre
- interversion de colonnes (ou de lignes)

C'est, du reste, à l'aide de telles opérations que Kailath démontre les formes de Hermite et de Smith-McMillan de toute matrice polynômiale.

Bien qu'un tel type de résultat soit déjà connu dans la littérature, il est significatif qu'il n'ait pas été adapté pour la conception de filtres ou leur implémentation comme ça a été le cas des matrices lossless. Dans son tutorial de 1990 [Vai2], Vaidyanathan pouvait même affirmer que la factorisation des matrices polynômiales FIR-inversibles était un problème ouvert, et il identifiait ainsi le problème: "…reduces essentially to one of characterizing causal unimodular matrices: one seeks to develop a structure whose multiplier parameters span all unimodular matrices"…

Donc, bien que l'on puisse utiliser des opérations qui s'appliquent à toutes les matrices polynômiales (non nécessairement FIR-inversibles) pour expliciter la factorisation des matrices unimodulaires en éléments simples, je vais utiliser directement la propriété que ces matrices unimodulaires sont FIR-inversibles. Cette méthode sera en particulier utile pour la démonstration que toute matrice FIR inversible peut s'écrire sous la forme d'un produit "matrice diagonale de délais"–"matrice unimoduaire en z"–"matrice unimodulaire en z^{-1}" (produit DUU: voir plus loin).

Par multiplication par un délai constant z^{-M}, on pourra se restreindre aux matrices unimodulaires qui s'écrivent

$$\mathbf{M}(z) = \sum_{n=0}^{N} \mathbf{M}_n z^n \qquad \text{(III.6)}$$

où ni $\mathbf{M}_0$ ni $\mathbf{M}_N$ ne sont nulles, c'est-à dire que dans la définition des matrices unimodulaires, on pourra se restreindre à l=0.

On a ici besoin d'un seul lemme

Lemme III.5 *Soit* $\mathbf{M}(z)$ *une matrice unimodulaire non constante. Alors on peut trouver deux vecteurs orthogonaux u et v, et un entier n tels que la matrice* $\mathbf{M'}(z)$ *définie par*

$$\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I} - z^n uv^T\right)$$

vérifie l'inégalité

$$\deg\mathrm{V}(\mathbf{M}') \le \deg\mathrm{V}(\mathbf{M}) - 1$$

Preuve

La démonstration sera ici un peu complexe, due à la prolifération des indices. Écrivons $\mathbf{M}$ et son inverse transposé $\tilde{\mathbf{M}}$ sous la forme

$$\mathbf{M}(z) = \begin{pmatrix} C_1(z)^T \\ C_2(z)^T \\ \vdots \\ C_D(z)^T \end{pmatrix} \quad \tilde{\mathbf{M}}(z) = \begin{pmatrix} \tilde{C}_1(z)^T \\ \tilde{C}_2(z)^T \\ \vdots \\ \tilde{C}_D(z)^T \end{pmatrix} \qquad \text{(III.7)}$$

Dans leur développement en puissances de z, les vecteurs polynômiaux se décomposent sous la forme

$$\begin{aligned} C_k(z) &= \sum_{n=0}^{N_k} C_{k,n} z^n \\ \tilde{C}_k(z) &= \sum_{n=0}^{\tilde{N}_k} \tilde{C}_{k,n} z^n \end{aligned} \qquad \text{(III.8)}$$

On a donc les D^2 relations suivantes, valables pour k,k'=1…D

$$C_{k'}(z)^T \tilde{C}_k(z) = \delta_{k-k'} \qquad \text{(III.9)}$$

En particulier, cette relation nous indique que pour tous k,k' tels que l'on ait soit $k \neq k'$, soit $N_{k'} \neq 0$, soit $\tilde{N}_k \neq 0$, les vecteurs $C_{k',N_{k'}}$ et $\tilde{C}_{k,\tilde{N}_k}$ sont orthogonaux. On va donc choisir les vecteurs u et v sous la forme

$$\begin{aligned} u &= C_{k_0, N_{k_0}} \\ v &= \lambda C_{k_1, N_{k_1}} \end{aligned} \qquad \text{(III.10)}$$

où λ est une constante non nulle que l'on va préciser. On opère en trois étapes

- prenons k_0 tel que

$$N_{k_0} = \max_{1 \le k \le D} N_k \qquad \text{(III.11)}$$

Notons que l'on a nécessairement $N_{k_0} \ge 1$, sinon $\mathbf{M}(z)$ et par suite $\mathbf{M}(z)$ serait une matrice constante.

- définissons les nombres s_k par $s_k = \min_{0 \le s} \left\{ s \text{ tel que } C_{k, N_k - s}^T u \ne 0 \right\}$. Si pour k donné on a $C_{k,n}^T u = 0$ quel que soit $n \le N_k$, alors par convention on posera $s_k = +\infty$. On a bien sûr, $s_k \ge 1$ pour tout k puisque $N_{k_0} \ge 1$. On pose ensuite

$$n = \min_{1 \le k \le D} s_k \qquad \text{(III.12)}$$

qui est donc toujours supérieur ou égal à 1. Il est d'autre part strictement inférieur à l'infini, sinon $\mathbf{M}$ ne serait pas inversible. On prend donc pour k_1 l'un des indices qui réalisent ce minimum, c'est-à dire tel que

$$s_{k_1} = n \qquad \text{(III.13)}$$

On est alors assuré que

i. pour tout $s<n$ on a $C_{k, N_k - s}^T u = 0$, ce qui entraîne l'inégalité sur les degrés

$$\deg \mathrm{M}\left[C_k(z)^T \left(\mathbf{I} - z^n u v^T \right) \right] \le \deg \mathrm{M}\left[C_k(z)^T \right]$$

pour tout indice $k=1 \ldots D$

ii. $C_{k_1, N_k - n}^T u \ne 0$

- posons enfin

$$\lambda = \frac{1}{C_{k_1, N_{k_1} - n}^T u} \qquad \text{(III.14)}$$

ce qui achève de définir v et nous assure que

$$\deg \mathrm{M}\left[C_{k_1}(z)^T \left(\mathbf{I} - z^n u v^T \right) \right] \le \deg \mathrm{M}\left[C_{k_1}(z)^T \right] - 1$$

Par construction on a donc $\deg\mathrm{V}\left[\mathbf{M}(z)\left(\mathbf{I}-z^{n}uv^{T}\right)\right]\le\deg\mathrm{V}\left[\mathbf{M}(z)\right]-1$, c'est-à dire ce qu'il fallait démontrer.

On peut maintenant énoncer le théorème de factorisation des matrices unimodulaires.

Théorème III.6 *Soit* $\mathbf{M}(z)$ *une matrice unimodulaire. Alors il existe*

- *un entier* M
- *une matrice constante inversible* $\mathbf{S}$
- *une suite d'entiers positifs* $n_1, n_2, \ldots n_K$
- *deux suites de vecteurs non nuls* $u_1, u_2, \ldots u_K$ *et* $v_1, v_2, \ldots v_K$ *tels que* $u_k^T v_k = 0$ *pour tout* k=1…K

tels que

$$\mathbf{M}(z)=z^{M}\mathbf{S}\left(\mathbf{I}+z^{n_1}u_1v_1^{T}\right)\left(\mathbf{I}+z^{n_2}u_2v_2^{T}\right)\ldots\left(\mathbf{I}+z^{n_K}u_Kv_K^{T}\right) \qquad \text{(III.15)}$$

Preuve

On prend pour M le plus bas degré de la matrice $\mathbf{M}(z)$ et dans ces conditions $z^{-M}\mathbf{M}(z)$ s'écrit sous la forme (III.6). La démonstration se fait alors par récurrence sur le degré vectoriel de $\mathbf{M}(z)$. En effet d'après le lemme III.5, on voit que l'on peut trouver un entier n ainsi que des vecteurs orthogonaux u et v, tels que la multiplication de $\mathbf{M}(z)$ par l'inverse de l'élément unimodulaire simple engendré par ce triplé (n,u,v) donne une matrice dont

- le développement polynômial est encore de la forme (III.6)
- le degré matriciel est inférieur ou égal à celui de $\mathbf{M}(z)$
- le degré vectoriel est strictement inférieur à celui de $\mathbf{M}(z)$

Bien évidemment le degré vectoriel d'une matrice de la forme (III.6) est toujours supérieur ou égal à zéro, ce qui permet d'affirmer que la récurrence est finie. Elle s'arrête d'ailleurs quand la matrice est constante.

On évalue donc la valeur maximale de K à $\deg\mathrm{V}(\mathbf{M}(z))$. Il est à noter que dans cette factorisation, les éléments simples pourraient être remplacés par des facteurs de la forme

$$\mathbf{I}+(z^{n}-1)uv^{T} \qquad \text{(III.16)}$$

qui sont particulièrement utiles quand on souhaite introduire un peu de régularité dans les filtres définis par la matrice polynômiale. Il faut enfin garder à l'esprit que cette factorisation est loin d'être unique, bien que dans le cas où D=2 (cas deux bandes) on puisse trouver une forme d'unicité. Dans le cas général, on voit ici qu'une matrice $\mathbf{M}(z)$ de degré matriciel N s'ex-

primera comme le produit d'au plus *ND* éléments unimodulaires simples: elle n'est donc pas non plus minimale.

3. Théorèmes généraux sur les matrices polynômiales

Sans même se restreindre aux matrices dont le déterminant est un délai, on peut avoir un certain nombre de résultats de factorisation sur les matrices polynômiales. Ce sont en particulier les formes de Hermite et de Smith-McMillan [Kai].

a. Forme de Hermite

Toute matrice polynômiale $\mathbf{M}(z)$ peut se mettre sous la forme

$$\mathbf{M}(z) = \mathbf{U}(z)\mathbf{T}(z)$$

où $\mathbf{U}(z)$ est une matrice unimodulaire, et $\mathbf{T}(z)$ une matrice triangulaire (supérieure ou inférieure). On peut ajouter des contraintes sur les degrés des éléments au-dessus ou au-dessous de la diagonale, par rapport au degré des éléments diagonaux, afin de rendre la matrice $\mathbf{T}(z)$ unique. La démonstration, qui se trouve dans Kailath [Kai], utilise de manière répétée la division euclidienne des plynômes, et par un premier processus itératif, on peut annuler toutes les composantes de la première colonne, sauf la première. On peut alors continuer avec une sous-matrice de taille inférieure pour obtenir finalement une matrice triangulaire supérieure. Ce résultat est valable, que les matrices soient carrées, ou rectangulaires, avec dans ce dernier cas une évidente extension de la définition des matrices triangulaires.

Dans le cas où la matrice $\mathbf{M}(z)$ est FIR-inversible les termes diagonaux de $\mathbf{T}(z)$ sont bien sûr des délais.

b. Forme de Smith-McMillan

Toute matrice polynômiale $\mathbf{M}(z)$ peut se mettre sous la forme [Kai],

$$\mathbf{M}(z) = \mathbf{U}(z)\mathbf{D}(z)\mathbf{U}'(z)$$

où $\mathbf{U}(z)$ et $\mathbf{U'}(z)$ sont des matrices unimodulaires, et $\mathbf{D}(z) = \mathrm{diag}\big(\lambda_1(z), \lambda_2(z), \ldots \lambda_D(z)\big)$ est une matrice diagonale dont les éléments vérifient la contrainte

$$\lambda_k(z) \text{ divise } \lambda_{k+1}(z)$$

ce qui permet d'assurer l'unicité de la matrice $\mathbf{D}(z)$. La démonstration dans Kailath, s'appuie encore une fois sur la division euclidienne, mais en combinant cette fois les opérations colonnes et lignes. Dans ce cas la matrice $\mathbf{D}$ est unique (à une permutation des éléments diagonaux près). Ce résultat est également valable pour des matrices rectangulaires, la matrice diagonale devenant rectangulaire, et les deux matrices unimodulaires ayant des tailles différentes. Là encore, si la matrice est FIR inversible, les éléments de $\mathbf{D}$ sont des délais.

4. Théorèmes sur les matrices FIR–inversibles

Si l'on impose aux matrices polynômiales d'appartenir à G, alors on peut obtenir des décompositions spécifiques. La première est la décomposition en un produit unimodulaire–paraunitaire (UP) et la seconde, en un produit diagonale-double unimodulaire (DUU).

a. Produit UP

Ce résultat a été développé indépendamment par [VC2: fin de l'article] et moi-même, bien que [VC1]. ne fasse pas mention de l'unicité de la décomposition

Théorème III.7 *Soit* $\mathbf{M}(z)$ *une matrice FIR inversible. Alors il existe une matrice unimodulaire* $\mathbf{U}(z)$ *et une matrice paraunitaire* $\mathbf{P}(z)$ *telles que*

$$\mathbf{M}(z) = \mathbf{U}(z)\mathbf{P}(z)$$

En outre, ces deux matrices sont uniques, à une rotation constante et un délai près.

Preuve

On va suivre le même cheminement que lors de la factorisation des matrices paraunitaires. Restreignons nous, tout d'abord aux matrices $\mathbf{M}(z)$ qui s'écrivent sous la forme

$$\mathbf{M}(z) = \sum_{n=0}^{N} \mathbf{M}_n z^n$$

On pose alors

$$\tilde{\mathbf{M}}(z) = \sum_{n=\tilde{M}}^{\tilde{N}} \tilde{\mathbf{M}}_n z^n$$

Si $\tilde{M} = 0$ alors $\mathbf{M}(z)$ est unimodulaire: il suffit donc de prendre $\mathbf{P}(z)$=$\mathbf{I}$. Bien évidemment, on ne pourrait pas avoir $\tilde{M} \geq 1$ puisqu'alors le terme constant du produit $\mathbf{M}\tilde{\mathbf{M}}^T$ serait nul au lieu d'égaler l'identité.

Supposons donc $\tilde{M} \leq -1$. Dans ce cas, on a la relation $\mathbf{M}_0\tilde{\mathbf{M}}_{\tilde{M}}^T = \mathbf{0}$ ce qui permet, à l'instar du lemme III.5 de définir un vecteur u tel que $\mathbf{M}_0 u = \mathbf{0}$. En posant

$$\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I} + (z^{-1} - 1)uu^T\right)$$

on a alors

- $\mathbf{M}'(z) = \sum_{n=0}^{N} \mathbf{M}'_n z^n$ c'est-à dire que le plus bas degré du développement polynômial de $\mathbf{M'}$ est ≥0 et son plus haut degré est inférieur à celui de $\mathbf{M}$

- $\mathrm{degD}(\mathbf{M}') = \mathrm{degD}(\mathbf{M}) - 1$

De son côté la matrice $\tilde{\mathbf{M}}$ se transforme en $\tilde{\mathbf{M}}' = \tilde{\mathbf{M}}\left(\mathbf{I} + (z-1)uu^T\right)$. On répète ainsi l'opération jusqu'à ce que le degré le plus bas de $\tilde{\mathbf{M}}'$ soit égal à zéro. Le nombre d'itérations est bien évidemment fini puisqu'à chacune on diminue le degré déterminant de $\mathbf{M}$ d'une unité. En fait on a exactement $\mathrm{degD}(\mathbf{M})$ itérations avant que les matrices $\mathbf{M}'$ et $\tilde{\mathbf{M}}'$ n'aient plus de terme de puissance négative. À la fin de ce processus, on a pu trouver une matrice paraunitaire $\mathbf{P}$ telle que $\mathbf{M}(z)\mathbf{P}(z)^{-1}$ soit unimodulaire.

Cette décomposition présente également une forme d'unicité. En effet, supposons que la matrice M puisse s'écrire sous deux formes distinctes

$$\mathbf{M}(z) = \mathbf{U}(z)\mathbf{P}(z) = \mathbf{U}'(z)\mathbf{P}'(z)$$

alors on aura nécessairement $\mathbf{U}'(z)^{-1}\mathbf{U}(z) = \mathbf{P}'(z)\mathbf{P}^{-1}(z)$. Or on sait que l'intersection entre matrices unimodulaires et paraunitaires se réduit aux produits de matrices de rotation constantes par un délai, ce qui montre l'unicité.

b. Produit DUU

Théorème III.8 *Soit* $\mathbf{M}(z)$ *une matrice FIR inversible. Alors il existe deux matrices unimodulaires* $\mathbf{U}(z)$ *et* $\mathbf{V}(z)$ *et une matrice diagonale* $\mathbf{D}(z)$ *composée de délais telles que*

$$\mathbf{M}(z) = \mathbf{D}(z)\mathbf{U}(z^{-1})\mathbf{V}(z)$$

Preuve

On va suivre ici les grandes lignes de la démonstration du lemme III.5, c'est-à dire décrire les matrices $\mathbf{M}$ et $\tilde{\mathbf{M}}$, qui vérifient par hypothèse $\mathbf{M}(z)\tilde{\mathbf{M}}(z)^T = \mathbf{I}$, sous forme de vecteurs lignes selon (III.7). On adapte évidemment les bornes de sommation de (III.8)

$$C_k(z) = \sum_{n=M_k}^{N_k} C_{k,n} z^n$$

$$\tilde{C}_k(z) = \sum_{n=\tilde{M}_k}^{\tilde{N}_k} \tilde{C}_{k,n} z^n$$

et l'on aura donc toujours (III.9). Toujours comme dans cette preuve, on va choisir u et v sous la forme (III.10) avec des paramètres que l'on détermine en trois étapes également. En fait, seule la première étape sera différente de la démonstration du lemme III.5. On va en effet choisir k_0 de telle sorte que

$$N_{k_0} + \tilde{N}_{k_0} \geq 1 \qquad \text{(III.17)}$$

ce qui assurera que $C^T_{k,N_k} C_{k_0,N_{k_0}} = 0$ pour toute valeur de k, et en particulier pour celle de k_1 qui détermine v. Les deux autres étapes seront identiques à celles du lemme III.5, ce qui montrera que

- si **M** est FIR inversible
- s'il existe k_0 tel que $N_{k_0} + \tilde{N}_{k_0} \geq 1$

alors on pourra trouver une matrice unimodulaire simple $\mathbf{U}(z) = \mathbf{I} + z^n uv^T$ telle que le degré vectoriel de $\mathbf{M}(z)\mathbf{U}(z)^{-1}$ soit strictement inférieur à celui de $\mathbf{M}(z)$, sans pour autant que le terme de plus bas degré du développement polynômial diminue. On peut recommencer cette opération —qui bien sûr ne peut pas être effectuée indéfiniment, puisque le degré vectoriel d'une matrice dont le terme de plus bas degré serait de la quantité s, doit être supérieur ou égal à sD— jusqu'à ce que notre hypothèse ne soit plus valide, c'est-à dire jusqu'à ce que

$$\forall k \in [1 \ldots D] \quad N_k + \tilde{N}_k = 0 \tag{III.18}$$

On ne peut en effet avoir $N_k + \tilde{N}_k < 0$ à cause de (III.9) qui comporte un second membre non nul quand $k{=}k'$. À cet instant on aura extrait une matrice $\mathbf{V}(z)$ de **M**, produit des éléments unimodulaires simples issus du précédent raisonnement par récurrence et nous nous retrouvons avec deux matrices **M'** et $\tilde{\mathbf{M}}'$

$$\mathbf{M}'(z) = \mathbf{M}(z)\mathbf{V}(z)^{-1}$$
$$\tilde{\mathbf{M}}'(z) = \tilde{\mathbf{M}}(z)\mathbf{V}(z)^T$$

transposées-inverses l'une de l'autre, et contraintes par la relation (III.18).

Posons maintenant

$$\mathbf{D}(z) = \mathrm{diag}\left(z^{N_1}, z^{N_2}, \ldots z^{N_D}\right) \tag{III.19}$$

Alors les matrices $\mathbf{D}(z)^{-1}\mathbf{M}'(z)$ et $\mathbf{D}(z)\tilde{\mathbf{M}}'(z)$ sont transposées-inverses l'une de l'autre, et d'autre part ne comportent dans leurs développements polynômiaux que des puissances négatives ou nulles de z (conséquence de (III.18)). Ceci implique qu'il existe une matrice unimodulaire U(z) telle que $\mathbf{D}(z)^{-1}\mathbf{M}'(z) = \mathbf{U}(z^{-1})$, c'est-à dire

$$\mathbf{M}(z) = \mathbf{D}(z)\mathbf{U}(z^{-1})\mathbf{V}(z)$$

ce qu'il fallait démontrer.

À la différence du produit UP, il n'y a pas ici d'unicité. Cette décomposition présente un avantage important par rapport à la précédente dans le cas de bancs de filtres uniformes —et

donc malheureusement pas dans le cas des bancs de filtres rationnels— puisque la matrice de délai n'a alors aucune influence sur les caractéristiques fréquentielles des filtres: elle se contente de rajouter un délai —éventuellement différent— sur chaque bande. Dans le cadre de la conception de filtres, on peut donc ne se préoccuper que du produit des deux matrices unimodulaires. Il se trouve qu'alors la longueur (i.-e. le nombre de coefficients matriciels) de ce produit est exactement égal à la somme des longueurs des deux matrices moins un, c'est-à dire le minimum possible. Ce n'est en général pas le cas dans la factorisation UP où par exemple, une matrice de degré N correspondant à des filtres symétriques peut nécessiter un produit de deux matrices de degré chacune égale à N, ce qui multiplie par deux le nombre d'inconnues.

c. Factorisation générale des matrices FIR–inversibles

Une fois ces décompositions en produit UP ou en produit DUU effectuées, on peut factoriser chaque matrice en éléments simples, ce qui donne les deux théorèmes ci-dessous

Théorème III.9 *Soit* $\mathbf{M}(z)$ *une matrice FIR–inversible. Alors il existe*

- *un entier* M
- *une matrice constante inversible* $\mathbf{S}$
- *une suite d'entiers positifs* $n_1, n_2, \dots n_{K_u}$
- *deux suites de vecteurs non nuls* $u_1, u_2, \dots u_{K_u}$ *et* $v_1, v_2, \dots v_{K_u}$ *tels que* $u_k^T v_k = 0$ *pour tout* k=$1 \dots K_u$
- *une suite de vecteurs orthonormés* $w_1, w_2, \dots w_{K_p}$

tels que

$$\mathbf{M}(z) = z^M \mathbf{S}\left(\mathbf{I} + z^{n_1} u_1 v_1^T\right)\dots\left(\mathbf{I} + z^{n_K} u_{K_u} v_{K_u}^T\right)\left(\mathbf{I} + (z-1) w_1 w_1^T\right)\dots\left(\mathbf{I} + (z-1) w_{K_p} w_{K_p}^T\right) \qquad \text{(III.20)}$$

Théorème III.10 *Soit* $\mathbf{M}(z)$ *une matrice FIR–inversible. Alors il existe*

- *une suite de délais* $d_1, d_2, \dots d_D$ *définissant une matrice diagonale* $\mathbf{D}(z) = \mathrm{diag}(z^{d_1}, z^{d_2}, \dots z^{d_D})$
- *une matrice constante inversible* $\mathbf{S}$
- *deux suites d'entiers positifs* $n_1, n_2, \dots n_K$ et $n'_1, n'_2, \dots n'_{K'}$
- *quatre suites de vecteurs non nuls* $u_1, u_2, \dots u_{K_u}$ *et* $v_1, v_2, \dots v_{K_u}$ *d'une part,* $u'_1, u'_2, \dots u'_{K'}$ *et* $v'_1, v'_2, \dots v'_{K'}$ *d'autre part, tels que* $u_k^T v_k = 0$ *pour tout* k=$1 \dots K$ *et* $u_k'^T v'_k = 0$ *pour tout* k=$1 \dots K'$

tels que

$$\mathbf{M}(z) = \mathbf{D}(z)\mathbf{S}\left(\mathbf{I} + z^{-n_1} u_1 v_1^T\right)\dots\left(\mathbf{I} + z^{-n_K} u_K v_K^T\right)\left(\mathbf{I} + z^{n'_1} u'_1 v_1'^T\right)\dots\left(\mathbf{I} + z^{n'_K} u'_{K'} v'^T_{K'}\right) \qquad \text{(III.21)}$$

5. Théorème de densité

On le voit tout de suite, la différence essentielle entre la factorisation des matrices unimodulaires, par rapport à celle des matrices paraunitaires est que les éléments unimodulaires simples sont de degré matriciel quelconque, et non pas 1. Le présent théorème montre que les éléments unimodulaires simple de degré n sont des limites de produit de n éléments unimodulaires simples de degré 1. Ceci permet de ne plus considérer comme paramètres que les vecteurs u et v et non plus l'entier n. Ce théorème sera en particulier utile pour la conception de filtres.

Théorème III.11 *Le sous-ensemble des matrices unimodulaires qui s'écrivent comme des produits de matrices unimodulaires simples de degré matriciel 1, est dense dans l'ensemble des matrices unimodulaires*

Preuve

Cela vient directement de la relation suivante que l'on peut aisément prouver en développant les produits

$$\mathbf{I} + z^{n+2} uv^T = \lim_{\varepsilon \to 0} \left[\left(\mathbf{I} + \frac{z}{\varepsilon} uv^T \right) \left(\mathbf{I} - \frac{\varepsilon^2}{\|u\|^2 \|v\|^2} z^n vu^T \right) \left(\mathbf{I} - \frac{z}{\varepsilon} uv^T \right) \right]$$

où u et v sont des vecteurs orthogonaux. Par récurrence on peut donc ramener tout élément simple de degré n à une matrice de degré 1 ou zéro, à l'aide de multiplications par des matrices de degré 1. Du théorème de factorisation des matrices unimodulaires, on tire alors immédiatement ce résultat de densité.

On peut voir d'après la démonstration qu'il faut exactement n produits de degré 1 pour obtenir un élément simple unimodulaire de degré n. Ce théorème nous incitera aussi à considérer les éléments unimodulaires simples de degré supérieur à 1 comme "exceptionnels" dans une factorisation, et donc qu'on pourra toujours considérer qu'une matrice FIR inversible ne s'écrit que comme des produits d'éléments de degré 1.

Un tel résultat est fort intéressant en simplifiant les procédures de conception de filtres. En effet, en se limitant aux facteurs de degré un on élimine un paramètre (entier) dans chaque section unimodulaire: on peut ainsi concentrer la recherche sur les vecteurs qui paramétrisent la factorisation.

D. Matrices rectangulaires

On peut avoir également besoin de résultats sur des matrices non carrées, en particulier sur des sous-matrices rectangulaires de matrices carrées. On va ainsi définir l'ensemble $G^{d\times D}$ constitué des matrices polynômiales de taille $d\times D$ de la façon suivante

$$\mathbf{M} \in G^{d\times D} \quad \Leftrightarrow \quad \exists \tilde{\mathbf{M}} \in G^{d\times D} \; \mathbf{M}(z)\tilde{\mathbf{M}}(z)^T = \mathbf{I}_d$$

Bien sûr ceci n'est possible que si $d \leq D$, et par la suite, on ne considèrera que $d<D$.

Dans le même esprit on définira l'ensemble $P^{d\times D}$, correspondant aux matrices paraunitaires dans le cas $d=D$, c'est-à dire vérifiant ici

$$\mathbf{M} \in P^{d\times D} \quad \Leftrightarrow \quad \begin{cases} \mathbf{M} \in G^{d\times D} \\ \mathbf{M}(z)\mathbf{M}(z^{-1})^T = \mathbf{I}_d \end{cases}$$

et l'ensemble $U^{d\times D}$, correspondant aux matrices unimodulaires dans le cas des matrices carrées, vérifiant

$$\mathbf{M} \in U^{d\times D} \quad \Leftrightarrow \quad (\exists l \in \mathbf{Z})\left(\exists \tilde{\mathbf{M}} \in G^{d\times D}\right) \text{ tel que } \begin{cases} z^{-l}\mathbf{M}(z) = \sum_{n\geq 0} \mathbf{M}_n z^n \\ z^{-l}\tilde{\mathbf{M}}(z) = \sum_{n\geq 0} \tilde{\mathbf{M}}_n z^n \end{cases} \text{ et } \mathbf{M}(z)\tilde{\mathbf{M}}(z^{-1})^T = \mathbf{I}_d$$

On va donc pouvoir énoncer des théorèmes de factorisation équivalents à ceux de la section précédente, applicables ici aux matrices rectangulaires de $G^{d\times D}$. De tels résultats ou leurs équivalents se trouvent par exemple dans [CV] (utilisant plutôt la forme de Smith-McMillan), et de manière parcellaire dans [DVN,VNDS] où il est indiqué comment factoriser un vecteur (matrice de taille 1xD) paraunitaire. Nous avons dû développer les résultats suivants car nous aurons effectivement besoin de factoriser des matrices rectangulaires de taille quelconque dans les bancs de filtres rationnels (afin de récupérer le filtre passe-haut à partir du passe-bas).

Théorème III.12 *Soit* $\mathbf{M}(z)$ *une matrice de* $P^{d\times D}$. *Il existe un entier* M, *une suite de vecteurs unitaires* $u_1, u_2, \ldots u_K$ *et une matrice* $\mathbf{R}$ *de taille* $d\times D$ *vérifiant* $\mathbf{R}\mathbf{R}^T = \mathbf{I}_d$ *tels que*

$$\mathbf{M}(z) = z^M \mathbf{R}\left(\mathbf{I} + (z-1)u_1 u_1^T\right)\left(\mathbf{I} + (z-1)u_2 u_2^T\right)\ldots\left(\mathbf{I} + (z-1)u_K u_K^T\right) \qquad \text{(III.22)}$$

Preuve

On va une nouvelle fois raisonner par récurrence, mais bien évidemment plus sur le déterminant de la matrice $\mathbf{M}$. Cependant on pourra encore suivre les grandes lignes des lemmes III.1 et III.2. Ainsi la contrainte $\mathbf{M}(z)\mathbf{M}(z^{-1})^T = \mathbf{I}_d$ implique-t-elle l'équation $\mathbf{M}_0\mathbf{M}_N{}^T = \mathbf{0}$ si l'on suppose que $\mathbf{M}$ n'est pas constante, que N est son degré matriciel et que par multiplication par un délai z^{-M}, on a ramené $\mathbf{M}$ à un développement polynômial ne comportant pas de puissance négative de z.

Soit donc un vecteur a de taille d tel que $\mathbf{M}_N{}^T a \neq 0$ et posons

$$u = \mathbf{M}_N{}^T a \qquad \text{(III.23)}$$

Dans ces conditions on a $\mathbf{M}_0 u = 0$. Puis on pose $\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I}_D + (z^{-1}-1)uu^T\right)$ qui

va donc, grâce à la propriété précédente, ne comporter que des puissances positives ou nulles de z

$$\mathbf{M}'(z) = \sum_{n=0}^{N} \mathbf{M}'_n z^n$$

où en particulier, $\mathbf{M}'_N = \mathbf{M}_N\left(\mathbf{I}_D - uu^T\right)$. Il est également clair que $\mathbf{M'}$ appartient encore à $P^{d\times D}$.

La propriété qui va nous permettre de raisonner par récurrence est la stricte décroissance de la dimension du noyau de $\mathbf{M}_N$

$$\dim \mathrm{Ker}\left(\mathbf{M}'_N\right) \le \dim \mathrm{Ker}\left(\mathbf{M}_N\right) - 1$$

En effet on observe que tout vecteur du noyau de $\mathbf{M}_N$ est orthogonal à u puisque si $v\in\mathrm{Ker}(\mathbf{M}_N)$ alors $u^Tv=a^T\mathbf{M}_Nv=0$. Donc tout vecteur de $\mathrm{Ker}(\mathbf{M}_N)$ appartient également à $\mathrm{Ker}(\mathbf{M}'_N)$. En outre u, qui n'appartient pas à $\mathrm{Ker}(\mathbf{M}_N)$ appartient en revanche à $\mathrm{Ker}(\mathbf{M}'_N)$ ce qui signifie que $\dim \mathrm{Ker}\left(\mathbf{M}'_N\right) \le \dim \mathrm{Ker}\left(\mathbf{M}_N\right) - 1$. On a même en fait l'égalité, mais c'est peu important pour notre démonstration.

Bien sûr, dès que $\dim \mathrm{Ker}\left(\mathbf{M}'_N\right) = 0$, $\mathbf{M}'_N$ devient nulle, ce qui signifie que le degré matriciel de $\mathbf{M'}$ est strictement inférieur à celui de $\mathbf{M}$. On arrête la récurrence quand $\mathbf{M'}$ est une matrice constante. On a ainsi pu prouver l'existence d'une matrice paraunitaire $\mathbf{P}(z)$ ainsi que d'une matrice constante $\mathbf{R}$ telles que $\mathbf{M}(z)\mathbf{P}(z)^{-1} = \mathbf{R}$. On vérifie sans mal que $\mathbf{R}\mathbf{R}^T = \mathbf{I}_d$ ce qui achève la démonstration.

Une importante conséquence de cette preuve est que, lors du processus de factorisation, tous les vecteurs u obtenus pour un même degré matriciel sont, d'une part en nombre inférieur à D-1, et d'autre part sont orthogonaux deux à deux (puisque tout nouveau vecteur u est orthogonal à $\mathrm{Ker}(\mathbf{M}_N)$ et que u appartient à $\mathrm{Ker}(\mathbf{M'}_N)$). Cette remarque signifie qu'une matrice $\mathbf{M}$ de degré matriciel N peut également être factorisée comme un produit de N facteurs paraunitaires de degré 1. En effet si l'on note $u_1, u_2, \ldots u_{D'}$ les vecteurs (avec $D'\le D$-1) associés à un même degré matriciel dans la factorisation, on a

$$\prod_{k=1}^{D'}\left(\mathbf{I} + (z-1)u_k u_k^T\right) = \left(\mathbf{I} + (z-1)\sum_{k=1}^{D'} u_k u_k^T\right)$$

Théorème III.13 *Soit* $\mathbf{M}(z)$ *une matrice de* $U^{d\times D}$. *Alors il existe*

- *un entier* M
- *une matrice constante* $\mathbf{S}$ *de taille* $d\times D$
- *une suite d'entiers positifs* $n_1, n_2, \ldots n_K$

- *deux suites de vecteurs non nuls* $u_1, u_2, \ldots u_K$ *et* $v_1, v_2, \ldots v_K$ *tels que* $u_K^T v_K = 0$ *pour tout* k=1…K

tels que

$$\mathbf{M}(z) = z^M \mathbf{S}\left(\mathbf{I} + z^{n_1} u_1 v_1^T\right)\left(\mathbf{I} + z^{n_2} u_2 v_2^T\right)\ldots\left(\mathbf{I} + z^{n_K} u_K v_K^T\right) \qquad \text{(III.24)}$$

Preuve

Comme dans le cas des matrices carrées, on ramène la matrice $\mathbf{M}(z)$ à des puissances positives de z par la multiplication d'un délai z^{-M}. On va essayer de suivre les étapes du lemme III.5. On pose tout d'abord (III.7) et (III.8) et de la relation de définition des matrices de $U^{d\times D}$ on obtient (III.9) pour $k,k'\in[1\ldots d]$. Par ces équations on va être amené à choisir u et v comme indiqué dans (III.10). Là encore, il s'agit de trouver k_0 et k_1, puis λ. On opère en trois étapes

- du fait que $C_k(z)^T \tilde{C}_k(z) = 1$ pour tout k=1…d, il est impossible que l'un quelconque des $C_k(z)$ ou $\tilde{C}_k(z)$ soit nul; donc si $\mathrm{degM}\left(C_k(z)\right) + \mathrm{degM}\left(\tilde{C}_k(z)\right) < 1$ pour tout k=1…d, alors les deux matrices $\mathbf{M}(z)$ et $\tilde{\mathbf{M}}(z)$ sont constantes et le travail est terminé. Sinon, il existe k_0 tel que $\mathrm{degM}\left(C_{k_0}(z)\right) + \mathrm{degM}\left(\tilde{C}_{k_0}(z)\right) \geq 1$ ce qui nous assure, d'après (III.9) que le vecteur u associé à cette valeur de k_0 sera bien orthogonal à tous les vecteurs C_{k,N_k} possibles

- on définit alors les nombres s_k comme dans le cas du lemme III.5, puis le minimum de tous ces s_k. Il y a alors une légère variation avec la démonstration des matrices carrées puisque l'on n'est alors plus sûr que n soit strictement inférieur à l'infini. Il peut en effet arriver que $C_k(z)^T u = 0$ pour tout valeur de k. Dans ce cas-là, on modifiera la matrice $\tilde{\mathbf{M}}(z)$ de la façon suivante

$$\tilde{C}_k(z) \rightarrow \begin{cases} \tilde{C}_k(z) & \text{si } k \neq k_0 \\ \tilde{C}_{k_0}(z) - z^{\tilde{N}_{k_0}} \tilde{C}_{k_0,\tilde{N}_{k_0}} & \text{si } k = k_0 \end{cases}$$

et l'on recommence les premières étapes de la démonstration (nouveau choix de u) puisqu'il est facile de vérifier que la nouvelle matrice $\tilde{\mathbf{M}}(z)$ vérifie encore l'équation $\mathbf{M}(z)\tilde{\mathbf{M}}(z) = \mathbf{I}_d$. Bien sûr, ce retour en arrière ne peut se faire indéfiniment puisqu'à chaque fois, le degré vectoriel de $\tilde{\mathbf{M}}(z)$ décroît. Il arrive ainsi nécessairement un moment où n est strictement inférieur à l'infini.

Le reste de la démonstration suit sans dévier le reste de la preuve du lemme III.5, et ainsi, par récurrence on en déduit le théorème III.13.

Théorème III.14 *Soit* $\mathbf{M}(z)$ *une matrice de* $G^{d\times D}$. *Alors il existe une matrice unimodulaire* $\mathbf{U}(z)$, *une matrice paraunitaire* $\mathbf{P}(z)$ *et une matrice constante de taille* $d\times D$ $\mathbf{S}$ *telles que*

$$\mathbf{M}(z) = \mathbf{S}\mathbf{U}(z)\mathbf{P}(z)$$

Preuve

Dans le même esprit que pour le théorème III.7 , on pose

$$\mathbf{M}(z) = \sum_{n=0}^{N} \mathbf{M}_n z^n$$
$$\tilde{\mathbf{M}}(z) = \sum_{n=\tilde{M}}^{\tilde{N}} \tilde{\mathbf{M}}_n z^n$$

et donc, si $\tilde{M} \le -1$ on aura $\mathbf{M}_0\tilde{\mathbf{M}}_{\tilde{M}}^T = 0$. Bien sûr dans l'autre cas, si $\tilde{M} \ge 0$ alors $\mathbf{M}$ appartient à $U^{d\times D}$, et en utilisant le théorème III.13 on voit que $\mathbf{P}(z)$=$\mathbf{I}$ ce qui termine la démonstration.

On choisit ensuite le vecteur $d\times 1$ a tel que $\tilde{\mathbf{M}}_{\tilde{M}}^T a \neq 0$ ce qui permet de définir le vecteur $D\times 1$ u sous la forme

$$u = \tilde{\mathbf{M}}_{\tilde{M}}^T a$$

Posant

$$\mathbf{M}'(z) = \mathbf{M}(z)\left(\mathbf{I} + (z^{-1} - 1)uu^T\right)$$
$$\tilde{\mathbf{M}}'(z) = \tilde{\mathbf{M}}(z)\left(\mathbf{I} + (z - 1)uu^T\right)$$

on a bien sûr toujours $\mathbf{M}'(z)\tilde{\mathbf{M}}'(z)^T = \mathbf{I}_d$ et d'autre part les plus petites puissances de z de $\mathbf{M}'$ et $\tilde{\mathbf{M}}'$ restent supérieures à celles de $\mathbf{M}$ et $\tilde{\mathbf{M}}$. Enfin on constate que, comme dans le théorème III.12, le vecteur u est orthogonal au noyau de $\tilde{\mathbf{M}}_{\tilde{M}}$ et comme on a $\tilde{\mathbf{M}}'_{\tilde{M}} = \tilde{\mathbf{M}}_{\tilde{M}}\left(\mathbf{I} - uu^T\right)$ il devient clair que

$$\dim \mathrm{Ker}\left(\tilde{\mathbf{M}}'_{\tilde{M}}\right) \le \dim \mathrm{Ker}\left(\tilde{\mathbf{M}}_{\tilde{M}}\right) - 1$$

On peut donc à nouveau raisonner par récurrence sur la dimension du noyau de $\tilde{\mathbf{M}}_{\tilde{M}}$. On démontre ainsi qu'après un nombre fini (inférieur ou égal à $M(D-1)$) de multiplications par des éléments paraunitaires simples on aboutit à un couple de matrices $\mathbf{M}'$ et $\tilde{\mathbf{M}}'$ vérifiant $\mathbf{M}'\tilde{\mathbf{M}}'^T = \mathbf{I}_d$ et ne comportant aucune puissance négative de z dans leurs développements polynômiaux respectifs. Ceci signifie que ces deux matrices appartiennent à U, ce qui avec l'aide du théorème III.13 achève de prouver notre théorème III.14.

Théorème III.15 *Soit* $\mathbf{M}(z)$ *une matrice de* $G^{d\times D}$. *Alors il existe deux matrices unimodulaires* $\mathbf{U}(z)$ *et* $\mathbf{V}(z)$, *une matrice diagonale de taille* $d\times d$ $\mathbf{D}(z)$ *composée de délais et une matrice constante* $\mathbf{S}$*telles que*

$$\mathbf{M}(z) = \mathbf{D}(z)\mathbf{S}\mathbf{U}(z^{-1})\mathbf{V}(z)$$

Preuve

On utilise la même technique que pour les théorèmes III.8 et III.13, c'est-à dire

- choix d'un k_0 tel que $\deg\mathrm{M}\left(C_{k_0}(z)\right) + \deg\mathrm{M}\left(\tilde{C}_{k_0}(z)\right) \geq 1$
- choix d'un k1 comme dans III.13, avec l'aide de nombres sk et n

ce qui permet de définir les vecteurs u, v et l'entier n tels que

$$\deg\mathrm{V}\left[\mathbf{M}(z)\left(\mathbf{I} - z^n uv^T\right)\right] \leq \deg\mathrm{V}\left[\mathbf{M}(z)\right] - 1$$

et permet de faire décroître le degré vectoriel de $\mathbf{M}(z)$ tant qu'il existe k_0 tel que $\deg\mathrm{M}\left(C_{k_0}(z)\right) + \deg\mathrm{M}\left(\tilde{C}_{k_0}(z)\right) \geq 1$ pour obtenir un couple de matrices $\mathbf{M}'(z)$ et $\tilde{\mathbf{M}}'(z)$ telles que

$$\mathbf{M}'(z) = \mathbf{M}(z)\mathbf{V}(z)^{-1}$$
$$\tilde{\mathbf{M}}'(z) = \tilde{\mathbf{M}}(z)\mathbf{V}(z)^{T}$$

où $\mathbf{V}(z)$ est unimodulaire. Dès que cette inégalité n'est plus vérifiée, on peut multiplier la matrice $\mathbf{M'}(z)$ à gauche par une matrice diagonale $\mathbf{D}(z)^{-1}$ de taille $d\times d$, composée de délais telle que $\mathbf{D}(z)^{-1}\mathbf{M}'(z)$ et $\mathbf{D}(z)\tilde{\mathbf{M}}'(z)$ ne contiennent plus, dans leurs développement polynômiaux que des puissances négatives de z. Cela signifie donc qu'il existe une matrice unimodulaire $\mathbf{U}(z)$ et une matrice constante $\mathbf{S}$ de taille $d\times D$ telles que $\mathbf{D}(z)^{-1}\mathbf{M}'(z) = \mathbf{S}\mathbf{U}(z^{-1})$, ce qui conduit au résultat énoncé par le théorème.

L'un des intérêts de ces résultats de factorisation de matrices rectangulaires sera pour nous lié au fait que nous ne connaîtrons en général que l'un des deux filtres qui composent notre banc de filtres rationnels. Notre tentation sera alors grande de factoriser la matrice rectangulaire correspondant à la branche associée à ce filtre en utilisant selon les cas les théorèmes III.12, III.13, III.14 ou III.15, pour en déduire l'autre filtre en complétant la matrice rectangulaire constante qui apparaît dans ces théorèmes de façon à la rendre rectangulaire. Cette méthode a été utilisée par Vaidyanathan et al. [VNDS] pour la conception d'un banc de N filtres orthogonal (i.-e. paraunitaire) à partir d'un seul filtre, obtenu d'une autre manière.

E. Utilité de la factorisation

Si d'un point de vue mathématique il est assez satisfaisant de pouvoir caractériser simplement le groupe des matrices FIR inversibles, il est également appréciable de pouvoir en tirer des conséquences pratiques.

1. Implantation numérique

Une première utilisation est l'implantation d'un banc de filtres biorthogonal sur un processeur de traitement de signal qui travaillerait en virgule fixe ou sur un circuit imprimé sans perdre la propriété de reconstruction parfaite. Dans le cas le plus général en effet, une matrice $\mathbf{M}(z)$ FIR invertible pourra s'écrire sous la forme

$$\mathbf{M}(z) = z^M \mathbf{S} \prod_{k=1}^{K_u} \left(\mathbf{I} - z^{n_k} u_k v_k^T \right) \prod_{k=1}^{K_p} \left(\mathbf{I} + (z-1) w_k w_k^T \right)$$

où $\mathbf{S}$ est une matrice constante inversible, u_k et v_k des vecteurs orthogonaux entre eux et w_k des vecteurs unitaires. Afin de rendre ces contraintes compatibles avec une implémentation en virgule fixe il suffit de voir que l'on peut réécrire cette formule sous la forme (la partie paraunitaire a été explicitée sous cette forme dans [Vai2])

$$\mathbf{M}(z) = z^M \mathbf{S}' \prod_{k=1}^{K_u} \left(a_k \mathbf{I} - z^{n_k} u_k (j_{k,1} \wedge j_{k,2} \wedge \dots j_{k,D-2} \wedge u_k)^T \right) \prod_{k=1}^{K_p} \left(w_k^T w_k \mathbf{I} + (z-1) w_k w_k^T \right)$$

où $\wedge$ est l'opérateur de produit vectoriel en dimension D, très utilisé en électromagnétisme pour D=3, et où maintenant les vecteurs u_k, $j_{k,1}, j_{k,2}, \dots j_{k,D-2}$, w_k et les réels a_k n'ont plus d'autre contrainte que de n'être pas nuls (voir P. P. Vaidyanathan pour la forme paraunitaire). Dans ces conditions il suffira d'approcher ces quantités, initialement réelles (lors du design) par des valeurs fractionnaires: on est assuré que la matrice résultante sera FIR inversible, et même mieux que cela, on pourra exprimer directement son inverse qui sera également implantable en virgule fixe.

Ajoutons que cette factorisation peut rendre plus économe, en termes de multiplications et additions, l'implémentation des bancs de filtres. C'est effectivement ce qui se passe dans le cas dyadique, comme on le verra dans la section consacrée aux treillis.

2. Conception de filtres

Un autre intérêt de ces factorisations est la conception de filtres pour les bancs de filtres. En effet, il s'agit là d'un problème très complexe dont on n'a de solution que pour des architectures très particulières comme les bancs de deux filtres paraunitaires [SB] ou pour des bancs de filtres modulés.

En fait, on cherchera souvent à minimiser, au sens d'une certaine norme, l'erreur entre notre banc de filtres et un banc de filtres idéal (en général constitué de filtres non réalisables). On ajoutera également certaines contraintes, comme la condition de reconstruction parfaite FIR, la phase linéaire des filtres, ou encore l'annulation des valeurs prises par les filtres en des va-

leurs particulières (comme la moitié de la fréquence d'échantillonnage, ou la composante continue). Le problème devient d'autant plus complexe à résoudre qu'il est hautement dépendant de la norme utilisée (cf chapitre VI).

Cependant, à l'aide des théorèmes de factorisation indiqués dans ce chapitre, on va pouvoir simplifier considérablement la principale des contraintes non-linéaires, celle de reconstruction parfaite FIR. Ainsi, au lieu d'écrire que $\det(\mathbf{M}(z))=z^n$, on explicitera $\mathbf{M}(z)$ à l'aide des vecteurs u, v et w issus de la factorisation. Dans ce cas, les contraintes sur les vecteurs sont beaucoup plus légères et sont aisément intégrables dans un programme de minimisation non linéaire [Vai2,VH]: on pourrait utiliser également la technique mise en œuvre dans [VNDS] (conception d'un filtre du banc de filtres, factorisation de ce filtre, puis optimisation des coefficients de la matrice constante) qui semble être efficace dans le cas paraunitaire.

La complexité du problème n'a bien entendu pas été éliminée par le fait de rendre explicite un problème implicite: il y a toujours un nombre extrêmement important de solutions minimales localement, ce qui rend les programmes de minimisation d'autant plus lourds et longs à exécuter.

3. Régularité

Un gros inconvénient de cette factorisation est son incapacité à prendre en compte simplement la régularité dans le cas de bancs de filtres à deux bandes.

En effet, si l'on impose que le filtre passe-bas d'un système à deux bandes soit multiple de $\frac{z^p-1}{z-1}$, ce qui se traduit sur les coefficients par

$$\sum_k g_{k+np} = C$$

pour tout entier k, et où C est une constante indépendante de k, alors on peut voir que la matrice (rectangulaire) polyphase $\mathbf{G}$ associée à G vérifie la condition suivante

$$(\underbrace{1,1,1,\ldots 1}_{q \text{ composantes}})\mathbf{G}(1) = C(\underbrace{1,1,1,\ldots 1}_{p \text{ composantes}})^T$$

ce qui s'exprime sous la forme très simple suivante

$$(\underbrace{1,1,1,\ldots 1}_{q \text{ composantes}})\mathbf{S} = C(\underbrace{1,1,1,\ldots 1}_{p \text{ composantes}})^T$$

quand on utilise par exemple la factorisation UP (III.20) et quand on modifie les éléments simples unimodulaires de la façon indiquée par (III.16).

Mais dès que l'on passe à des ordres de régularité supérieurs, il est impossible de séparer l'influence de chaque vecteur, ce qui rend la factorisation très inadaptée à ces cas.

F. Structures en treillis

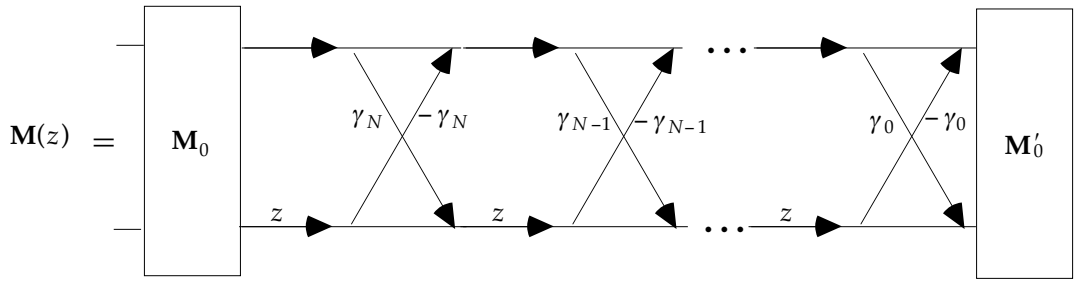
Dans la mesure où une matrice peut se mettre sous la forme d'un produit de matrices simples, on peut immédiatement en tirer une version graphique, sous forme de treillis cascadé, comme c'est le cas pour les matrices paraunitaires [VH]. Il suffit pour cela d'écrire graphiquement l'équivalent d'une matrice paraunitaire simple et d'une matrice unimodulaire simple. Voyons plus précisément ce qui se passe dans le cas d'une matrice de dimension 2.

1. Matrices paraunitaires

Comme on le sait [VH], on peut ramener un produit d'éléments simples paraunitaires sous la forme suivante, ne faisant apparaître qu'un seul paramètre pour chaque élément

$$\mathbf{M}(z) = \mathbf{M}'_0 \begin{pmatrix} 1 & -\gamma_0 \\ \gamma_0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z \end{pmatrix} \begin{pmatrix} 1 & -\gamma_1 \\ \gamma_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z \end{pmatrix} \cdots \begin{pmatrix} 1 & -\gamma_K \\ \gamma_K & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & z \end{pmatrix} \mathbf{M}_0$$

où $\mathbf{M}_0$ est orthogonale et $\mathbf{M'}_0$ est une similitude, i.-e. vérifie $\mathbf{M}'_0 \mathbf{M}'^{T}_0 = \mathrm{Cte} \times \mathbf{I}$. Sous forme graphique cette égalité devient



Un des avantages d'une telle présentation tient au fait que le nombre de multiplications et d'additions est réduit. En effet, s'il y a N sections, décrivant des filtres passe-haut et passe-bas de longueur $2N+2$, chaque section ne contribuera que pour 2 additions et deux multiplications pour deux échantillons, soit au total $N+2$ additions et $N+4$ multiplications par échantillon, alors qu'il aurait théoriquement été nécessaire d'effectuer environ $2N$ additions et multiplications par échantillon. Une division par deux du nombre d'opérations…

2. Matrices unimodulaires

On peut dans le même esprit obtenir un treillis réduisant le nombre d'opérations dans le cas unimodulaire. Il suffit pour cela de remarquer que la matrice unimodulaire simple peut s'écrire sous la forme

$$\mathbf{I} + z^n u v^T = \mathbf{P}^T \mathbf{A} \begin{pmatrix} 1 & z^n \\ 0 & 1 \end{pmatrix} \mathbf{A}^{-1} \mathbf{P}$$

où **P** est une matrice orthogonale et **A** une matrice de la forme

$$\mathbf{A} = \begin{pmatrix} a & x \\ 0 & 1 \end{pmatrix}$$

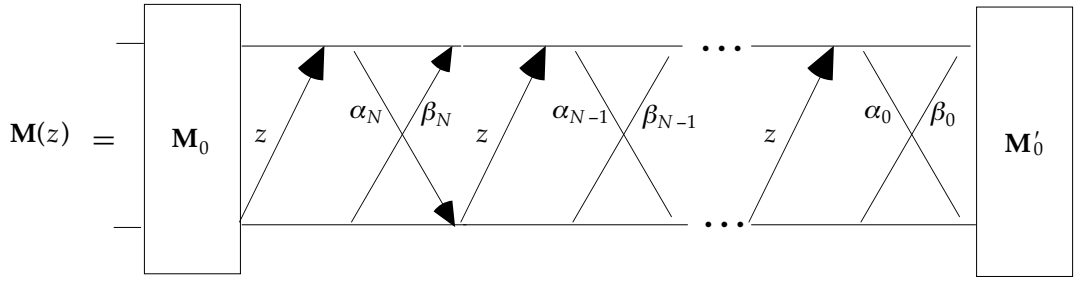
sachant que $a = \|u\|\|v\|$. x représente ici un paramètre que nous allons adapter pour simplifier l'expression finale. Le produit de matrices unimodulaires va faire apparaître des termes de la forme

$$\mathbf{A}'^{-1}\mathbf{P}'\mathbf{P}^T\mathbf{A}\begin{pmatrix} 1 & z^n \\ 0 & 1 \end{pmatrix}$$

Une étude relativement simple de $\mathbf{A}'^{-1}\mathbf{P}'\mathbf{P}^T\mathbf{A}$ montre qu'on peut l'écrire sous la forme (à une constante multiplicative près) $\begin{pmatrix} 1 & \beta \\ \alpha & 1 \end{pmatrix}$ sauf cas exceptionnel où l'on devra l'écrire sous la forme $\begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$ ce cas exceptionnel ne pouvant se produire quand le degré des éléments unimodulaires est 1, ce qui est en pratique toujours vrai puisqu'on considèrera essentiellement la forme dense de la factorisation. On pourra donc, dans la majorité des cas, écrire la factorisation des matrices unimodulaires de dimension 2 de la façon dense suivante

$$\mathbf{M}(z) = \mathbf{M}'_0\begin{pmatrix} 1 & \beta_0 \\ \alpha_0 & 1 \end{pmatrix}\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & \beta_1 \\ \alpha_1 & 1 \end{pmatrix}\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}\cdots\begin{pmatrix} 1 & \beta_N \\ \alpha_N & 1 \end{pmatrix}\begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}\mathbf{M}_0$$

ce qui se représente graphiquement par



À nouveau le nombre d'opérations pour deux échantillons s'élève, par section unimodulaire, à 2 multiplications et 3 additions, ce qui conduit, pour un filtre de longueur $2N+2$ à $3/2N+2$ additions et $N+4$ multiplications par échantillon.

3. Treillis UP et DUU

On peut joindre les deux précédents treillis pour obtenir le treillis de matrices unimodulaires et paraunitaires. Malheureusement, on n'est pas assuré de l'économie de la transformation. Par contre, la décomposition DUU apporte plus d'assurances à ce sujet, d'autant qu'ici, la matrice de délais ne modifie pas les caractéristiques fréquentielles des filtres considérés. On

pourra donc plutôt préférer la décomposition DUU, aussi bien pour l'implémentation que pour la conception des filtres. La factorisation s'écrira alors

$$\mathbf{M}(z) = \begin{pmatrix} z^n & 0 \\ 0 & z^m \end{pmatrix} \mathbf{M}'_0 \begin{pmatrix} 1 & \beta'_0 \\ \alpha'_0 & 1 \end{pmatrix} \begin{pmatrix} 1 & z^{-1} \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & \beta'_{N'} \\ \alpha'_{N'} & 1 \end{pmatrix} \begin{pmatrix} 1 & z^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \beta_0 \\ \alpha_0 & 1 \end{pmatrix} \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & \beta_N \\ \alpha_N & 1 \end{pmatrix} \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix} \mathbf{M}_0$$

et le treillis graphique s'en déduit automatiquement.

G. Résumé du chapitre

Cette partie est un peu à part dans le travail sur les bancs de filtres rationnels itérés. Elle s'applique en fait à tout banc de filtres, itéré ou non, voire issu de branches généralisées. Son but dans un cadre de traitement de signal est de proposer un outil pour la conception de filtres et pour l'implémentation. La factorisation des matrices polyphases d'un banc de filtres permet en effet de simplifier les procédures de conception de filtres, et d'autre part en permet l'implémentation à la fois rapide —du moins dans le cas dyadique— et en virgule fixe —évitant entre autres les erreurs d'arrondi—. L'originalité du travail présenté dans ce chapitre tient essentiellement à la factorisation des matrices FIR inversibles non nécessairement orthogonales —c'est-à dire le cas général—, un résultat peu connu dans la littérature de traitement de signal.

Les résultats classiques dans le cas orthogonal [Vai2] ont été précisés, certains ont également été étendus —factorisation des matrices rectangulaires, mais voir aussi [CV]—. Enfin une version treillis de la factorisation générale dans le cas dyadique est exhibée, présentant des avantages similaires aux treillis "lossless".

IV. Itérations

Afin d'implémenter une transformation à Q-constant, on veut construire directement un banc de filtres rationnel à N branches. Il s'agit là d'une tâche assez difficile dans la mesure où l'on doit choisir une structure dont l'inverse est miroir de l'analyse. Il se trouve qu'alors, l'itération d'un banc de deux filtres permet d'obtenir un banc de filtres de taille aussi grande que voulue, tout en conservant des propriétés simples de reconstruction: c'est ce qui est fait couramment dans le cas dyadique [RV,Ma2,KB].

L'itération se fait sur une seule des deux bandes, la branche passe-bas, ce qui permet d'obtenir une transformation à Q-constant. L'un des exemples qui nous intéressera le plus dans le but d'une application pratique sera le cas du codage de sons: le système auditif humain semblant analyser les signaux sonores sur des bandes d'un tiers d'octave, on voudra pouvoir réaliser une transformation en tiers d'octave, ce qui signifie qu'il faudra trouver des entiers p et q tels que $\left(\frac{p}{q}\right)^3$ soit suffisamment proche de 2. Un calcul rapide montre que p/q=5/4 est un bon candidat. On verra cependant que la taille des bandes critiques (voir chapitre VII) impose plutôt p/q=6/5 puisque l'analyse du système auditif est légèrement plus fine que le tiers d'octave [ZF].

Le problème des itérations est qu'elles sont en général très instables. C'est parfaitement naturel: si l'on prend une fonction presque égale à 1 sur une très grande partie de son domaine de définition, la mettre à la puissance N va faire apparaître des différences avec 1, d'autant plus importantes que N est plus grand. Or justement, les itérations se comportent comme une forme de produit —il s'agit plutôt d'un produit de matrices (cas dyadique, voir [DauL2]) comme on le verra dans le chapitre V—.

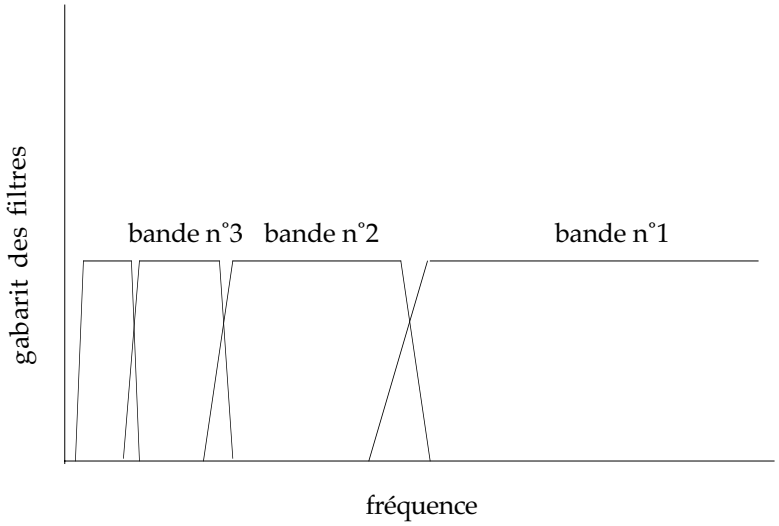
Un tel comportement n'est bien sûr pas désiré puisque l'on veut que le banc de filtres itéré soit précisément très sélectif. On verra qu'une réponse à ce problème consiste à étudier certaines fonctions que l'on peut construire à partir du filtre passe-bas, et en particulier, leur régularité. Ce chapitre est dédié à la mise en évidence de ces fonctions dont on va voir qu'elles sont limites de schémas discrets itérés.

Le problème avait auparavant été soulevé dans la littérature par Kovačević et Vetterli [KV1,KV3], qui avaient conclu qu'il n'existait pas de fonction limite associée à l'itération du filtre passe-bas FIR, comme c'est le cas dans les itérations en octave. La situation est en fait plus compliquée: Kovačević et Vetterli cherchaient une seule fonction et ses versions translatées, alors qu'en réalité, il existe un nombre infini —dénombrable— de fonctions "presque" translatées les unes par rapport aux autres. Qui plus est, Kovačević et Vetterli, au lieu de suivre exactement la technique utilisée dans les bancs de filtres dyadiques, avaient plutôt mis en évidence la fonction moyenne —définie plus loin dans ce chapitre— des fonctions limites, dont les schémas discrets ne convergent pas ponctuellement mais au sens "faible" des distributions.

La plus grosse partie de ce chapitre sera ainsi consacré à un sujet —la mise en évidence des fonctions limites et quelques unes de leurs propriétés— qui a donné lieu à publication dans une revue [Blu1]: on sera donc plus rapide pour les démonstrations qui seront fréquemment renvoyées à cet article.

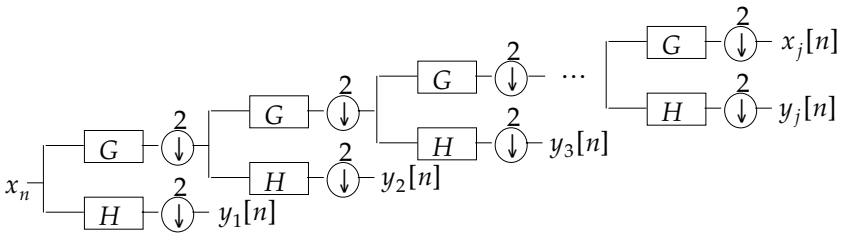
A. Cas Dyadique

Comme il peut être difficile de suivre ce chapitre si l'on ne connaît pas le cas dyadique, on va en rappeler les caractéristiques de façon succincte. Ce cas où p=2 et q=1 est bien sûr bien mieux référencé dans la littérature que le cas rationnel —lacune qui a d'ailleurs justifié le présent travail—. Itérer un banc de deux filtres sous-échantillonnés par deux, sur le seul passe-bas conduit à un banc de filtres non uniforme dont les facteurs d'échantillonnage sont des puissances croissantes de 2. Cette opération doit permettre d'obtenir une décomposition du signal d'entrée en bandes de largeurs inégales comme indiqué par le schéma ci-dessous



et surtout permet d'envisager un aussi grand nombre d'itérations que cela est souhaitable.

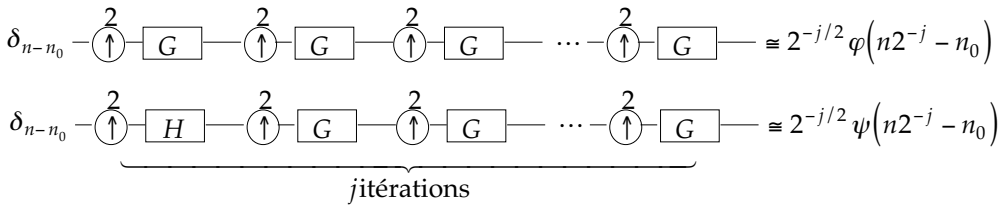
Le résultat fondamental concernant cette architecture est tiré de [Ma1] où il est montré qu'un banc de filtres itéré en octaves implémente une transformée en ondelettes discrète de facteur d'échelle 2. Plus précisément, si G et H sont les filtres orthogonaux passe-bas et passe-haut de l'analyse ci-dessous



il existe deux fonctions φ (“père” [Mey1]) et ψ (“mère” [ibid.]) construites à partir de G et H telles que si l'on pose $x(t) = \sum_n x_n \varphi(n-t)$ alors

$$x_j[n] = 2^{-j/2} \int x(t) \varphi\left(n - 2^{-j} t\right) dt$$
$$y_j[n] = 2^{-j/2} \int x(t) \psi\left(n - 2^{-j} t\right) dt$$

φ et ψ sont obtenues à travers un processus itératif infini (voir par exemple [Ri1,DauL1])



Elles sont appelées pour cette raison “fonctions limites”. Il faut tout de suite noter la dépendance particulière de la limite quand on change la valeur de n_0: le résultat est une fonction translatée de la même quantité, une invariance qui ne sera plus vraie quand on étendra cette définition au cas rationnel.

Les fonctions limites vérifient les équations suivantes (la première est dénommée “équation de changement d'échelle” ou “two-scale difference equation” [DauL1])

$$\varphi(x) = \sqrt{2} \sum_n g_n \varphi(2x - n)$$
$$\psi(x) = \sqrt{2} \sum_n h_n \varphi(2x - n)$$

ce qui permet de démontrer que le processus de construction est toujours convergent, au moins au sens des distributions pourvu que $G(1) = \sqrt{2}$ ici [DauL1]. On peut étendre ce résultat directement au cas biorthogonal: cela nécessite une formulation légèrement plus complexe puisqu'il faut alors faire intervenir la fonction père de reconstruction qui est différente de celle d'analyse.

Cette interprétation du banc de filtres itéré est capitale car elle fait le lien entre une transformation continue qui a été bien étudiée mathématiquement et une transformation discrète. Celle-ci présente pour cette raison un intérêt semblable à celui de la transformée de Fourier pour l'analyse et la synthèse de signaux réels. Notons également que l'ondelette d'analyse (et de synthèse si l'on considère un banc de filtres à reconstruction FIR parfaite) est parfaitement bien localisée en temps puisque, c'est l'une des propriétés des fonctions issues de schémas FIR itérés, cette fonction est à support borné.

En fait, il est devenu très vite nécessaire de s'intéresser à la régularité de ces fonctions limites car dans le cas général, même si les filtres générateurs sont bien sélectifs en fréquence, φ et ψ sont des distributions dont le comportement fréquentiel est extrêmement erratique. Ceci signifie que pour un nombre d'itérations suffisamment grand, le banc de filtres itéré ne présente pas non plus de bonnes propriétés fréquentielles.

Deux méthodes ont été utilisées pour qualifier la régularité des fonctions limites, l'une dans l'espace de Fourier (régularité au sens de Sobolev [Dau1]) et l'autre directement dans l'espace de définition des fonctions limites (régularité au sens de Hölder [DauL1,DauL2,Ri1]). De ces deux approches, la plus précise est la seconde qui bénéficie en outre d'un algorithme efficace de calcul [Ri1]. On retiendra cependant que, de manière grossière, la régularité de φ et ψ est déterminée par le nombre de facteurs en $(1+z)$ dans le polynôme passe-bas $G(z)$. L'analyse montre en effet que ce facteur, dit de "régularité", agit comme un intégrateur.

Tous ces résultats vont maintenant être étendus au cas rationnel, et l'on verra que, mise à part la perte d'invariance en translation, la généralisation prend une forme assez naturelle. La partie consacrée à la régularité sera reportée au chapitre V, également consacré au calcul de l'erreur induite par la non invariance en translation.

B. Forme des filtres itérés

N itérations du banc de filtres d'analyse définissent le banc de $N+1$ filtres donné en figure 4

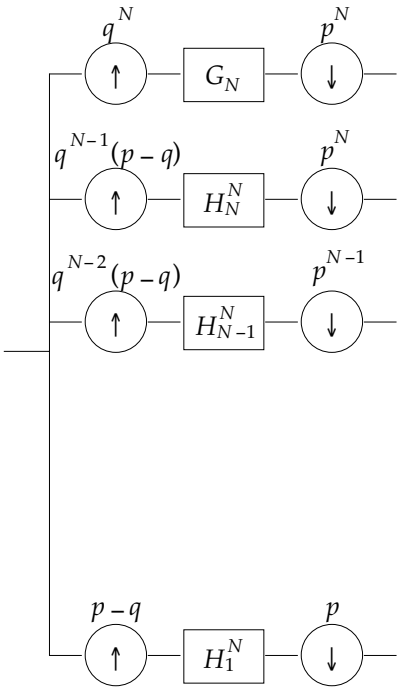


figure 4

Ce banc de filtres est composé de N filtres passe-bande que l'on note H_k^N, et d'un filtre passe-bas G_N qui sont définis par les relations suivantes

$$G_N(z) = G\left(z^{p^{N-1}}\right)G\left(z^{qp^{N-2}}\right)\ldots G\left(z^{q^{N-2}p}\right)G\left(z^{q^{N-1}}\right) \qquad \text{(IV.1)}$$

$$H_k^N(z) = G_{k-1}\left(z^{p-q}\right)H\left(z^{p^{k-1}}\right) \qquad \text{(IV.2)}$$

pour k=1…N. Ces relations sont aisément démontrées en utilisant la loi de composition des branches rationnelles; en particulier, on a $H_1^N = H$ pour tout N. On voit ainsi l'importance prépondérante du filtre G qui est le seul à être effectivement itéré, pour donner G_N qui obéit à l'équation de récurrence suivante

$$G_{N+N'}(z) = G_N\left(z^{p^{N'}}\right)G_{N'}\left(z^{q^N}\right) \qquad \text{(IV.3)}$$

sachant bien sûr que $G_0 = 1$ et $G_1 = G$.

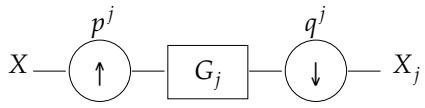
Dans une situation idéale, les filtres H_k^N sont passe-bande et G_N est passe-bas, comme on l'a dit plus haut. Cependant, il n'est pas évident de savoir comment choisir G et H afin de s'approcher de cette situation idéale. On verra qu'une description du banc de filtres à l'aide de fonctions à temps continu permet d'introduire de nouvelles propriétés, assez peu naturelles du point de vue discret d'un banc de filtres, mais qui sont étroitement reliées à cette préoccupation. Ces propriétés ont pour nom la régularité et l'amnésie, mais auparavant, il faut introduire les fonctions limites associées à ce banc de filtres itérés.

C. Fonctions limites

Il y a au moins deux méthodes pour construire des fonctions associées à un banc de filtres itérés. La première met en jeu des suites discrètes censées converger par densité [Ri1], alors que la deuxième construit des fonctions destinées à s'approcher le plus possible d'une certaine fonction limite [Dau1].

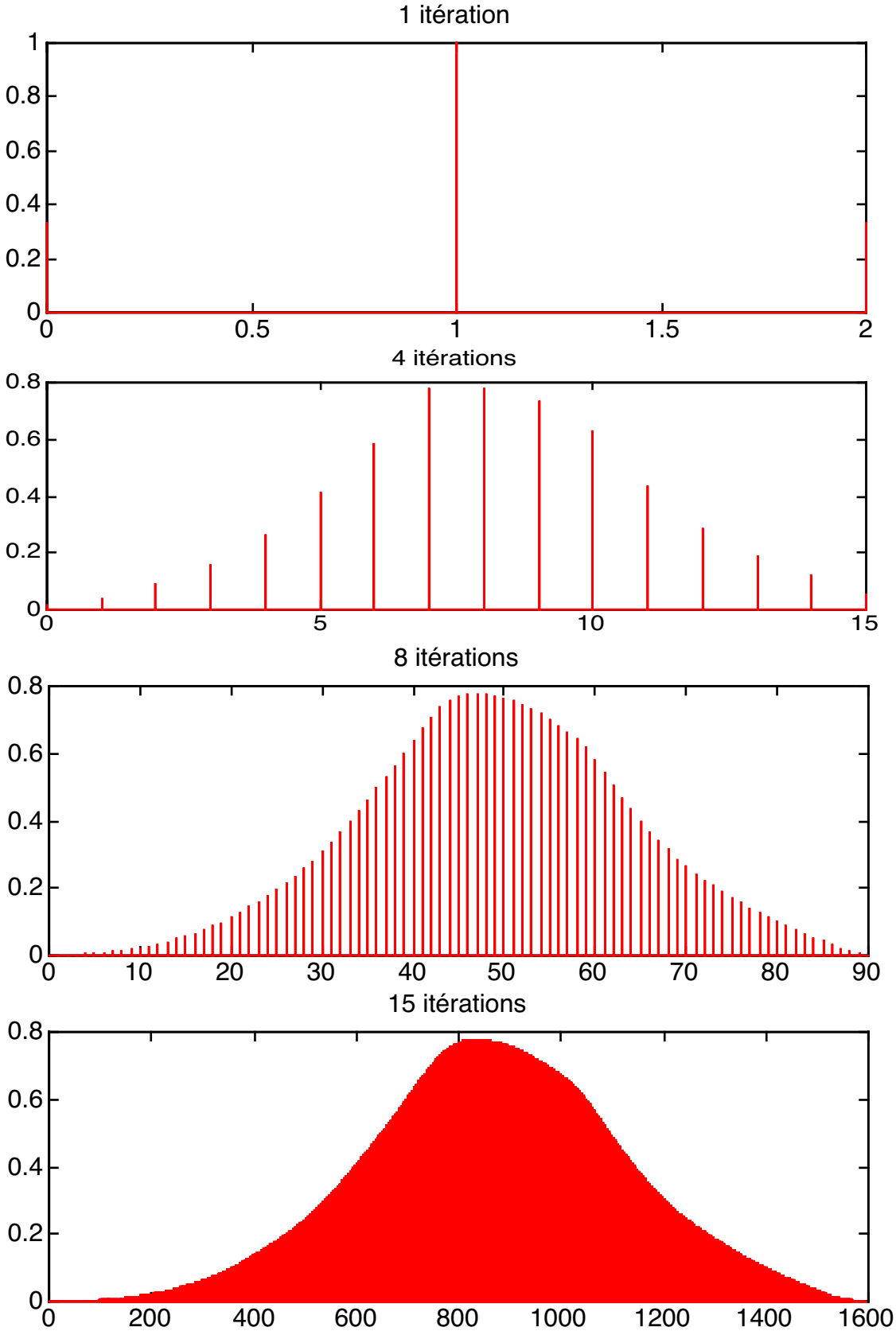
1. Convergence des schémas discrets

La première se base sur la partie synthèse, ou reconstruction: on itère un filtre interpolant, et c'est la manière la plus simple d'obtenir les fonctions limites. Décrivons-là maintenant: soit un signal d'entrée X constitué d'un unique délai $X(z) = z^s$. Après j itérations de la branche p/q on obtient, dans un domaine temporel suréchantillonné de $(p/q)^j$, le signal X_j selon le schéma

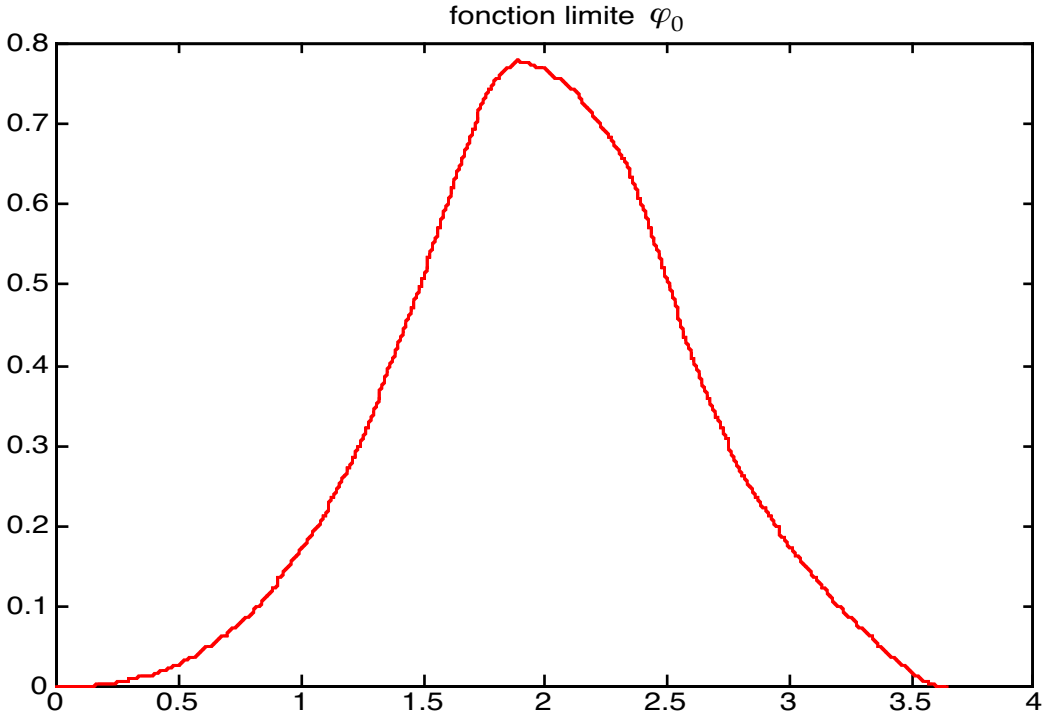


ce qui donne les tracés suivants, en choisissant

$$\frac{p}{q} = \frac{3}{2} \quad s = 0 \quad \text{et} \quad G(z) = \frac{1}{3}\left(\frac{z^3 - 1}{z - 1}\right)^2$$


1 itération
4 itérations
8 itérations
15 itérations

Si l'on répète le processus à l'infini, la forme de la frontière du graphe se stabilise, et si on la trace en fonction du temps, dans le domaine non sur-échantillonné, on obtient (ici ∞=25 itérations) la fonction limite d'indice 0



Au vu de ces graphes, il n'est pas nécessaire d'expliquer pourquoi on dit que la construction de cette fonction (de la variable continue) suit un schéma discret. On verra plus loin les conditions nécessaires pour que ce genre de processus converge, et dans le chapitre V des conditions suffisantes de convergence vers des fonctions de régularité fixée (voir aussi [BR,RB]). Mathématiquement, pour chaque s entier, on définit une fonction φ_s telle que

$$\lim_{j\to\infty} \sup_k \left| \varphi_s\left(k\frac{q^j}{p^j}\right) - g_j[kq^j - sp^j] \right| = 0 \qquad \text{(IV.4)}$$

et c'est en ce sens-là, c'est-à-dire uniformément et ponctuellement que l'on considère la convergence. On dit alors que le polynôme itéré G_j converge fortement vers les fonctions φ_s.

Au delà du plaisir de faire apparaître un objet exotique, il est en fait assez justifié de s'intéresser à l'amplitude des coefficients de G_j car c'est celle-ci qui donnera des indications sur la stabilité de la reconstruction d'un système d'analyse-synthèse après quantification des sous-bandes issues de l'analyse. On peut donc considérer utile l'étude d'une fonction qui synthétise l'information sur les coefficients du filtre itéré. Mais il y a plus: cette fonction va se comporter comme une fonction d'interpolation pour le système d'analyse-synthèse. On démontrera cette importante propriété plus loin.

2. Convergence des schémas continus

Une deuxième façon de faire apparaître des fonctions limites est issue de la partie analyse. Soit $x(t)$ un signal réel à bande limitée, et donc obéissant à la formule d'interpolation de Nyquist (fréquence d'échantillonnage $F=1/T$) pour les échantillons x_n et sa formule inverse

$$\begin{aligned} x(t) &= \sum_n x_n \chi(Ft - n) \\ x_n &= \int x(uT)\chi(u - n)du \end{aligned} \qquad \text{(IV.5)}$$

Après j itérations, on a

$$\begin{aligned} x_j[n] &= \int x(uT)\varphi_{j,n}\left(\tfrac{q^j}{p^j}u\right)du \\ \text{où} \quad \varphi_{j,n}(u) &= \sum_k g_j[np^j - kq^j]\chi\left(\tfrac{p^j}{q^j}u - k\right) \end{aligned} \qquad \text{(IV.6)}$$

Ce sont les fonctions $\varphi_{j,n}$ qui vont converger vers des fonctions $\phi_n(u)$ quand j tend vers l'infini. Mathématiquement, on n'a pas besoin de convergence forte ici. À la place, on utilise la convergence au sens des distributions [GM]: pour toute fonction $f(u)$ à support compact et indéfiniment dérivable

$$\lim_{j\to\infty}\int f(t)\left[\phi_n(t) - \varphi_{j,n}(t)\right]dt = 0 \qquad \text{(IV.7)}$$

On dit alors que le polynôme itéré G_j converge au sens des distributions, ou faiblement, vers les distributions ϕ_s.

Note 1: indéfiniment dérivables car ayant un support fréquentiel borné, les signaux $x(uT)$ ne sont pas à support temporel compact, mais on verra que les fonctions, ou distributions φ_n le sont, ce qui signifie que la définition (IV.7) s'applique en fait à toute fonction indéfiniment dérivable, sans restriction de support.

Note 2: On n'est pas assuré que s'il y a convergence pour toute fonction f C^∞ à support compact alors l'objet limite est bien une distribution vérifiant en particulier la propriété de "continuité" [GM] qui doit la caractériser

$$\left|\int f(t)\phi_n(t)dt\right| \le C \max_{j=0\ldots N}\sup_t\left|f^{(j)}(t)\right|$$

où C est une constante qui ne dépend que du support de f (cas général pour une distribution).

a. Lien entre les deux types de convergence

Constatons tout de suite que, si la convergence au sens des distributions est ponctuelle et uniforme, ce qui est bien plus restrictif, alors le schéma discret associé au même filtre converge au

sens fort et l'on a en outre $\check{\varphi}_n(u) = \varphi_{-n}(-u)$. C'est assez facile à prouver grâce aux propriétés interpolantes de la fonction $\chi(u)$ puisqu'on obtient alors

$$\varphi_{j,n}\left(k\frac{q^j}{p^j}\right) = g_j[np^j - kq^j]$$

ce qui permet de passer d'une définition de convergence à l'autre, mais sous les contraintes d'uniformité et de ponctualité uniquement. En fait, des deux définitions de convergence, la première, c'est-à dire la convergence au sens fort, entraîne la deuxième, au sens des distributions.

Lemme *Une façon équivalente d'écrire la convergence du polynôme G_j est la suivante: pour toute fonction* f(u) *à support compact et indéfiniment dérivable*

$$\lim_{j\to\infty}\left[\frac{q^j}{p^j}\sum_k f\left(k\frac{q^j}{p^j}\right)g_j[kq^j - np^j] - \int f(t)\varphi_n(t)dt\right] = 0 \qquad \text{(IV.9)}$$

Preuve

Bien sûr, il n'y aurait rien à démontrer si f était à support fréquentiel borné puisqu'on aurait alors une égalité pour j suffisamment grand. C'est donc cette propriété d'interpolation exacte qui s'étend au cas des fonctions C^∞ à support compact. On va d'abord démontrer que pour tout entier N et pour toute fonction C^∞ à support compact $f(t)$, il existe une constante C_N telle que

$$\left|\int f(t)\chi\left(\frac{p^j}{q^j}t - k\right)dt - \frac{q^j}{p^j}f\left(k\frac{q^j}{p^j}\right)\right| \le C_N\frac{q^{jN}}{p^{jN}}$$

Pour cela on réécrit le membre de gauche dans l'espace de Fourier

$$\int f(t)\chi\left(\frac{p^j}{q^j}t - k\right)dt - \frac{q^j}{p^j}f\left(k\frac{q^j}{p^j}\right) = \frac{q^j}{p^j}\int \hat{f}(\nu)\left[\hat{\chi}\left(\frac{q^j}{p^j}\nu\right) - 1\right]e^{2i\pi k\frac{q^j}{p^j}\nu}d\nu$$

Comme f est au moins N fois dérivable et est à support compact S, on a la majoration suivante pour la transformée de Fourier de f

$$\left|\hat{f}(\nu)\right| \le \frac{\max_t\left|\partial^N f(t)\right|}{\left|2\pi\nu\right|^N}S$$

et comme $\hat{\chi}(\nu)$ est l'indicatrice de l'intervalle [-1/2,1/2], on obtient finalement après intégration

$$\left|\int f(t)\chi\left(\frac{p^j}{q^j}t - k\right)dt - \frac{q^j}{p^j}f\left(k\frac{q^j}{p^j}\right)\right| \le \underbrace{\frac{S\max_t\left|f^{(N)}(t)\right|}{(N-1)\pi^N}}_{C_N}\frac{q^{jN}}{p^{jN}}$$

D'autre part, les coefficients de $G_j(z)$ ainsi que sa longueur de ne peuvent pas croître plus qu'exponentiellement: on peut donc trouver N tel que

$$\frac{q^{jN}}{p^{jN}} \sum_k \left| g_j[kq^j - np^j] \right| \underset{j\to\infty}{\to} 0$$

d'où le résultat.

Un à-côté de la démonstration est qu'il n'est en fait pas nécessaire d'avoir une fonction indéfiniment dérivable, mais cependant suffisamment régulière pour contrebalancer l'éventuelle divergence des coefficients g_j quand j tend vers l'infini.

3. Fonctions passe-haut

Si le filtre passe-bas permet de définir des fonctions "pères" [Mey1], le filtre passe-haut permet lui-aussi de définir des fonctions "mères". Il s'agit cette fois, par définition, de la limite des schémas discrets associés au filtre $G_{j-1}(z^{p-q})H(z^{p^{j-1}})$. On peut voir facilement en développant l'expression que les fonctions limites associées sont directement liées aux fonctions passe-bas par la relation

$$\psi_n(t) = \sum_k h_{k(p-q)-np} \varphi_k\left(\frac{p}{q}t\right) \qquad \text{(IV.10)}$$

Cette équation pourra d'ailleurs être comparée, pour la forme, à l'équation de changement d'échelle (IV.19) que nous démontrerons plus loin.

La série de fonctions ainsi définie jouera le rôle véritable d'ondelettes, ou plutôt de pseudo-ondelettes puisqu'elles ne vérifient pas la condition d'invariance temporelle dans cette transformation.

D. Propriétés

Ces fonctions que nous venons d'introduire présentent un certain nombre de caractéristiques qui soit en restreignent l'intérêt (régularité finie, perte de la propriété d'invariance en translation), soit à l'inverse les mettent en valeur (support borné, équation de changement d'échelle).

1. Support

Afin de simplifier les énoncés on définit la localisation d'une suite de fonctions

Définition IV.1 *On dit qu'une suite de fonctions f_n est localisée autour de n si et seulement s'il existe deux réels a et b tels que le support de chaque fonction f_n soit contenu dans [a+n,b+n].*

On va voir que, justement, les fonctions limites sont localisées autour de n: c'est une conséquence du fait que G est FIR. En effet, si l'on pose $G(z) = \sum_l^L g_n z^n$ alors on aura

$$G_j(z) = \sum_{l\frac{p^j-q^j}{p-q}}^{L\frac{p^j-q^j}{p-q}} g_j[n] z^n$$

La somme (IV.9) dont la convergence définit une fonction limite n'engage donc que les indices k tels que

$$n + \frac{l}{p-q}\frac{p^j-q^j}{p^j} \le k\frac{q^j}{p^j} \le n + \frac{L}{p-q}\frac{p^j-q^j}{p^j}$$

ce qui signifie qu'un produit scalaire avec la fonction limite n'engagera que les valeurs de la variable compris entre $n+l/(p-q)$ et $n+L/(p-q)$. On a ainsi

$$\text{support}(\varphi_n) \subset \left[n + \frac{l}{p-q}, n + \frac{L}{p-q} \right]$$

En fait cette inclusion est stricte en général [Blu1], et des valeurs plus fines du support peuvent être trouvées. On observe alors que la largeur du support n'est pas la même pour toutes les fonctions, illustrant ainsi la propriété d'amnésie qui sera décrite plus loin. Voir [Blu1] pour plus de détails.

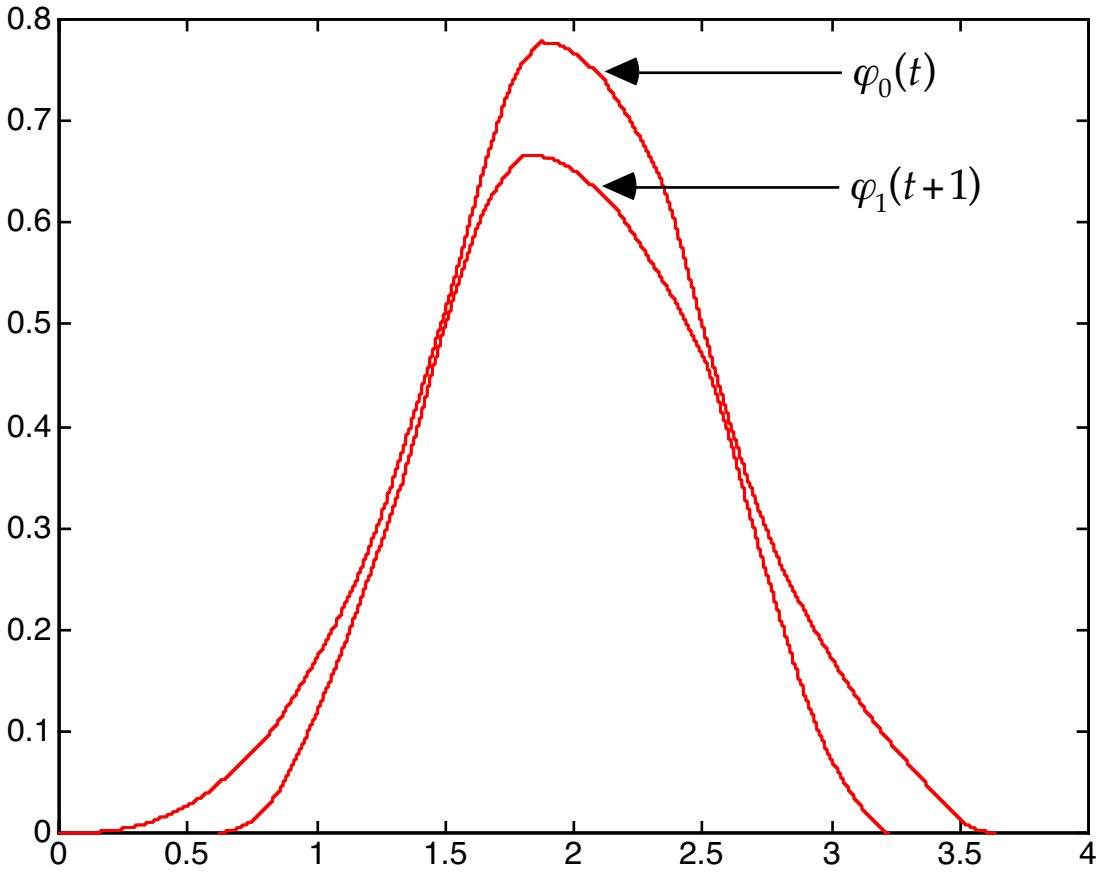
Il est à remarquer que, si l'on note $[a_n,b_n]$ le support de la fonction φ_n, alors ces deux suites de réels vérifient les relations

$$\left.\begin{aligned} \tfrac{p}{q} a_n &= a_{\lambda_0(n)} \\ \tfrac{p}{q} b_n &= b_{\lambda_1(n)} \end{aligned}\right\} \text{ avec } \begin{cases} \lambda_0(n) = E\left(\frac{l+q-1+np}{q}\right) \\ \lambda_1(n) = E\left(\frac{L+np}{q}\right) \end{cases} \qquad \text{(IV.10)}$$

Cette propriété sera définie plus loin (partie sur les valeurs particulières) comme la propriété d'invariance d'échelle pour les suites.

2. Amnésie

Par ce terme on désigne la perte d'invariance temporelle des fonctions de base φ_n. Ceci apparaît directement sur l'exemple ci-dessous



En fait on peut démontrer qu'il est nécessaire que le filtre G soit IIR pour avoir une amnésie nulle [CD,KV1].

3. Fonction moyenne

On a vu que les fonctions limites φ_n sont supportées par des intervalles de la forme I+n, où l'on pose bien sûr I=[$l/(p\text{-}q)$,$L/(p\text{-}q)$], ce qui signifie que les fonctions $\varphi_n(t+n)$ sont toutes supportées par l'intervalle I qui ne dépend pas de n. Dans le cas dyadique, on sait que toutes ces fonctions sont identiques, et d'autre part, on verra que sous certaines conditions, dans le cas rationnel également, les différences peuvent être minimes, atténuant ainsi les effets dûs à l'amnésie $\varphi_n(t+n) \cong \varphi_{n'}(t+n')$. On peut donc essayer de s'intéresser à une sorte de fonction moyenne qui pourrait être définie par une expression de la forme

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \varphi_n(t+n)$$

Il se trouve qu'une telle expression a un sens dès que la convergence des fonctions se fait au sens des distributions *de manière uniforme pour tous les indices n*. En d'autre termes, cela signifie que

$$\forall f \in C_0^\infty \quad \lim_{j\to\infty} \sup_n \left| \langle \varphi_n^j - \varphi_n, f \rangle \right| = 0$$

On va démontrer ce résultat, ainsi que quelques corollaires associés, en 3 étapes.

Lemme IV.2 *Si G(1)=p, le produit infini*

$$\prod_{j\geq 0}\frac{1}{p}G\left(e^{-2i\pi\nu q^j/p^{j+1}}\right) \qquad (IV.11)$$

est convergent, et définit la transformée de Fourier d'une distribution $\varphi(t)$ à support compact vérifiant l'équation de changement d'échelle

$$\varphi(t)=\frac{1}{q}\sum_k g_k\varphi\left(\frac{p}{q}t-\frac{k}{q}\right) \qquad (IV.12)$$

La transformée de Fourier vérifie l'équation

$$\hat{\varphi}(\nu)=\frac{1}{p}G\left(e^{-2i\pi\frac{\nu}{p}}\right)\hat{\varphi}\left(\nu\frac{q}{p}\right) \qquad (IV.13)$$

Preuve

Ce produit infini est évidemment convergent: pour la preuve on peut se reporter à Daubechies et Lagarias [DauL1]. Il définit donc une certaine fonction f de ν qui vérifie l'équation suivante

$$f(\nu)=\frac{1}{p}G\left(e^{-2i\pi\nu\frac{1}{p}}\right)f\left(\frac{q}{p}\nu\right)$$

On peut trouver une fréquence ν_0 telle que f soit bornée sur l'intervalle $[-\nu_0,\nu_0]$. En effet comme G est un polynôme de degré fini, au voisinage de ν=0 il existe ν_0 telle que $|G(e^{-2i\pi\nu})|$ soit croissante, ou bien décroissante, pour tout $0\leq\nu\leq\nu_0$. Dans le cas où $|G(e^{-2i\pi\nu})|$ est décroissante, alors $|f|$ est également décroissante sur $0\leq\nu\leq p\nu_0$ et donc $|f|\leq 1$ sur cet intervalle. Dans l'autre cas, où $|G(e^{-2i\pi\nu})|$ est croissante, alors $|f|$ l'est également sur $0\leq\nu\leq p\nu_0$, ce qui entraîne $|f(\nu)|\leq|f(\nu_0)|$ sur le même intervalle. Le même raisonnement peut être fait pour les valeurs négatives de ν. Tout ceci indique donc qu'il existe un intervalle $[-\nu_0,\nu_0]$ sur lequel la fonction f est bornée. Cette propriété est indispensable pour montrer que f ne peut, en outre, croître plus rapidement qu'un polynôme de degré fixe en ν pour toute valeur de ν. Pour voir cela, il suffit de noter que l'on a l'inégalité suivante

$$\forall\nu\in\mathbf{R}\quad |f(\nu)|\leq C\left|\frac{\nu}{\nu_0}\right|^{\frac{\log C}{\log(p/q)}}\sup_{-\nu_0\leq\nu'\leq\nu_0}|f(\nu')|$$

$$\text{où}\quad C=\frac{1}{p}\sup_{0\leq\theta\leq 2\pi}|G(e^{i\theta})|$$

qui vient directement de l'itération de l'équation fonctionnelle vérifiée par *f*.

f fait donc partie de la classe des distributions tempérées [GM] dont on peut prendre la transformée de Fourier: on peut ainsi définir la fonction φ par $\hat{\varphi} = f$. D'après l'équation fonctionnelle vérifiée par *f*, il est clair que φ vérifie pour sa part l'équation (IV.12).

Enfin, pour prouver que φ est bien à support compact, il suffit d'écrire que

$$\hat{\varphi}(\nu) = \lim_{j\to\infty} \frac{1}{p^j} G_j\left(e^{-2i\pi\nu/p^j}\right)\chi\left(\nu\frac{q^j}{p^j}\right)$$

où χ est la fonction d'interpolation idéale définie en introduction (sinus cardinal). Sa transformée de Fourier est donc la fonction caractéristique de l'intervalle [-0.5,0.5]. En ramenant cette limite en variable temporelle

$$\varphi(t) = \lim_{j\to\infty} \frac{1}{q^j} \sum_k g_j[k]\chi\left(\frac{p^j}{q^j}t - \frac{k}{q^j}\right)$$

ce qui n'a de signification qu'au sens des distributions bien sûr (et non de façon forte). On constate alors simplement que le support des différentes fonctions qui convergent vers φ, reste borné entre $l/(p\text{-}q)$ et $L/(p\text{-}q)$ à cause de la longueur de *Gj*.

Lemme IV.3 *La suite des coefficients de G_j converge au sens des distributions vers la fonction φ, c'est-à dire*

$$\langle\varphi, f\rangle = \lim_{j\to\infty} \frac{1}{p^j} \sum_k g_j[k] f\left(\frac{k}{p^j}\right)$$

pour toute fonction f C^∞ à support compact. Cependant, les $g_j[k]$ ne peuvent pas converger ponctuellement, tant que G est FIR.

Preuve

Posons que le support de φ soit contenu dans l'intervalle [0,*T*]. Alors on peut développer *f* en série de Fourier

$$f(t) = \sum_k f_k e^{2i\pi k\frac{t}{T}}$$

en outre, cette convergence est absolue, et comme pour tout *N*, on peut trouver une constante C_N, telle que pour tout $k\neq 0$ on a

$$|f_k| \leq \frac{C_N}{|k|^N}$$

on a également la convergence absolue des dérivées

$$\partial^N f(t) = \sum_k f_k \partial^N e^{2i\pi k \frac{t}{T}}$$

Grâce à la propriété de continuité des distributions, on en déduit donc que

$$\langle \varphi, f \rangle = \sum_k f_k \langle \varphi, e^{2i\pi k \frac{t}{T}} \rangle$$

où les coefficients de f_k ne croissent pas plus vite qu'un certain polynôme, ce qui entraîne que la convergence est uniforme. D'autre part, pour toute fréquence ν, on a

$$\hat{\varphi}(-\nu) = \langle \varphi, e^{2i\pi\nu} \rangle = \lim_{j\to\infty} \frac{1}{p^j} \sum_k g_j[k] e^{2i\pi\nu \frac{k}{p^j}}$$

la convergence étant uniforme pour tout ν appartenant à un intervalle de la forme [-A,A]. Cette double uniformité entraîne (on omet ici le passage un peu technique de la preuve)

$$\begin{aligned}\langle \varphi, f \rangle &= \lim_{j\to\infty} \frac{1}{p^j} \sum_k \sum_{k'} f_{k'} g_j[k] e^{2i\pi \frac{kk'}{Tp^j}} \\ &= \lim_{j\to\infty} \frac{1}{p^j} \sum_k g_j[k] f\left(\frac{k}{p^j}\right)\end{aligned}$$

c'est-à dire ce que l'on voulait démontrer. La convergence ne se fait bien évidemment jamais au sens fort, car si c'était le cas, on aurait

$$\lim_{j\to\infty} \sup_k \left| g_j[k] - \varphi\left(\frac{k}{p^j}\right) \right| = 0$$

ce qui entraînerait en particulier

$$\lim_{j\to\infty} \sup_k \left| g_j[kq^j - np^j] - \varphi\left(\frac{kq^j - np^j}{p^j}\right) \right| = 0$$

et donc $\varphi_n(t) = \varphi(t-n)$, ce qui est impossible, on l'a vu, lorsque G est FIR.

Ce résultat est en fait très instructif et est la cause de l'interprétation de Kovačević et Vetterli. Ils avaient en effet affirmé dans un premier temps [KV1] en utilisant un filtre particulièrement régulier que les schémas rationnels convergeaient vers une fonction limite, puis sont revenus sur leur opinion, en observant le comportement oscillant des coefficients du polynôme G_j. En fait plutôt que $g_j[k]$, ils auraient dû observer $g_j[kq^j - np^j]$ (qui converge pour chaque n), et d'autre part les $g_j[k]$ qu'ils avaient tracés eux-même convergent… au sens faible

des distributions certes, mais ce sens est suffisant si l'on s'intéresse aux schémas d'analyse où toutes les quantités apparaissent sous le signe d'intégration.

Théorème IV.4 *Supposons que G_j converge au sens des distributions de manière uniforme vers les distributions φ_n, alors la limite*

$$\lim_{N_1,N_2\to\infty} \frac{1}{N_1+N_2+1} \sum_{k=-N_1}^{N_2} \varphi_n(t+n)$$

a un sens et constitue une nouvelle définition de la fonction φ introduite dans le lemme 1

$$\hat{\varphi}(\nu) = \frac{1}{p} G\left(e^{-2i\pi\nu\frac{1}{p}}\right) \frac{1}{p} G\left(e^{-2i\pi\nu\frac{q}{p^2}}\right) \frac{1}{p} G\left(e^{-2i\pi\nu\frac{q^2}{p^3}}\right) \dots \frac{1}{p} G\left(e^{-2i\pi\nu\frac{q^{j-1}}{p^j}}\right) \dots \qquad \text{(IV.14)}$$

Preuve

On va obtenir trois majorations qui permettront de démontrer ce résultat. En utilisant l'identité

$$G(z) = (M'+M) \sum_{n=-Mq^j}^{M'q^j-1} \sum_k g_j[kq^j - np^j] z^{kq^j-np^j}$$

et en posant $M_1 = E(N_1 / q^j)$ et $M_2 = E(N_2 / q^j)$, on obtient

$$\sum_{-N_1}^{N_2} \hat{\varphi}_n^j(\nu) e^{2ni\pi\nu} = (M_2+M_1)\frac{q^j}{p^j} G_j\left(e^{-2i\pi\nu/p^j}\right) \hat{\chi}\left(\frac{q^j}{p^j}\nu\right) + \sum_{\substack{-N_1 \le n < -q^j M_1 \\ \text{ou} \\ q^j M_2 \le n \le N_2}} \hat{\varphi}_n^j(\nu) e^{2ni\pi\nu}$$

d'où la première majoration

$$\left| \frac{1}{N_1+N_2+1} \sum_{-N_1}^{N_2} \hat{\varphi}_n^j(\nu) e^{2ni\pi\nu} - \frac{1}{p^j} G_j\left(e^{-2i\pi\nu/p^j}\right) \hat{\chi}\left(\frac{q^j}{p^j}\nu\right) \right| \le \frac{q^j}{N_1+N_2+1} A_j(\nu)$$

A_j ne dépendant ni de N_1 ni de N_2, mais seulement de j.

La deuxième inégalité s'obtient en écrivant la condition de convergence uniforme sur tous les indices n, ce qui donne, dans l'espace de Fourier

$$\left| \hat{\varphi}_n^j(\nu) - \hat{\varphi}_n^r(\nu) \right| \le \varepsilon_j(\nu)$$

où ε_j ne dépend pas de n.

Enfin la dernière majoration est simplement la convergence du produit infini vers φ

$$\left|\frac{1}{p^j}G_j\left(e^{-2i\pi\nu/p^j}\right)\hat{\chi}\left(\frac{q^j}{p^j}\nu\right)-\hat{\varphi}(\nu)\right|\le\eta_j(\nu)$$

Mises bout à bout, ces majorations donnent

$$\left|\frac{1}{N_1+N_2+1}\sum_{-N_1}^{N_2}\hat{\varphi}_n(\nu)e^{2ni\pi\nu}-\hat{\varphi}(\nu)\right|\le\frac{q^j}{N_1+N_2+1}A_j(\nu)+\varepsilon_j(\nu)+\eta_j(\nu)$$

et donc

$$\limsup_{N_1,N_2\to\infty}\left|\frac{1}{N_1+N_2+1}\sum_{-N_1}^{N_2}\hat{\varphi}_n(\nu)e^{2ni\pi\nu}-\hat{\varphi}(\nu)\right|\le\varepsilon_j(\nu)+\eta_j(\nu)$$

On peut maintenant faire tendre j vers l'infini, en observant que ce paramètre n'apparaît pas dans le membre de gauche. En conséquence, la limite supérieure du terme de gauche (qui est positif) est nulle, et donc la limite simple est également nulle, ce qui prouve à la fois l'existence de la limite, et la valeur de celle-ci.

Il faut noter que la contrainte supplémentaire que nous avons imposée, c'est-à dire l'uniformité de la convergence au sens des distributions entraînera également la même uniformité pour les suites dérivées.

Enfin, si la convergence se fait au sens fort donc nécessairement uniformément, il existe une suite discrète qui converge vers cette fonction moyenne.

Théorème IV.5 *Si G_j converge fortement vers les fonctions φ_n, alors la suite de polynômes G'_j définie par*

$$G'_j(z)=\frac{1}{q^j}\frac{z^{q^j}-1}{z-1}G_j(z)$$

converge fortement vers la suite de fonctions $\varphi(x-n)$

Preuve

On utilise un résultat qui sera démontré au chapitre V (lemme V.3) et qui établit que si G_j converge au sens fort vers la série de fonctions φ_n, alors il existe un réel α strictement positif et une constante C tels que $\left|g_j[n+q^j]-g_j[n]\right|\le C\frac{q^{j\alpha}}{p^{j\alpha}}$. Ceci entraîne que $\left|g'_j[n+1]-g'_j[n]\right|\le C\frac{q^{j\alpha}}{p^{j\alpha}}$. Fixons donc un réel x, et soit une suite d'entiers k_j telle que $\left|p^jx-k_j\right|\le A$. Après quelques manipulations, on vérifie qu'il existe une constante C' telle que

$$\left|g'_{j'}[k_{j'}] - g'_j[k_j]\right| \le C' \frac{q^{\alpha \min(j,j')}}{p^{\alpha \min(j,j')}}$$

ce qui montre que la suite $g'_j[k_j]$ est de Cauchy, et donc converge. Cette convergence (ponctuelle, mais dont on montre facilement l'uniformité) se fait vers une quantité qui ne dépend que de x (et non de la suite k_j) et dont on va maintenant déterminer qu'il s'agit de la fonction moyenne définie plus haut.

En effet considérons la suite de fonctions $\varphi'_j(x) = \sum_k g'_j[k]\chi(p^j x - k)$ où χ est la fonction indicatrice de l'intervalle [0 1]. Elle converge uniformément comme on vient de le voir. En passant dans l'espace de Fourier

$$\hat{\varphi}'_j(\nu) = \frac{1}{q^j} \frac{1 - e^{-2i\pi\nu q^j / p^j}}{1 - e^{-2i\pi\nu / p^j}} \frac{G_j\left(e^{-2i\pi\nu / p^j}\right)}{p^j} \hat{\chi}\left(\nu / p^j\right)$$

qui converge ponctuellement (dans l'espace de Fourier) vers $\hat{\varphi}'(\nu)$. Cette convergence est en fait uniforme à cause du support borné des fonctions φ'_j et de leur limite. On en déduit $\lim_{j\to\infty} \varphi'_j = \varphi$. La propriété d'interpolation de c nous permet enfin d'écrire que

$$\lim_{j\to\infty} \left| g'_j[kq^j - np^j] - \varphi\left(k \frac{q^j}{p^j} - n\right) \right| = 0$$

ce qu'il fallait démontrer.

On peut souhaiter calculer la fonction moyenne. Il se trouve que les suites discrètes que l'on vient de décrire et qui convergent ponctuellement vers φ nécessitent le calcul de tous les coefficients du filtre G_j. Il s'agit là d'un nombre qui est trop grand dès que l'on souhaite avoir une bonne précision sur cette fonction.

À la différence du cas dyadique, il n'existe pas de possibilité de calculer [Ri1,DauL2] (par résolution de système linéaire) la valeur exacte de cette fonction en des points donnés, qui permettraient, par application récursive de l'équation de changement d'échelle de déduire d'autres valeurs jusqu'à obtenir un ensemble dense de points.

On peut cependant calculer *exactement* les moments de cette fonction, c'est-à dire les nombres

$$I_n = \int x^n \varphi(x) dx$$

En effet, on sait déjà que I_0=1 et l'on peut facilement vérifier que les nombres I_n sont reliés par l'équation de récurrence

$$I_s = \frac{1}{p^s - q^s} \sum_{k=0}^{s-1} \binom{s}{k} q^k \gamma_{s-k} I_k$$

si l'on pose $\gamma_s = \frac{1}{p}\sum_k k^s g_k$. On verra que du fait qu'ils suivent la même équation de récurrence, ces nombres I_n sont identiques au nombres α_s définis plus bas (théorème IV.6). À partir de ces valeurs, on peut calculer les coefficients du développement de la fonction moyenne sur des polynômes de Legendre. Si en effet on définit les polynômes suivants

$$L_n(z) = A_n \partial^n \left[(b-z)(z-a)\right]^n \tag{IV.15}$$

où $a = \frac{l}{p-q}$ et $b = \frac{L}{p-q}$ sont les extrémités du support de φ, et A_n est tel que $\int_a^b L_n^2 = 1$ (un calcul simple montre qu'alors $A_n^{-2} = \frac{n!^2}{2n+1}(b-a)^{2n+1}$) alors dès que la fonction moyenne sera continue, on pourra écrire

$$\varphi(x) = \sum_{n\geq 0}\left(\int L_n \varphi\right) L_n(x) \tag{IV.16}$$

pour tout x appartenant au support de φ, la série étant absolument convergente. Les polynômes L_n sont des polynômes de Legendre, orthogonaux sur $[a,b]$, c'est-à dire vérifiant $\int_a^b L_n L_{n'} = \delta_{n-n'}$. Bien sûr les produits scalaires entre la fonction φ et les divers polynômes s'obtiennent exactement à l'aide des nombres α_n. En outre, plus la fonction est régulière plus la convergence de la série est rapide.

Une autre manière, aussi efficace, de calcul de cette fonction moyenne est tout simplement le développement en série de Fourier sur le support de φ. En effet si $x\in[a,b]$ alors si φ est continue

$$\varphi(x) = \frac{1}{b-a}\sum_n \hat{\varphi}\left(\frac{n}{b-a}\right) e^{2i\pi\frac{n}{b-a}(x-a)} \tag{IV.17}$$

Là encore, si la régularité de la fonction est grande, cette série converge d'autant plus rapidement. L'intérêt de cette formulation est qu'il suffit d'évaluer la transformée de Fourier de φ en un nombre limité de points (d'autant qu'en général, dans les cas qui nous intéressent, $\hat{\varphi}$ est essentiellement concentré dans l'intervalle fréquentiel [-1/2,1/2]).

4. Condition nécessaire de convergence forte

Comme dans le cas dyadique, la convergence discrète, qui correspond, on l'a vu, à une convergence fonctionnelle ponctuelle et uniforme, va impliquer une certaine factorisation du polynôme générateur [Ri1]. Dans le cas dyadique, il était ainsi nécessaire que $G(-1)=0$ et $G(1)=2$, et sa généralisation est ici

$$\begin{aligned} G\left(e^{2i\pi k/p}\right) &= 0 \quad \text{pour } k = 1\ldots p-1 \\ G(1) &= p \end{aligned} \tag{IV.18}$$

Pour la preuve voir [Blu1].

5. Équation de changement d'échelle

Dans le cas dyadique, les schémas itérés mettaient en évidence une fonction dont la version à une échelle deux fois plus grossière, se déduisait par combinaison linéaire d'elle-même à l'échelle standard —*two-scale difference equation*— [DauL1,DauL2]. On obtient dans le cas rationnel quelque chose d'équivalent

$$\varphi_n(t) = \sum_k g_{kq-np}\varphi_k\left(\frac{p}{q}t\right) \qquad \text{(IV.19)}$$

La preuve est ici immédiate puisque les fonctions φ_n^j vérifient la relation

$$\varphi_n^{j+1}(t) = \sum_k g_{kq-np}\varphi_k^j\left(\frac{p}{q}t\right)$$

ce qui montre (IV.19).

6. Dérivation/Intégration

Supposons que l'on puisse écrire

$$G(z) = \frac{q}{p}\frac{z^p-1}{z^q-1}H(z)$$

ce qui signifie en fait que G admet pour zéros les valeurs $e^{2i\pi k/p}$ pour k=1…p-1, c'est-à dire les sinusoïdes de fréquence multiple de la fréquence d'échantillonnage. Notons φ_n les distributions limites engendrées par le polynôme G. Alors le polynôme H engendre des distributions $\varphi_n^\bullet$ qui sont reliées aux φ_n par la relation

$$\partial\varphi_n = \varphi_n^\bullet - \varphi_{n+1}^\bullet$$

La preuve est donnée dans [Blu1]. Le fait que la multiplication du polynôme générateur par le facteur $R(z) = \frac{q}{p}\frac{z^p-1}{z^q-1}$ équivale à une intégration des fonctions limites nous incitera à dénommer désormais cette fraction "facteur de régularité" qui se réduit dans le cas dyadique au polynôme z+1 [Ri1,DauL2]. On peut alors voir que la fonction moyenne φ associée à φ_n est reliée à la fonction moyenne $\varphi^\bullet$ associée à $\varphi_n^\bullet$ par la relation classique du cas dyadique [Ri1]

$$\partial\varphi(x) = \varphi^\bullet(x) - \varphi^\bullet(x-1)$$

7. Combinaisons linéaires

Les fonctions φ_n introduites plus haut engendrent, ainsi que nous le verrons, un espace vectoriel V_0 à partir duquel on peut bâtir une analyse multirésolution. Si l'on change de base dans V_0, l'analyse ne change bien évidemment pas. Par contre, les nouvelles fonctions de base, que l'on notera f_n, ne vérifieront plus nécessairement une équation de changement d'échelle de la forme (IV.19). Néanmoins, si l'on impose que le passage entre les deux bases se fasse sous forme de filtrage

$$f_n = \sum_k h_{n-k}\varphi_k \tag{IV.20}$$

alors ces nouvelles fonctions obéiront elles aussi à une équation d'échelle du type (IV.19). Le polynôme G' intervenant dans l'équation d'échelle pour les f_n s'écrira alors

$$G'(z) = \frac{H(z^{-p})}{H(z^{-q})}G(z) \tag{IV.21}$$

Preuve

Définissons $\tilde{H}$ par $\tilde{H}(z)H(z) = 1$. On peut alors écrire en inversant l'équation (IV.20) que

$$\varphi_n = \sum_k \tilde{h}_{n-k} f_k$$

que l'on peut insérer dans (IV.19) en ayant également fait apparaître le filtre H

$$f_n(t) = \sum_{k_1} h_{n-k_1} \sum_{k_0} g_{k_0 q - k_1 p} \sum_{k} \tilde{h}_{k_0 - k} f_k\left(\frac{p}{q}t\right)$$

Il est maintenant facile de vérifier (en utilisant la propriété de composition des branches) que l'expression

$$\sum_{k_1} h_{n-k_1} \sum_{k_0} g_{k_0 q - k_1 p} \tilde{h}_{k_0 - k}$$

peut se mettre sous la forme g'_{kq-np}, le filtre G' étant défini par (IV.21).

On n'a cependant aucune assurance que les schémas discrets engendrés par G' soit convergents. Cependant dans les cas où àla fois G et G' sont FIR et engendrent fortement des fonctions limites, on peut être plus précis. Ainsi, on observe que les fonctions limites associées au filtre G' ne seront en général pas égales aux fonctions f_n. Si $H(1)\neq 0$, les fonctions limites φ'_n associées à G' seront reliées à f_n par

$$f_n = H(1)\varphi'_n \tag{IV.22}$$

Par contre, si H s'annule en 1, posons

$$H(z) = (1 - z^{-1})^N H_0(z)$$

alors on observe que les schémas discrets engendrés par le filtre G' ne convergent plus au sens fort, puisque $G'(1)\neq p$. Cependant, le filtre $G''(z)=pG'(z)/G'(1)$ engendre des fonctions limites que l'on notera également φ'_n qui seront reliées aux fonctions f_n par la relation

$$f_n = H_0(1)\partial^N \varphi'_n \tag{IV.23}$$

Preuve

D'après (IV.20) les fonctions f_n sont engendrées par des schémas discrets donnés par les coefficients du filtre $H(z^{-p^j})G_j(z)$ qui égale $H(z^{-q^j})G'_j(z)$ d'après (IV.21). Si H ne s'annule pas en 1, comme $(z^{q^j}-1)G'_j(z)$ tend vers zéro quand j tend vers l'infini on a clairement $H(z^{-p^j})G_j(z) \cong H(1)G'_j(z)$ d'où le résultat.

Si $H(1)=0$, alors $H(z^{-p^j})G_j(z) = \frac{p^{jN}}{q^{jN}} H(z^{-q^j})G''_j(z)$ qui est donc équivalent, quand j tend vers l'infini à $H_0(1)\frac{p^{jN}}{q^{jN}}(1-z^{q^j})^N G''_j(z)$. Les coefficients de ce filtre tendent alors vers $H_0(1)\partial^N \varphi'_n$ s'où le résultat.

8. Sommes remarquables

Les fonctions φ_n engendrées par l'itération de schémas en p/q sont loin d'être quelconques. Dès que la convergence est forte, elles vérifient l'égalité suivante

$$\sum_n \varphi_n = 1$$

qui est le pendant direct de la relation bien connue dans le cas dyadique $\sum_n \varphi(t-n) = 1$. Cette propriété peut être étendue en fonction du nombre de facteurs de régularité du filtre générateur ainsi que le théorème suivant l'indique.

Théorème IV.6 *Supposons que G_j converge au moins au sens des distributions et que G comporte N facteurs de régularité*

$$G(z) = \left(\frac{q}{p}\frac{z^p-1}{z^q-1}\right)^N G^N(z)$$

Alors il existe des constantes α_s pour s=0..N-1 telles que

$$\left.\begin{aligned}\sum_n (x-n)^s \varphi_n(x) &= \alpha_s \\ \sum_n (x-n)^s \varphi(x-n) &= \alpha_s\end{aligned}\right\} \quad \text{pour } s = 0..N-1$$

Preuve

Soit $s \le N-1$ et f une fonction test (indéfiniment dérivable à support compact) alors

$$\begin{aligned}\left\langle (x-n)^s \varphi_n, f\right\rangle &= \left\langle \varphi_n, (x-n)^s f\right\rangle \\ &= \lim_{j\to\infty} \frac{q^j}{p^{j(s+1)}} \sum_k g_j[kq^j - np^j](kq^j - np^j)^s f\left(k\frac{q^j}{p^j}\right)\end{aligned}$$

Du fait que f est à support compact et que φ_n est localisée autour de n, la somme $\sum_n (x-n)^s \varphi_n$ est finie ainsi que la somme sur n du membre de gauche. On peut

donc intervertir sans scrupule l'opérateur de sommation et la limite. D'autre part, G est multiple de $\left(\frac{z^p-1}{z-1}\right)^N$ ce qui implique que G_j est multiple de $\left(\frac{z^{p^j}-1}{z-1}\right)^N$. Il est facile de démontrer que cela équivaut aux égalités suivantes

$$\sum_n (k+np^j)^s g_j[k+np^j] = \frac{1}{p^j}\sum_n n^s g_j[n]$$

pour tout s=0..N-1 et tout k=1..p^j. On aura donc finalement

$$\left\langle \sum_n (x-n)^s \varphi_n, f \right\rangle = \alpha_s \int f$$

où l'on a posé $\alpha_s = \lim_{j\to\infty} p^{-j(s+1)} \sum_n n^s g_j[n]$, c'est-à dire ce que l'on voulait démontrer concernant les fonctions φ_n.

Quant à la fonction φ, on va passer dans l'espace de Fourier. En effet, à l'aide d'une somme de Poisson, on montre facilement que $\sum_n (x-n)^s \varphi(x-n) = (2i\pi)^{-s} \sum_n \partial^s \hat{\varphi}(n) e^{2i\pi nx}$ au sens des distributions. D'autre part, du fait que $G_j\left(e^{-2i\pi\nu/p^j}\right)$ s'annule N fois pour toutes les fréquences entières non multiples de p^j, on en déduit que $\hat{\varphi}$ s'annule N fois pour toutes les fréquences entières non nulles et donc que

$$\sum_n (x-n)^s \varphi(x-n) = (-2i\pi)^{-s} \partial^s \hat{\varphi}(0)$$

Il s'agit maintenant de démontrer l'identité entre le membre de droite de cette équation et α_s. Il suffit de réécrire α_s sous la forme

$$\alpha_s = (-2i\pi)^{-s} \lim_{j\to\infty} p^{-j}\, \partial^s G_j\left(e^{-2i\pi\nu/p^j}\right)\Big|_{\nu=0}$$

et d'utiliser la formule de récurrence (IV.13) itérée d'où

$$(-2i\pi)^{-s} \partial^s \hat{\varphi}(0) = \sum_{k=0}^{s} \binom{s}{k} \frac{q^{jk}}{p^{jk}} \left[(-2i\pi)^{-k} \partial^k \hat{\varphi}(0)\right] \left[(-2i\pi)^{-s+k} p^{-j}\, \partial^{s-k} G_j\left(e^{-2i\pi\nu/p^j}\right)\Big|_{\nu=0}\right]$$

qui donne le résultat attendu en passant à la limite quand j tend vers l'infini.

Propriété IV.7 *Posons $\gamma_s = \frac{1}{p}\sum_n n^s g_n$. Les nombres α_s peuvent être calculés par récurrence à l'aide de la formule suivante*

$$\begin{cases} \alpha_0 = 1 \\ \alpha_s = \dfrac{1}{p^s - q^s} \displaystyle\sum_{k=0}^{s-1} \binom{s}{k} q^k \gamma_{s-k} \alpha_k \end{cases} \qquad (IV.24)$$

pour s=1..N-1.

Preuve

En utilisant (IV.13).

Le théorème IV.6 montre en particulier que si *G* contient *N* ordres de régularité, alors tous les polynômes de degré inférieur ou égal à *N*-1 appartiendront à l'espace $\mathrm{Vect}\{\varphi_n\}_{n\in\mathbb{Z}}$ engendré par les fonctions limites, et dont nous parlerons un peu plus dans la partie sur l'analyse multirésolution. Une autre conséquence de ce théorème sera de montrer que plus les fonctions limites sont régulières, plus on peut les approcher rapidement par des suites discrètes convenablement interpolées (théorème V.4).

9. Moments

Si le facteur $\frac{z^p-1}{z-1}$ conduit à des sommes particulières, le facteur dual $\frac{z^q-1}{z-1}$ conduit à des moments fixés pour toutes les fonctions limites.

Théorème IV.8 *Supposons que les schémas discrets convergent au sens des distributions, alors si G contient N facteurs* $\frac{z^q-1}{z-1}$ *les fonctions limites vérifient*

$$\int (t-n)^s \varphi_n(t) dt = \alpha_s$$

où α_s *est la suite définie plus haut (IV.24).*

Preuve

Si $\left(\frac{z^q-1}{z-1}\right)^N$ divise G, alors $\left(\frac{z^{q^j}-1}{z-1}\right)^N$ divise G_j, ce qui se traduit sur les coefficients du filtre itéré par

$$\sum\nolimits_k (kq^j + n)^s g_j[kq^j + n] = C_{j,s}$$

quel que soit *s*=0…*N*-1, et où $C_{j,s}$ sont des constantes indépendantes de *n* qui varie de 0 à q^j-1□□□□. D'autre part, la convergence faible des schémas discrets nous indique que, pour tout *s*=0…*N*-1

$$\frac{q^j}{p^j} \sum\nolimits_k (k\frac{q^j}{p^j} - n)^s g_j[kq^j - np^j] \xrightarrow[j\to\infty]{} \int (t-n)^s \varphi_n(t) dt$$

qui est donc une quantité indépendante de *n*. En particulier, pour s=0, on obtient

$\int \varphi_n = 1$. On vérifie alors directement sur l'équation de changement d'échelle la valeur exacte de ces moments.

Ce résultat est utile car il montre que l'amnésie des fonctions limites peut être rendue insensible pour certains signaux —ici les polynômes de degré inférieur à N-1—. Bien sûr, cela ne signifie pas que pour d'autres signaux l'amnésie n'ait aucune conséquence sur le processus d'échantillonnage.

10. Valeurs particulières/Interpolation

À la différence du cas dyadique [DauL2,Ri1], on ne peut pas connaître les valeurs exactes prises par les fonctions limites sur un ensemble dense de points. On se souvient en effet que dans le cas dyadique, on accède à la connaissance de la valeur de l'unique fonction limite en les points entiers à l'aide de la résolution d'un système linéaire. On en déduit alors la valeur de la fonction en les points demi-entiers, puis par itération de l'équation de changement d'échelle, en tous les points de la forme $k2^{-n}$ qui est bien sûr dense dans $\mathbf{R}$.

Ceci n'est donc pas possible dans le cas rationnel à cause de la perte de la propriété d'invariance temporelle. Cependant, on peut accéder aux valeurs $\varphi_n(0)$ de façon semblable à celle qui permet dans le cas dyadique d'obtenir les valeurs $\varphi(n)$. On observe simplement que si l'on fait t=0 dans (IV.19) en faisant varier n on écrit un ensemble fermé et fini (donné par le support maximal des fonctions limites) d'équations qui peuvent se traduire sous la forme $\mathbf{A}\Phi=\Phi$. Une dernière équation est disponible qui permet de résoudre ce système linéaire: il s'agit de la somme $\sum_n \varphi_n(0) = 1$ due au fait que l'on suppose la convergence forte, et donc sa conséquence (IV.18).

En analysant plus en détail les raisons pour lesquelles dans le cas dyadique on peut accéder ainsi à la fonction limite sur un nombre dense de valeurs réelles, on constate que ce n'est pas tant l'invariance par translation qui importe mais plutôt l'invariance par changement d'échelle de la suite initiale de points (ici, les nombres entiers).

a. Suites à invariance d'échelle

En effet, le système d'équations est initialement fermé car la multiplication de tout nombre entier par deux est également un nombre entier. Cette propriété peut se généraliser au cas rationnel: on définit alors les nombres réels a_n tels que la multiplication d'un quelconque de ces nombres par p/q fasse encore partie de ces nombres.

Définition IV.9 *Soit une suite $(a_n)_{n\in\mathbf{Z}}$ de réels. Elle sera dite à invariance d'échelle si et seulement si la nouvelle suite $\left(\frac{p}{q}a_n\right)_{n\in\mathbf{Z}}$ est contenue dans $(a_n)_{n\in\mathbf{Z}}$*

En d'autre termes, on pourra définir une fonction d'entiers λ telle que

$$\frac{p}{q}a_n = a_{\lambda(n)} \qquad \text{(IV.25)}$$

On supposera cette suite ordonnée. Afin de ne pas trop s'éloigner du cas dyadique on peut rajouter la contrainte suivante

$$\limsup_n |a_n - n| < \infty \qquad \text{(IV.26)}$$

En fait les deux contraintes (IV.25) et (IV.26) définissent de manière unique la suite de réels a_n. En effet, on déduit de (IV.26)

$$\left|a_{\lambda^j(n)} - \lambda^j(n)\right| \le \text{Constante}$$

pour tout n entier, avec la notation évidente $\lambda^j(n) = \underbrace{\lambda(\lambda(\ldots\lambda(n)))}_{j \text{ fois}}$. On obtient alors

$$\left|a_n - \frac{q^j}{p^j}\lambda^j(n)\right| \le \frac{q^j}{p^j}\,\text{Cte}$$

c'est-à dire $a_n = \lim_{j\to\infty} \frac{q^j}{p^j}\lambda^j(n)$. Bien sûr, toutes les fonctions $\lambda(n)$ ne conviennent pas pour définir cette suite de réels, cependant on peut démontrer que la condition suivante

$$\limsup_n \left|\lambda(n) - \frac{p}{q}n\right| < \infty \qquad \text{(IV.27)}$$

est nécessaire et suffisante pour , d'une part l'existence de la suite a_n, et d'autre part le fait que cette suite vérifie les conditions (IV.25) et (IV.26). Une forme simple pour λ peut être $\lambda(n) = E\left(\frac{p}{q}n\right)$ ou encore les formes λ_0 et λ_1 données par (IV.10)

Une fois une telle suite définie, (IV.19) s'écrit

$$\varphi_n(a_s) = \sum_k g[kq - np]\varphi_k(a_{\lambda(s)})$$

pour s et n entiers. Cet ensemble d'équations est fermé et permet, a priori, avec l'aide des équations $\sum_n \varphi_n(a_s) = 1$ de déterminer les quantités $\varphi_n(a_s)$, puis par application répétée de l'équation de changement d'échelle (IV.19) de déterminer les quantités $\varphi_n(\frac{q^j}{p^j}a_s)$, c'est-à dire, de façon semblable au cas dyadique, de préciser la valeur des fonctions limites sur un ensemble de points dense dans **R**.

Le problème est ici que les équations sont en nombre infini et sans une hypothèse particulière, la solution n'est pas accessible simplement.

b. Fonctions limites d'interpolation

Une hypothèse qui marche consiste à imposer

$$\varphi_n(a_s) = \delta_{n-s}$$

c'est-à dire que les fonctions φ_n sont des fonctions d'interpolations en les points a_s. Pour simplifier l'expression du résultat on suppose en outre que p-q=1. De la condition d'interpolation on déduit

$$g[q\lambda(s) - np] = \delta_{n-s} \tag{IV.28}$$

pour tout n et s. Le cas $\lambda(n) = E\left(\frac{p}{q}n\right)$ conduit alors à la solution

$$G(z) = z^{-q}\,\frac{z^p - 1}{z - 1}\left(1 + (z-1)F(z^p)\right) \tag{IV.29}$$

où F est une série quelconque. Inversement, étant donné un filtre vérifiant (IV.29) c'est-à dire (IV.28), on a par itération de (IV.28) les égalités suivantes

$$g_j[q^j\lambda^j(s) - np^j] = \delta_{n-s}$$

pour tout j, s et n entier. Cela implique, à la limite quand j tend vers l'infini que $\varphi_n(a_s) = \delta_{n-s}$, pourvu que les fonctions soient continues en ces points.

En fait ce n'est pas le cas puisque, pour la fonction d'entiers λ choisie on a $a_0 = a_1 = 0$ ce qui implique la discontinuité des fonctions limites φ_0 et φ_1 en 0. Partout ailleurs cependant, il n'y a pas d'obstacle à la convergence des suites discrètes.

Plutôt que d'imposer une interpolation exacte, on peut imposer une forme d'invariance temporelle

$$\varphi_n(a_s) = \rho_{n-s} \tag{IV.30}$$

où ρ est une suite de réels. Le problème consistant à trouver les filtres FIR G imposant cette nouvelle contrainte revient en fait à celui de l'interpolation stricte, une fois que l'on a remarqué qu'il existe une suite de coefficients h_n tels que

$$f_n = \sum\nolimits_k h_{n-k}\varphi_k \quad \text{et} \quad f_n(a_s) = \delta_{n-s}$$

En effet, ρ est une suite finie puisque G est FIR. On peut donc sans difficulté trouver les coefficients du développement en série de l'inverse du polynôme qui lui est associé et qui définit la suite h_n. On sait d'ailleurs que les fonctions f_n suivent une équation de changement d'échelle, avec pour filtre générateur G' donné par(IV.21). De (IV.29) on déduit la forme générale (pour le choix de λ) du filtre G

$$G(z) = z^{-q}\,\frac{R(z^{-p})}{R(z^{-q})}\,\frac{z^p - 1}{z - 1}\left[1 + (z-1)F(z^p)\right]$$

où R est le polynôme associé à la suite ρ_n. Les fonctions associées ne sont bien sûr pas plus régulières que celles concernant l'interpolation exacte.

c. Exemple des fonctions de Haar généralisées

L'exemple le plus simple de telles fonctions d'interpolation est engendré par le polynôme

$$G(z) = \frac{z^p - 1}{z - 1}$$

dont les schémas discrets convergent vers les fonctions indicatrices des intervalles $[a_n, a_{n+1}]$, les a_n étant engendrés par $\lambda(n) = E\left(\frac{pn+q-1}{q}\right)$.

E. Analyse multirésolution

Cette suite de fonctions engendrées par le filtre passe-bas (aussi bien celui d'analyse que celui de synthèse) constitue très naturellement la base d'une analyse multirésolution telle qu'elle a été définie dans le chapitre I dès que les fonctions sont suffisamment régulières pour être de carré intégrables. En effet, l'équation de changement d'échelle (IV.19) nous assure de l'invariance d'échelle associée à l'espace V_0 défini par les fonctions limites.

Pour voir plus précisément ce qu'il se passe, on va supposer que la convergence vers les fonctions limites se fait au sens fort et l'on va devoir modifier légèrement le banc de filtres: en effet, si le banc de filtres de synthèse est exactement l'inverse du banc de filtres d'analyse, alors les deux filtres passe-bas ne peuvent pas converger simultanément pour de simples raisons de normalisation. En effet, on a alors $\sum_{k=0}^{p-1} G(z)\tilde{G}(ze^{2i\pi\frac{k}{p}}) = pq$ d'après (II.9) d'où $G(1)\tilde{G}(1) = pq \neq p^2$ (la convergence forte a entraîné $\tilde{G}(e^{2i\pi\frac{k}{p}}) = 0$ pour k=1..p-1) comme cela serait nécessaire. On va donc supposer que la composition du banc de filtres d'analyse suivie de celui de synthèse égale p/q fois l'identité.

Une fois ceci posé, les deux filtres passe-bas d'analyse G et de synthèse $\tilde{G}$ vont engendrer deux séries de fonctions φ_n et $\tilde{\varphi}_n$, tandis que les deux passe-haut vont engendrer deux séries de pseudo-ondelettes ψ_n et $\tilde{\psi}_n$, vérifiant les quatre relations

$$\begin{aligned} \varphi_n(t) &= \sum_k g_{kq-np}\varphi_k\left(\frac{p}{q}t\right) \quad \tilde{\varphi}_n(t) = \sum_k \tilde{g}_{kq-np}\tilde{\varphi}_k\left(\frac{p}{q}t\right) \\ \psi_n(t) &= \sum_k h_{k(p-q)-np}\varphi_k\left(\frac{p}{q}t\right) \quad \tilde{\psi}_n(t) = \sum_k \tilde{h}_{k(p-q)-np}\tilde{\varphi}_k\left(\frac{p}{q}t\right) \end{aligned} \qquad \text{(IV.31)}$$

D'autre part, du fait que le banc de filtres construit à partir de $\tilde{G}$ et $\tilde{H}$ est inverse —à p/q près— de celui construit à partir de G et H, on peut démontrer les deux relations suivantes

$$\begin{aligned} \frac{p}{q}\varphi_n\left(\frac{p}{q}t\right) &= \sum_k \tilde{g}_{kp-nq}\varphi_k(t) + \sum_k \tilde{h}_{kp-n(p-q)}\psi_k(t) \\ \frac{p}{q}\tilde{\varphi}_n\left(\frac{p}{q}t\right) &= \sum_k g_{kp-nq}\tilde{\varphi}_k(t) + \sum_k h_{kp-n(p-q)}\tilde{\psi}_k(t) \end{aligned} \qquad \text{(IV.32)}$$

Maintenant si l'on pose

$$V_N = \underset{n\in\mathbf{Z}}{\text{Vect}}\left\{\varphi_n\left(\frac{p^N}{q^N}t\right)\right\}$$
$$W_N = \underset{n\in\mathbf{Z}}{\text{Vect}}\left\{\psi_n\left(\frac{p^N}{q^N}t\right)\right\}$$

alors on peut vérifier, grâce à (IV.31) la condition d'imbrication des espaces V_N puis que $V_N = V_{N-1} + W_{N-1}$ grâce à (IV.32). Le fait que cette somme d'ensemble soit directe, c'et-à dire à intersection nulle résulte des résultats de biorthonormalité entre les quatre fonctions que l'on va démontrer dans laprochaine section.

1. Biorthonormalité

Les fonctions limites associées à un banc de filtres à reconstruction parfaite vérifient, pourvu que la convergence des schémas discrets soit forte les quatre relations de biorthonormalité

$$\begin{aligned}\int \varphi_{-n}(-t)\tilde{\varphi}_{n'}(t)dt &= \delta_{n-n'} \quad \int \varphi_{-n}(-t)\tilde{\psi}_{n'}(t)dt = 0\\ \int \psi_{-n}(-t)\tilde{\psi}_{n'}(t)dt &= \delta_{n-n'} \quad \int \psi_{-n}(-t)\tilde{\varphi}_{n'}(t)dt = 0\end{aligned} \tag{IV.33}$$

Ceci permettra a priori, sous réserve de convergence, de décomposer les fonctions de L^2 sur des séries de pseudo-ondelettes sous la forme

$$f(t) = \sum_{j,n} \psi_n\left(\frac{p^j}{q^j}t\right)\frac{p^j}{q^j}\int f(\tau)\tilde{\psi}_{-n}\left(-\frac{p^j}{q^j}\tau\right)d\tau \tag{IV.34}$$

une relation semblable pouvant être obtenue en intervertissant les positions de ψ_n et $\tilde{\psi}_n$.

2. Régularité

Quand les filtres se correspondent dans un schéma d'analyse-synthèse, si l'un d'eux contient des facteurs de régularité on peut en déduire un certain nombre de conséquences sur le banc de filtres opposé, comme on peut le voir avec le théorème suivant.

Théorème IV.10 *Soient G,H et $\tilde{G},\tilde{H}$ deux bancs de filtres se correspondant dans un schéma d'analyse-synthèse. Si G admet N facteurs de régularité et $\tilde{G}$ $\tilde{N}$, alors*

- *G admet $\tilde{N}$ facteurs* $\frac{z^q-1}{z-1}$
- *$\tilde{G}$ admet N facteurs* $\frac{z^q-1}{z-1}$
- *H admet $\tilde{N}$ facteurs* $z^{p-q}-1$
- *$\tilde{H}$ admet N facteurs* $z^{p-q}-1$
- *les bancs de filtres*

$$\left(\frac{q}{p}\frac{z^p-1}{z^q-1}\right)^{s-N}G(z),(z^{p-q}-1)^{N-s}H(z) \text{ et } \left(\frac{q}{p}\frac{z^p-1}{z^q-1}\right)^{N-s}\tilde{G}(z),(z^{p-q}-1)^{s-N}\tilde{H}(z)$$

sont inverses l'un de l'autre, ceci pour tout s=0..N+N̄

Preuve

Par récurrence. En utilisant les relations (II.9), on démontre facilement qu'un facteur de régularité sur le passe-bas de synthèse entraîne les divisibilités indiquées dans le théorème. Toujours à l'aide de (II.9), on vérifie qu'il est possible de transférer le facteur de régularité $\frac{q}{p}\frac{z^p-1}{z^q-1}$ de $\tilde{G}$ à G.et que cela entraîne un transfert de facteur $z^{p-q}-1$ de H vers $\tilde{H}$. On peut alors recommencer le même raisonnement pour les facteurs de régularité suivants du filtre passe-bas de synthèse.

En particulier, N facteurs de régularité à la synthèse vont entraîner sur les moments des fonctions limites d'analyse les résultats suivants

$$\int t^s \varphi_n = C_s \text{ indépendant de } n$$
$$\int t^s \psi_n = 0$$

pour tout s=0…N-1. Il est parfois très utile d'avoir des moments nuls pour les fonctions mères.

D'un autre côté, on peut affirmer qu'une synthèse ne sera pas régulière —plutôt: ne contiendra pas de facteur de régularité— dès que le filtre d'analyse ne contiendra pas de facteurs $\frac{z^q-1}{z-1}$.

F. Lien Banc de filtres/Ondelettes

On va maintenant voir que la transformée continue sous-jacente au banc de filtres itéré en fraction d'octave n'est pas loin d'une transformée en ondelettes pour l'analyse, et un développement en série d'ondelettes pour la synthèse. En fait on refait, pour les bancs de filtres rationnels, le travail qui avait été réalisé par Mallat [Ma1] pour montrer l'équivalence de l'approche banc de filtres et ondelettes. La différence tient ici essentiellement à l'amnésie des fonctions limites.

1. À l'analyse

Si au lieu d'interpoler le signal d'entrée à l'aide de la fonction de Nyquist nous décidons de l'interpoler à l'aide des fonctions de reconstruction $\tilde{\varphi}_n$, c'est-à dire

$$x(t) = \sum_n x_n \tilde{\varphi}_n(t) \qquad \text{(IV.35)}$$

alors, en supposant que $G(1)=p$ on aura les formules simples suivantes pour les sorties du banc de filtre itéré

$$\begin{aligned} x_j[n] &= \int x(t)\varphi_{-n}\left(-\frac{q^j}{p^j}t\right)dt \\ y_j[n] &= \int x(t)\psi_{-n}\left(-\frac{q^j}{p^j}t\right)dt \end{aligned} \qquad \text{(IV.36)}$$

Supposons un instant que l'amnésie des fonctions limites soit nulle, alors les sorties s'écriraient sous la forme

$$\begin{aligned} x_j[n] &= \int x(t)\varphi\left(n - \frac{q^j}{p^j}t\right)dt \\ y_j[n] &= \int x(t)\psi\left(n\frac{q}{p-q} - \frac{q^j}{p^j}t\right)dt \end{aligned} \qquad \text{(IV.37)}$$

(on peut vérifier que la condition amnésie=0 s'écrit différemment pour le passe-bas et le passe-haut [Blu1]: dans le premier cas l'unité de translation est 1 alors que dans le deuxième cas c'est $q/(p\text{-}q)$) où l'on reconnaît immédiatement pour le passe-haut une transformée en ondelettes. On voit donc l'intérêt de minimiser l'amnésie des fonctions mises en jeu. Il se trouve que si $p\text{-}q=1$, minimiser l'amnésie de la fonction père minimise automatiquement l'amnésie de la fonction mère, ce qui n'est pas le cas pour $p\text{-}q\neq 1$. On considérera donc le plus souvent $p\text{-}q=1$ dans cette thèse, d'autant plus que dans ce cas la conception du filtre passe-haut est simplifiée.

2. À la synthèse

On constate là-encore le même phénomène: si l'on note

$$\begin{aligned} x_j(t) &= \sum_n \frac{q^j}{p^j} x_j[n]\phi_n\left(\frac{q^j}{p^j}t\right) \\ y_j(t) &= \sum_n \frac{q^j}{p^j} y_j[n]\psi_n\left(\frac{q^j}{p^j}t\right) \end{aligned}$$

alors on aura la formule de reconstruction

$$x(t) = x_N(t) + \sum_{j=1}^{N} y_j(t)$$

qui peut aussi se voir comme une série d'ondelettes dès que l'amnésie est suffisamment faible, à laquelle se joint un terme passe-bas correctif. Bien sûr, si la décomposition ou le nombre d'itération était allé jusqu'à l'infini on aurait écrit une formule du type (IV.34).

G. Résumé du chapitre

On a montré ici comment un banc de filtres rationnel itéré peut engendrer comme dans le cas dyadique des fonctions limites, qui en retour permettent d'interpréter de nouvelle façon les sorties du banc de filtres. Un tel résultat était effectivement nouveau au moment où l'article [Blu1] fut publié, même si les analogies avec le cas dyadique ne manquent pas. On a donc décrit en détails les fonctions limites et leurs propriétés, leurs déficiences parmi lesquelles la plus notable —et originale par rapport au cas dyadique— est la non invariance par translation, et leur lien avec le banc de filtres. Ce qui nous a amené à introduire la transformée en pseudo-ondelettes que nous avions déjà évoqué dans de cadre de l'analyse multirésolution définie au chapitre I.

V. Régularité – Amnésie

Ce chapitre est consacré à l'étude précise des fonctions obtenues à l'aide de l'itération d'un schéma en p/q. La partie consacrée à la régularité (essentiellement une extension des notions de continuité et dérivation avec des exposants non entiers) est très étroitement liée aux précédents résultats d'Olivier Rioul dans le cas des schémas dyadiques [Ri1,Ri2]. Il s'agit en effet d'une application des techniques utilisées dans [Ri1] au cas rationnel, développées en parallèle puis en coopération par Olivier Rioul et moi-même [BR,RB]. À l'aide de relations nouvelles, les théorèmes ont même été rendus plus précis et les démonstrations plus globales.

Auparavant, Daubechies et Lagarias dans une série d'articles SIAM [DauL1,DauL2] avaient démontré l'intérêt d'utiliser la régularité au sens de Hölder plutôt que la régularité de Sobolev pour analyser les caractéristiques des fonctions issues des équations de changement d'échelle.

On verra plus tard que l'étude précise de cette régularité conditionnera un bon estimateur de l'erreur de translation, ou amnésie. D'autre part, la régularité (prise cette fois au sens plus restrictif de continuité, dérivation, etc…) est intimement liée au comportement des filtres itérés, ce qui peut justifier son inclusion lors de la conception de filtres.

Dans tout ce chapitre, on considérera des filtres dont les itérations convergent fortement vers un ensemble de fonctions φ_n, c'est-à dire telles que

$$\lim_{j\to\infty} \sup_{n,k} \left| \varphi_n\left(k\frac{q^j}{p^j}\right) - g_j[kq^j - np^j] \right| = 0$$

On sait que cela implique en particulier (IV.18) que $G(z)$ puisse s'écrire sous la forme

$$G(z) = \frac{q}{p}\frac{z^p - 1}{z^q - 1} G^1(z)$$

où G^1 est FIR comme $G(z)$ et vérifie $G^1(1) = p$. Une autre conséquence de cette contrainte est que, si les φ_n sont continues de façon uniforme pour tout n, ce que l'on peut traduire par $\lim_{h\to 0} \sup_n |\varphi_n(x+h) - \varphi_n(x)| = 0$, alors

$$\lim_{j\to\infty} \sup_{k,n} \left| g_j[kq^j - np^j] - g_j[(k-1)q^j - np^j] \right| = 0$$

ce qui, compte tenu du fait que G_j^1 est de longueur proportionnelle à p^j, équivaut à

$$\lim_{j\to\infty} \sup_{k} \frac{q^j}{p^j} \left| g_j^1[k] \right| = 0$$

On va par la suite ne s'intéresser aux propriétés des fonctions φ_n que de manière globale. On sera ainsi à même à définir un ordre de régularité minimal pour toutes les fonctions φ_n, étant entendu que ces fonctions n'auront pas nécessairement toutes le même exposant de Hölder.

Avant d'aborder le chapitre de la convergence, on aura en particulier besoin du lemme technique suivant. Celui-ci nous sera utile aussi bien pour estimer des ordres de régularité, que pour le calcul de l'amnésie, et nous permettra enfin, de lier directement la régularité des fonc-

tions limites à la vitesse de convergence de fonctions interpolantes vers ces mêmes fonctions limites. Son intérêt est montrer l'effet *mécanique* des facteurs de régularité, puis de la régularité elle-même sur l'interpolation des fonctions limites: ce résultat qui s'applique également au cas dyadique n'a pas, à ma connaissance, été encore indiqué dans la littérature.

Lemme V.1 *Supposons que G comporte N facteurs de régularité, et posons* $G(z)=\left(\frac{q}{p}\frac{z^p-1}{z^q-1}\right)^s G^s(z)$ *où s=0..N. Soient également* v_n *une suite de fonctions localisées autour de n vérifiant*

$$\sum_n (x-n)^s v_n(x)=0 \qquad (V.1)$$

pour tout s=0..N-1 et les suites w_n, v_n^s *et* w_n^s *qui s'en déduisent*

$$w_n(x)=\sum_k g_j[kq^j-np^j]v_k\left(\frac{p^j}{q^j}x\right)$$

$$v_n^s=\sum_{k\geq 0}\binom{k+s-1}{k}v_{k+n} \qquad (V.2)$$

$$w_n^s=\sum_{k\geq 0}\binom{k+s-1}{k}w_{k+n} \qquad (V.3)$$

pour s=0..N. On a alors les identités suivantes

$$v_n^{s+1}-v_{n+1}^{s+1}=v_n^s \qquad (V.4)$$

$$w_n^{s+1}-w_{n+1}^{s+1}=w_n^s \qquad (V.5)$$

$$w_n^s(x)=\frac{q^{js}}{p^{js}}\sum_k g_j^s[kq^j-np^j]v_k^s\left(\frac{p^j}{q^j}x\right) \qquad (V.6)$$

En outre les suites de fonctions v_n^s *et* w_n^s *sont localisées autour de n.*

Preuve

Les deux premières égalités se vérifient directement par la définition des suites (V.2) et (V.3). Pour la troisième égalité, on démontre d'abord que $\sum_n (x-n)^s w_n(x)=0$ pour s=0..N-1. Il suffit pour cela d'écrire

$$\begin{aligned}\sum_n (x-n)^s w_n(x) &= \sum_{l=0}^{s}\sum_{k,n}\binom{s}{l}(x-k\tfrac{q^j}{p^j})^l(k\tfrac{q^j}{p^j}-n)^{s-l}g_j[kq^j-np^j]v_k\left(\tfrac{p^j}{q^j}x\right)\\ &= p^{-j(s-l+1)}\sum_{l=0}^{s}\sum_{k,n}\binom{s}{l}n^{s-l}g_j[n](x-k\tfrac{q^j}{p^j})^l v_k\left(\tfrac{p^j}{q^j}x\right)\\ &= 0\end{aligned}$$

où l'on a en particulier utilisé la propriété $\sum_n g_j[kq^j-np^j](kq^j-np^j)^s = p^{-j}\sum_n g_j[n]n^s$, conséquence du fait que G comporte N facteurs de régularité.

On démontre ensuite que les fonctions w_n^s et v_n^s sont localisées autour de n. Prenons le cas de v_n^s. À cause de la propriété de sommation des fonctions v_n, on a

$$\sum_k \frac{(k+s-1)(k+s-2)..(k+1)}{(s-1)!} v_{k+n} = 0$$

Supposons donc que le support de v_n soit contenu dans $[a+n,b+n]$. Alors, si $x \le a+n$ la somme ne contient que des termes nulles. Si $x \ge b+n$, alors la somme pour $k \ge 0$ se transforme en somme pour tout k qui est également nulle comme on vient de le voir. En conséquence le support de v_n^s est contenu dans $[a+n,b+n]$. v_n^s est donc localisée autour de n. De même pour w_n^s une fois qu'on aura remarqué que, par sa définition, w_n est également localisée autour de n.

Après ces préparatifs, on peut démontrer (V.6) par récurrence sur s. La formule étant vraie pour $s=0$, supposons qu'elle le soit pour s donné. Alors grâce aux égalités $v_n^{s+1} - v_{n+1}^{s+1} = v_n^s$, $w_n^{s+1} - w_{n+1}^{s+1} = w_n^s$ et $\frac{q^j}{p^j}\left(g_j^{s+1}[k] - g_j^{s+1}[k-p^j]\right) = g_j^s[k] - g_j^s[k-q^j]$, puis en utilisant le fait que toutes les fonctions sont localisées autour de n on en déduit la formule à l'ordre $s+1$.

A. Convergence

Ainsi qu'on l'a vu au chapitre IV, la convergence forte des suites discrètes $g_j[kq^j - np^j]$ implique une propriété de divisibilité du polynôme $G(z)$, que l'on peut également écrire à l'aide des coefficients du polynôme $\forall k \in \mathbf{Z} \quad \sum_n g[k+np] = 1$. On va donc désormais supposer que G comporte $N \ge 1$ facteurs de régularité, et poser comme dans le lemme V.1

$$G(z) = \left(\frac{q}{p}\frac{z^p-1}{z^q-1}\right)^s G^s(z) \qquad \text{(V.7)}$$

pour $s=0..N$. On écrira aussi

$$G(z) = \sum_{k=l}^{L} g_k z^k \qquad \text{(V.8)}$$

afin de fixer les indices extrêmes du filtre. Enfin, on sera fréquemment amené à considérer les fonctions limites associées aux filtres dérivés (V.7): on les notera φ_n^s tandis que les fonctions moyennes seront désignées par φ^s.

1. Conditions

Bien que nécessaire, la propriété $\forall k \in \mathbf{Z} \quad \sum_n g[k+np] = 1$ est loin d'être suffisante pour assurer la convergence forte des schémas discrets. Cependant si l'on impose en outre aux fonctions limites d'être continues, on obtient alors la nouvelle condition nécessaire suivante

$$\lim_{j\to\infty} \frac{q^j}{p^j} \left| G_j^1 \right|_\infty = 0 \tag{V.9}$$

On va voir qu'en fait cette condition ainsi que (IV.18) sont suffisantes (et nécessaires donc) pour assurer la convergence forte des suites discrètes vers des fonctions continues.

a. Une interpolation privilégiée

Pour cela, plutôt que d'approcher les fonctions limites à l'aide de suites discrètes, on va reprendre la formulation de la convergence au sens des distributions qui définit les fonctions $\varphi_{j,n}$ en introduisant une fonction interpolante à support compact χ

$$\varphi_{j,n}(x) = \sum_k g_j[kq^j - np^j] \chi\left(\frac{p^j}{q^j} x - k\right) \tag{V.10}$$

Si χ est la fonction indicatrice de l'intervalle [0,1[on vérifie que l'on retrouve les fonctions discrètes qui conduisent à $\varphi_n(k\frac{q^j}{p^j}) \cong g_j[kq^j - np^j]$.

Toutes les fonctions χ ne vont pas conduire à une convergence ponctuelle de la série de fonctions, mais suivant le nombre de facteurs de régularité de *G*, on pourra définir une interpolante χ telle que, non seulement la convergence des suites $\varphi_{j,n}$ sera uniforme vers φ_n, mais également celle des dérivées. Cette interpolation nous permettra de déduire de manière très simple les théorèmes estimant la régularité minimale des fonctions limites φ_n.

Reprenant les notations du théorème IV.6, imposons à l'interpolante χ de vérifier les *S* conditions suivantes

$$\sum_n (x-n)^s \chi(x-n) = \alpha_s \tag{V.11}$$

pour *s*=0..*S*-1. Le théorème suivant montre comment construire toutes les interpolantes vérifiant ces sommes.

Théorème V.2 *Définissons les nombres β_n^s par le développement en série*

$$\left(\frac{z}{e^z - 1}\right)^S = \sum_{n \geq s} \frac{\beta_n^S}{n!} z^n$$

alors toute interpolante χ à support compact vérifiant (V.11) pourra s'écrire dans l'espace de Fourier sous la forme

$$\hat{\chi}(\nu) = \left(\frac{1-e^{-2i\pi\nu}}{2i\pi\nu}\right)^S \hat{u}(\nu) \qquad (V.12)$$

où u est une distribution à support compact dont les S premiers moments sont donnés par

$$\int x^s u(x)dx = \sum_{k=0}^{s} \binom{s}{k} \beta_{s-k}^S \alpha_k \qquad (V.13)$$

pour s=0..S-1. Inversement, toute distribution u à support compact vérifiant les conditions (V.13) permettra de construire, à l'aide de (V.12) une interpolante à support compact vérifiant (V.11).

Preuve

Soit donc χ une interpolante vérifiant les sommes (V.11). Comme $\sum_n \chi(x-n) = 1$ on peut définir la distribution à support compact χ^1 par l'équation $\chi^1(x) = \sum_{n\geq 0} \partial\chi(x-n)$, ou de manière implicite $\chi^1(x) - \chi^1(x-1) = \partial\chi(x)$. On vérifie alors que les S sommes (V.11) entraînent que χ^1 vérifie S-1 égalités du type (V.11) avec des constantes différentes. Par dérivation de (V.11) on obtient en effet les S-1 égalités $\sum_n P_s(x-n)\chi^1(x-n) = (s+1)\alpha_s$ pour s=0..S-2, où l'on a posé $P_s(x) = (x+1)^{s+1} - x^{s+1}$. Ces S-1 polynômes engendrant l'ensemble des polynômes de degré inférieur ou égal à S-2 on en déduit que la fonction χ^1 vérifie bien S-1 égalités du type de (V.11).

Ce raisonnement peut aisément être poursuivi par récurrence: on définit alors une suite de distributions à support compact χ^s s=0..S (bien sûr $\chi^0 = \chi$) qui vérifient chacune S-s égalités du type (V.11). L'équation de passage de χ^s à χ^{s+1} est alors $\chi^{s+1}(x) - \chi^{s+1}(x-1) = \partial\chi^s(x)$. On peut alors exprimer simplement χ en fonction de χ^S en passant dans l'espace de Fourier. Il vient en effet

$$\hat{\chi}(\nu) = \left(\frac{1-e^{-2i\pi\nu}}{2i\pi\nu}\right)^S \hat{\chi}^S(\nu)$$

On identifie alors u à χ^S. Comme u est à support compact, sa transformée de Fourier est indéfiniment dérivable; on a d'ailleurs $\partial^S\hat{u}(0) = (-2i\pi)^s \int x^s\chi(x)dx$. D'autre part l'identité suivante

$$\begin{aligned}\sum_n (x-n)^s \chi(x-n) &= (-2i\pi)^s \sum_n \partial^s\hat{\chi}(n) \\ &= (-2i\pi)^s\, \partial^s\hat{\chi}(0) \\ &= \alpha_s\end{aligned}$$

est vraie pour tout s=0..S. On peut alors identifier les coefficients du développe-

ment de McLaurin jusqu'à l'ordre S-1 des deux membres de l'équation $\left(\frac{2i\pi\nu}{1-e^{-2i\pi\nu}}\right)^S \hat{\chi}(\nu) = \hat{u}(\nu)$: on obtient directement (V.13).

Inversement si les équations (V.13) sont vérifiées alors nécessairement les S premières puissances du développement de McLaurin de $\hat{\chi}(\nu)$ sont fixées de manière unique, et comme son expression impose que $\hat{\chi}(n) = 0$ pour tout n non nul, χ vérifie nécessairement les S égalités (V.11). La fonction ainsi construite est bien évidemment à support compact puisque $\left(\frac{1-e^{-2i\pi\nu}}{2i\pi\nu}\right)^S$ est la transformée de Fourier de la fonction B-spline d'ordre S, à support compact.

On pourra en particulier prendre pour u une somme de dérivées de masses de Dirac en zéro, dont la transformée de Fourier serait exactement un polynôme en ν vérifiant les conditions (V.13). L'interpolante ainsi construite sera alors polynômiale par morceaux et discontinue en les nombres entiers. On pourra aussi, si l'on souhaite avoir une fonction suffisamment régulière, choisir pour u une somme de masses de Dirac en les points entiers avec des coefficients tels que les conditions (V.13) soient vérifiées: l'interpolante χ sera alors $\mathcal{C}^{S-1}$.

Revenant à la convergence des suites discrètes, on va maintenant énoncer le lemme suivant qui montre que si convergence forte il y a, alors cette convergence est nécessairement exponentielle

Lemme V.3 *Supposons que G vérifie (IV.18). Les deux propositions suivantes sont équivalentes*

- $\lim_{j\to\infty} \frac{q^j}{p^j}\left|G_j^1\right|_\infty = 0$
- *Il existe une quantité α strictement positive et une constante C, telles que*

$$\left(\frac{q}{p}\right)^j \left|G_j^1\right|_\infty \le C\left(\frac{q}{p}\right)^{j\alpha}$$

Preuve

Il suffit bien sûr de démontrer l'implication dans un seul sens, l'autre étant triviale puisque α est strictement positif.

Si A est la longueur du filtre G^1, alors d'après l'équation de récurrence $G_{j+j_0}^1(z) = G_{j_0}^1(z^{q^j})G_j^1(z^{p^{j_0}})$ on a

$$\left|G_{j+j_0}^1\right|_\infty \le A\left|G_{j_0}^1\right|_\infty \left|G_j^1\right|_\infty$$

Il suffira donc de choisir j_0 tel que $A\left|G_{j_0}^1\right|_\infty \le a < 1$ pour avoir une convergence exponentielle de G_j^1 vers zéro $\left|G_j^1\right|_\infty \le Ca^{\frac{j}{j_0}}$. Une valeur possible de α est par exemple $\alpha = \frac{\log(a)}{j_0 \log(q/p)}$.

Ce résultat étonnant, dû à Olivier Rioul et qui caractérise les schémas itérés va maintenant nous être utile pour démontrer la convergence des suite discrètes vers des fonctions continues.

Théorème V.4 *Les conditions (V.9) et (IV.18) sont nécessaires et suffisantes à la convergence forte des suites discrètes vers des fonctions continues et bornées de façon uniforme* φ_n.

En outre, choisissant l'interpolante définie dans le théorème V.2 pour S=N, si les conditions (V.9) et (IV.18) sont vérifiées, alors

- *il existe une constante* C_N *et un réel positif* α *tels que* $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty \le C_N \frac{q^{j\alpha}}{p^{j\alpha}}$ *tendant vers zéro au moins aussi rapidement que* $\frac{q^j}{p^j}\left|G_j^1\right|_\infty$
- *il existe une constante V telle que*

$$\sup_{x,n}\left|\varphi_{j,n}(x)-\varphi_n(x)\right| \le V\frac{q^{j\alpha}}{p^{j\alpha}} \qquad (V.14)$$

Cette constante V peut être estimée par la formule suivante

$$V = C_N \frac{2^N p^\alpha}{p^\alpha - q^\alpha}\sup_x \sum_k \left|v_k^N(x)\right| \qquad (V.15)$$

Preuve

On vérifie d'abord par récurrence qu'il existe des constantes non nulles $C_1,\dots C_N$ telles que $C_N \frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty \le \dots \le C_2 \frac{q^{2j}}{p^{2j}}\left|G_j^2\right|_\infty \le C_1 \frac{q^j}{p^j}\left|G_j^1\right|_\infty$. Pour cela on utilise le fait que $(z^{p^j}-1)G_j^{s+1}(z) = \frac{p^j}{q^j}(z^{q^j}-1)G_j^s(z)$ et que le degré du filtre G_j^s est inférieur à une constante multipliée par p^j. La convergence de $\frac{q^j}{p^j}\left|G_j^1\right|_\infty$ vers zéro entraîne donc celle de $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$ à une vitesse au moins égale.

Les conditions (V.9) et (IV.18) sont nécessaires comme on l'a vu plus haut. Pour montrer qu'elles sont suffisantes, choisissons comme interpolante une fonction χ continue vérifiant les sommes (V.11) avec *S=N*. Si l'on pose

$$v_n(x) = \chi(x-n) - \sum_k g[kq-np]\chi\left(\frac{p}{q}x-k\right)$$
$$w_n(x) = \varphi_{j+1,n}(x) - \varphi_{j,n}(x)$$

alors ces fonctions vérifient les hypothèses du lemme V.1. Il suffit en effet de vérifier que $\sum_n (x-n)^s v_n(x) = 0$, ce qui se démontre de la même manière que dans le lemme V.1 alors qu'il s'agissait de prouver la relation de somme des fonctions w_n. La différence est ici que les fonctions $\chi(x-n)$ obéissent à (V.11) et non pas (V.1).

On arrive cependant au résultat, grâce au théorème IV.7 qui donne la relation de récurrence vérifiée par les nombres α_s.

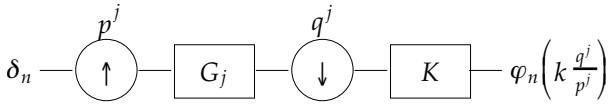
De (V.6) pour $s=N$ on déduit $\left|w_n^N(x)\right| \le C\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$ avec $C = \sup_x \sum_k \left|v_k^N(x)\right|$. Avec l'aide de (V.5), en posant $C' = 2^N C$ on a

$$\left|\varphi_{j+1,n}(x) - \varphi_{j,n}(x)\right| \le C'\frac{q^{j\alpha}}{p^{j\alpha}}$$

Cette différence tend exponentiellement vers zéro: la suite de fonctions $\varphi_{j,n}$ est donc de Cauchy. Elle converge alors uniformément (exponentiellement) vers une fonction au moins continue puisque l'interpolante est elle-même continue. L'erreur entre la fonction limite φ_n ainsi construite et $\varphi_{j,n}$ va alors être majorée par une constante V multipliée par $\frac{q^{j\alpha}}{p^{j\alpha}}$. L'expression précise de cette constante est donc

$$V = C_N \frac{2^N p^\alpha}{p^\alpha - q^\alpha} \sup_x \sum_k \left|v_k^N(x)\right|$$

Ce résultat est très intéressant: on verra en effet que dans les cas qui nous concernent, la vitesse de convergence vers zéro de la quantité $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$ est proportionnelle à $\frac{q^{j\alpha}}{p^{j\alpha}}$ où α est l'ordre de régularité au sens de Hölder des fonctions limites qui peut être supérieur à 1. Une manière de calculer numériquement les fonctions limites peut donc être d'utiliser cette propriété pour limiter le nombre d'itérations. Ainsi, si l'on définit le filtre $K(z)$ par $K(z) = \sum_n \chi(n) z^n$ le schéma suivant nous donnera accès aux valeurs de $\varphi_n\left(k\frac{q^j}{p^j}\right)$ avec une précision de l'ordre de $\frac{q^{j\alpha}}{p^{j\alpha}}$, qui peut évidemment être bien plus grande que si l'on utilise pour χ l'indicatrice de l'intervalle [0 1[



Bien entendu, un tel lien entre efficacité de l'interpolation et régularité est également applicable au cas dyadique: on se souvient que dans ce cas, le filtre $K(z) = \sum_n \varphi(n) z^n$ permet d'obtenir les valeurs *exactes* de la fonction limite aux résolutions 2^{-j} [DauL2,Ri1]. Dans le résultat présent on va cependant plus loin, puisque l'on donne d'autres interpolantes qui permettent d'avoir une convergence liée à la régularité: il semble que de telles interpolantes n'aient jusqu'à présent jamais été décrites dans la littérature, même dans le cas dyadique. Ce résultat va d'ailleurs nous permettre de donner une nouvelle interprétation de la régularité, et donc des indications plus précises sur l'utilité (controversée) de la régularité dans les bancs de filtres.

Le théorème suivant précise la forme de tout polynôme $K(z)$.

Théorème V.5 *Soient les quantités* γ_n^s *définies par*

$$\log^s(1-z) = \sum_{n\geq s} \gamma_n^s z^n$$

et soit $K^0(z)$ *un filtre de degré N-1*

$$\begin{aligned} K^0(z) &= \sum_{n=0}^{N-1} \kappa_n (1-z)^n \\ &= \sum_{n=0}^{N-1} \kappa_n^0 z^n \end{aligned}$$

tel que les coefficients κ_n *soient donnés par*

$$\kappa_n = \sum_{k=0}^{n} \tfrac{1}{k!} \gamma_n^k \alpha_k$$

Supposons enfin que $\frac{q^{jN}}{p^{jN}} \left| G_j^N \right|_\infty$ *tende vers zéro quand j tend vers l'infini. Alors il existe une constante V et un réel* $\alpha>0$ *tels que*

$$\left| \varphi_n\left(k\tfrac{q^j}{p^j}\right) - \sum_{k'} \kappa_{k-k'}^0 g_j[k'q^j - np^j] \right| \leq V \tfrac{q^{j\alpha}}{p^{j\alpha}}$$

D'autre part, tout autre polynôme K issu d'une interpolante χ *et vérifiant (V.11) obéit à l'équation de congruence*

$$K(z) \equiv K^0(z) \left(\mathrm{mod}\,(1-z)^N \right)$$

Preuve

D'après le théorème précédent, si l'interpolante χ vérifie les N identités (V.11) alors la convergence quand j tend vers l'infini des fonctions $\varphi_{j,n}$ vers les fonctions φ_n se fait à une vitesse imposée par $\frac{q^{jN}}{p^{jN}} \left| G_j^N \right|_\infty$, et si on l'évalue aux points $x = k\frac{q^j}{p^j}$ (V.14) devient

$$\left| \varphi_n\left(k\tfrac{q^j}{p^j}\right) - \sum_{k'} \chi(k-k') g_j[k'q^j - np^j] \right| \leq V \tfrac{q^{j\alpha}}{p^{j\alpha}} \qquad \text{(V.16)}$$

On peut choisir $\chi = \varphi$ puisque la fonction moyenne vérifie les sommes (V.11). Intéressons nous plus précisément au polynôme $\Phi(z) = \sum_n \varphi(n) z^n$. En posant $z = e^{-2i\pi\nu}$, il vient $\Phi(e^{-2i\pi\nu}) = \sum_n \hat{\varphi}(\nu+n)$. G contenant N facteurs de régularité, on peut donc écrire $\hat{\varphi}(\nu) = \left(\frac{1-e^{-2i\pi\nu}}{2i\pi\nu}\right)^N u(\nu)$, d'où

$$\Phi(e^{-2i\pi\nu}) = \phi(\nu) + (1-e^{-2i\pi\nu})^N \sum_{n\neq 0} \frac{u(\nu+n)}{\left[2i\pi(\nu+n)\right]^N}$$

que l'on veut évaluer entre -0.5 et +0.5. Comme $\left|(z^{q^j}-1)^N G_j\right|_\infty \le 2^N C_N \frac{q^{j\alpha}}{p^{j\alpha}}$ et donc contribue de la même manière à (V.16) que son second membre, on pourra déduire de Φ tout multiple du polynôme $(z\text{-}1)^N$. Le polynôme résiduel, de degré N-1 sera obtenu en développant Φ en série de McLaurin, c'est-à dire d'après l'expression ci-dessus $\Phi(e^{-2i\pi\nu}) = \sum_{k=0}^{N-1} \frac{\partial^k \phi(0)}{k!} \nu^k + O(\nu^N)$. En posant $z = 1-e^{-2i\pi\nu}$ on est conduit à

$$\Phi(1-z) = \sum_{k=0}^{N-1} \frac{\partial^k \phi(0)}{(-2i\pi)^k k!} \log^k(1-z) + O(z^N)$$

et en réarrangeant les termes (et en se souvenant que $\alpha_k = (-2i\pi)^{-k}\partial^k\phi(0)$), on arrive alors à $\Phi(z) \equiv K^0(z) \left(\mathrm{mod}(1-z)^N\right)$ ce qui suffit à prouver la première partie du théorème.

Si l'on prend maintenant une autre interpolante, alors on montre facilement que (V.11) implique $\partial^s\chi(0) = \partial^s\phi(0)$ pour s=0..N-1. En poursuivant le même raisonnement que plus haut, on tombe donc sur le même polynôme modulo $(1\text{-}z)^N$, ce qu'il fallait démontrer.

B. Régularité

On définit en prélude la définition de régularité d'ordre α d'une fonction [DauL1,Ri1], afin d'en trouver des équivalents sur les suites qui convergent vers ces fonctions.

Définition V.6 *Une fonction f de la variable réelle x est régulière d'ordre α au sens de Hölder au voisinage de x_0 si et seulement si*

- *f admet N dérivées bornées en x_0, où $N=E(\alpha)$*
- $\sup_{-1\le h\le 1} \left|\frac{\partial^N f(x_0+h) - \partial^N f(x_0)}{h^{\alpha-N}}\right| < +\infty$

Si la fonction vérifie cette condition pour tout x_0, on écrira que f est $\mathcal{C}^\alpha$

Il est à noter que, quand on se restreint aux nombres entiers N, les ensembles $\mathcal{C}^N$ contiennent plus de fonctions que les ensembles C^N des fonctions N fois continûment dérivables: il y a là en effet toute la différence entre une fonction continue et une fonction bornée.

1. Ordres de régularité théoriques

La démonstration du théorème V.4 nous fournit un bon point de départ pour estimer la régularité au sens de Hölder d'une fonction limite.

Théorème V.7 *Supposons que les conditions (V.9) et (IV.18) soient vérifiées. Soient* $\alpha>0$ *et une constante* C *tels que*

$$\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty \le C\frac{q^{j\alpha}}{p^{j\alpha}} \qquad (V.17)$$

alors la suite de fonctions φ_n *est au moins régulière d'ordre* α.

Preuve

En utilisant l'interpolation (V.10) avec l'interpolante χ définie par le théorème V.2 et $S \ge \max(\alpha+1, N)$, on peut comme dans le théorème V.4 poser

$$v_n(x) = \chi(x-n) - \sum_k g[kq-np]\chi\left(\frac{p}{q}x-k\right)$$
$$w_n(x) = \varphi_{j+1,n}(x) - \varphi_{j,n}(x)$$

qui vérifient les conditions du lemme V.1. On a donc

$$w_n^N(x) = \frac{q^{jN}}{p^{jN}}\sum_k g_j^N[kq^j - np^j]v_k^N\left(\frac{p^j}{q^j}x\right)$$

Supposons que G_j^N soit majoré comme dans (V.17), alors en posant $N' = -E(-\alpha)-1$ (c'est-à dire le plus grand entier *strictement inférieur* à α) on peut dériver N' fois cette équation pour obtenir la majoration $\left|\partial^{N'}w_n^N(x)\right| \le C\frac{q^{j(N-N')}}{p^{jN-N')}}\left|G_j^N\right|_\infty \le C'\frac{q^{j(\alpha-N')}}{p^{j(\alpha-N')}}$ qui entraîne la convergence uniforme des fonctions continues $\partial^{N'}\varphi_{j,n}$, cette convergence se faisant donc vers les dérivées des fonctions limites $\partial^{N'}\varphi_n$: les fonctions φ_n sont N' fois continûment dérivables. En posant maintenant

$$v_n(x) = \chi(x-n) - \varphi_n(x)$$
$$w_n(x) = \varphi_{j,n}(x) - \varphi_n(x)$$

on a alors

$$\partial^{N'}w_n^N(x) = \frac{q^{j(N-N')}}{p^{j(N-N')}}\sum_k g_j^N[kq^j - np^j]\partial^{N'}v_k^N\left(\frac{p^j}{q^j}x\right)$$

c'est-à dire

$$\left|\partial^{N'}\varphi_{j,n}(x) - \partial^{N'}\varphi_n(x)\right| \le C\frac{q^{j(N-N')}}{p^{j(N-N')}}\left|G_j^N\right|_\infty$$

Pour montrer que φ_n est $\mathcal{C}^\alpha$, on utilise l'inégalité triangulaire qui donne $\left|\partial^{N'}\varphi_n(x+h) - \partial^{N'}\varphi_n(x)\right| \le C'\frac{q^{j(\alpha-N')}}{p^{j(\alpha-N')}} + \left|\partial^{N'}\varphi_{j,n}(x+h) - \partial^{N'}\varphi_{j,n}(x)\right|$. Or $\varphi_{j,n}(x)$ est au

moins $N'+1$ fois dérivable (à cause de notre choix pour S) et on a $\left|\partial^{N'+1}\varphi_{j,n}(x)\right| \le C'' \frac{q^{j(\alpha-N'-1)}}{p^{j(\alpha-N'-1)}}$ ce qui montre que

$$\left|\partial^{N'}\varphi_n(x+h) - \partial^{N'}\varphi_n(x)\right| \le C' \frac{q^{j(\alpha-N')}}{p^{j(\alpha-N')}}\left(1 + C|h|\frac{p^j}{q^j}\right)$$

Le premier membre ne dépendant pas de j, il suffira de prendre j de telle sorte que $|h|\frac{p^j}{q^j}$ soit inférieur à une constante donnée. La fonction $\partial^{N'}\varphi_n(x)$ sera alors automatiquement $\mathcal{C}^{\alpha-N'}$ ce qu'il fallait démontrer.

Ce résultat simplifie celui obtenu par Rioul dans le cas dyadique [Ri1]. En effet, il était apparu une différence de traitement selon que l'estimation de régularité était un nombre entier ou non. Il avait alors été nécessaire de parler de fonctions "presque $\mathcal{C}^\alpha$" dans ce premier cas. Le résultat prouvé ici est que ces fonctions "presque $\mathcal{C}^\alpha$" sont en fait $\mathcal{C}^\alpha$.

À la différence du cas dyadique (voir la condition de "stabilité" définie dans [Ri1]), les estimations de régularité peuvent difficilement être prouvées optimales. Cependant, sous certaines conditions qui vont être exposées dans le théorème suivant, ceci peut être prouvé.

Théorème V.8 *Soit G normalisé à G(1)=p et comportant N facteurs de régularité. Supposons qu'il existe $\tilde{G}$ comportant $\tilde{N}$ facteurs de régularité et tel que G et $\tilde{G}$ vérifient les 4 conditions suivantes*

c1 *G, $\tilde{G}$ forme un couple analyse-synthèse, ce qui se traduit par*

$$\sum_k \tilde{g}[kq - np]\,g[kq - n'p] = \frac{p}{q}\delta_{n-n'}$$

c2 *il existe un entier N_0 tel que*

$$\frac{q^{j(N-N_0)}}{p^{j(N-N_0)}}\left|G_j^N\right|_\infty \underset{j\to\infty}{\to} 0$$
$$\frac{q^{j(\tilde{N}+N_0)}}{p^{j(\tilde{N}+N_0)}}\left|\tilde{G}_j^{\tilde{N}}\right|_\infty \underset{j\to\infty}{\to} 0$$

c3 $\sup_j \left|G_j^N\right|_\infty = +\infty$

c4 $N + \tilde{N} \ge 1$

alors on a les propriétés suivantes

- *les deux filtres convergent au sens des distributions vers des fonctions φ_n et $\tilde{\varphi}_n$*
- *les deux propositions suivantes sont équivalentes*

  i φ_n *est* $\mathcal{C}^\alpha$

i i il existe une constante C telle que $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right| \le C\frac{q^{j\alpha}}{p^{j\alpha}}$

Preuve

Par *c4*, l'un des deux filtres comporte nécessairement au moins un facteur de régularité. Alors en sommant *c1* pour n=0..q-1 et n'=0..q-1, il vient $G(1)\tilde{G}(1) = p^2$ d'où $\tilde{G}(1) = p$.

Observons également que du fait de la convergence forte des schémas issus de $\tilde{G}^{N_0}$ nécessairement $N_0 \ge -\tilde{N}+1$. De même on constate que $N_0 \le N-1$.

Dans ces conditions les filtres $R(z)^{-N_0}G(z)$ et $R(z)^{N_0}\tilde{G}(z)$ engendrent par convergence forte des fonctions limites au moins continues. Ceci entraîne automatiquement la convergence au sens des distributions des schémas dérivés ou intégrés, et donc ceux qui définissent φ_n et $\tilde{\varphi}_n$. D'autre part, d'après le théorème IV.10, les filtres $R(z)^{-N_0}G(z)$ et $R(z)^{N_0}\tilde{G}(z)$ forment également un couple analyse-synthèse, et donc les fonctions limites associées $\varphi_n^{N_0}$ et $\tilde{\varphi}_n^{-N_0}$ sont biorthonormales.

Comme nous avons éventuellement plusieurs choix pour N_0, nous prendrons le plus grand qui permette à la condition *c2* d'être vérifiée. L'estimation de régularité de $\varphi_n^{N_0}$ sera alors $0<\alpha_0\le 1$. Supposons maintenant que $\varphi_n^{N_0}$ soit $\mathcal{C}^\alpha$ avec bien sûr $\alpha\ge\alpha_0$. Deux cas se présentent:

- $\alpha \le 1$, alors en intégrant l'équation (IV.19) itérée contre $\tilde{\varphi}_k^{-N_0}$ on obtient l'égalité suivante

$$\int \varphi_n^{N_0}\left(\frac{q^j}{p^j}x\right)\tilde{\varphi}_k^{-N_0}(x)dx = g_j^{N_0}[kq^j - np^j] \qquad \text{(V.18)}$$

  Remarquons que la convergence forte de $R(z)^{N_0}\tilde{G}(z)$ implique que ce filtre contient encore au moins un facteur $\frac{z^q-1}{z-1}$, d'où $\int \tilde{\varphi}_k^{-N_0} = 1$. On en déduit alors l'égalité

$$\int\left(\varphi_n^{N_0}\left(\frac{q^j}{p^j}x\right) - \varphi_n^{N_0}\left(\frac{q^j}{p^j}k\right)\right)\left(\tilde{\varphi}_k^{-N_0}(x) - \tilde{\varphi}_{k+1}^{-N_0}(x)\right)dx = g_j^{N_0}[kq^j - np^j] - g_j^{N_0}[(k+1)q^j - np^j]$$

  qui implique l'existence d'une constante C telle que $\left|(z^{q^j}-1)G_j^{N_0}\right|_\infty \le C\frac{q^{j\alpha}}{p^{j\alpha}}$ pour tout j. Cette inégalité implique, on le sait, $\frac{q^{j(N-N_0)}}{p^{j(N-N_0)}}\left|G_j^N\right|_\infty \le C\frac{q^{j\alpha}}{p^{j\alpha}}$ qui signifie que l'estimateur de régularité α_0 est supérieur à α, et finalement lui est égal, d'où l'équivalence entre les deux propositions.

- $\alpha > 1$, alors $\varphi_n^{N_0+1}$ est continue de façon uniforme pour tout n tandis que $\tilde{\varphi}_n^{-N_0-1}$ est encore plus régulière. On vérifie d'autre part aisément que ces deux suites de fonctions sont encore biorthonormales. Si $N_0+1<N$ alors

on a toujours $\int \phi_n^{-N_0-1} = 1$ ce qui, à l'aide d'une démonstration identique au premier point, implique que les schémas issus de G^{N_0+1} convergent fortement. On se retrouve donc dans le cas précédent. Par contre si $N_0 + 1 = N$, on a seulement (V.18). On utilise alors l'hypothèse *c3* qui conduit à l'incompatibilité entre la divergence du terme de droite avec le fait que le terme de gauche est uniformément borné. Cette dernière situation est donc impossible.

Ce théorème est très intéressant car il affirme que les estimateurs de régularité extraits des suites discrètes sont optimaux sous certaines conditions. Détaillons un peu plus ces conditions:

- la première nécessite que le filtre G qui nous intéresse puisse être partie d'un système de bancs de filtres d'analyse-synthèse FIR. On sait que ce n'est pas toujours vérifié, en particulier si G peut s'écrire sous la forme $G'(z)H(z^p) / H(z^q)$ où G' et H sont FIR avec $H(1)$=1. Dans ce cas précis, on peut cependant s'en sortir car la régularité des fonctions limites associées à G' est la même que celle de nos fonctions φ_n. Ceci dit, dans un objectif de traitement de signal nous serons amenés à considérer persque exclusivement des bancs de filtres FIR à reconstruction parfaite.
- la seconde impose en pratique que les schémas itérés d'analyse et de synthèse convergent. C'est là encore souhaitable en général puisque cela semble être la propriété minimale pour que les filtres itérés passe-bas G_j et $\tilde{G}_j$ soient à la fois suffisamment sélectifs en fréquence, et pour que leur réponse impulsionnelle ne contienne pas de coefficients "explosifs". Étant donné un filtre G, une méthode pour trouver le filtre $\tilde{G}$ vérifiant cette condition pourra être d'imposer un certain nombre de facteurs $\frac{z^q-1}{z-1}$: il se trouve que cette méthode simple donne d'assez bons résultats. On n'a bien sûr aucun problème si le filtre G est orthonormal (ou lossless, ou paraunitaire), ce qui est le cas que nous considèrerons principalement dans les cas de conception de filtre.
  Il est cependant clair que tous les filtres ne contenant au plus qu'un ordre de régularité et aucun facteur $\frac{z^q-1}{z-1}$, et ceux ne comportant aucun ordre de régularité avec au plus un facteur $\frac{z^q-1}{z-1}$, échapperont aux conditions de ce théorème, et que par conséquent leur régularité ne sera pas prouvée optimale.
- la troisième est une propriété qui semble toujours vérifiée dans les cas qui nous intéressent, c'est-à dire si c1 est vérifiée (on peut trouver des contre-exemples sinon) et si le filtre ne contient plus aucun facteur de régularité. Elle n'est cependant pas démontrée ici, mais est assez facile à mettre en évidence, à l'aide d'une extraction de valeurs propres, comme on le verra dans la partie qui va suivre.
- la quatrième, couplée avec le fait que G(1)=1 nous permet de ne considérer que les schémas convergents.

Pour terminer avec les ordres de régularité théoriques on montre une petite propriété qui indique que dès le plus petit s tel que $\left|G_j^s\right|_\infty$ tend vers l'infini, toutes les estimations issues de $\frac{q^{js'}}{p^{js'}}\left|G_j^{s'}\right|_\infty$ pour $s \le s' \le N$ donnent le même résultat.

Propriété V.9 *Supposons qu'il existe une constante C et un réel positif a tels que* $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty \le C\frac{q^{j\alpha}}{p^{j\alpha}}$ *alors pour tout* $s \le N$ *il existe une constante Cs telle que*

$$\frac{q^{js}}{p^{js}}\left|G_j^s\right|_\infty \le C_s \frac{q^{j\min(\alpha,s)}}{p^{j\min(\alpha,s)}}$$

Preuve

Démontrons ce résultat pour $\alpha \le 1$ en utilisant à nouveau l'interpolante χ qui vérifie (V.11). Choisissons également une fonction continue à support compact qui obéisse aux N relations suivantes

$$\sum_n \int n^s \chi(x-n)\not{\chi}(x)dx = \delta_s \qquad \text{(V.9)}$$

pour s=0..N-1. On peut aisément voir que ces contraintes équivalent à fixer les N premiers moments de $\not{\chi}$ puisque on a alors

$$\sum_{k=0}^{s}\binom{s}{k}(-1)^k \alpha_k \int x^{s-k}\not{\chi}(x)dx = \delta_s$$

pour s=0..N-1. Ceci nous assure que le polynôme $U(z)$ défini par ses coefficients $u_n = \int \chi(x)\not{\chi}(x-n)dx$ peut s'écrire sous la forme $U(z) = 1 + (z-1)^N V(z)$. En effet, on a bien d'une part, $U(1)$=1 (puisque a0=1) et d'autre part $\sum_n n^s u_n = 0$ pour tout s non nul (d'après (V.9)).

Maintenant, en utilisant (V.10pol) on peut écrire

$$\int \not{\chi}(x-k)\varphi_{j,n}\left(\frac{q^j}{p^j}x\right)dx = \sum_{k'} g_j[k'q^j - np^j]u_{k-k'}$$

On reconnaît que le membre de droite diffère de $g_j[kq^j - np^j]$ d'une constante multipliée par $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$ grâce à la divisibilité de $U(z)$-1. De même, on a prouvé dans le théorème V.4 que φ_n et $\varphi_{j,n}$ diffèrent également d'une constante multipliée par $\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$. D'où

$$\left|\int \not{\chi}(x-k)\varphi_n\left(\frac{q^j}{p^j}x\right)dx - g_j[kq^j - np^j]\right| \le C\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty$$

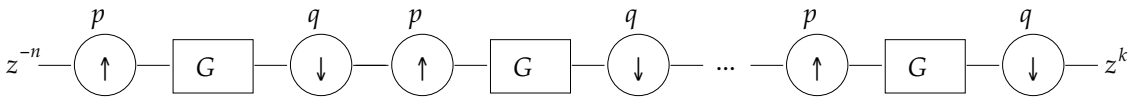
et si α est la valeur estimée de régularité pour φ_n, on en obtient facilement

$$\left|g_j[(k+1)q^j - np^j] - g_j[kq^j - np^j]\right| \le C\frac{q^{j\alpha}}{p^{j\alpha}}$$

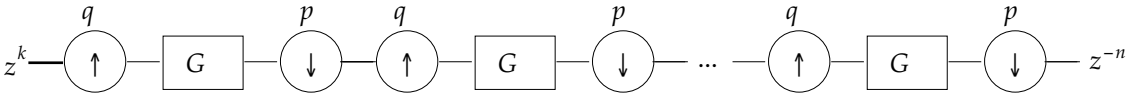
c'est-à dire finalement que $\frac{q^j}{p^j}\left|G_j^1\right|_\infty \le C\frac{q^{j\alpha}}{p^{j\alpha}}$. On déduit alors l'inégalité du théorème par un résultat déjà démontré dans le théorème V.4, à savoir l'existence de constantes C_s telles que $C_N\frac{q^{jN}}{p^{jN}}\left|G_j^N\right|_\infty \le \ldots \le C_2\frac{q^{2j}}{p^{2j}}\left|G_j^2\right|_\infty \le C_1\frac{q^j}{p^j}\left|G_j^1\right|_\infty$. Si α>1 alors on pose $N' = -1 - E(-\alpha)$ qui est donc supérieur ou égal à 1. On peut alors appliquer le résultat du théorème au filtre $G^{N'}$ puisque la régularité estimée est alors $\alpha - N' \le 1$. Les filtres G^s pour $s \le N'$ convergent donc fortement et sont donc bornés. On en déduit le résultat annoncé

2. Un produit de matrices

Le résultat clé de ce chapitre est que l'ensemble des éléments $g_j[kq^j - np^j]$ qui approchent $\varphi_n(kq^j / p^j)$ est accessible à travers un produit de matrices carrées de dimension finie, comme cela se passe dans le cas dyadique [DauL1,DauL2]. En effet on sait que ces quantités résultent de l'itération d'un schéma en p/q de la forme



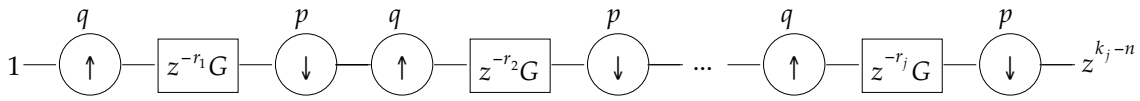
ce qui conduit à un produit itéré de matrices de taille non finie. On peut cependant observer qu'en renversant les taux d'interpolation et d'échantillonnage p et q



on obtient alors les mêmes quantités. Il faut noter que le filtre G peut être considéré comme ayant sa plus petite puissance de z nulle, c'est-à dire l=0: cela n'influe bien évidemment pas sur $\left|G_j\right|_\infty$. De manière plus précise, si l'on définit par récurrence les suites k_n et r_n par

$$\begin{cases} k_0 = k \\ k_{n+1} = E\left(qk_n / p\right) \\ r_{n+1} = -qk_n + pk_{n+1} \end{cases}$$

on peut alors faire migrer k de la gauche du schéma vers la droite, ce qui donne



Étant maintenant donné que le signal d'entrée est une impulsion à l'instant 0, on peut facilement voir que les opérateurs mis bout à bout correspondent à des matrices de taille finie $A \times A$. Il y a en tout p matrices différentes correspondant aux restes r_n. Le calcul de la taille de ces matrices donne

$$A = E\left(\frac{L-q}{p-q}\right) \tag{V.20}$$

si L est le degré du filtre. Ces p matrices que l'on notera $\mathbf{T}_{r=0..p-1}$ auront pour coefficients

$$\mathbf{T}_r\big|_{n,k} = g[np - kq + r + a(p-q)] \tag{V.21}$$

pour n et k compris entre 0 et A. On peut résumer ces observations dans le théorème ci-dessous

Théorème V.10 *Soit G tel que l=0, et* $\mathbf{T_r}$ *les p matrices définies par (V.21), alors pour tout j≥1 et n entiers, soit* $g_j[n] = 0$*, soit il existe une suite d'entiers* r_1*,* r_2*,..* r_j *et k compris entre 0 et p-1 telle que*

$$g_j[n] = e_k^T \mathbf{T}_{r_j} \mathbf{T}_{r_{j-1}} \dots \mathbf{T}_{r_1} e_0$$

où les vecteurs e_i *sont définis par* $e_i = (0,0,..\underset{\overset{\uparrow}{i}}{1},0,..)^T$.

Cet important résultat qui est une généralisation au cas p/q des matrices qui apparaissaient déjà dans le cas dyadique, va nous permettre d'obtenir des majorants et des minorants pour la régularité des fonctions limites. Dans certains cas, on pourra même obtenir l'ordre exact de régularité.

3. Estimateurs de régularité

On va d'abord exposer les résultats que l'on peut obtenir avec les matrices, puis on montrera l'intérêt qu'il peut y avoir à itérer brutalement les schémas discrets. On obtiendra ainsi un algorithme qui généralise celui indiqué par Rioul dans le cas dyadique et dont les propriétés de convergence permettent d'obtenir de bonnes majorations à peu de frais.

a. À partir des matrices

Choisissons une norme vectorielle et désignons la par N: ce peut être par exemple

$$N(u) = \|u\|_\alpha \overset{\Delta}{=} \sqrt[\alpha]{\sum_k |u_k|^\alpha}$$

De cette définition nous induisons la norme matricielle définie par $N(\mathbf{M}) = \sup_{N(u)=1} N(\mathbf{M}u)$. On a alors le résultat suivant qui découle directement du théorème V.11

Théorème V.11 *Les deux quantités suivantes sont équivalentes quand j tend vers l'infini*

$$\sup_{0\le r_1,r_2,\dots r_j\le p-1} N\left(\prod_{k=1}^{j}\mathbf{T}_{r_k}\right) \quad \text{et} \quad \left|G_j\right|_\infty$$

On définira alors la fonction R par $R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1}) = \lim_{j\to\infty}\sup_{0\le r_1,r_2,\dots r_j\le p-1} N\left(\prod_{k=1}^{j}\mathbf{T}_{r_k}\right)^{\frac{1}{j}}$

Ce résultat conduit immédiatement à une borne inférieure sur $R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})$.

Théorème V.12 *Définissons par* $\rho(\mathbf{M})$ *la plus grande valeur propre de la matrice* $\mathbf{M}$. *Fixons un entier n≥1, on a alors la borne inférieure suivante pour* $R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})$

$$\max_{0\le r_1,r_2,\dots r_n\le p-1}\rho(\mathbf{T}_{r_1}\mathbf{T}_{r_2}\dots\mathbf{T}_{r_n})^{\frac{1}{n}} \le R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})$$

Pour la démonstration il suffit de considérer les vecteurs propres associés aux valeurs propres maximales ce qui conduit directement au résultat. On constate en général que cette estimation minimale (correspondant à un ordre de régularité maximal) semble être égale à la valeur optimale dès les plus petites valeurs de *n*: il est parfois possible de le démontrer en utilisant les majorations qui viennent plus loin, cependant aucune démonstration globale n'existe pour l'instant de cette observation empirique, même dans le cas dyadique…

En utilisant une propriété algébrique de la norme induite sur les matrices, c'est-à dire $N(\mathbf{MM}') \le N(\mathbf{M})N(\mathbf{M}')$ on obtient la majoration suivante

Théorème V.13 *Fixons un entier n. On a alors la majoration suivante pour* $R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})$

$$R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1}) \le \max_{0\le r_1,r_2,\dots r_n\le p-1} N(\mathbf{T}_{r_1}\mathbf{T}_{r_2}\dots\mathbf{T}_{r_n})^{\frac{1}{n}}$$

D'autre part, si l'inégalité devient égalité, alors il existe une constante C non nulle telle que

$$\left|G_j\right|_\infty \le C\,R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})^j$$

On utilisera cette formule essentiellement dans le cas où $N = \|\ \|_2$ pour les vecteurs, car alors $N(\mathbf{M}) = \sqrt{\rho(\mathbf{M}^T\mathbf{M})}$. Dans le but de rapprocher les minorants des majorants on peut utiliser la propriété que $R(\mathbf{P}^{-1}\mathbf{T}_0\mathbf{P},\mathbf{P}^{-1}\mathbf{T}_2\mathbf{P},\dots\mathbf{P}^{-1}\mathbf{T}_{p-1}\mathbf{P}) = R(\mathbf{T}_0,\mathbf{T}_2,\dots\mathbf{T}_{p-1})$ pour toute matrice inver-

sible mais le travail dans ce cas reste très artisanal, en particulier quand les filtres sont de grande longueur. Enfin, il y a également la possibilité de passer aux transposées: on constate en effet que $R(\mathbf{T}_0^T, \mathbf{T}_1^T, \dots \mathbf{T}_{p-1}^T) = R(\mathbf{T}_0, \mathbf{T}_2, \dots \mathbf{T}_{p-1})$

b. À partir des itérations

Les itérations donnent un autre moyen d'obtenir simplement une majoration de $R(\mathbf{T}_0, \mathbf{T}_2, \dots \mathbf{T}_{p-1})$: c'est la source de l'algorithme rapide d'estimation d'un ordre de régularité minimal dans le cas dyadique [Ri1]. Tout découle en fait de l'inégalité suivante

$$\left|G_{j+l}\right|_\infty \le \left|G_l\right|_\infty \max_k \sum_{k'} \left|g_j[kq^j - k'p^j]\right|$$

que l'on démontre aisément en écrivant l'équation itérée qui définit les coefficients de G_{j+l}.

Théorème V.14 *On a la majoration suivante pour* $R(\mathbf{T}_0, \mathbf{T}_2, \dots \mathbf{T}_{p-1})$

$$R(\mathbf{T}_0, \mathbf{T}_2, \dots \mathbf{T}_{p-1}) \le \left(\max_k \sum_{k'} \left|g_n[kq^n - k'p^n]\right| \right)^{\frac{1}{n}}$$

où n est un entier quelconque. D'autre part, si l'inégalité devient égalité, alors il existe une constante C non nulle telle que

$$\left|G_j\right|_\infty \le C\, R(\mathbf{T}_0, \mathbf{T}_2, \dots \mathbf{T}_{p-1})^j$$

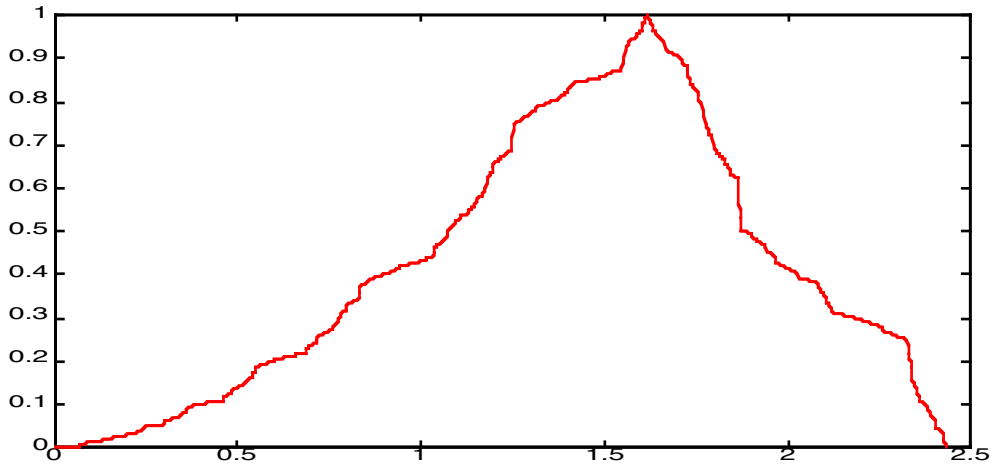
Cette estimation est fréquemment de très bonne qualité. En fait, il est assez facile de voir qu'elle correspond à celle obtenue dans le théorème V.11 pour la norme $\|\ \|_\infty$ pour la matrice transposée (c'est à dire $N(\mathbf{M}) = \sup_{\|u\|_\infty = 1} \left\|\mathbf{M}^T u\right\|_\infty$) à l'aide de la formule d'équivalence donnée dans le théorème V.10 (voir [Ri1] pour le cas dyadique). Son intérêt résulte donc essentiellement dans le fait qu'il existe un algorithme, naturellement rapide pour calculer cette estimation, plutôt que de faire des produits de matrices à tour de bras. Cependant, si p a une valeur importante, comme on devra calculer environ Lp^n termes, la valeur admissible pour n devra être diminuée afin de ne pas consommer trop de temps et donc la précision de l'estimation en sera affectée: c'est là une importante contrainte qui était négligeable dans le cas dyadique et qui devient très lourde dans le cas rationnel. On sera ainsi amené à considérer d'autres estimateurs à base de normes matricielles plutôt qu'à base d'itération des schémas discrets. Il est enfin à noter que tous ces estimateurs tendent, quand n tend vers l'infini, vers la valeur idéale de R.

4. Exemples

Pour fixer les idées, on va donner quelques exemples simples où parfois, la régularité peut être calculée exactement.

a. Exemple n°1

Prenons pour commencer le filtre $G(z) = \frac{1}{3}(1+z+z^2)^2$. Les estimations minimale et maximale (ici la norme $\|\ \|_2$ avec n=4 dans le théorème V.13) concordent pour donner α_0=1, c'est-à dire que la suite de fonctions va être au moins à dérivées bornées, ou $\mathcal{C}^1$. Pour être sûr que l'ordre de régularité effectif est bien 1 (on ne peut pas ici appliquer le théorème V.8) on va vérifier qu'il existe au moins une dérivée de fonction limite qui soit discontinue: cela revient à montrer que l'une des fonctions limites engendrées par $G'(z) = \frac{1}{2}(1+z+z^2)(1+z)$ est discontinue. Supposons que toutes les fonctions f_n associées à ce filtre soit continues, alors comme G' contient encore un facteur de régularité on doit avoir $f_{-1}(0) + f_{-2}(0) = 1$ (toutes les autres fonctions f_n s'annulent en 0). Or si l'on fait un calcul plus précis des supports de ces deux fonctions on constate que 0 est extrémité de support pour chacune et donc, par continuité on doit avoir $f_{-1}(0) = 0$ et $f_{-2}(0) = 0$, d'où la contradiction. La suite de fonctions engendrée par G est donc *exactement* $\mathcal{C}^1$. Si l'on trace les fonctions f_n, on a cependant quelques raisons d'être déçu car, à part f_{-1} et f_{-2}, les autres fonctions sont bien continues. Voici par exemple f_0



La méthode ne permet donc pas de déterminer la régularité maximale d'une fonction limite particulière.

b. Exemple n°2

Rajoutons un degré de régularité avec le filtre $G(z) = \frac{1}{9}(1+z+z^2)^3$. Ici l'ordre de régularité maximal et celui issu de l'algorithme itératif coïncident (en fait dès la première itération on obtient l'optimum), ce qui montre que l'ordre de régularité optimum au sens des suites discrètes est $\alpha_0 \cong 2.71$. En fait, on pourrait facilement démontrer directement d'après la suite discrète itérée que $\left|G_j^3\right|_\infty = \left(\frac{9}{8}\right)^j$ ce qui nous permet de dire $\alpha_0 = 3 - \log(9/8)/\log(3/2)$. Pour démontrer que cette estimation est optimale, il suffit de trouver un filtre de reconstruction correspondant

à G tel que $R(z)^2\tilde{G}(z)$ engendre des fonctions au moins continues (théorème V.8). Après quelques tâtonnements, on trouve que le filtre suivant

$$\tilde{G}(z) = \tfrac{1}{256}(1+z)^6(-9+54z-135z^2+204z^3-261z^4+306z^5-261z^6+204z^7-135z^8+54z^9-9z^{10})$$

vérifie les conditions. En effet la régularité minimale de $R(z)^2\tilde{G}(z)$ (à l'aide de l'algorithme sur 8 itérations) est d'environ 0.65, alors que celle de $R(z)^{-2}G(z)$ est d'environ 0.71. Il est à noter qu'ici l'algorithme, aussi bien pour le calcul de la régularite de $R(z)^{-2}G(z)$ que pour celle de $R(z)^2\tilde{G}(z)$ est meilleur que les estimations issues de la norme $\|\ \|_2$ qui donnait respectivement 0.64 et 0.36.

c. Exemple n°3

Prenons pour G le polynôme donné par les coefficients suivants

n	g_n
0	2.126034814375624e-01
1	6.466859128409039e-01
2	9.514392716433173e-01
3	7.443981487857338e-01
4	1.698106680714083e-01
5	-1.663391902634467e-01
6	-1.405050493102101e-01
7	7.738418411914632e-13
8	3.139649953321556e-02

dans le cas p/q=3/2. Ce filtre a été engendré à l'aide de l'algorithme présenté dans le chapitre suivant. Il comporte un seul facteur de régularité, et est orthonormal (ce qui nous assure de l'optimalité de l'estimation de régularité). La régularité maximale du filtre est d'environ 0.46, alors que la régularité minimale, à l'aide de l'algorithme et de la norme 2 sont respectivement de 0.12 (sur 8 itérations) et 0.11, des quantités assez proches. On peut en conclure que malgré nos efforts la détermination de la régularité "exacte" reste encore peu performante: on peut en effet s'attendre à ce que celle-ci soit assez proche de la régularité maximale calculée.

C. Amnésie

L'amnésie, ou "shift error" est une conséquence du fait que l'on ne peut imposer aux fonctions limites d'être à invariance de translation: $\varphi_n(x+n) \neq \varphi_{n'}(x+n')$ dans le cas général. On définit alors quantitativement l'amnésie ε d'une suite de fonctions à l'aide de sa fonction moyenne par la formule

$$\varepsilon = \sup_{x,n}\left|\varphi_n(x) - \varphi(x-n)\right| \qquad \text{(V.22)}$$

On peut définir de manière alternative une amnésie en utilisant la norme L^2 plutôt que la norme L^∞, sous la forme

$$\eta = \limsup_{N\to\infty} \sqrt{\tfrac{1}{2N}\sum_{n=-N}^{N-1} \left\|\varphi_n(x) - \varphi(x-n)\right\|_2^2} \qquad (V.23)$$

Cela ne pose pas de problème de trouver une majoration de η par ε: on a alors par Cauchy-Schwartz $\eta \le \sqrt{\frac{L-l}{p-q}}\varepsilon$ où *L-l* représente la longueur du support de φ. Par contre, une majoration de ε par η semble être beaucoup plus ardue à trouver: je n'ai pas de solution pour l'instant…

À quoi peut-on comparer ε et η? Ce qui est important n'est en effet pas la valeur brute de ces quantités mais leur valeur relative, par exemple $\varepsilon / \|\varphi\|_\infty$ et $\eta / \|\varphi\|_2$. On peut essayer de calculer les deux normes de φ, mais en fait il suffit de les majorer. On trouve ainsi que $\|\varphi\|_\infty = \limsup_{j\to\infty} \left|\frac{1}{q^j}\frac{z^{q^j}-1}{z-1}G_j\right|_\infty$ qui est lui-même supérieur ou égal à $1/\Lambda$ puisque $G_j(1) = p^j$ ce qui signifie qu'il suffira de comparer $\Lambda\varepsilon$ à 1.

De même, comme $\int\varphi = 1$, on a (par Cauchy-Schwartz) $\|\varphi\|_2 \ge \Lambda^{-1/2}$ et il suffira donc, ici, de comparer $\Lambda^{-1/2}\eta$ à 1. Dans le cas $\|\varphi\|_2 \cong 1$ qui nous concernera le plus souvent (cas des fonctions orthonormées à faible amnésie: voir le théorème V.20) on pourra cependant se permettre de comparer η directement à 1.

Il reste donc à estimer proprement les valeurs de ε et η, et c'est-à dire en fait la valeur de ε, puisque l'on peut utiliser la relation $\sqrt{\Lambda}\,\varepsilon \ge \eta$ pour majorer η. De façon assez étonnante, on peut calculer *exactement* la valeur de η pour tous les filtres possibles, ainsi que celle de ε pour une certaine classe de filtres. Cependant, comme ces expressions exactes ne mettent pas aisément en évidence les dépendances de l'amnésie, il est utile de donner également des majorations de ε.

1. Estimateurs

L'amnésie est en générale difficile d'accès direct car elle impose alors le calcul d'un très grand nombre de termes (il faut estimer toutes les fonctions limites φ_n, et non plus une seule). La version discrète de la définition (V.22) est donnée par le théorème suivant.

Théorème V.15 *Si les suites discrètes convergent fortement, alors on a*

$$\varepsilon = \lim_{j\to\infty}\left|\left(1 - \tfrac{1}{q^j}\tfrac{z^{q^j}-1}{z-1}\right)G_j(z)\right|_\infty \qquad (V.24)$$

En outre, la convergence de $\left(1 - \frac{1}{q^j}\frac{z^{q^j}-1}{z-1}\right)G_j(z)$ *vers* ε *est exponentielle de raison* $\frac{q^\alpha}{p^\alpha}$ *si G est régulier d'ordre* $\alpha<1$ *et de raison* $\frac{q}{p}$ *si* $\alpha\ge1$

La preuve en est immédiate puisque l'on sait que $\frac{1}{q^j}\frac{z^{q^j}-1}{z-1}G_j(z)$ converge vers $\varphi(x)$.

a. Majorations de ε

On va d'abord cerner la valeur de ε. On pourrait essayer d'obtenir des estimations à partir de (V.24), hélas la convergence de l'expression est trop lente pour le nombre d'éléments calculés: en effet le nombre de coefficients de G_j qui doivent tous être calculés est proportionnel à p^j, alors que dans le même temps la convergence des schémas discrets ne peut se faire qu'à une vitesse maximale qui est plafonnée par $\frac{q^j}{p^j}$. Ainsi, pour avoir une précision de l'ordre de h il est nécessaire de calculer un nombre de coefficients de l'ordre de $h^{-\frac{\log p}{\log p/q}}$ ce qui devient très vite exorbitant, d'autant plus lorsque p/q est proche de 1. Bien sûr, si l'on s'arrête à un faible ordre d'itérations (V.24) conduit à une estimation trop grossière pour être exploitée.

i. Cas du produit de deux filtres

On peut tout de même, à partir de (V.24) énoncer un résultat qui montre comment estimer l'amnésie associée à un filtre G à partir de celle associée à l'un des diviseurs de G.

Théorème V.16 *Supposons que G comporte au moins deux facteurs de régularité. Soient G' et G'' tels que $G(z) = \frac{1}{p}G'(z)G''(z)$ et tels que les schémas discrets associés à chaque filtre convergent fortement. Notons L', L'' les longueurs des filtres G', G'' et φ'_n, φ''_n les fonctions limites associées à ces filtres. On a alors les inégalités suivantes*

$$\begin{aligned}\varepsilon(G) &\le \frac{\min(L',L'')}{p-q}\sup_n\left\|\varphi''_n\right\|_\infty \varepsilon(G')\\ &\le \frac{\min(L',L'')}{p-q}\sup_n\left\|\varphi'_n\right\|_\infty \varepsilon(G'')\end{aligned}$$

Preuve

En utilisant le théorème précédent V.15, on a

$$\left|\left(1-\tfrac{1}{q^j}\tfrac{z^{q^j}-1}{z-1}\right)G'_j(z)\tfrac{1}{p^j}G''_j(z)\right|_\infty \le \tfrac{p^j-q^j}{p-q}\max(L',L'')\left|\tfrac{1}{p^j}G''_j(z)\right|_\infty\left|\left(1-\tfrac{1}{q^j}\tfrac{z^{q^j}-1}{z-1}\right)G'_j(z)\right|_\infty$$

qui conduit immédiatement au résultat.

Ce théorème nous sera utile en particulier quand on cherchera à montrer qu'une multiplication suffisante de facteurs $\frac{z^p-1}{z-1}$ rend l'amnésie aussi faible que l'on souhaite. Éventuellement, d'un point de vue pratique, il sera relativement efficace d'estimer l'erreur de translation en isolant les termes $\left(\frac{z^p-1}{z-1}\right)^N\left(\frac{z^q-1}{z-1}\right)^{N'}$ du filtre G, et de calculer l'amnésie correspondant à ces facteurs simples. Il reste encore à calculer l'amnésie pour certains filtres (par exemple, les facteurs de régularité), ce qui va maintenant être développé.

ii. Estimation par les fonctions limites

Une autre façon de calculer l'amnésie ε, plus précise que (V.24) si la fonction est très régulière est d'utiliser une interpolante qui permette d'approcher les fonctions limites à la vitesse définie par leur régularité. C'est ce qui motive le théorème suivant

Théorème V.17 *Choisissons comme interpolante la fonction moyenne φ, alors en utilisant la majoration (V.14) du théorème V.4 l'amnésie des fonctions limites est bornée par les formules*

$$\varepsilon \le \sup_{\substack{x \\ 0\le n\le q^j-1}} \left|\varphi_{j,n}(x+n) - \tfrac{1}{q^j}\sum_{k=0}^{q^j-1}\varphi_{j,k}(x+k)\right| + 2V\frac{q^{j\alpha}}{p^{j\alpha}}$$

$$\varepsilon \le \sup_{\substack{x \\ 0\le n\le q^j-1}} \left|\varphi_{n}(x+n) - \tfrac{1}{q^j}\sum_{k=0}^{q^j-1}\varphi_{k}(x+k)\right| + 4V\frac{q^{j\alpha}}{p^{j\alpha}}$$

(la première fomule fait apparaître des fonctions interpolées, et la seconde des fonctions limites). En outre la constante V sera reliée à la série de constantes ε_j^s calculées plus bas (V.27) par

$$V \le C_N \frac{2^N p^\alpha}{p^\alpha - q^\alpha}(\Lambda - N)\varepsilon_1^N$$

Preuve

Montrons d'abord que $\left|\varphi(x) - \tfrac{1}{q^j}\sum_{k=0}^{q^j-1}\varphi_{j,k}(x+k)\right| \le V\frac{q^{j\alpha}}{p^{j\alpha}}$. On a en effet

$$e_N = \left|\varphi(x) - \tfrac{1}{N}\sum_{k=0}^{N-1}\varphi_{k}(x+k)\right|$$

qui tend vers zéro quand N tend vers l'infini. On en déduit facilement que

$$\left|\varphi(x) - \tfrac{1}{N}\sum_{k=0}^{N-1}\varphi_{j,k}(x+k)\right| \le e_N + V\frac{q^{j\alpha}}{p^{j\alpha}}$$

D'autre part on a la relation d'invariance $\varphi_{j,k+q^j}(x+k+q^j) = \varphi_{j,k}(x+k)$ qui entraîne

$$\lim_{N\to\infty}\tfrac{1}{N}\sum_{k=0}^{N-1}\varphi_{j,k}(x+k) = \tfrac{1}{q^j}\sum_{k=0}^{q^j-1}\varphi_{j,k}(x+k)$$

d'où l'assertion. Les deux inégalités du théorème résultent alors d'une simple inégalité triangulaire.

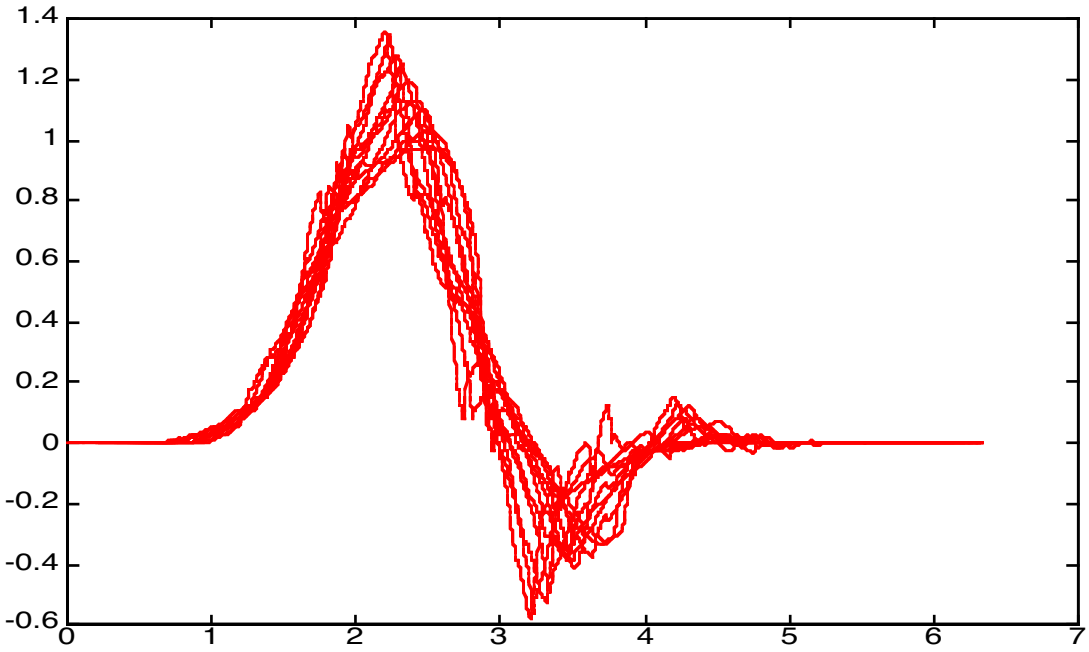
Pour prouver la majoration de la constante V il suffit de revenir à l'expression (V.15). On peut en effet facilement démontrer que $\partial^N v_n^N = \varphi^N(x-n) - \sum_k g^N[kq-np]\varphi^N(\frac{p}{q}x-k)$ ce qui s'écrit fréquentiellement

$$(2i\pi\nu)^N \hat{v}_n^N(\nu) = e^{-2ni\pi\nu}\hat{\varphi}^N(\tfrac{q}{p}\nu)\tfrac{1}{p}\sum_{k=1}^{q-1} G(e^{-2i\pi(\frac{\nu}{p}+\frac{k}{q})})e^{-2i\pi nk\frac{p}{q}}$$

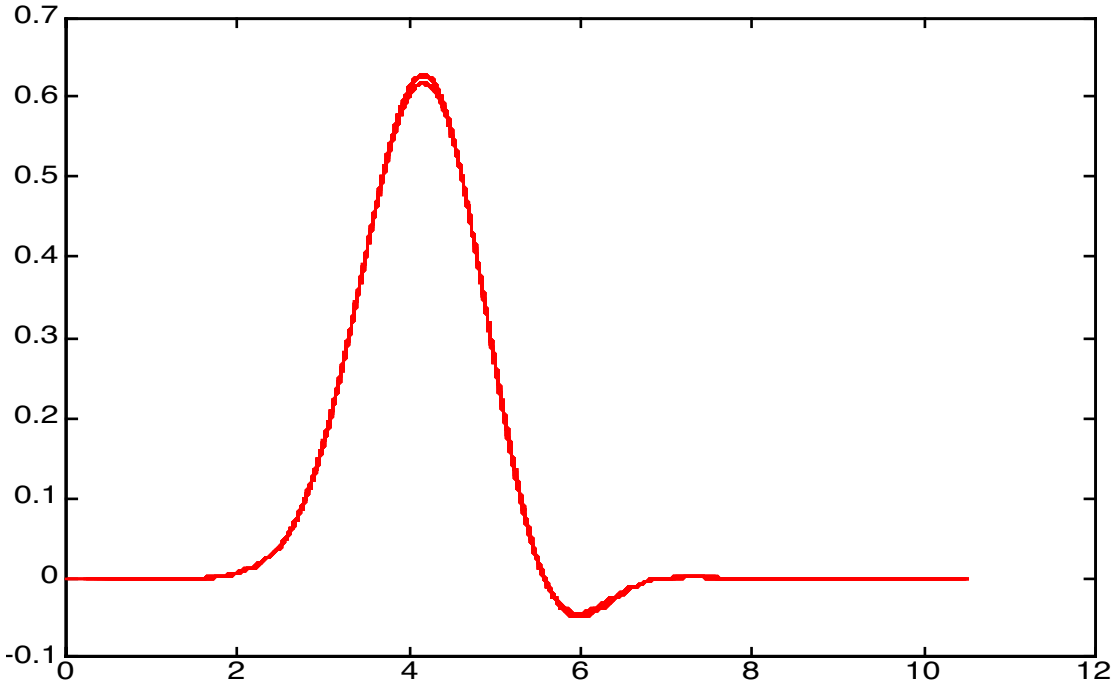
Ce résultat est particulièrement intéressant en ce qu'il justifie que l'on calcule seulement un nombre fini (q^j, d'autant plus petit que la suite de fonctions est plus régulière, et donc converge plus rapidement) de fonctions limites pour estimer directement leur amnésie. C'est par ailleurs l'attitude qui a été retenue ici pour estimer graphiquement l'erreur de translation des fonctions limites.

iii. En tenant compte de la régularité

Le lien entre facteurs de régularité et amnésie apparaît assez facilement. Dans [Blu1] on a ainsi mis en évidence la baisse de l'amnésie d'un ensemble de fonctions après multiplications du filtre générateur par un certain nombre de facteurs $\frac{z^p-1}{z-1}$. Le résultat est moins évident si l'on s'intéresse à la régularité elle-même, bien que l'on puisse donner un exemple de cette influence avec le filtre $G(z) = \frac{1}{16}(z+1)^4(z^2+z+1)(-z+2)$ dont on a ici tracé 11 fonctions limites, ramenées à la même abscisse $\varphi_n(x+n)$ où n=0..10



Si l'on trace les fonctions associées au filtre $G(z) = \frac{1}{3^4}(z^2+z+1)^5(-z+2)$ dont les dérivées quatrièmes ne sont autres que des combinaisons linéaires des fonctions précédentes on obtient une importante réduction de l'amnésie



On aurait d'ailleurs pu vérifier que l'amnésie "visuelle" de chaque série de fonctions correspondant aux filtres $G(z) = \frac{1}{2^n 3^{4-n}}(z+1)^n(z^2+z+1)^{5-n}(-z+2)$ (c'est-à dire en gros à toutes les dérivées jusqu'à l'ordre 4 de de la série de fonctions tracées pour n=0) diminue avec n.

On peut donc soupçonner que la présence de facteurs de régularité dans le filtre générateur permette de simplifier les expressions qui donnent accès à l'amnésie. Ce n'est en fait que partiellement vrai, mais ce lien mérite d'être explicité.

Reprenant les notations du lemme V.1 posons $v_n(x) = \varphi(x-n) - \varphi_n(x)$ et $\varepsilon^s_{j,n} = v^s_n - w^s_n$. On va réintroduire le paramètre j du lemme V.1 qui était muet dans w^s_n. Les estimations seront plus fines quand j sera plus grand, mais on pourra bien sûr se limiter à j=1. On peut vérifier que grâce aux propriétés de récurrence vérifiées par les dérivées, on a

$$\partial^s v^s_n(x) = \varphi^s(x-n) - \varphi^s_n(x) \qquad \text{(V.25)}$$

$$\partial^s \varepsilon^s_{j,n}(x) = \varphi^s(x-n) - \sum_k g^s_j[kq^j - np^j]\varphi^s\left(\frac{p^j}{q^j}x - k\right) \qquad \text{(V.26)}$$

cette dernière équation ne dépendant que d'une seule fonction (et non plus d'une infinité comme dans V.22): elle peut relativement facilement se laisser calculer si l'on remarque qu'il suffit de l'estimer pour n=0..q^j-1 grâce à la relation d'invariance $\partial^s \varepsilon^s_{j,n+q^j}(x) = \partial^s \varepsilon^s_{j,n}(x - q^j)$ où j et fixé bien sûr. Notre but est bien sûr d'estimer la valeur supérieure de $v_n = v^0_n$. Pour cela, on va utiliser directement la convergence forte des schémas discrets. Posons tout d'abord

$$\varepsilon^s_j = \sup_{x,n} \left|\varepsilon^s_{j,n}(x)\right| \qquad \text{(V.27)}$$

Grâce aux relations de récurrence qui régissent $\varepsilon_{j,n}^s$ on a $\varepsilon_j^0 \le 2^s \varepsilon_j^s$ et inversement, comme le support des fonctions $\varepsilon_{j,n}^s$ est $\frac{L-l}{p-q} - s$ (directement d'après (V.26)), on a $\varepsilon_j^s \le \Lambda(\Lambda-1)\ldots(\Lambda-s+1)\varepsilon_j^0$ où $\Lambda = \frac{L-l}{p-q}$ est la taille du support de φ. On peut donc calculer $\varepsilon_{j,n}^s$ à partir d'échantillons de sa transformée de Fourier par la formule

$$\varepsilon_{j,n}^s(x) = \frac{1}{\Lambda-s} \sum_k \hat{\varepsilon}_{j,n}^s\left(\frac{k}{\Lambda-s}\right) e^{2i\pi k \frac{x}{\Lambda-s}} \tag{V.28}$$

pour tout x appartenant au support de $\varepsilon_{j,n}^s$, et dans la mesure où l'on a accès à la transformée de Fourier de $\varepsilon_{j,n}^s$

$$\hat{\varepsilon}_{j,n}^s(\nu) = e^{-2ni\pi\nu} \frac{\hat{\varphi}^s\left(\frac{q^j}{p^j}\nu\right)}{(2i\pi\nu)^s} \left[\sum_{k=1}^{q^j-1} \frac{1}{p^j} G_j^s\left(e^{-2i\pi\left(\frac{\nu}{p^j}+\frac{k}{q^j}\right)}\right) e^{-2i\pi k \frac{p^j}{q^j}} \right] \tag{V.29}$$

(noter que la somme entre [] ne compte pas k=0). On va maintenant exprimer l'amnésie en fonction de ces quantités ε_j^s.

Lemme V.18 *Supposons que* $\frac{q^{js}}{p^{js}} \left| G_j^s \right|_\infty \le C_s \frac{q^{j\alpha_s}}{p^{j\alpha_s}}$ *où* $\alpha_s > 0$ *puisque les schémas discrets issus de G convergent fortement. On a alors l'inégalité suivante*

$$\left| w_n^s(x) \right| = \left| v_n^s(x) - \varepsilon_{j,n}^s(x) \right| \le C_s \frac{q^{j\alpha_s}}{p^{j\alpha_s} - q^{j\alpha_s}} (\Lambda - s) \varepsilon_j^s$$

Preuve

On utilise pour cela le lemme V.1. On peut plus précisément inverser l'équation déduite de (V.6) qui donne $\varepsilon_{j,n}^s$ en fonction de v_n^s sous la forme

$$v_n^s(x) = \sum_{l \ge 0} \sum_k \frac{q^{ljs}}{p^{ljs}} g_{lj}^s [kq^{lj} - np^{lj}] \varepsilon_{j,k}^s \left(\frac{p^{lj}}{q^{lj}} x \right)$$

ce qui, en incluant la valeur du support de $\varepsilon_{j,n}^s$ (tempéré par le fait que la continuité de cette fonction implique qu'elle s'annule aux bornes de son support) conduit immédiatement à la majoration annoncée.

Nous pouvons maintenant donner un majorant de l'erreur de translation.

Théorème V.19 *En utilisant les notations du lemme V.18 si les suites discrètes engendrées par* G_j *convergent fortement alors l'amnésie des fonctions limites est majorée de la façon suivante*

$$\varepsilon \le \varepsilon_j^0 + \min_{1 \le s \le N} \left[C_s \frac{2^s q^{j\alpha_s}}{p^{j\alpha_s} - q^{j\alpha_s}} (\Lambda - s) \varepsilon_j^s \right]$$

Preuve

On utilise le lemme V.18 ainsi que l'équation de récurrence de w_n^s pour obtenir, quel que soit s=1..N

$$\varepsilon \le \varepsilon_j^0 + C_s \frac{2^s q^{j\alpha_s}}{p^{j\alpha_s} - q^{j\alpha_s}} (\Lambda - s)\varepsilon_j^s$$

et on a bien sûr identifié ε à $\sup_{n,x} \left| u_n^0(x) \right|$.

Dans ce théorème, on peut en particulier choisir N=1 et comme $\varepsilon_j^1 \le \Lambda \varepsilon_j^0$ on obtient une majoration qui ne dépend plus que de ε_j^0

$$\varepsilon \le \left[1 + C_1 \frac{2q^{j\alpha_1}}{p^{j\alpha_1} - q^{j\alpha_1}} \Lambda^2 \right] \varepsilon_j^0 \qquad \text{(V.30)}$$

La valeur de ce type d'estimateur tient surtout par la possibilité de démontrer des résultats généraux de convergence mettant en évidence les dépendances de l'amnésie. On peut cependant en retirer des valeurs assez précises si l'on fait j>1. On voit en effet que $\lim_{j\to\infty} \varepsilon_j^0 = \varepsilon$ et que la majoration du théorème V.19 quantifie l'erreur entre ε et ε_j^0: cette erreur tend vers zéro quand j tend vers l'infini, d'autant plus rapidement que les fonctions limites sont plus régulières.

Bien sûr, quand j augmente, la complexité du calcul de ε_j^s en est augmentée. Ceci justifie que l'on s'intéresse à obtenir des résultat exacts, indépendants du nombre d'itérations: c'est ce que l'on va voir dans la section suivante.

b. Calculs exacts

On peut trouver une formulation simple, à l'aide du polynôme itéré G_j, de l'amnésie au sens L^2 η.

LemmeV.20 *Supposons que les schémas discrets convergent au sens fort, alors on a*

$$\eta^2 = \limsup_{j\to\infty} \frac{1}{p^j} \int_0^1 \left| G_j(e^{-2i\pi\nu}) \right|^2 d\nu - \|\varphi\|_2^2 \qquad \textit{(V.31)}$$

Preuve

On a l'identité

$$\frac{1}{N} \sum_{n=0}^{N-1} \left\| \varphi(x) - \varphi_n(x+n) \right\|_2^2 = \frac{1}{N} \sum_{n=0}^{N-1} \|\varphi_n\|_2^2 - \|\varphi\|_2^2$$

et si les schémas discrets convergent fortement vers les fonctions φ_n, on peut alors affirmer que

$$\|\varphi_n\|_2^2 = \frac{q^j}{p^j}\sum_k \left|g_j[kq^j - np^j]\right|^2 + O(\frac{q^{j\alpha}}{p^{j\alpha}})$$

où α est le paramètre indiquant la vitesse de convergence des suites discrètes. On en déduit alors que

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1}\|\varphi(x) - \varphi_n(x+n)\|_2^2 = \frac{1}{p^j}\sum_k \left|g_j[k]\right|^2 - \|\varphi\|_2^2 + O(\frac{q^{j\alpha}}{p^{j\alpha}})$$

On pourrait d'ailleurs voir que la limite inférieure suit une relation semblable. À la limite (indépendante de N) quand j tend vers l'infini, on obtient le résultat désiré, compte tenu de l'identité $\sum_k \left|g_j[k]\right|^2 = \int_0^1 \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu$.

Ce petit résultat va maintenant nous être utile pour montrer comment calculer exactement l'amnésie L^2.

Théorème V.21 *Supposons que les schémas discrets convergent au sens fort, alors posant* $\Gamma(z) = \frac{1}{p}G(z)G(z^{-1})$ *les suites discrètes issues de ce nouveau filtre convergent vers des fonctions limites* Φ_n *dont on notera la moyenne par* Φ. *On a alors la relation simple*

$$\eta^2 = \Phi_0(0) - \Phi(0) \qquad (V.32)$$

Preuve

Bien sûr la convergence ne pose aucun problème puisque $\left|\Gamma_j^1\right|_\infty \le C\left|G_j\right|_\infty\left|G_j^1\right|_\infty$. Il suffit alors de reprendre (V.31) et de noter que, d'une part $\left|\hat{\varphi}(\nu)\right|^2 = \hat{\Phi}(\nu)$ d'où

$$\|\varphi\|_2^2 = \int\left|\hat{\varphi}(\nu)\right|^2 d\nu = \int\hat{\Phi}(\nu)d\nu = \Phi(0)$$

et d'autre part

$$\int_0^1\left|G_j(e^{-2i\pi\nu})\right|^2 d\nu = \int_0^1\Gamma_j(e^{-2i\pi\nu})d\nu = \gamma_j[0]$$

qui tend, quand j tend vers l'infini vers $\Phi_0(0)$.

Dans la pratique $\Phi_0(0)$ ne pose aucun problème de calcul puisque l'on peut y accéder par résolution d'un système linéaire ainsi que cela a été exposé dans le chapitre précédent sur les valeurs particulières des fonctions limites. Pour calculer $\Phi(0)$ on écrit l'identité

$$\Phi(0) = \|\varphi\|_2^2 = \frac{1}{\Lambda}\sum_n\left|\hat{\varphi}\left(\frac{n}{\Lambda}\right)\right|^2$$

où Λ est la taille du support de φ. Cette expression est alors aisément calculable puisque

chaque terme de la somme est le résultat d'un produit infini à convergence exponentielle. Si l'on ajoute qu'en général la fonction φ est très sélective en fréquence autour de [-1/2,1/2] on voit qu'il suffit de calculer un nombre de termes de l'ordre de Λ. Enfin, quand on tronque cette somme, le résultat obtenu est une majoration de η ce qui est bien suffisant en général.

Dans le cas particulièrement intéressant où le filtre G est paraunitaire la formule (V.32) devient encore plus simple

$$\eta^2 = 1 - \|\varphi\|_2^2 \qquad \text{(V.33)}$$

puisque l'on a, pour toute valeur de n $\|\varphi_n\|_2^2 = 1$.

On peut également dans certains cas, accéder à la valeur exacte de ε. Ce résultat ne présente pas l'universalité de celui pour η, mais présente tout de même l'intérêt d'être appliqué à un grand nombre de cas non triviaux.

Théorème V.22 *Supposons que G soit symétrique et puisse s'écrire sous la forme* $G(z) = z^N P(z)\overline{P}(z^{-1})$. *Supposons également que p-q=1 et que les schémas discrets issus de G convergent au sens fort, alors on a*

$$\varepsilon = \varphi_{-N}(0) - \varphi(N) \qquad \textit{(V.34)}$$

Preuve

Remarquons tout d'abord que $\hat{\varphi}$ est de module intégrable puisque l'on a $\varphi(N) = \int |\hat{\varphi}(\nu)| d\nu$, ce qui nous permettra d'affirmer que la suite de fonctions φ_{ν_0} définies par

$$\varphi_{\nu_0}(x) = \int_{-\frac{\nu_0}{2}}^{\frac{\nu_0}{2}} \hat{\varphi}(\nu) e^{2i\pi\nu x} d\nu$$

converge uniformément vers $\varphi(\nu)$ quand ν_0 tend vers l'infini. Par ailleurs, on a la limite suivante

$$\int_{-\frac{\nu_0}{2p^j}}^{\frac{\nu_0}{2p^j}} G_j(e^{-2i\pi\nu}) e^{2i\pi p^j \nu x} d\nu \xrightarrow[j\to\infty]{} \int_{-\frac{\nu_0}{2}}^{\frac{\nu_0}{2}} \hat{\varphi}(\nu) e^{2i\pi\nu x} d\nu \qquad \text{(V.35)}$$

Pour cela il suffit de se rappeler que le produit infini (IV.14) qui définit la transformée de Fourier de φ est uniformément convergent sur tout intervalle de la forme $[-\nu_0/2, \nu_0/2]$, d'où la convergence, uniformément pour tout x, de l'expression ci-dessus.

Intéressons nous donc à la quantité ε_{ν_0} définie par

$$\varepsilon_{\nu_0} = \sup_{n,x} \left| \varphi_n(x) - \varphi_{\nu_0}(x-n) \right|$$

Cette suite de nombres converge vers ε quand ν_0 tend vers l'infini du fait de la convergence uniforme de φ_{ν_0} vers φ.

Grâce à (V.35), on peut écrire ε_{ν_0} comme une limite sur j sous la forme

$$\varepsilon_{\nu_0} = \lim_{j\to\infty} \sup_{k,n} \left| \int_{\frac{\nu_0}{2p^j}}^{1-\frac{\nu_0}{2p^j}} G_j(e^{-2i\pi\nu}) e^{2i\pi\nu(kq^j - np^j)} d\nu \right|$$

en utilisant également l'hypothèse de convergence forte des suites discrètes. La forme particulière de G nous permet de réécrire cette formule sous la forme

$$\begin{aligned}
\varepsilon_{\nu_0} &= \lim_{j\to\infty} \int_{\frac{\nu_0}{2p^j}}^{1-\frac{\nu_0}{2p^j}} \left| P_j(e^{-2i\pi\nu}) \right|^2 d\nu \\
&= \lim_{j\to\infty} \int_{\frac{\nu_0}{2p^j}}^{1-\frac{\nu_0}{2p^j}} G_j(e^{-2i\pi\nu}) e^{2i\pi\nu N(p^j - q^j)} d\nu \\
&= \varphi_{-N}(0) - \varphi_{\nu_0}(N)
\end{aligned}$$

On en déduit alors immédiatement le résultat.

Il peut sembler que la contrainte de forme de G est trop sévère. Cependant un certain nombre de polynômes la vérifient, comme par exemple

$$G(z) = p\left(\frac{1}{p}\frac{z^p - 1}{z - 1}\right)^{2M} \left(\frac{1}{q}\frac{z^q - 1}{z - 1}\right)^{2M'}$$

quelles que soient les valeurs de M, M', p et q. On peut alors, en combinant le théorème sur l'amnésie d'un produit de polynômes et le résultat (V.34) obtenir une estimation de l'amnésie au sens L^∞ du polynôme total.

2. Dépendances de l'amnésie

Deux propriétés qui sont partiellement liées ont une influence bénéfique sur la réduction de l'amnésie: la sélectivité de l'ondelette limite et la régularité. La sélectivité est une notion floue qui prend en compte à la fois l'atténuation dans la bande atténuée et la largeur de la bande de transition, sachant que dans la bande passante le filtre est censé être approximativement constant. Ainsi, à même atténuation, un filtre sera plus sélectif qu'un autre si sa bande de transition est plus étroite. De même, à bande de transition égale, un filtre sera plus sélectif qu'un autre si son atténuation est supérieure. Par contre, on pourra difficilement comparer des filtres d'atténuation et de bande de transition différentes (sauf cas trivial…). Cependant, on

considère souvent qu'une atténuation de 30 à 40 dB est nécessaire pour un système de traitement de signal efficace, et qu'au-delà de 40 dB, la valeur de l'atténuation n'a plus beaucoup d'importance: la bande de transition devient alors primordiale. On peut alors définir la sélectivité de la façon suivante: avant 35 dB d'atténuation, c'est l'atténuation seule qui compte, et après c'est la bande de transition.

La propriété de régularité et celle de sélectivité se recouvrent partiellement puisqu'en général, une plus grande régularité correspond à une plus forte sélectivité. Cela vient simplement du fait qu'une fonction limite de régularité α vérifie en fréquence une inégalité de la forme

$$\left|\nu^{\alpha}\hat{\varphi}(\nu)\right| \le C$$

et donc tend d'autant plus vite vers zéro quand ν tend vers l'infini, ce qui tend à réduire la bande de transition.

Cependant, comme on le verra plus loin, aucune de ces deux propriétés seule n'est suffisante pour faire baisser l'amnésie.

a. Sélectivité

Supposons que $\hat{\varphi}$ soit de régularité supérieure à 1 (et donc de module intégrable), alors en reprenant (V.29) on peut écrire

$$\varepsilon_1^0 \le \frac{q-1}{q}\int\left|G(e^{-2i\pi\frac{\nu+k}{q}})\right|\left|\hat{\varphi}(\nu)\right|d\nu(x)$$

ε_1^0 est lié à ε par la majoration (V.30). On constate alors que si G (passe-bas) est suffisamment sélectif en fréquence, alors φ sera également sélective en fréquence et ε_1^0 pourra être rendu aussi petit que l'on souhaite. En effet, si G est très sélectif autour de $[-a,a]$, alors $\hat{\varphi}$ sera très sélectif autour de $[-ap,ap]$ (d'après l'équation fréquentielle de $\hat{\varphi}$). Plus précisément, on peut assez facilement voir que si

$$a \le \frac{1}{p+q}$$

alors pour n donné soit $G(e^{-2i\pi\frac{\nu+k}{q}})$ soit $\hat{\varphi}(\nu)$ sera dans sa bande atténuée. À la limite des filtres idéaux, on aurait ainsi $\varepsilon_1^0=0$. On peut donc soupçonner que la sélectivité du filtre (ajouté au fait que les fonctions limites sont au moins continûment dérivables) permet de rendre l'amnésie plus faible.

On va le voir sur un exemple. Soit donc

$$G(z) = \frac{1}{p}\left(\frac{z^p-1}{z-1}\right)^2 G'(z)^S$$

où le filtre passe-bas G' vérifie $G'(1)=1$, a tous ses coefficients positifs et est de longueur finie

L'. Alors plus S est grand, plus le filtre G est sélectif. Plus précisément, si $\sup_{1/(p+q)\le\nu\le 1/2}\left|G'(e^{-2i\pi\nu})\right| \le A < 1$, le filtre G est de longueur $2p+L'S-2$ et a une atténuation de pA^S sur la bande indiquée. D'autre part, comme on a imposé la positivité des coefficients de G', on peut facilement montrer que

$$\left|G_j^1\right|_\infty \le 1$$

et donc que φ est au moins $\mathcal{C}^1$ avec pour constante C=1. En outre $\left|G'(e^{-2i\pi\nu})\right| \le 1$ ce qui implique $\left|\hat{\varphi}(\nu)\right| \le \frac{|\sin(\pi\nu)|^2}{|\pi\nu|^2}$ qui est donc de module intégrable. On peut alors majorer ε_1^0 à l'aide de (V.29) par

$$\begin{aligned}\varepsilon_1^0 &\le p\frac{q-1}{q}\max_{1\le k\le q-1}\int\left|G'(e^{-2i\pi\frac{\nu+k}{q}})G'(e^{-2i\pi\frac{\nu}{p}})\right|^S\left|\frac{\sin(\pi\nu)}{\pi\nu}\right|^2 d\nu\\ &\le p\frac{q-1}{q}A^S\end{aligned}$$

De cette majoration, on déduit

$$\Lambda\varepsilon \le MS^3A^S \qquad \text{(V.36)}$$

où M est une constante ne dépendant que de L', de p et de q (la quantité Λ est celle définie au début de la partie "Amnésie" et indique la taille du support de φ). On constate immédiatement la décroissance au moins exponentielle de cette quantité vers zéro quand S tend vers l'infini, ce qui montre que des filtres suffisamment sélectifs peuvent minimiser l'amnésie.

La situation est cependant plus complexe, comme on peut le voir sur le graphique suivant

On a calculé à l'aide de l'algorithme exposé dans le chapitre VI des filtres orthogonaux à atténuation "maximale" (en tout cas, proche de l'optimum). La longueur des filtres ainsi que le paramètre d'échelle ont été fixés à 20 et 3/2 respectivement, tandis qu'on imposait un facteur de régularité. Le seul paramètre qui a été retenu comme variable est la fréquence début de la bande atténuée. On constate alors que l'amnésie au sens L^2 (calculée à l'aide de la formule (V.32)) présente un minimum pour ν=0.22, ce qui correspond à une bande de transition assez large (environ 0.054 pour une bande passante de 0.16). L'atténuation de son côté ne présente bien sûr pas de minimum en ce point: elle est strictement décroissante quand la largeur de la bande de transition augmente…

Il n'y a donc pas de relation directe entre sélectivité et amnésie.

b. Régularité

Là encore l'influence de la régularité sur l'amnésie n'est pas totale. Ce qui est vrai, et qui d'ailleurs était décrit dans l'article, c'est que si G est un filtre donnant lieu à convergence forte de suites discrètes, alors en multipliant par un nombre suffisamment grand de facteurs $\frac{z^p-1}{z-1}$ —directement liés comme on le sait à la régularité des fonctions limites associées— l'amnésie des fonctions limites peut être rendue aussi petite que l'on veut. Cela vient bien sûr directement du fait que l'amnésie liée au filtre $\left(\frac{z^p-1}{z-1}\right)^s$ tend vers zéro exponentiellement quand s tend vers l'infini (à l'aide de (V.36)) et du théorème V.16. Cependant, on doit attribuer cela au fait que le filtre $\left(\frac{z^p-1}{z-1}\right)^s$ devient très sélectif, et non au fait qu'il s'agit d'un facteur de régularité, c'est-à dire que n'importe quel facteur suffisamment sélectif se comportera de la même manière.

En réalité, on ne peut pas démontrer que la régularité fait *automatiquement* diminuer l'amnésie car c'est faux… il suffit en effet de prendre les filtres

$$G(z)=\tfrac{1}{9}\left(1+z+z^2\right)^3 \quad \text{et} \quad G^1(z)=\tfrac{1}{6}\left(1+z+z^2\right)^2(1+z)$$

qui se correspondent à travers la multiplication par un facteur de régularité. Le premier filtre, bien que plus régulier que le second d'une unité —les fonctions limites de G^1 sont combinaisons linéaires des dérivées des fonctions de G— a une amnésie au sens L^2 plus grande (η=0.0215) que G^1 (η=0.021). Cette différence peut paraître faible mais est en fait amplifiée quand on augmente le nombre de facteurs de régularité possible: ceci est illustré dans le graphique suivant où l'on a tracé l'amnésie au sens L^2 calculée pour la série de filtres

$$G^k(z)=\tfrac{1}{3^{8-k}2^k}\left(1+z+z^2\right)^{9-k}(1+z)^k \qquad \text{pour } k=0..8$$

dont les graphes se trouvent page suivante.

Là aussi, ces filtres se correspondent à travers la multiplication de facteurs de régularité. On constate un minimum de l'amnésie, non pas pour k=0, mais pour k=2: le rapport entre l'amnésie pour k=0 et celle pour k=2 atteint presque 10.

Il reste cependant vrai que la régularité est globalement bénéfique pour l'amnésie, comme le montre le reste de la courbe. Un tel résultat doit être considéré comme empirique, même si en travaillant un peu les nombres ε_j^s (V.27) on pourrait peut-être arriver à mettre en évidence cette propriété (cela ne serait de toute façon pas suffisant, les résultats obtenus étant seulement des majorations).

Même si la régularité n'a pas une influence déterminante sur l'amnésie, il n'en reste pas moins qu'une forte régularité implique une convergence beaucoup plus rapide des schémas discrets —adéquatement interpolés— vers les fonctions limites. C'est là que se situe son intérêt: si en effet, on peut estimer que les fonctions limites sont atteintes en une itération, l'espace engendré par ces fonctions pourra être considéré comme composé de seulement q différentes fonctions translatées d'un multiple de q. Cela vient bien sûr de la relation

$$\varphi_n(x) \cong \sum_k g[kq - np]\varphi\left(\frac{p}{q}x - k\right)$$

Si deux itérations sont nécessaires, on aura alors q^2 différentes fonctions translatées de multiples de q^2. Ceci bien sûr facilite grandement le calcul de l'amnésie, et c'est d'ailleurs ce qui est exprimé par le théorème V.17.

En particulier, si l'on n'a besoin que d'une itération pour atteindre les fonctions limites, on aura $\varepsilon \cong \varepsilon_1^0$ dont on a vu, dans le paragraphe consacré plus haut à la sélectivité, qu'il est minimisé quand le filtre G est suffisamment atténué sur la bande $[1/(p+q)\ 1/2]$. Ceci montre l'interdépendance des deux propriétés: sélectivité et régularité, aucune des deux ne l'emportant seule pour minimiser l'erreur de translation ε.

3. Exemples

On va donner deux exemples qui permettent de mesurer la qualité des estimations, par rapport à l'amnésie L^2 et éventuellement par rapport à la véritable valeur de ε.

Pour cela, reprenons les mêmes filtres que dans les exemples sur la régularité.

a. Exemple n°1

Soit le filtre $G(z) = \frac{1}{3}(1 + z + z^2)^2$. On peut calculer alors exactement l'amnésie à l'aide du théorème V.22 ce qui donne ε=0.257 et au sens L^2 on aurait η=0.082.

Pour estimer à l'aide des nombres ε_s l'amnésie L^∞, on démontre d'abord aisément que $\left|G_j^2\right| \le \frac{3^j}{2^j}$ et que $\left|G_j^1\right| \le 1$. L'estimation minimale en une itération donne alors pour résultat $\varepsilon \le 0.75$. En 4 itérations on obtient $\varepsilon \le 0.39$

b. Exemple n°2

Si l'on prend le filtre $G(z) = \frac{1}{9}(1 + z + z^2)^3$ alors on a η=0.022. À l'aide de ε_1^0 et ε_1^3 on trouve $\varepsilon \le 0.09$. Si l'on pousse les itérations jusqu'à l'ordre 3, le calcul de ε_3^0 et ε_3^3 nous permet alors d'affiner notre estimation à $\varepsilon \le 0.039$

D. Conséquences sur les filtres itérés

Il apparaît difficile de relier de prime abord ces notions continues d'amnésie et surtout de régularité aux propriétés d'un banc de filtres itéré un nombre fini de fois. La propriété la plus importante que l'on réclame en général d'un banc de filtres est une forte sélectivité des filtres qui le composent afin de bien séparer les différentes composantes fréquentielles du signal initial et qui sont souvent les éléments pertinents d'information subjective: c'est en particulier le cas de l'oreille humaine qui met dans un même sac tout ce qui se trouve dans un même tiers d'octave [ZF] (voir chapitre VII).

En fait l'apport principal de la régularité viendra du fait que les fonctions limites sont atteintes plus rapidement, permettant de caractériser le banc de filtres itéré à partir des premières itérations. Ce résultat est déjà partiellement indiqué dans [Ri1] mais ne met pas en évidence de différence de convergence entre une régularité d'ordre 1 et une régularité plus grande. Il se trouve en effet que si l'on ne prête pas garde à la continuité des fonctions limites, l'itération dégrade indéfiniment la réponse fréquentielle [Dau2].

De même, l'amnésie est fortement liée à la sélectivité fréquentielle du filtre itéré comme on va le démontrer. Enfin, si les filtres conduisent à des fonctions très régulières et d'amnésie suffisamment faible, le banc de filtres se comportera dès les premières itérations comme une transformée en ondelettes, ce qui facilite bien sûr l'interprétation des coefficients de sortie du banc de filtres.

1. Sélectivité

Exprimons tout d'abord ce que l'on entend par sélectivité fréquentielle pour une branche rationnelle itérée. Pour cela, on considère un signal x_n stationnaire à l'entrée d'un banc de

filtres rationnel itéré et l'on note $x_j[n]$ la $j^{\text{ème}}$ sortie passe-bas c'est-à dire $x_j[n] = \sum_k g_j[np^j - kq^j]x_k$.

a. Définition de la sélectivité

En utilisant (II.12) on a

$$\sum_{n=0}^{q^j-1} \left\langle \left| x_j[n] \right|^2 \right\rangle = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| G_j(e^{-2i\pi\nu}) \right|^2 R(e^{-2i\pi q^j \nu}) d\nu \qquad \text{(V.37)}$$

où l'on reconnaît dans $R(e^{-2i\pi\nu})$ la densité spectrale de puissance du signal d'entrée. Pour la sélectivité du banc de filtres d'analyse, on impose que chaque branche j ne laisse passer que les fréquences comprises entre $-\frac{q^j}{2p^j}\nu_0$ et $\frac{q^j}{2p^j}\nu_0$ et que d'autre part, à l'intérieur de la bande passante $-\frac{\nu_p}{2} \le \nu \le \frac{\nu_p}{2}$ (où $\nu_p \le \nu_0$), les filtres G_j soient approximativement constants en norme (en tous cas, il ne doivent pas s'y annuler).

Notons qu'en toute rigueur on n'a pas besoin de privilégier autant les fréquences autour de zéro puisque le terme qui dépend de la densité spectrale de puissance de x est répété q^j fois dans l'intégrale (V.37): on pourrait tout aussi bien décider que le filtre G_j soit passe-bande.

Idéalement, on souhaiterait avoir $\nu_0 = 1$ mais cela répercuterait de très mauvaises propriétés sur le filtre de reconstruction. Alors, en effet, l'analyse comporterait des "trous fréquentiels" qui entraîneraient une forte instabilité du système d'analyse-synthèse et qui ne pourraient être résolus que par l'utilisation d'un filtre de longueur infinie. On a donc toujours $\nu_0 > 1$ ce qui entraîne un certain recouvrement entre bandes. Ce recouvrement doit bien sûr être assez faible afin de pouvoir bénéficier des propriétés de séparation fréquentielle de ce type de transformation.

Supposons donc que le signal x_n ait une densité spectrale de puissance nulle sur l'intervalle $\left[-\frac{q^j}{2p^j}\nu_0, \frac{q^j}{2p^j}\nu_0\right]$. Dire que la branche j est sélective au sens L^2 signifie que $\left\langle \left| x_j[n] \right|^2 \right\rangle$ doit être petit par rapport à ce qu'il aurait pu être si toute sa puissance avait été contenue, au contraire dans l'intervalle $\left[-\frac{q^j}{2p^j}\nu_0, \frac{q^j}{2p^j}\nu_0\right]$.

On peut se limiter pour x_n à une forme de bruit blanc filtré. Mathématiquement donc, la propriété de sélectivité à toutes les échelles j impose l'existence d'une constante σ, petite devant 1, telle que

$$\sigma^2 = \limsup_{j\to\infty} \frac{\int_{\frac{\nu_0}{2p^j}}^{\frac{1}{2}} \left| G_j(e^{-2i\pi\nu}) \right|^2 d\nu}{\int_0^{\frac{\nu_0}{2p^j}} \left| G_j(e^{-2i\pi\nu}) \right|^2 d\nu} \qquad \text{(V.38)}$$

Comme le produit infini (IV.11) est toujours convergent dès que $G(1)=p$, on peut donc définir une fonction $\phi(\nu)$ —correspondant à une distribution dans le domaine temporel—. Si cette fonction n'est pas de carré intégrable, alors on peut démontrer que σ n'est pas finie. En effet comme

(IV.11) converge uniformément sur tout intervalle fréquentiel, $\hat{\varphi}(\nu)$ est au moins localement de carré intégrable. À partir de (V.38), fixons deux fréquences $\nu_1 > \nu_0$ on a

$$\limsup_{j\to\infty} \frac{\int_0^{\frac{\nu_1}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu}{\int_0^{\frac{\nu_0}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu} \le 1+\sigma^2$$

qui donne donc

$$\frac{\int_0^{\frac{\nu_1}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu}{\int_0^{\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu} \le 1+\sigma^2$$

Maintenant, si l'on augmente la valeur de ν_1, comme $\hat{\varphi}(\nu)$ n'est pas de carré intégrable, le premier membre va augmenter sans limite, et donc $\sigma = +\infty$.

Par suite, si $\hat{\varphi}(\nu)$ n'est pas de carré intégrable, on n'a aucune chance que des itérations répétées du schéma rationnel conduisent à des filtres sélectifs. C'est donc une condition nécessaire que l'on supposera désormais.

b. Calcul de σ

Le théorème suivant permet de calculer la valeur de σ.

Théorème V.23 *La constante σ définie par (V.38) est donnée par la formule suivante*

$$\sigma^2 = \frac{\int_{|\nu|\ge\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu + \eta^2}{\int_{|\nu|\le\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu} \qquad (V.39)$$

Si les schémas discrets ne convergent pas fortement η sera donné par la formule (V.31) plutôt que (V.23).

Preuve

Le résultat vient en fait directement du lemme V.20 puisqu'on peut écrire

$$\sigma^2 = \limsup_{j\to\infty} \left[\frac{p^{-j}\int_0^1 \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu - \|\varphi\|_2^2}{p^{-j}\int_{|\nu|\le\frac{\nu_0}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu} + \frac{\|\varphi\|_2^2 - p^{-j}\int_{|\nu|\le\frac{\nu_0}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu}{p^{-j}\int_{|\nu|\le\frac{\nu_0}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu} \right]$$

qui converge vers (V.39) quand j tend vers l'infini.

Ce théorème fait ainsi le lien entre certaines propriétés des fonctions limites et le paramètre de sélectivité du banc de filtres d'analyse. On constate ainsi l'importance primordiale qui est donnée à l'amnésie qui indépendamment de la largeur de la bande atténuée, donne une valeur minimale à la sélectivité du banc de filtres. On a ainsi $\sigma \geq \eta / \|\varphi\|_2$. Il est clair qu'il ne sert à rien de minimiser la sélectivité fréquentielle de la fonction moyenne: en deçà d'une valeur fixée par l'amnésie, le paramètre σ ne sera plus modifié. Dans le cas d'un filtre orthonormé, la formule (V.39) se simplifie en

$$\sigma^2 = \frac{1}{\int_{|\nu|\leq\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu} - 1 \qquad \text{(V.40)}$$

Dans ce dernier résultat, on a implicitement supposé $G(1)=p$ —c'était imposé dans (V.31)— ce qui implique que $\frac{z^p-1}{z-1}$ divise G, puisque la relation d'orthonormalité

$$\sum_{k=0}^{p-1} \left| G(e^{-2i\pi\frac{\nu+k}{p}}) \right|^2 = p^2$$

l'impose. Si l'on veut que $G(1)\neq p$, alors on doit revoir les formules de la façon suivante: on peut toujours définir une fonction limite moyenne φ à l'aide du produit infini (IV.11) en substituant à G le filtre $pG(z)/G(1)$. On aura alors

$$\begin{aligned} 1+\sigma^2 &= \limsup_{j\to\infty} \frac{p^j G(1)^{-2j} \int_0^1 \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu}{p^j G(1)^{-2j} \int_{|\nu|\leq\frac{\nu_0}{2p^j}} \left|G_j(e^{-2i\pi\nu})\right|^2 d\nu} \\ &= \frac{\limsup_{j\to\infty} p^{2j} G(1)^{-2j}}{\int_{|\nu|\leq\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu} \end{aligned}$$

qui n'est pas borné puisque $G(1)<p$ d'après notre nouvelle hypothèse. Ceci montre donc que dans le cas de filtres orthonormés, il est indispensable que ceux-ci comportent au moins un facteur de régularité.

Le résultat ci-dessus est en fait plus instructif puisqu'il nous indique jusqu'à quel ordre d'itération on peut aller pour ne pas trop dégrader la sélectivité du filtre itéré. En effet, si l'on se fixe une sélectivité de l'ordre de 35 dB, sachant que $\int_{|\nu|\leq\frac{\nu_0}{2}} \left|\hat{\varphi}(\nu)\right|^2 d\nu \cong \|\varphi\|_2^2 \cong 1$ pour des filtres suffisamment sélectifs, alors il faut que j vérifie une inégalité de la forme (après simplifications utilisant le fait que $G(1)$ est proche de p)

$$j \leq \frac{p^2 10^{-3.5}}{\sum_{k=1}^{p-1} \left|G(\xi^k)\right|^2}$$

où $\xi = e^{2i\pi/p}$. Ainsi dans le cas dyadique, si l'on souhaite faire 10 itérations (10 octaves per-

mettant de décrire l'étendue de la sensibilité de l'oreille humaine) il sera nécessaire que $G(-1)$ soit inférieur à 39 dB.

2. Régularité seule

On va maintenant s'intéresser aux implications de la régularité pour le banc de filtres.

a. Nécessité de la convergence forte à la synthèse

On a vu lors du calcul sur la sélectivité qu'il était alors nécessaire que la fonction moyenne soit de carré intégrable. Sans cela, on pouvait prédire une explosion des caractéristiques fréquentielles du filtre itéré.

En fait on doit être encore plus exigeant à la synthèse. On doit en effet imposer que le filtre itéré de synthèse ne donne en sortie qu'un signal passe-bas, de bande passante d'autant plus faible que le nombre d'itérations est élevé. Ceci implique en particulier que quel que soit le filtre passe-haut s'annulant à la fréquence 0 et mis en série après le filtre itéré, les échantillons de la sortie finale de ce système seront de plus en plus petits —jusqu'à atteindre 0— quand augmente le nombre d'itérations, devant ceux que l'on observe sans le filtre passe-haut. On vérifie aisément que, si H est ce filtre passe-haut, cette condition s'écrit mathématiquement

$$\frac{\left|G_j(z)H(z^{q^j})\right|_\infty}{\left|G_j\right|_\infty} \xrightarrow[j\to\infty]{} 0$$

En particulier, le filtre $H(z)=z-1$ convient: on retombe alors exactement sur la condition de convergence forte des suites discrètes, dont on sait qu'elle conduit à des fonctions limites continues —en fait au moins $\mathcal{C}^\alpha$ avec $\alpha>0$—. Par contre, il est bien clair que l'on n'a pas besoin, à l'aide de ce seul argument d'ordres de régularité plus élevés.

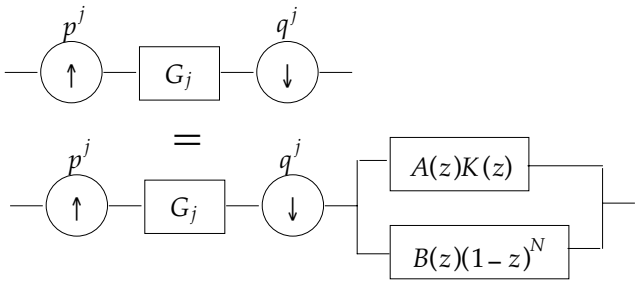
b. Une convergence plus rapide

On va maintenant voir l'intérêt d'avoir plus de régularité que la simple continuité. On a en effet montré dans le théorème V.4 qu'une forte régularité implique une convergence d'autant plus rapide des schémas discrets correctement interpolés. On sait ainsi que que les valeurs $\varphi(n)$ de la fonction moyenne en les points entiers déterminent un filtre K tel que $\sum_{k'} \varphi(k-k')g_j[kq^j - np^j]$ tende vers $\varphi_n\left(k\frac{q^j}{p^j}\right)$ à la vitesse de $\frac{q^{j\alpha}}{p^{j\alpha}}$ quand j tend vers l'infini.

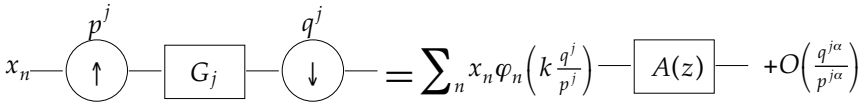
Commençons à la synthèse pour montrer comment se manifeste la régularité sur la sortie du banc de filtres, et supposons donc que le filtre de synthèse G comporte N facteurs de régularité. Comme $K(1) = \sum_n \varphi(n) = 1 \neq 0$ les filtres K et $(z-1)^N$ sont premiers entre eux. On peut donc écrire la relation de Bezout suivante

$$A(z)K(z) + B(z)(z-1)^N = 1$$

définissant deux polynômes A et B uniques vérifiant $\deg(A) \le N-1$ et $\deg(B) \le \deg(K)-1$. Graphiquement cela se traduira par



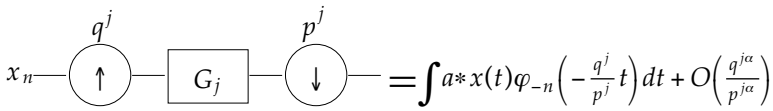
Il est alors facile de voir que grâce à la régularité du filtre



ce qui signifie que la sortie passe-bas d'un banc de filtres de synthèse n'est rien d'autre qu'une série de pseudo-ondelettes "pères" filtrée, à une erreur près d'autant plus petite que la régularité est plus grande. Il ne pose aucune difficulté d'établir le même résultat pour les branches passe-bande, puisque l'opérateur passe-bande s'applique alors, pour la synthèse, sur la partie gauche du graphique.

Bien sûr on aurait pu, sans utiliser de relation de Bezout, démontrer un résultat très analogue où cette fois le filtre A aurait été remplacé par l'inverse de $K(z)$. Mais c'est ici un important avantage d'avoir un filtre de longueur finie plutôt que de longueur infinie —dont par ailleurs, on ne sait même pas s'il est stable—.

De la même manière, à l'analyse, on aura l'équivalence graphique suivante



en supposant x_n interpolé par la fonction de Nyquist χ pour donner $x(t)$, et où $a(t)$ est la version "continue" —c'est-à dire $a(t) = \sum_k a_k \delta(t-k)$ les a_k étant les coefficients de $A(z)$—. On obtiendrait une formule semblable pour chaque sortie passe-bande, ce qui montre que le banc de filtres régulier d'analyse se comporte comme la mise en série d'un filtre RIF, suivi d'une transformée en pseudo-ondelettes

En fait il n'y a pour l'instant que peu d'intérêt à ce qu'on atteigne les fonctions limites en quelques itérations… sauf si l'on rajoute des propriétés aux filtres. En effet, si la régularité d'ordre 1 nous assure que la transformation discrète va atteindre un état limite pour un suffisamment grand nombre d'itérations (et donc il ne va pas y avoir "explosion" du filtre passe-bas itéré), les ordres plus élevés de régularité nous assurent que cet état limite sera atteint plus rapidement. On peut ainsi imaginer atteindre la limite, à peu de chose près en une seule

itération, ce qui simplifie de beaucoup l'étude du banc de filtres itéré, ramenant son étude à celle d'une conception de filtres classique.

c. Influence sur la sélectivité de la fonction moyenne

On peut également voir un intérêt à imposer une forte régularité aux fonctions limites et donc à la fonction moyenne dans la mesure où la régularité se traduit en partie fréquentiellement par une plus rapide décroissance du spectre de la fonction considérée. Bien sûr, cela ne nous assure pas que la fonction limite sera sélective, mais cela va dans le même sens.

d. Moments nuls pour la pseudo-ondelette duale

Comme on l'a déjà indiqué (voir le théorème IV.10), les facteurs de régularité —et non plus la régularité— à la synthèse entraînent mécaniquement autant de moments nuls pour les fonctions passe-haut limites, ou plutôt des facteurs (z-1) pour les filtres passe-haut. Cette dernière propriété est indispensable dans un but de codage quand les signaux analysés ont un spectre très fortement concentré autour de la fréquence zéro, comme les images: il faut en effet impérativement empêcher le moindre transfert de l'énergie de la fréquence continue vers des sorties passe-bande, sinon elles masqueraient les autres composantes spectrales, plus faibles, et obligeraient, dans un but de qualité, une quantification beaucoup plus fine.

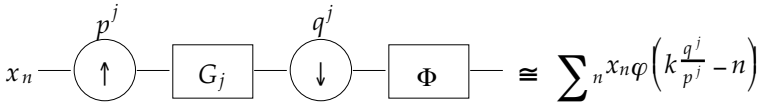
3. Amnésie seule

Comme on l'a vu plus haut (théorème V.23), une faible amnésie est une condition nécessaire —mais pas suffisante— de forte sélectivité fréquentielle, en particulier dans le cas orthonormal où $\int \varphi^2 < 1$.

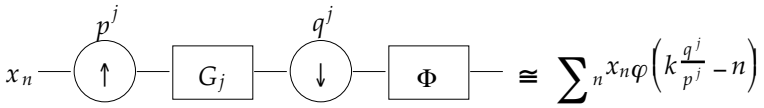
Une autre conséquence d'une faible amnésie réside dans l'inégalité suivante

$$\left|\varphi\left(k\frac{q^j}{p^j}-n\right)-\sum_{k'} g_j[k'q^j-np^j]\varphi(k-k')\right| \le \varepsilon\left(1+\frac{L}{p-q}\left|G_j\right|_\infty\right)$$

c'est-à dire que là encore, sans aucune régularité cette fois —autre que la continuité assurant que $\left|G_j\right|_\infty$ reste borné—, on a l'équivalence suivante, *ceci étant indépendant de l'échelle j considérée*



le filtre $\Phi(z)$ s'écrivant sous la forme $\Phi(z)=\sum_n \varphi(n)z^n$. Cette relation, écrite à la synthèse peut également être obtenue à l'analyse, d'où cette fois



L'utilité de ces relations tient dans le fait que le système discret d'analyse-synthèse se comporte comme la version discrète d'un système continu invariant dans le temps. En particulier, un signal sinusoïdal se transformera en un autre signal sinusoïdal à travers une branche rationnelle. C'est donc là l'intérêt d'un banc de filtres à faible amnésie.

4. Régularité+amnésie

Si à la régularité on ajoute une faible amnésie alors le système d'analyse se comportera comme la composition d'un préfiltrage FIR suivi d'une transformée en ondelettes de facteur d'échelle fractionnaire. À la synthèse, on aura une décomposition en série d'ondelettes suivie d'un postfiltrage FIR.

Ce type de résultat a bien sûr des implications autant en terme d'interprétation du banc de filtres que dans le but d'implémenter une transformée en ondelettes de rapport fractionnaire. En effet dans ce dernier cas, on pourrait appliquer des techniques semblables à celles qui ont été mises en œuvre dans l'article de Rioul et Duhamel [RD1] sur l'implémentation de transformations en ondelettes.

5. Sélectivité du filtre passe-bas+régularité

Enfin, si le filtre passe-bas est suffisamment régulier pour converger en une itération, comme on l'a vu dans la partie consacrée aux dépendances de l'amnésie, la sélectivité du filtre va entraîner une amnésie d'autant plus faible des fonctions limites, ce qui permettra de déduire les mêmes conclusions que dans le paragraphe précédent.

E. Résumé du chapitre

On s'est ici consacré aux deux caractéristiques les plus importantes des fonctions limites engendrées par un banc de filtres rationnel itéré.

La régularité est fréquemment la propriété que l'on demande immédiatement à des fonctions, et il se trouve qu'il n'est pas facile d'avoir des fonctions limites continues: si l'on n'y prend pas garde, on récupère en général des objets pathologiques dont on n'est même pas sûr, comme c'était par contre le cas avec les itérations en octave, qu'ils représentent des distributions… On s'est donc penché sur les méthodes permettant d'estimer la régularité des fonctions limites: celles-ci sont très semblables à celles qui ont cours dans le cas dyadique, du moins en ce qui concerne la régularité au sens de Hölder , puisqu'il est apparu que l'amnésie empêchaient d'étendre les techniques utilisées pour calculer la régularité au sens de Sobolev —rapidité de décroissance du spectre des fonctions—. Un nouveau résultat concerne le lien direct entre la régularité globale des fonctions limites et la vitesse de convergence des schémas discrets correctement interpolés. Ce résultat avait été masqué dans le cas dyadique, probablement à cause de l'existence de schémas d'interpolation *exacts* permettant d'obtenir les échantillons de la fonction limite.

Enfin, on s'est intéressé au calcul de l'amnésie dans la mesure où l'on sait que plus elle est faible plus notre transformation s'approche d'une transformation en ondelettes. Des estimations sont donc données, certaines étant même exactes, d'autres tirant profit de la régularité. Je m'étais en effet rendu compte [Blu1] que celle-ci est influencée par la régularité. En fait, on

montre maintenant que le résultat est plus complexe et que l'amnésie est en fait très liée à la sélectivité du banc de filtres.

Le chapitre se termine sur une étude précise de l'interprétation des propriétés de ces fonctions à temps continu dans les bancs de filtres d'analyse et de synthèse qui sont bien sûr à temps discret, et donc sur l'intérêt d'avoir une forte régularité et une faible amnésie pour un banc de filtres itéré.

VI. Conception de filtres

Après avoir décrit dans les précédents chapitres un certain nombre de propriétés souhaitables pour un banc de filtres itéré en fraction d'octave, on doit maintenant aborder le problème de la synthèse de filtres contraints par ces propriétés mises ensemble. On sait en effet résoudre les problèmes indépendamment, par exemple si l'on veut

- de la régularité, on multiplie le polynôme passe-bas par un suffisamment grand nombre de facteurs

$$\frac{z^p - 1}{z - 1}$$

  (ce qui est équivalent à imposer autant de facteurs de régularité $\frac{z^p-1}{z^q-1}$) ou bien l'on modifie les coefficients du filtre en estimant à chaque fois la régularité (à l'aide des résultats du chapitre V)

- de la sélectivité, selon la définition que l'on utilise de l'atténuation on calcule le filtre à l'aide d'un algorithme de Remez (sens L^∞) [PM] ou par une simple résolution de système linéaire (sens L^2)

- une faible amnésie, on impose de la régularité et l'on ajoute de la sélectivité, ou bien on modifie les coefficients du filtre tout en calculant à chaque étape l'amnésie L^2, à l'aide de (V.32)

- un système à reconstruction parfaite FIR aussi bien à l'analyse qu'à la synthèse, on imposera à la matrice polyphase du système à deux bandes d'être de déterminant égal à z^n. Ceci peut s'obtenir de deux manières différentes

  - soit on choisit le filtre passe-bas d'analyse, duquel on peut déduire un filtre passe-bas de synthèse qui soit FIR. De là, on obtient directement les filtres passe-haut d'analyse et de synthèse
  - soit on écrit la matrice polyphase du système comme le produit d'une matrice unimodulaire et d'une matrice paraunitaire qui se factorisent d'après les théorèmes du chapitre III sur la factorisation.

- de la paraunitarité, on utilise également les résultats du chapitre III sur la factorisation des matrices polynômiales

On a ici volontairement exclu des propriétés utiles la symétrie du filtre puisque celle-ci, découlant de la contrainte de phase linéaire n'a pas de sens pour un opérateur qui ne transforme pas une sinusoïde en une sinusoïde. Tout au plus peut-on noter que la symétrie du filtre passe-bas associée à une faible erreur de translation conduit à un opérateur proche de la phase linéaire.

Quand elles sont imposées ensemble, les propriétés énoncées plus haut rendent particulièrement ardue la conception de filtres. C'est le but de ce chapitre d'évaluer ces difficultés et de proposer des algorithmes pour les contourner.

Il faut noter cependant que dans ce domaine, à la différence de ce qui se passe pour les résultats sur les fonctions limites, il existe dans la littérature des algorithmes qui permettent de calculer des bancs de filtres rationnels [NBS1,NBS2,KV3]. Dans le premier cas [NBS1,NBS2] il s'agit de minimiser une fonctionnelle complexe dans laquelle sont prises en compte aussi bien l'atténuation, la sélectivité ou la platitude des filtres que la reconstruction. Dans ces conditions la reconstruction n'est pas parfaite (on se fixe au départ les degrés des filtres du banc d'analyse et de synthèse) et la procédure itérative paraît assez lourde. Elle s'applique en échange à tout type de filtres. Les résultats indiqués dans [NBS1] pour le cas 3/2 donnent un banc de filtres biorthogonal dont les caractéristiques paraissent bonnes, mais j'avoue ne pas en avoir implémenté l'algorithme. Dans le deuxième cas [KV1], il s'agit de conception de filtres orthogonaux utilisant la factorisation des matrices paraunitaires [VNDS].

A. Complexité du problème

C'est l'introduction de la condition de reconstruction parfaite FIR qui rend le problème hautement non-linéaire. En effet, en termes d'équations cette condition impose que les filtres passe-bas d'analyse et synthèse G et $\tilde{G}$ vérifient la première série d'équations (II.9)

$$G(z)\tilde{G}(ze^{-2i\pi s/q}) = q\delta_s + zP_{s,1}(z^p) + z^2P_{s,2}(z^p) + \ldots z^{p-1}P_{s,p-1}(z^p) \qquad \text{(VI.1)}$$

pour s=0…q-1, et où les $P_{s,k}$ sont des polynômes. Ainsi, en fixant les degrés de liberté de G et $\tilde{G}$ par $G(z) = \sum_{k=l}^{L} g_k z^k$ et $\tilde{G}(z) = \sum_{k=\tilde{l}}^{\tilde{L}} \tilde{g}_k z^k$ le nombre d'équations de contraintes N_e est

$$N_e = q\left[E\left(\frac{L+\tilde{L}}{p}\right) + E\left(-\frac{l+\tilde{l}}{p}\right)\right]$$

Ces équations sont bien évidemment quadratiques en $(G,\tilde{G})$ (c'est-à dire homogènes de degré 2 en ces variables). Si l'on souhaite éliminer $N_e - 1$ inconnues de ces équations il restera à la fin une seule équation homogène en ces $\tilde{L} + L - \tilde{l} - l - N_e + 3$ inconnues et de degré 2^{N_e}.

La minimisation d'une certaine fonctionnelle décrivant la sélectivité du couple d'analyse-synthèse conduira donc à un très grand nombre de minima locaux (au moins 2^{N_e} dans le cas général), ce qui pose bien sûr des problèmes très difficiles à surmonter pour la conception de filtres.

Imposer la paraunitarité revient à diviser le nombre d'inconnues par deux, ainsi que (approximativement) le nombre d'équations, puisqu'il est facile de vérifier que les équations (VI.1) deviennent alors globalement symétriques en z (pour s=0 c'est évident et si $s\neq 0$ changer z en $1/z$ revient à changer s en q-s). Cette remarque sera d'ailleurs le point central de l'algorithme de conception de filtres paraunitaires qui sera développé plus loin.

On peut se poser la question suivante: à largeur de bande de transition donnée $\delta\nu$, quelle taille de filtre sera-t-il nécessaire d'imposer pour atteindre une atténuation suffisante, disons 30 dB, dans la bande atténuée? Dans le cas d'une conception de filtre agissant directement —et non pas comme ici dans le domaine suréchantillonné— sur le signal, la réponse est que l'on doit compter sur un filtre de longueur approximativement égale à $1/\delta\nu$.

Dans notre cas, le filtre H agissant directement sur le signal, cette conclusion tient toujours. Par contre, cela signifie que le filtre passe-bas G qui agit dans le domaine suréchantillonné par q doit avoir une bande de transition q fois plus étroite afin de respecter la même contrainte. Il sera en conséquence q fois plus long, c'est-à dire de taille approximativement égale à $q/\delta\nu$. Ainsi dans le cas 5/4, si l'on veut concevoir un banc de deux filtres tel que la bande de transition du filtre passe-haut soit de largeur inférieure au quart de la largeur de sa bande passante dans [0 1/2] soit $\delta\nu \leq 0.025$, il sera nécessaire d'envisager un filtre passe-bas de taille $4/0.025$ c'est-à dire 160. Il faut donc s'attendre à avoir des filtres passe-bas assez longs sans pour autant que la sélectivité fréquentielle de l'analyse dépasse une niveau moyen…

Cela aura également des conséquences assez néfastes sur le délai de la transformation, pouvant les rendre impropres au codage des conversations, par exemple.

B. Solutions classiques

On sait résoudre les problèmes de la minimisation L^2 et L^∞ de filtres orthonormés avec adjonction de régularité dans le cas entier, c'est-à dire q=1. Le cas biorthogonal reste encore largement moins bien exploré, malgré l'utilisation des techniques de décomposition des matrices polynômiales donnant par exemple des filtres symétriques [NV].

Le cas le plus intéressant est le cas dyadique dont on va rappeler les solutions, sans cependant inclure d'hypothèse de régularité. Dans tous les cas, on part de l'équation d'analyse-synthèse

$$G(z)G(z^{-1}) = 1 + zR(z^2) \qquad \text{(VI.2)}$$

où R est un filtre symétrique vérifiant plus précisément $R(z^{-1}) = zR(z)$ et tel que $1 + e^{2i\pi\nu}R(e^{4i\pi\nu}) \geq 0$ pour toute valeur de ν. La raison pour laquelle on va pouvoir résoudre nos problèmes de minimisation est que l'on rend linéaire, en reportant sur R les calculs, un problème qui était quadratique initialement. Le fait d'ajouter de la régularité dans cette équation n'en changera pas le caractère linéaire, c'est pourquoi on se restreindra ici aux cas de régularité zéro.

1. Norme L^∞: Smith et Barnwell [SB]

Étant donnée une fréquence ν_0, il s'agit de trouver G minimisant $\max_{\nu_0 \leq \nu \leq \frac{1}{2}} \left|G(e^{2i\pi\nu})\right|$. Grâce à (VI.2) on peut se ramener à trouver le polynôme symétrique R minimisant $1 + e^{2i\pi\nu}R(e^{4i\pi\nu})$ —noter qu'il n'y a plus de valeur absolue— pour $\nu_0 \leq \nu \leq 1/2$ sous la contrainte de positivité de

$1 + e^{2i\pi\nu} R(e^{4i\pi\nu})$. Ce problème est ainsi directement linéaire et peut être résolu par une technique de programmation linéaire.

En fait, dans ce cas on bénéficie de l'algorithme d'échange des zéros "remez", mis au point par Parks et McClellan [PM], plus rapide encore que la programmation linéaire. On peut en effet facilement voir que le polynôme symétrique F minimisant le critère de Tchebitcheff

$$\max\left(\max_{0 \le \nu \le 1-\nu_0} \left| |F(e^{2i\pi\nu})| - 1 \right|, \max_{\nu_0 \le \nu \le \frac{1}{2}} |F(e^{2i\pi\nu})| \right)$$

vérifie nécessairement $F(z) + F(-z) = \text{Constante}$, c'est-à dire peut s'écrire sous la forme $F(z) = \text{Cte} + zR(z^2)$. Le filtre G idéal sera alors de la forme $G(z)G(z^{-1}) = b\big(a + F(z)\big)$ où a est la constante minimale assurant la positivité du second membre sur le cercle unité, et b le facteur de normalisation positif assurant (VI.2).

La version de cette solution en rajoutant des facteurs de régularité a été mise au point par Olivier Rioul [RD2].

2. Norme L^2

La solution au critère L^2, c'est-à dire trouver G minimisant

$$\int_{\nu_0}^{\frac{1}{2}} \left| G(e^{2i\pi\nu}) \right|^2 d\nu$$

sous la contrainte de reconstruction parfaite (VI.2) se ramène également à un problème de programmation linéaire [Ri4]. En effet, on peut écrire

$$\left| G(e^{2i\pi\nu}) \right|^2 = 1 + \sum_{k=0}^{N} a_k \cos\big(2\pi(2k+1)\nu\big)$$

où les a_k sont nos nouvelles inconnues vérifiant la positivité du membre de droite pour toute valeur de ν. Le problème revient donc à trouver ces nombres qui minimisent

$$1 - \nu_0 - \sum_{k=0}^{N} a_k \frac{\sin\big(2\pi(2k+1)\nu_0\big)}{2\pi(2k+1)}$$

sous les contraintes

$$1 + \sum_{k=0}^{N} a_k \cos\big(2\pi(2k+1)\nu\big) \ge 0 \quad \forall \nu \in [0,1[$$

ce qui constitue un problème éminemment linéaire, dont la programmation ne pose pas de gros problème.

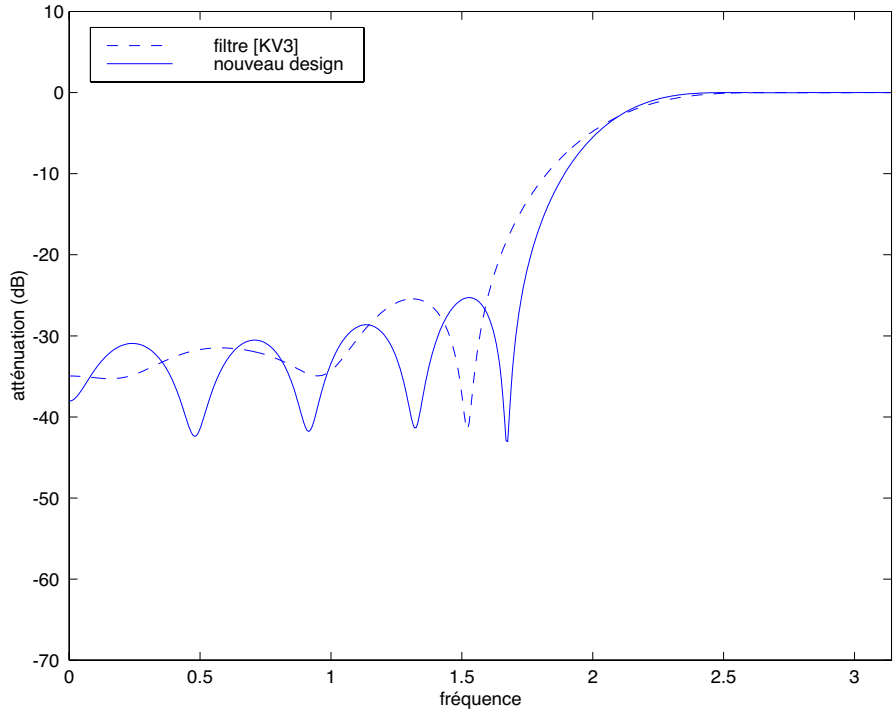
En fait, concernant ces deux méthodes, il faut ajouter au coût de la conception du filtre $G(z)G(z^{-1})$ le coût de l'extraction des racines de ce polynôme symétrique afin de récupérer G. Il

se trouve que plus le filtre est long, plus il faudra de précision à l'extraction de racines sous peine d'avoir une relation de reconstruction trop approximative. Les algorithmes qui vont être présentés maintenant (factorisation, [blu3]) ne présentent pas cet inconvénient puisqu'ils donnent directement le filtre G.

3. Algorithme par factorisation de matrices

C'est la méthode la plus générale, mais également la plus lourde et longue. Elle permet de s'affranchir des équations non-linéaires de reconstruction parfaite à l'aide de la paramétrisation adéquate issue des théorèmes sur la factorisation des matrices polynômiales FIR inversibles (cf chapitre III). Les paramètres du filtre sont alors indépendants les uns des autres —du moins tant qu'on n'impose pas d'autre contrainte, comme la régularité par exemple—, la complexité étant ainsi reportée directement sur la fonctionnelle à minimiser.

Cette méthode a été appliquée dans [KV3] pour le cas de filtres orthonormaux. Le procédé est en fait assez lourd à cause du nombre de minima locaux. En outre, le résultat n'est, semble-t-il pas meilleur, que l'algorithme qui va être développé dans cette partie [Blu3] si l'on compare sur l'exemple qui est donné. En effet, on peut concevoir avec l'algorithme développé plus loin dans ce chapitre un filtre passe-bas de longueur 30, correspondant à un passe-haut de longueur 15 dont l'atténuation est semblable à celle de [KV3] (30 dB) mais avec une bande de transition plus étroite. Rappelons que le passe-bas et le passe haut de [KV3] sont de longueur 32 et 15. Le filtre passe-haut obtenu est donné ci-après dans les unités de [KV3] pour comparaison.



C. Un algorithme direct

Cet algorithme ne s'applique qu'au cas des systèmes paraunitaires, et à l'usage on verra qu'il ne converge que pour des ordres de régularité inférieurs au égaux à 1 (sauf dans le cas dyadique…). Il s'agit d'un algorithme itératif qui minimise à chaque étape une atténuation

du filtre obtenu au sens L^2[Blu3]. Le résultat, ainsi qu'on le verra, est d'excellente qualité, même si le filtre finalement obtenu ne minimise pas l'atténuation au sens L^2.

1. Description

On part de la constatation déjà faite dans la partie sur la complexité du problème de minimisation: les équations (VI.1) sont redondantes quand on considère des filtres orthonormaux. On peut les écrire sous la forme "coefficients" suivante

$$\sum_k g[kq - n_0'p]g[kq - n_0p - spq] = \delta_s\delta_{n_0'-n_0} \qquad \text{(VI.3)}$$

pour tout s entier et n_0,n'_0=0..q-1. Cette dégénérescence du système d'équations peut être levée en ne considérant qu'une partie de celles-ci, c'est-à dire celles vérifiant

$$\forall s > 0\ \forall n_0, n_0' = 0..q-1 \text{ et si } s = 0\ \forall n_0 \le n_0' = 0..q-1$$

On vérifie aisément que ce nouveau système équivaut au précédent: on a ainsi écrit seulement la moitié des équations du système.

L'idée de l'algorithme est la suivante: étant donné le filtre $G_{n\text{-}1}$ on détermine le filtre G_n vérifiant les contraintes de "demi reconstruction" suivantes

$$\sum_k g_n[kq - n_0'p]g_{n-1}[kq - n_0p - spq] = \delta_s\delta_{n_0'-n_0} \begin{vmatrix} \forall s > 0\ \forall n_0, n_0' = 0..q-1 \\ \text{si } s = 0\ \forall n_0 \le n_0' = 0..q-1 \end{vmatrix} \qquad \text{(VI.4)}$$

ainsi qu'éventuellement d'autres contraintes —régularité par exemple—, et minimisant la fonctionnelle

$$J_0 = \int_{\nu_0}^{1-\nu_0} \left|G_n(e^{-2i\pi\nu})\right|^2 d\nu \qquad \text{(VI.5)}$$

Une fois ce filtre G_n obtenu, on calcule l'erreur de reconstruction du filtre G=G_n dans les équations (VI.3), et on recommence en substituant G_n à $G_{n\text{-}1}$ jusqu'à ce que cette erreur soit aussi petite que souhaitée. L'algorithme s'écrit donc de la manière condensée suivante

$$G_n = F(G_{n-1})$$

et s'il converge, ce sera vers un filtre G vérifiant G=F(G) qui sera ainsi orthonormé. Il s'agit donc d'un algorithme de point fixe, dont on sait que s'il converge, en général, sa vitesse de convergence sera exponentielle (c'est seulement dans le cas où, en le point de convergence, certaines valeurs propres —de modules nécessairement inférieurs ou égaux à 1— de la matrice différentielle de F seraient de module 1 que la convergence pourrait être plus lente).

En fait, quand on suit cette procédure, on observe une oscillation entre deux valeurs limites, c'est-à dire que l'on tend vers un couple de filtres G et G' vérifiant G=F(G') et G'=F(G). Il est facile de voir que ce couple est un couple valide d'analyse-synthèse. Cependant, comme nous

souhaitons avoir des filtres orthonormés, on modifiera la procédure itérative sous la forme suivante

$$G_n = \tfrac{1}{2}\big(G_{n-1} + F(G_{n-1})\big)$$

qui cette fois convergera bien vers un filtre orthonormé.

Il se trouve que la fonction F est facile à calculer dans la mesure où les contraintes ((VI.4) et éventuellement la régularité) imposées au problème de minimisation sont linéaires, et puisque la fonctionnelle (VI.5) est quadratique, donc conduisant, par différentiation, à un opérateur linéaire. On obtient donc simplement G_n à l'aide de la résolution d'un système linéaire.

Un gros avantage de ce type d'algorithmes est d'être modulable: on peut rajouter autant de contraintes linéaires que l'on souhaite, ainsi que de termes quadratiques dans la fonctionnelle à minimiser, sans perdre la simplicité de la résolution. On peut entre autres insérer des poids dans la fonctionnelle (VI.5) afin de minimiser de façon plus spécifique certains intervalles fréquentiels.

2. Implémentation

Trois parties peuvent être distinguées dans cette procédure

- choix d'un polynôme initial G_0
- implémentation d'une itération
- arrêt des itérations

En fait l'initialisation ne pose aucun problème: à moins de choisir précisément un polynôme qui rende singulier le système d'équations à résoudre lors de la première itération, l'expérience a montré que l'algorithme est totalement insensible à ce paramètre, ce qui est d'un intérêt non négligeable.

Si maintenant on note $G_{n\text{-}1}$ le vecteur des coefficients du filtre —et non plus sa forme en z— à l'itération n-1, et $\boldsymbol{\Gamma}_n$ le filtre dont on se propose de minimiser l'atténuation sous contraintes, on pourra écrire (VI.5) comme

$$J_0 = \boldsymbol{\Gamma}_n^T \mathbf{A}_0 \boldsymbol{\Gamma}_n \qquad \text{(VI.6)}$$

où $\mathbf{A}_0$ est la matrice symétrique suivante

$$\mathbf{A}_0[k',k] = \begin{cases} 1 - 2\nu_0 & \text{si } k = k' \\ -\dfrac{\sin(2\pi(k'-k)\nu_0)}{\pi(k'-k)} & \text{si } k \neq k' \end{cases} \qquad \text{(VI.7)}$$

pour k,k'=0..L, où l'on suppose que L est le degré du filtre.

D'un autre côté, les contraintes de demi-reconstruction (VI.4) vont pouvoir s'écrire sous la forme

$$\mathbf{S}_{n-1}\Gamma_n = C \tag{VI.8}$$

où $\mathbf{S}_{n\text{-}1}$ est une matrice ne dépendant que du filtre $G_{n\text{-}1}$ et C est un vecteur constant, correspondant au second membre de (VI.4).

À cela on peut rajouter des équations correspondant au nombre de facteurs de régularité du filtre limite. Si l'on impose N facteurs de régularité, on peut s'écrire

$$\mathbf{T}\Gamma_n = 0 \tag{VI.9}$$

où la matrice $\mathbf{T}$ a pour composantes

$$\mathbf{T}[s+kp,n] = \begin{cases} -\dfrac{n^k}{p} & \text{si } n \not\equiv s \bmod(p) \\ \left(1-\dfrac{1}{p}\right)n^k & \text{si } n \equiv s \bmod(p) \end{cases} \tag{VI.10}$$

pour k=0..N-1, s=0..p-1 et n=0..L.

En introduisant les multiplicateurs de Lagrange λ et μ, on doit donc minimiser une fonctionnelle de la forme $J_1 = \Gamma_n^T\mathbf{A}_0\Gamma_n - 2\lambda^T(\mathbf{S}_{n-1}\Gamma_n - C) - 2\mu^T\mathbf{T}\Gamma_n$. En fait, pour éviter certains problèmes numériques—la matrice $\mathbf{A}_0$ a des valeurs propres très proches de zéro pour L suffisamment grand—, il est nécessaire de rajouter un autre terme quadratique qui vienne "répéter" la contrainte de demi reconstruction. En la chargeant d'un poids α, on obtient finalement la fonctionnelle suivante à minimiser

$$J = \Gamma_n^T\mathbf{A}_0\Gamma_n + \alpha\left|\mathbf{S}_{n-1}\Gamma_n - C\right|^2 - 2\lambda^T(\mathbf{S}_{n-1}\Gamma_n - C) - 2\mu^T\mathbf{T}\Gamma_n \tag{VI.11}$$

La valeur α=1 a jusqu'à présent donné toute satisfaction.

La différencielle de (VI.11), égalée à 0 nous donne les équations

$$\begin{aligned} \left(\mathbf{A}_0 + \alpha\mathbf{S}_{n-1}^T\mathbf{S}_{n-1}\right)\Gamma_n &= \mathbf{S}_{n-1}^T(\alpha C + \lambda) + \mathbf{T}^T\mu \\ \mathbf{S}_{n-1}\Gamma_n &= C \\ \mathbf{T}\Gamma_n &= 0 \end{aligned}$$

ce qui conduit finalement à une expression que l'on me pardonnera de ne pas expliciter ici, du fait de sa longueur qui n'ajoute rien à la compréhension du problème.

On obtient alors G_n par

$$G_n = \tfrac{1}{2}\left(G_{n-1} + \Gamma_n\right) \tag{VI.12}$$

dont on calcule l'erreur de reconstruction

$$e_n = \max_{s,n_0,n_0'} \left| \delta_s \delta_{n_0'-n_0} - \sum\nolimits_k g_n[kq - n_0'p] g_n[kq - n_0 p - spq] \right| \qquad \text{(VI.13)}$$

On arrête les itérations quand e_n devient inférieur à une constante très petite. Dans le programme écrit en matlab, je me suis arrêté à 10^{-10} (il s'agit donc d'un bruit de reconstruction inférieur à -200 dB!), sachant que pour des filtres de longueur suffisamment grande (supérieure à 200) il peut être utile —afin de calculer proprement le filtre passe-haut associé— d'aller jusqu'à 10^{-14}.

3. Convergence

Il n'est pas question ici de démontrer que cet algorithme converge. Le problème n'a pour l'instant pas reçu de réponse de ma part. Empiriquement, on constate que l'on peut itérer jusqu'à des valeurs d'erreur de l'ordre de la sensibilité de l'ordinateur, ce qui signifie qu'il converge informatiquement…

Ce résultat est certes très surprenant quand on prend en compte le nombre de minima que peut comporter la fonctionnelle à minimiser: on sait d'ailleurs que dans le cas dyadique la minimisation au sens L^2, ou au sens L^∞ conduisent en général, pour un filtre de degré L, à 2^L solutions ayant la même performance. La seule chose que l'on puisse dire, est que la solution limite de l'algorithme n'est pas minimale au sens L^2 ou L^∞ comme on le verra plus loin: elle est simplement très proche de la solution de la norme L^2 à phase minimale, dans le cas dyadique, sans lui être égale.

Le "miracle" de la convergence s'arrête quand on impose plus d'un facteur de régularité, sauf dans le cas dyadique où l'on peut même calculer les filtres orthonormés de Daubechies (de longueur minimale pour un nombre de facteurs de régularité donné). Il semble également que quand l'atténuation du filtre limite dépasse une valeur de l'ordre de 80 à 100 dB, l'algorithme ait plus de mal à converger, sans doute pour cause de problèmes numériques.

En tout état de cause, la convergence est assez générale et se fait, comme prévu de façon exponentielle. Il est ainsi nécessaire de faire entre 40 et 100 itérations pour obtenir une erreur de reconstruction inférieure à 10^{-10}, ce qui, suivant la taille des filtres signifie entre quelques secondes et quelques minutes pour calculer un filtre sur un Macintosh Quadra 700.

Enfin l'initialisation ne joue aucun rôle, même si un "mauvais" G_0 va retarder de quelques itérations le résultat.

4. Calcul du filtre passe-haut

Comme dans le cas dyadique, on peut mettre au point une formule donnant directement le filtre passe-haut à partir du passe-bas, c'est le résultat (II.11). Cependant, comme cette méthode donne un filtre passe-haut trop grand, il est plus avantageux de décomposer en éléments simples la matrice polyphase $\mathbf{G}$ issue du filtre passe-bas: on obtiendra alors une matrice rectangulaire constante de taille $q \times p$ $\mathbf{M}_0$ et et K éléments paraunitaires simples de la forme (III.3), dont le produit donnera $\mathbf{G}$

$$\mathbf{G} = \mathbf{M}_0 \mathbf{P}_0 \mathbf{P}_1 \ldots \mathbf{P}_{K-1}$$

On orthogonalisera alors la matrice $\mathbf{M}_0$ en prenant le produit vectoriel normalisé des q lignes de cette matrice, l'arbitraire étant ici réduit au signe, ce qui n'a pas de conséquence sur les caractéristiques du filtre passe-haut. On pourra donc en déduire la matrice polyphase totale du système, et par voie de conséquence le filtre passe-haut.

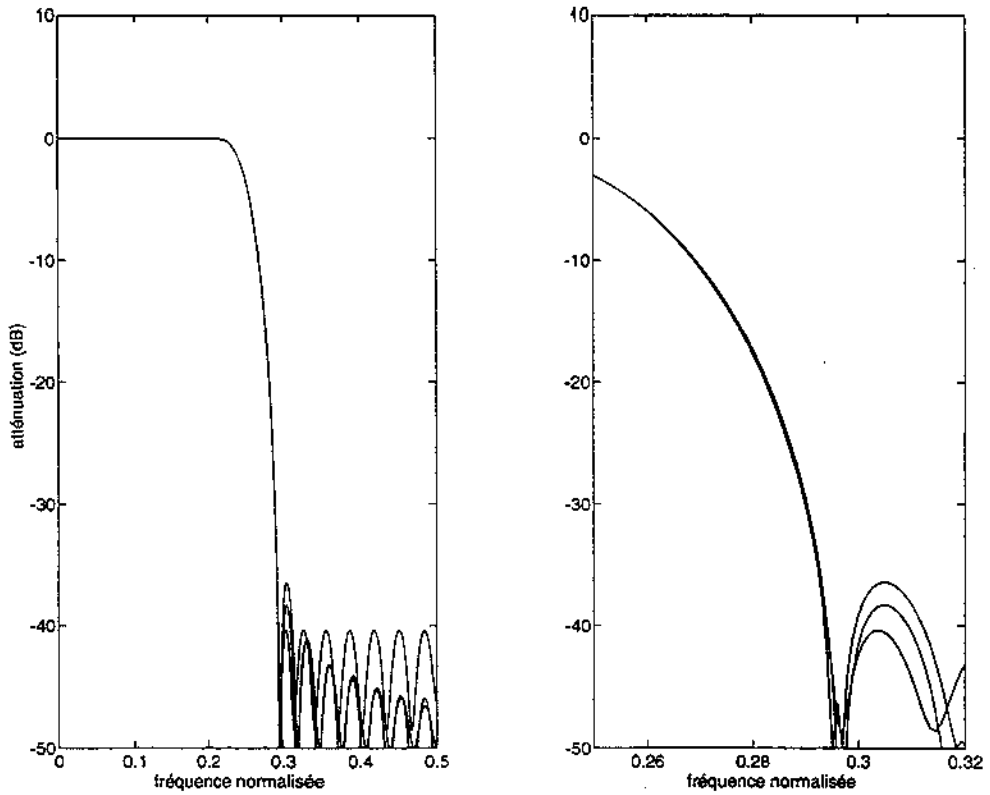
Le seul point un peut délicat concerne donc la décomposition de la matrice **G** en éléments simples, dans la mesure où notre filtre G ne vérifie pas exactement la condition de reconstruction parfaite. Cependant une erreur de 10^{-10} est apparue jusque là totalement indolore (du moins pour des filtres de longueur < 150) sur cette décomposition.

5. Résultats

a. Comparaison avec le cas dyadique

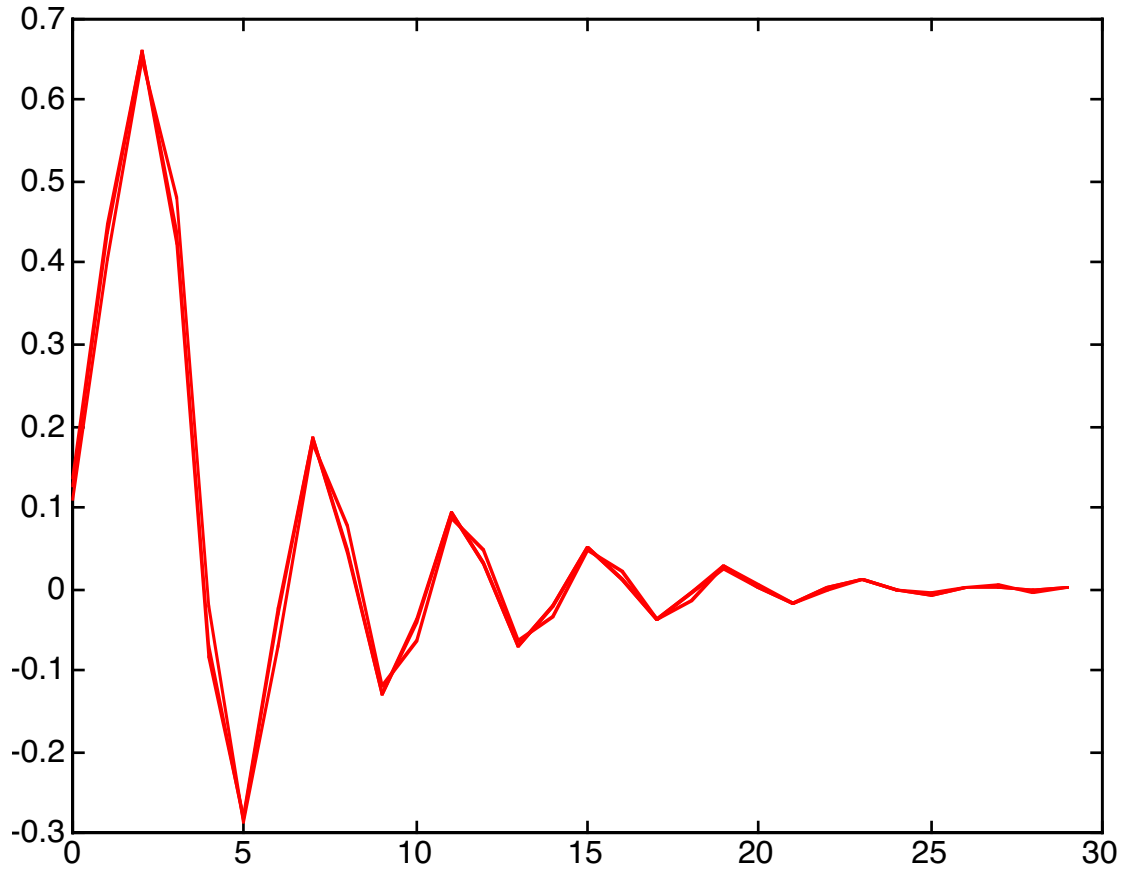
On va d'abord voir que les résultats ne sont pas ridicules dans le cas dyadique par rapport aux solutions indiquées dans la section "solutions classiques". La conclusion est ici que pour l'atténuation au sens L^∞ l'algorithme semble meilleur que la minimisation au sens L^2, et est bien sûr moins bon, mais de seulement quelques dB (3 ou 4 selon les cas), que la solution de Tchebytcheff.

L'exemple ci-dessous a été obtenu pour une valeur de ν_0=0.2902 dans l'algorithme et pour un filtre de degré 29. On a également calculé la solution de Smith et Barnwell, où l'on a fait ν_0=0.294, ainsi que la solution de la norme L^2 où ν_0=0.29: il était bien sûr nécessaire de changer la fréquence de début de la bande atténuée dans les fonctionnelles à minimiser pour pouvoir comparer les filtres au sens de la norme L^∞.



Les résultats on été indiqués sur une même courbe grossie au voisinage de la bande de transition (à gauche et à droite). On constate ainsi que la performance de l'algorithme se situe entre celle de la norme L^∞ et de la norme L^2. Dès le deuxième rebond, l'algorithme (ainsi que la norme L^2) sont meilleurs que la norme L^∞ pour atteindre plus de 6 dB à la fréquence de Nyquist.

Finalement, quand on compare les coefficients du filtre obtenu avec ceux issus de la minimisation L^2 et L^∞, on constate que ceux qui s'approchent le plus de notre résultat sont les filtres à phase minimale, c'est-à dire ayant toutes leurs racines à l'extérieur du —ou sur le— cercle unité comme on le voit sur la figure suivante

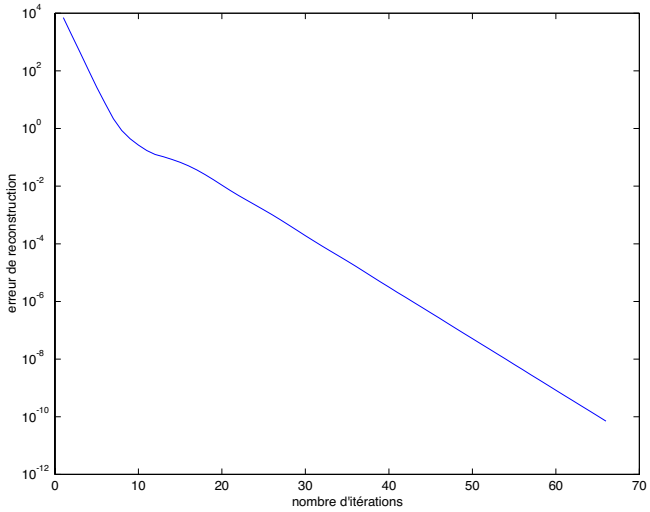


où l'on a tracé les coefficients du filtre issus de l'algorithme et ceux des filtres à phase minimale L^2 et L^∞

b. Vitesse de convergence

Ainsi qu'on l'a dit, un tel algorithme de point fixe, s'il converge, doit en général le faire à une vitesse exponentielle et c'est effectivement ce que l'on observe. Cette vitesse observée varie relativement peu quand on change les paramètres (longueur du filtre, fréquence de début de la bande atténuée, facteurs p et q=p-1, ajout d'un facteur de régularité). On peut empiriquement voir que cette erreur vérifie la relation approchée: erreur $\propto 0.7^{\text{itération}}$.

Ceci est illustré sur le graphique suivant qui correspond au calcul d'un filtre de degré 74, p/q=5/4, un facteur de régularité et une fréquence de début de la bande atténuée égale à 0.12.

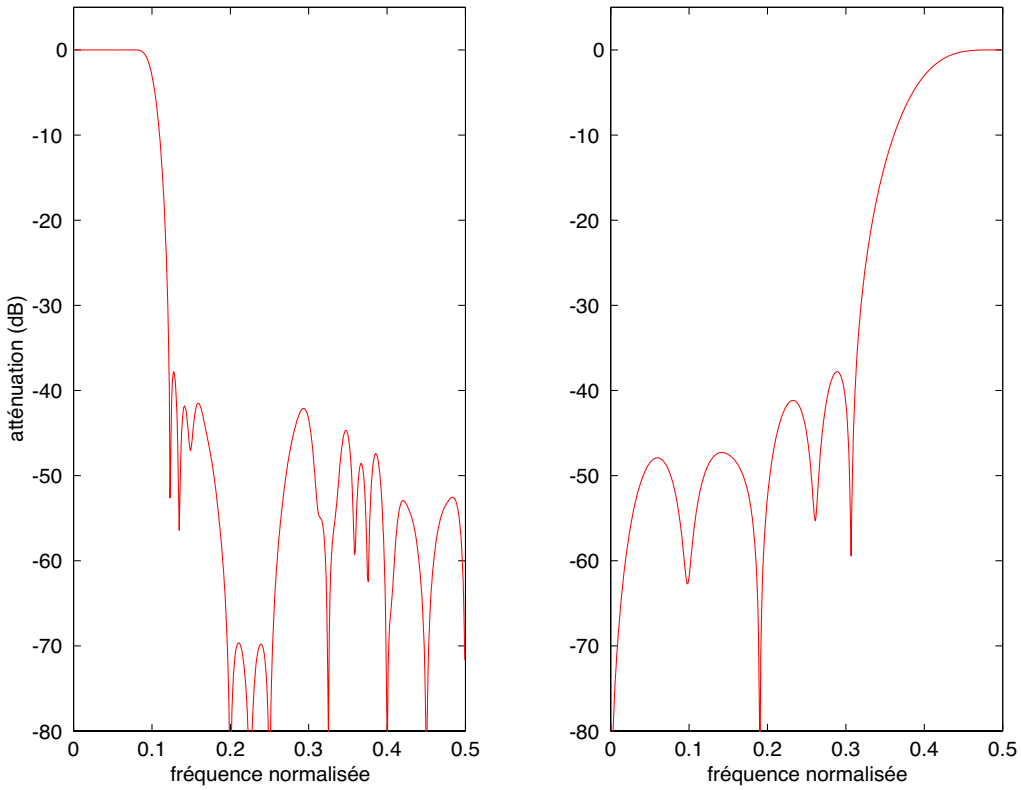


c. Exemples de filtres

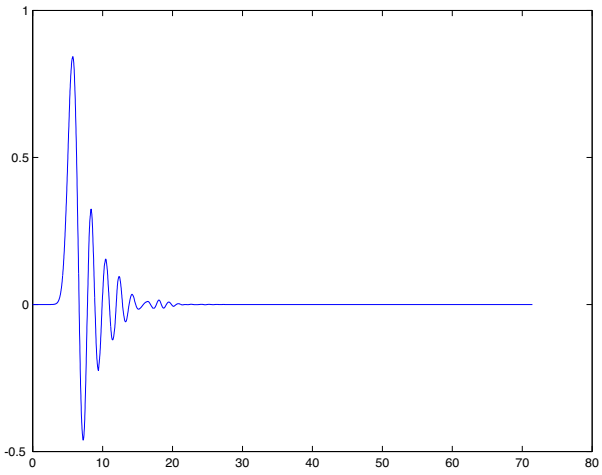
On revient au filtre de degré 74 correspondant aux paramètres suivants: p/q=5/4, ν_0=0.12, 1 facteur de régularité. L'algorithme nous donne les coefficients suivants pour le filtre passe-bas, dont on déduit le passe-haut

coeff	filtre passe-bas	filtre passe-haut	coeff	filtre passe-bas
0	2.866680073140252e-02	3.043765937829996e-03	38	3.731207278697348e-02
1	8.307565196587664e-02	9.942848160474251e-03	39	4.728203502094577e-02
2	1.833576579070865e-01	-2.047556246692463e-02	40	3.602311786843183e-02
3	3.354816584645757e-01	7.570553244406282e-03	41	1.574360861104994e-02
4	5.137042528543457e-01	3.259562165862628e-02	42	-9.680485355962614e-03
5	6.822439599643735e-01	-8.755951062208622e-02	43	-2.697942620785776e-02
6	8.086477715339978e-01	1.097308220904675e-01	44	-2.772919758379506e-02
7	8.505576725731043e-01	-5.964769926994694e-02	45	-1.968940860352614e-02
8	7.827068808419096e-01	-8.023030731132601e-02	46	-4.559662996115635e-03
9	6.067203729481050e-01	2.718519567549769e-01	47	1.101036127966241e-02
10	3.572660979326854e-01	-4.450758009655909e-01	48	1.582640054850962e-02
11	7.902667407726788e-02	5.226622138998910e-01	49	1.527267178780935e-02
12	-1.693179286316963e-01	-4.887432288100219e-01	50	7.585766307220270e-03
13	-3.290198162002506e-01	3.655573204208893e-01	51	-2.655450184545246e-03
14	-3.652956400613206e-01	-2.081658554505313e-01	52	-6.458787761064165e-03
15	-2.929498990147362e-01	8.495292883143664e-02	53	-8.722054415781014e-03
16	-1.454633472562613e-01	-2.048872555470213e-02	54	-6.038875205716814e-03
17	3.077450190143085e-02	2.625153807275018e-03	55	3.402342009871640e-04
18	1.676233150074401e-01	-1.464943314660177e-04	56	1.557713567652133e-03
19	2.276171580255633e-01		57	3.559488977883038e-03
20	2.044569579983413e-01		58	3.455585915535573e-03
21	1.210708483495994e-01		59	-1.898646179552138e-05
22	-4.058173403702415e-05		60	-1.995847761173937e-04
23	-1.014372128476964e-01		61	-8.584682574928117e-04
24	-1.513122086465713e-01		62	-1.410231969686158e-03
25	-1.441423693182781e-01		63	1.533002932531624e-13
26	-8.490765081731394e-02		64	1.113764773213890e-05
27	-1.908637445935444e-03		65	1.099927474046723e-04
28	6.933134546690430e-02		66	3.401160641877199e-04
29	1.061776416223847e-01		67	-6.256672849226116e-14
30	9.637934431277774e-02		68	-3.870370590600021e-17
31	5.428804864884577e-02		69	-6.138045684219960e-06
32	-2.890901807673164e-03		70	-4.357796576629327e-05
33	-5.115155834853503e-02		71	1.508208832645783e-14
34	-7.195946474420029e-02		72	1.846555513594087e-17
35	-6.162024136811259e-02		73	2.132991819395787e-17
36	-3.087866232336277e-02		74	2.431828925181656e-06
37	6.875063066508935e-03			

Les filtres correspondants sont tracés ci-dessous. La convergence vers une erreur inférieure à 10^{-10} a été obtenue en 66 itérations, à partir d'un filtre initial quelconque (créé par une fonction aléatoire…). L'amnésie du filtre est évaluée à 0.029, c'est-à dire une contribution à la sélectivité σ de 30 dB environ.



Enfin, on a tracé la fonction φ: on constate que malgré sa longueur de 75, son support pratique est plus proche de 15. On connaissait déjà cette particularité dans le cas dyadique avec, par exemple les ondelettes orthonormées de Daubechies. Ceci constitue un réel problème dans la mesure où une telle longueur de filtre correspond à un délai de transformation assez élevé (par exemple, dans le cas d'un signal sonore où l'on analyserait les fréquences de 500 Hz à 16000 Hz, le délai sera de 74×32 échantillons, soit 74 ms ce qui limite les applications interactives), et il est désagréable de savoir que ce délai est gaspillé aux quatre/cinquièmes par la faiblesse des valeurs de la fonction limite.



On pourrait penser qu'un tel défaut pourrait être levé par l'utilisation d'autres filtres, en particulier non orthonormaux. Un calcul heuristique simple montre qu'il n'en est malheureusement rien. En effet, supposons que l'on souhaite obtenir, à l'issue du processus d'itérations une fonction moyenne φ de support effectif L. On souhaite bien sûr avoir le meilleur comportement fréquentiel possible pour cette fonction passe-bas, c'est-à dire qu'elle sera de module approximativement 1 sur sa bande passante et observera une bande de transition de l'ordre de $1/L$. D'après (IV.13), la fonction $\frac{1}{p}G(e^{-2i\pi\nu/p})$ devra obéir aux mêmes contraintes, ce qui signifie que la bande de transition du filtre G sera en fait de $1/pL$ ce qui est bien plus étroit. Le filtre G devra donc être de longueur approximativement pL afin que φ ait les bonnes caractéristiques définies ci-dessus. En conséquence, le support de φ sera en réalité de l'ordre de pL bien que son support effectif soit p fois plus petit…

On peut voir cela d'une autre manière encore: le banc de filtres itéré n'est pas la manière la plus économique —en terme de longueur de filtre— d'obtenir un banc de filtres. Prenons en effet le passe-bas et ses itérés G_j. D'après l'équation de récurrence (IV.3), si G_j a une bande passante de $[-\frac{1}{2p^j}, \frac{1}{2p^j}]$ et une bande de transition de taille $\delta\nu_j$ alors G_{j+1} aura une bande passante de $[-\frac{1}{2p^{j+1}}, \frac{1}{2p^{j+1}}]$ et une bande de transition de taille $\delta\nu_{j+1} = \frac{1}{p}\delta\nu_j$. Ainsi, si G de longueur L a une bande de transition de $\delta\nu$ —et vérifie toutes les bonnes conditions de régularité et d'amnésie— on aura $\delta\nu_j \cong p^{-j+1}\delta\nu$. La longueur L_j du filtre itéré G_j sera, elle, de $L(p^j\text{-}q^j)$. On sait que si le filtre est bien optimisé, on a $L\delta\nu \cong 1$ mais on constate que $L_j\delta\nu_j \cong pL\delta\nu$ pour j suffisamment grand, ce qui signifie que le filtre itéré est p fois trop grand, ce qui évidemment induit un délai p fois trop important…

On touche donc là un des problèmes importants dus aux bancs de filtres itérés: le manque d'efficacité fréquentielle impliquant pour une sélectivité donnée un nombre plus important de calculs à effectuer et surtout un délai plus grand. Il n'en reste pas moins qu'il s'agit là de la manière la plus simple de réaliser des transformations non uniformes.

6. Remarques

Les solutions obtenues par cet algorithme sont comme on l'a vu à phase minimale quand on reste dans le cas dyadique. Cela semble également être le cas pour d'autres valeurs de p et q, mais on ne se hasardera pas à essayer de le démontrer: l'expérience apprend simplement que les racines des polynômes obtenus semblent être sur, ou à l'extérieur du cercle unité.

On constate par ailleurs que quelle que soit la longueur souhaitée pour le filtre, on obtient toujours à la fin le filtre dont la longueur est le multiple de p immédiatement inférieur ou égal à la longueur requise. En pratique donc, on limite le degré du filtre à être un multiple de p moins 1.

Par ailleurs, dans la mesure où les programmes d'extraction de racines deviennent lents et imprécis quand il y a trop de racines à retrouver, on peut tout simplement utiliser notre algorithme pour obtenir des filtres de très bonne qualité dans le cas dyadique, ceci d'autant plus que l'on peut alors rajouter des facteurs de régularité à volonté. La conception étant, comme on l'a vu, orientée vers la minimisation de l'erreur de reconstruction, on pourra ainsi obtenir de bons filtres avec une erreur de reconstruction inférieure par exemple, à 10^{-10}.

a. Degré minimum

On peut s'intéresser aux filtres de degré minimum permettant de remplir les conditions de régularité associées à l'orthonormalité. Dans le cas dyadique, on peut traiter ce cas complètement, ce qui conduit aux ondelettes orthonormées de Daubechies [Dau1]. Le problème est malheureusement rendu ici beaucoup plus ardu du fait qu'il n'existe plus *une seule équation* (VI.2) définissant l'orthonormalité, mais exactement q de ce genre (premier système d'équations de (II.9)). Je n'ai pour l'instant pas de solution à ce problème, et ne peux même pas répondre à la question de l'existence de tels filtres *à coefficients réels*.

On peut cependant déterminer le degré minimum du polynôme candidat-solution à ce problème. C'est tout simplement celui qui impose l'égalité entre le nombre de coefficients du filtre et celui des contraintes. Dans ce cas l'algorithme donne tout le temps la même solution, indépendamment de la fréquence ν_0 donnée (bien sûr, cette remarque n'a d'importance pratique dans l'algorithme que lorsque q=1, seul cas où l'on peut imposer plus d'un facteur de régularité).

Voyons donc quelle est cette valeur minimale. Soit N le degré de G et faisons le compte des équations non redondantes du premier système d'équations de (II.9). Comme on le sait, à cause de la symétrie $z \text{—>} z^{-1}$, on ne doit considérer que les équations correspondant à des puissances positives de z. D'autre part, ces équations ne spécifient que les valeurs des coefficients des puissances de z multiples de p, c'est-à dire que, comme la puissance la plus élevée de $G(z)G(z^{-1}e^{2i\pi s/q})$ est N, on a $E(N/p)$+1 équations quadratiques pour chaque valeur de s. La régularité de son côté impose $K(p\text{-}1)$ équations linéaires, soit au total

$$qE\left(\frac{N}{p}\right) + q + K(p-1) \text{ équations}$$

dont on peut espérer l'indépendance. Le nombre des inconnues est lui de N+1, ce qui implique que le degré N minimum satisfaisant ces équations doit obéir à la double inégalité suivante

- $N+1 \geq qE\left(\frac{N}{p}\right) + q + K(p-1)$: nombre d'équations inférieur ou égal au nombre d'inconnues
- $N < qE\left(\frac{N-1}{p}\right) + q + K(p-1)$: si le degré est N-1, le système est surdéterminé

ce qui implique les égalités suivantes

$$\begin{aligned} N+1 &= qE\left(\frac{N}{p}\right) + q + K(p-1) \\ &= qE\left(\frac{N-1}{p}\right) + q + K(p-1) \end{aligned}$$

En développant les inégalités qui définissent la fonction partie entière, on trouve que N peut prendre plusieurs valeurs. Définissons en effet λ par l'équation $q\lambda = N+1-K(p-1)$ alors les conditions que l'on a imposées équivalent à

$$E\left(\frac{K(p-1)+p-q-1}{p-q}\right) \le \lambda \le E\left(\frac{K(p-1)+p-2}{p-q}\right)$$

Comme on s'intéresse au degré minimum, on a finalement

$$N = K(p-1)-1+qE\left(\frac{K(p-1)+p-q-1}{p-q}\right) \qquad \text{(VI.14)}$$

et dans le cas p-q=1 que l'on considérera en général, cela se simplifie en

$$N = Kpq-1 \qquad \text{(VI.15)}$$

On retrouve bien sûr le résultat connu du cas dyadique.

D. Résumé du chapitre

Étant donné la faible publication qui a jusqu'à présent entouré les bancs de filtres rationnels, il n'existait pas jusqu'à présent d'algorithme spécifique de conception de filtres pour ce cas particulier, ce qui a motivé la mise au point de l'algorithme que j'ai proposé [Blu3]. Ce chapitre est donc essentiellement dévolu à la description, les indications et contre-indications de cet algorithme. Il serait bien sûr faux de dire que certaines techniques plus classiques ne permettent pas de concevoir de bancs de filtres rationnels. Cependant ces techniques —factorisation des matrices par exemple— sont beaucoup plus complexes à mettre en œuvre et nécessitent des programmes de minimisation suffisamment efficaces pour éviter la multitude de minima locaux qui peuvent les piéger. Finalement, de tels inconvénients deviennent probablement rédhibitoires quand il s'agit de concevoir les filtres de longueur très importante —par exemple de longueur 200—, nécessaires pour des facteurs d'échelle p/q proches de 1, ce qui n'est pas le cas pour l'algorithme présenté ici qui s'avère extrêmement robuste et rapide. Sur la seule comparaison que l'on puisse réellement faire, il donne également de meilleurs résultats.

VII. Application au codage de sons

Le but, dans ce dernier chapitre, est de montrer ce que l'on peut faire avec un banc de filtres itéré en fractions d'octave. Le choix de l'application se porte tout naturellement sur le codage de sons dans la mesure où l'on sait que l'oreille tend à décomposer le son qui lui parvient en bandes fréquentielles de largeurs inégales —bande critiques de Bark— [ZF]. En fait la largeur de ces bandes, pour des fréquences supérieures à environ 500 Hz, croît proportionnellement avec la fréquence de telle sorte que trois bandes critiques mises bout à bout occupent une largeur de bande d'une octave.

On précisera d'abord certains résultats relatifs aux propriétés psychoacoustiques de l'oreille qui ont déjà permis [Mah,MPC,DLR,ST,JB], par l'exploitation principale du phénomène de "masquage" fréquentiel, de proposer des algorithmes de codage de son haute fidélité à 128 kbit/s, voire 64 kbit/s sans détérioration subjective, soit un facteur de compression de 4 à 8, ceci à 32 kHz (à 44 et 48 kHz on obtient des taux de compression encore plus élevés). On verra que l'on peut utiliser d'autres propriétés de l'oreille, liées cette fois non pas aux caractéristiques mécaniques de la membrane basilaire mais plutôt aux capacités de décharge des neurones du nerf auditif.

On étudiera donc un modèle simplifié de transformation et codage des signaux sonores par l'oreille [YWS]: il ne s'agira pas bien sûr de reproduire exactement la réponse de l'oreille à tout signal, mais plutôt d'en dégager quelques grandes lignes qui seront particulièrement pertinentes en termes de réduction de débit. On verra enfin comment implémenter ce modèle à l'aide d'un banc de filtres rationnel itéré ainsi que le potentiel que l'on peut en retirer.

A. Résultats de psychoacoustique (tirés de [ZF])

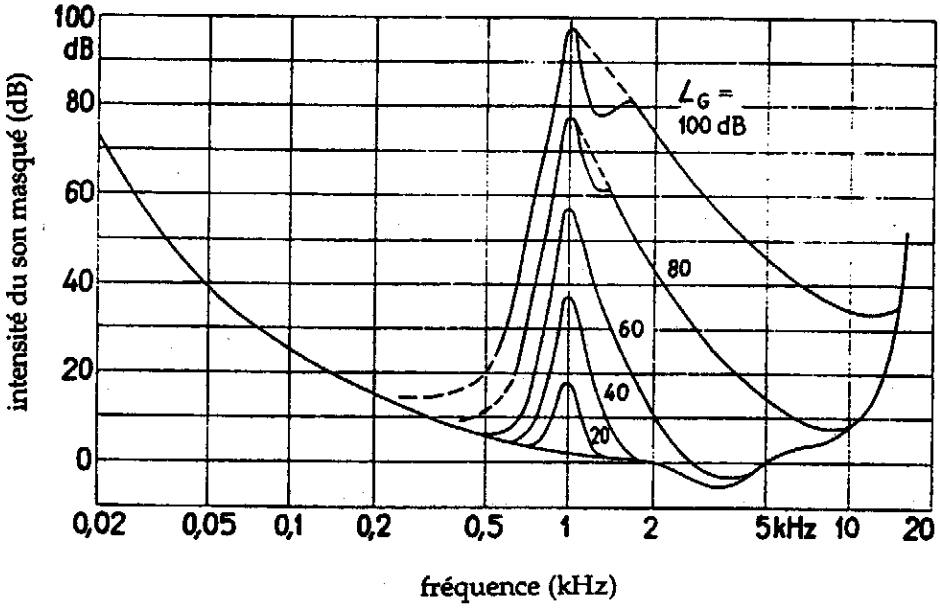
On va ici rappeler quelques résultats importants décrits par Zwicker et Feldtkeller concernant la perception auditive.

1. Bandes critiques

Exposons tout d'abord cette particularité qu'à l'oreille d'estimer la puissance sonore sur des bandes de largeur précise. Comme on le sait, l'oreille est susceptible de discerner jusqu'à une pression d'intensité aussi faible que $2\ 10^{-5}$ Pascal pour un son d'une fréquence de 1 kHz. Il se passe alors le phénomène suivant à cette fréquence précise: si au lieu de présenter un son pur, on en présente deux, de même intensité, à des fréquences différentes, alors la puissance du seuil de détection — c'est-à dire le carré de la pression— est divisé par deux si les deux fréquences se trouvent à l'intérieur d'une certaine bande de fréquence, alors que si elles sont trop séparées, ce seuil devient le minimum des seuils associés à chacune. En d'autre termes, à l'intérieur d'une certaine bande de fréquence, que l'on appelle bande critique, les deux sons sont traités ensemble, alors que s'ils sont trop séparés ils sont traités indépendamment (addition des puissances). On peut établir en chaque point de l'ensemble des fréquences une bande critique: on constate [ZF] que les bandes critiques sont approximativement de taille constante en dessous de 500 Hz et de taille proportionnelle à la fréquence au dessus de 500 Hz. Le facteur de proportionnalité entre les fréquences centrales est alors d'environ $2^{1/3}$ c'est-à dire que les bandes critiques décomposent le spectre auditif en tiers d'octave. De façon plus précise, ce facteur de proportionnalité est en fait assez proche du rapport d'entiers $6/5$.

2. Masquage fréquentiel

Quand on s'échappe des seuils de détection absolus, les bandes critiques apparaissent de façon plus précise sous forme de courbes de masquage. On constate en effet que la présence d'un son A peut empêcher d'entendre un autre son B, pourvu que l'intensité de ce dernier soit inférieure à une certaine quantité dépendant de l'intensité du son A et de sa fréquence. Ce seuil de détection "masqué" est en général bien supérieur au seuil de détection absolu de B. C'est cet effet que l'on appelle masquage d'un son par un autre, et l'on peut, en faisant varier la fréquence du son B, obtenir une courbe, dite de masquage dont cinq exemplaires pour des intensités différentes d'un son A à 1 kHz (il s'agit en fait d'un bruit de largeur de bande 160 Hz) sont données ci-dessous (extrait de [ZF])



Cette courbe met en évidence le fait que le masquage est maximal et approximativement constant dans la bande critique du son A, puis décroît, rapidement pour les fréquences inférieures à celle de A, et plus lentement pour les fréquences supérieures.

En fait, la présentation exposée ici est assez simpliste puisqu'elle ne tient pas compte du fait que le système auditif est capable de détecter avec une grande efficacité des sinusoïdes dans du bruit, et que la forme constante des courbes de masquage indiquée ici ne correspond pas au masquage d'un son pur par un autre son pur.

3. Masquage temporel

De la même manière qu'un son peut masquer l'autre en fréquence, un son peut en masquer un autre s'il le précède, ou même s'il lui est postérieur. À ce sujet, les résultats sont plus complexes que pour le cas fréquentiel car on a beaucoup plus de paramètres à manier, dans la mesure où les signaux qui nous intéressent sont alors transitoires, et non stationnaires.

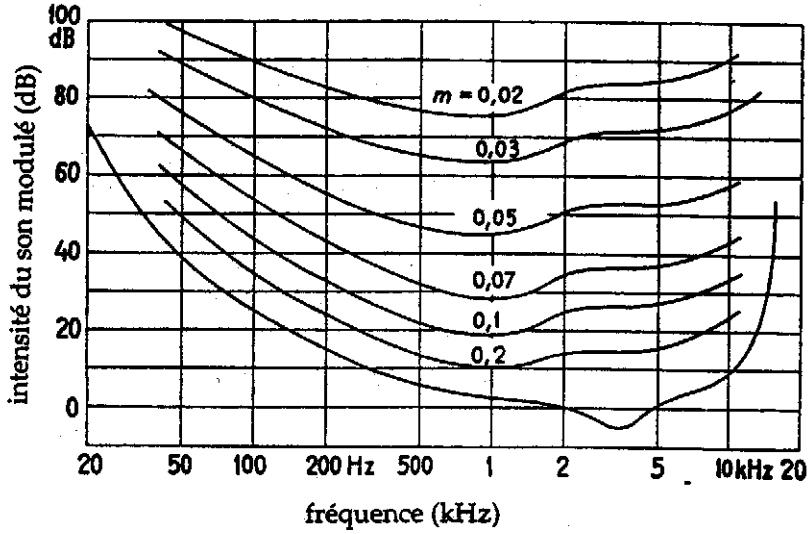
4. Sensibilité fréquentielle

En schématisant encore une fois les résultats réels de [ZF], on peut estimer la sensibilité fréquentielle de l'oreille, c'est à dire sa capacité à comprendre que deux sons sont très proches. L'une des expériences menée par [ZF] montre que la sensibilité de l'oreille pour un son de 70 dB est quasiment constante à 2 Hz pour des fréquences inférieures à 500 Hz, puis croît linéairement avec la fréquence avec une pente d'environ 0.0035.

Bien sûr ces seuils s'élèvent quand l'intensité du son baisse.

5. Quanta d'intensité perçue

Comme tout objet physique, l'oreille n'a pas une sensibilité infinie ce qui signifie en particulier qu'elle ne pourra pas détecter de variation d'amplitude d'un signal sinusoïdal inférieure à une certaine quantité, dépendant du niveau du signal. Il s'agit là d'un effet non linéaire dont on pourra tirer profit par la quantification lors de la compression. Pour évaluer cette sensibilité, [ZF] ont fait des expériences mesurant le seuil de détection de la modulation en amplitude d'une sinusoïde, c'est-à dire qu'ils ont cherché les valeurs $m(f,A)$ minimales telles que la modulation $A\big(1+m(f,A)\cos(\omega t)\big)\cos(ft)$ soit sensible (ω a été choisi égal à 4 Hz). Ce qui donne les courbes suivantes [ZF]



On observe ainsi qu'à la fréquence la plus sensible, 1 kHz, la fonction de modulation m varie de 0.2 à 0.01 ce qui donne sur une échelle de 100 dB un maximum de 160 niveaux discernables.

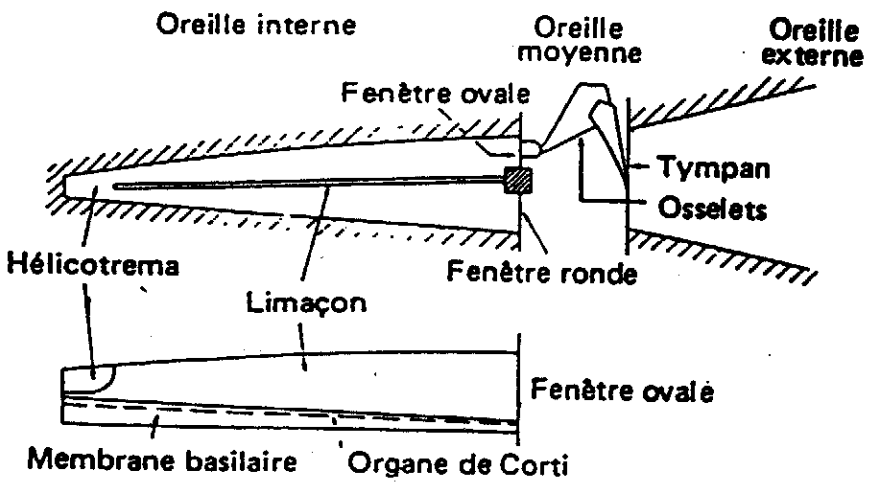
6. Effets non-linéaires

Il existe bien évidemment des effets non linéaires dans l'oreille qui rendent son étude plus complexe. Le plus simple est bien sûr le seuil absolu de détection en fonction de la fréquence, et participe au fait que pour certains phénomènes —par exemple effet de masque—, l'augmentation du niveau sonore donne une courbe qui n'est pas simplement décalée de celle à un niveau plus bas.

Il y a aussi le fait que l'oreille détecte mieux les sons sinusoïdaux purs que les bruits, et a tendance à en créer des harmoniques. Le problème dans l'étude des effets non-linéaires de la perception auditive tient au fait que ces effets semblent être hautement variables en fonction des cobayes utilisés (chapitre XX) et ne permettent donc pas d'étude quantitative globale.

B. Description de l'oreille interne

Après les résultats issus de tests d'audition, nous nous intéressons maintenant directement à la physiologie de l'oreille. Le but sera de mettre en évidence la partie transformation, la partie quantification et la partie codage, sachant qu'une part nécessairement non négligeable du traitement de l'information restera cachée dans l'activité cérébrale. Une fois ces parties identifiées, on proposera dans la section suivante un modèle relativement simple qui tentera de faire converger à la fois les caractéristiques physiologiques et les résultats de psychoacoustique. Ci-dessous, tiré de [ZF] on a reproduit une figure simplifiée

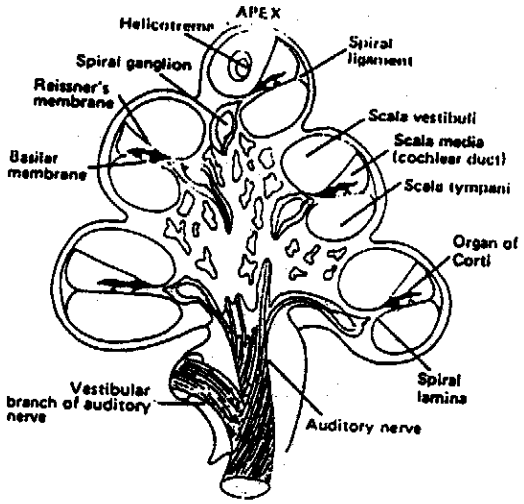


indiquant le trajet d'une onde sonore dans l'oreille: elle est d'abord captée par le pavillon (oreille externe), puis met en vibration le tympan qui lui-même met en branle les osselets (oreille moyenne) dont le rôle est de transformer les ondes de pression gazeuses en ondes de pression acqueuse dans l'oreille interne: cette dernière partie de l'oreille baigne en effet dans un liquide lymphatique.

L'oreille interne comporte deux parties dont l'une —le vestibule—, seulement osseuse, est plutôt reliée à l'équilibre, alors que l'autre —la cochlée, ou limaçon—, à la fois osseuse et membraneuse est directement impliquée dans le processus auditif. On ne s'intéressera donc désormais qu'à la cochlée et ses constituants.

1. La cochlée

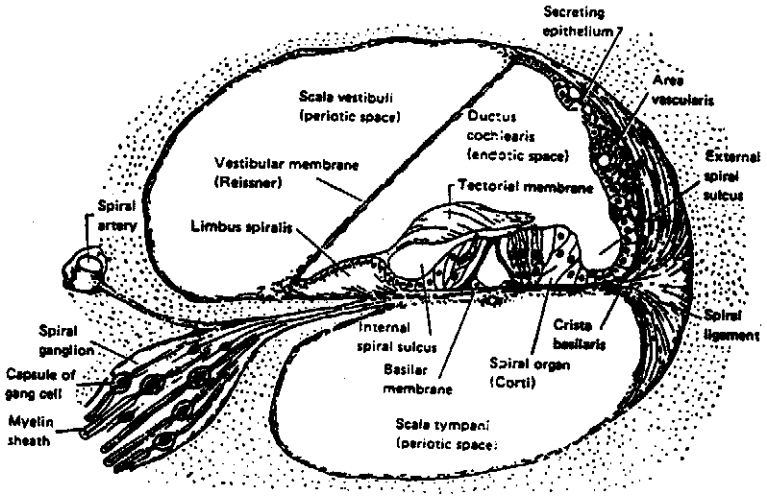
Dans la figure reproduite plus haut, la cochlée avait été étirée pour indiquer ses constituants internes. En fait, elle a la forme d'un limaçon osseux d'une longueur de 35 mm, comme indiqué sur le graphique ci-dessous (extrait de [Zem])



C'est à l'intérieur du conduit cochléaire que l'on trouve les "senseurs de pression" qui sont à la base du phénomène de l'audition.

2. Les membranes basilaire et tectorielle

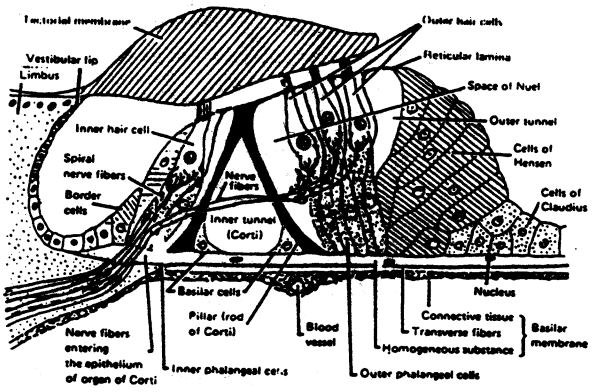
On a représenté la cochlée en coupe transversale (tiré de [Zem]). Elle fait apparaître en particulier les membranes basilaire et tectorielle, l'organe de Corti qui supporte les cellules ciliées ainsi qu'une branche du nerf auditif.



La membrane basilaire fait l'interface entre l'organe de Corti et la paroi osseuse de la cochlée, alors que la membrane tectorielle se place au dessus de l'organe de Corti, semblant protéger les cellules ciliées. Soumise aux variations de pression, la membrane basilaire verra se développer des ondes de propagation qui s'étaleront sur toute sa longueur. Le fait que la largeur de la membrane augmente quand on s'approche du sommet du limaçon pourra expliquer la forme de l'analyse fréquentielle effectuée par l'oreille. En effet, on constate que si l'on fait varier la fréquence du son excitant, le lieu du maximum d'amplitude de l'onde de propagation se déplace de façon proportionnelle le long de la membrane basilaire.

3. L'organe de Corti et les cellules ciliées

L'organe de Corti est décrit dans la figure suivante (tiré de [Zem])



qui met en évidence la présence des cellules ciliées. Celles-ci sont accrochées à l'organe de Corti et, pour certaines d'entre elles, à la membrane tectorielle. Ces cellules sont composées de deux types, celles qui sont les plus proches de l'axe du limaçon et que l'on appelle internes, et celles qui sont au delà du pilier et sont appelées cellules ciliées externes. Ces cellules sont reliées au nerf auditif par des neurones, et l'on peut estimer que ce sont elles qui mesurent les variations de pression soit par leur sensibilité aux actions d'extension quand elles sont accrochées à la membrane tectorielle, soit directement par leur mouvement dans le liquide perturbé. Au total il y a environ 3500 rangées, composées de 3 cellules ciliées externes et d'une cellule ciliée interne, réparties de manière uniforme sur l'organe de Corti [Zem].

Jusqu'au niveau des cellules ciliées, on peut considérer que nous sommes dans la phase transformation de notre signal sonore, une transformation qui sera considérée comme linéaire dans le modèle qui sera exposé plus loin.

4. Le nerf auditif et ses fibres

Les cellules ciliées transforment apparemment l'énergie des actions mécaniques auxquelles elles sont soumises en énergie électrique, récupérée ensuite par les neurones du nerf auditif. Des mesures de potentiel électrique ont indiqué comment les cellules ciliées et le nerf auditif codent le signal sous forme de potentiels d'action. Il s'agit là de brusques décharges électriques dont la fréquence est étroitement lié à l'intensité du signal à coder. Au repos cependant, des décharges continuent à se produire à un rythme bien sûr plus faible, que l'on appelle cadence d'activité spontanée.

Au total, il y a environ 30000 fibres dans le nerf auditif de l'homme [Del], soit environ deux par cellule ciliée. On peut les ranger en trois classes [Del]: celles à faible cadence spontanée (<0.5 décharge par seconde) qui constituent environ 15% du nombre total de fibres, celles à cadence moyenne (entre 0.5 et 18 décharges par seconde) représentant 25% de l'ensemble, et celles à cadence élevée (>18 décharges par seconde) majoritaires à 60%. La cadence maximale est de l'ordre de 1000 à 2000 décharges par seconde, à cause des propriétés réfractaires des neurones [Del].

Ces fibres ont la particularité de répondre différemment suivant la fréquence du signal excitant: elles présentent une fréquence caractéristique pour laquelle elles sont beaucoup plus sensibles et dépassent rapidement leur seuil d'activité minimal. On peut alors tracer des courbes d'accord qui permettent de mesurer la sensibilité d'une fibre en fonction de la fré-

quence. On constate expérimentalement qu'une fibre de fréquence caractéristique donnée innervera précisément le lieu d'amplitude maximal de la membrane basilaire pour cette fréquence [Del]. Il y a donc dans l'oreille un double système de sélectivité fréquentielle.

La caractéristique principale des fibres du nerf auditif reste dans la nature aléatoire de leur réponse: la seule information pertinente est leur taux de décharges et il ne faut pas chercher de corrélation entre l'apparition d'une décharge sur deux fibres distinctes. On ne va pas rentrer dans le détail de la suppression de son, ou de l'adaptation [Del]: la réponse d'une fibre est fortement non-linéaire et adaptative ce qui peut expliquer certains phénomènes non-linéaires dans la perception auditive.

C. Modélisation du traitement du son par l'oreille interne

Après filtrage et amplification dans l'oreille externe, le son met en mouvement le tympan et les parties osseuses mobiles de l'oreille moyenne. Ces mouvements sont communiqués au fluide dans lequel baigne l'oreille interne: des ondes de pression se propagent alors depuis la fenêtre ronde jusque vers l'hélicotrème , c'est-à dire depuis l'entrée large de la cochlée jusque vers son sommet. Ces ondes mettent en mouvement les membranes basilaire et tectorielle qui tapissent l'intérieur du limaçon. Les cellules ciliées, soumises à la fois aux vibrations de la membrane basilaire (et pour certaines également à celles de la membrane tectorielle) et aux ondes de pression du liquide cochléaire, en récupèrent l'information sous forme de courants ioniques qui seront transformés en série de potentiels d'action ("spikes") par les neurones qui leur sont reliés. Ces potentiels seront enfin véhiculés jusqu'au cerveau par les fibres du nerf auditif [YWS]. De toutes ces étapes, la première nous intéressera plus particulièrement car on pourra la mettre directement en relation avec le phénomène de masquage fréquentiel décrit plus avant. Les deux autres étapes pourront être plus sommairement traduites par une quantification non-linéaire et éventuellement un sous-échantillonnage.

1. Le mouvement de la membrane basilaire

Il a en fait été constaté que le déplacement de la membrane basilaire est extrêmement faible [ZF] par rapport à ses dimensions —de la taille de quelques diamètres d'atomes!— ce qui autorise à considérer que la fonctionnelle qui relie les variations de pression $\Delta p(t)$ du fluide à l'entrée de l'oreille interne aux mouvements basilaires $\Delta x(l,t)$ est linéaire, se limitant au premier ordre [YWS]. Une telle hypothèse est remise en cause quand on considère des signaux de très faible intensité car alors la cochlée devient active afin de les amplifier [YWS]: cette hypothèse restera cependant valable pour les intensités que nous considérerons ici.

La fonctionnelle qui nous intéresse est donc de la forme (l étant le lieu de la membrane basilaire)

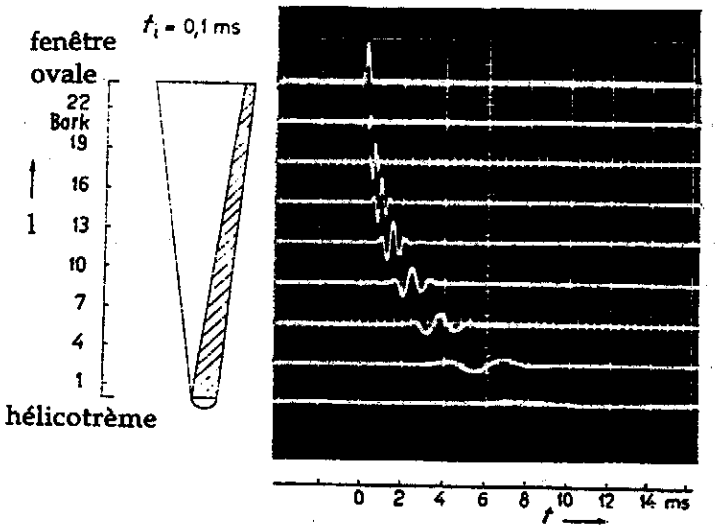
$$\Delta x(l,t) = \int K(l,t,t')\Delta p(t')dt'$$

en admettant en outre que seule la direction longitudinale de la cochlée mérite d'être prise en compte. Le fait que la membrane basilaire soit essentiellement passive —donc à réponse inva-

riante dans le temps— indique que l'opération linéaire indiquée plus haut est un filtrage temporel

$$\Delta x(l,t) = \int K(l, t-t')\Delta p(t')dt'$$

Par ailleurs, le destin d'une impulsion sonore sur la membrane basilaire ressemble fortement à une même fonction simplement dilatée et mise à l'échelle, au moins pour les fréquences élevées, comme on peut le voir sur cette figure, tirée de [ZF]



La transformation prend donc la forme d'une transformation en ondelettes

$$\Delta x(l,t) = b(l)\int \psi\left(\frac{t'-t}{a(l)}\right)\Delta p(t')dt' \qquad \text{(VII.1)}$$

Pour des fréquences inférieures à 500 Hz, il semble cependant que la transformation prenne plutôt la forme d'une transformée de Fourier à court terme. L'opérateur idéal devrait donc glisser continûment d'une transformation en ondelettes vers une transformation de type Fourier. On ne va désormais plus s'intéresser qu'aux fréquences supérieures à 500 Hz, qui par ailleurs contribuent de la façon la plus importante au débit d'une source sonore HiFi.

2. Seuillages/Quantification et codage

La formule (VII.1) ne donne que la valeur du déplacement en un point de la membrane basilaire. Que fait ensuite l'oreille de cette information?

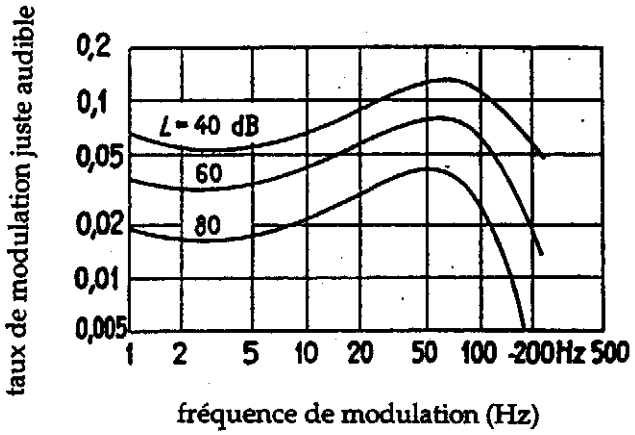
Comme on l'a vu plus haut, la réponse des cellules ciliées à un stimulus n'est pas totalement déterministe, mais comporte une part importante d'aléatoire. La précision viendra plutôt du fait qu'un nombre important de cellules ciliées réagissent au même stimulus, ce qui permet de faire une moyenne sur le nombre de réalisations de la variable aléatoire: c'est cette moyenne qui véhicule l'information "déterministe" sur le stimulus [Zem].

Tout se passe donc comme si une cellule ciliée seule ne permettait donc d'obtenir qu'une parcelle d'information, de la même manière que le bit ne prend sa signification que dans l'octet

dont il fait partie. La précision d'un tel système de senseurs dépend donc du nombre de cellules ciliées qui "tâtent" une fréquence donnée. On peut répondre à cette question en observant que l'oreille distingue effectivement 620 échelons de hauteur [ZF], d'où en moyenne 22 cellules ciliées par échelon et finalement environ 50 fibres du nerf auditif par échelon de fréquence.

Une particularité de ce type de codage est que pour les faibles amplitudes, le débit d'information délivré par les neurones est plus lent que pour les amplitudes plus élevées. D'après les graphiques donnés dans [Del] on remarque qu'un taux de décharges de 600 par seconde s'observe pour des signaux d'amplitude plutôt forte. Donc, afin de pouvoir discerner une centaine de quanta d'intensités différentes il va falloir que le cerveau collecte les informations à une vitesse inférieure à $600\text{x}50/100 \cong 300$ Hz.

Bien sûr ce calcul est seulement indicatif, puisque l'on a fait une suite de suppositions assez fausses comme celle concernant l'uniformité sur la cochlée de la répartition des fibres du nerf auditif de taux de décharges fixé. Il n'en reste pas moins que cette estimation est confirmée par une observation indépendante tirée de [ZF] où les auteurs donnent la forme du seuil différentiel d'intensité juste audible pour 1 kHz et à des intensités différentes (40, 60 et 80 dB), en fonction du taux de modulation



Bien que le but des auteurs ait été de justifier le choix de 4 Hz comme taux de modulation idéal, ces courbes peuvent aussi nous enseigner d'autres choses. On observe en effet qu'au-delà de 200 Hz, le signal modulé en amplitude est perçu comme une somme de —trois— sinusoïdes, alors qu'au voisinage de 4 Hz il est perçu comme la variation d'amplitude d'une sinusoïde unique. Cela fixe la limite extrême de la capacité du système auditif à poursuivre les variations d'un signal sonore à 200 Hz pour une fréquence de 1 kHz, ce qui n'est pas si éloigné de l'estimation donnée plus haut de 300 Hz.

Le fait que pour des faibles intensités le codeur auditif envoie moins d'information que pour les fortes intensités suggère un moyen simple de tirer profit de cette particularité dans un codeur numérique: le sous-échantillonnage comme fonction de l'intensité.

3. Seuil de détection et masquage fréquentiel

On part de la formule (VII.1) en supposant donc qu'*in fine*, la détection d'un son équivaut à la détection de ce son dans une partie significative de l'ensemble des $\Delta x(l,t)$, et que les variations de pression du liquide à l'entrée de l'oreille interne sont proportionnelles à ce son. Notons

que l'on peut supposer que le maximum de la transformée de Fourier de ψ est atteint pour la valeur 1 (sans unité) et que ce maximum est de module 1: $|\hat{\psi}(1)| = 1$.

Détection d'un son pur

Supposons que $\Delta p(t) = \Re\left(Se^{2i\pi ft}\right)$, alors on a

$$\Delta x(l,t) = \Re\left(a(l)b(l)\hat{\psi}\left(a(l)f\right)Se^{2i\pi ft}\right)$$

qui reste donc encore un son pur de même fréquence que le signal d'entrée. Ce signal sera détecté si et seulement si on peut trouver un lieu l tel que l'énergie moyenne de ce signal soit supérieure au seuil de détection local $\varepsilon(l)$. En fait, à cause de la sélectivité fréquentielle de ψ, le lieu principal de détection se trouve au voisinage de $l = a^{-1}(1/f)$. Le seuil de détection en fonction de la fréquence se décrit donc par l'inégalité

$$|S| \geq \sqrt{2} f \frac{\varepsilon\left(a^{-1}(1/f)\right)}{b\left(a^{-1}(1/f)\right)}$$

que l'on peut donc faire "coller" aux résultats de psychoacoustique afin d'obtenir le rapport ε/b.

Masquage d'un son pur par un autre

On suppose ici que $\Delta p(t) = \Re\left(S_0 e^{2i\pi f_0 t}\right) + \Re\left(S_1 e^{2i\pi f_1 t}\right)$ afin d'étudier le masquage du son pur de fréquence f_0 par celui de fréquence f_1. On suppose donc que $|S_0| >> |S_1|$. En outre, on va maintenant introduire une hypothèse de sensibilité, à savoir qu'un déplacement $\Delta x(l,t)$ au lieu l ne sera distingué d'un déplacement $\Delta x'(l,t)$ au même lieu qu'à la condition que l'énergie de la différence des deux déplacements soit suffisamment grande

$$\left|E(\Delta x' - \Delta x)\right| \geq \eta\left(l, E(\Delta x)\right)$$

En fait, on constate que η est approximativement linéaire (sensibilité logarithmique) et l'on peut ré-écrire cette inégalité comme

$$\left|E(\Delta x' - \Delta x)\right| \geq \eta(l) E(\Delta x)$$

où l'on peut considérer que η varie lentement en fonction de $E(\Delta x)$.

Dans le cas d'une somme de deux sons purs, les lieux principaux l_0 et l_1 à étudier sont ceux correspondant à f_0 et f_1. On obtient, pour les déplacements en l_0 et l_1

$$\Delta x(l_0,t) = \frac{b(l_0)}{f_0}\Re\left(\hat{\psi}(1)S_0 e^{2i\pi f_0 t} + \hat{\psi}\left(f_1/f_0\right)S_1 e^{2i\pi f_1 t}\right)$$
$$\Delta x(l_1,t) = \frac{b(l_1)}{f_1}\Re\left(\hat{\psi}\left(f_0/f_1\right)S_0 e^{2i\pi f_0 t} + \hat{\psi}(1)S_1 e^{2i\pi f_1 t}\right)$$

et il y aura donc masquage du son 1 par le son 0 dès que les deux inégalités suivantes sont simultanément vérifiées

$$\left|\psi^{\wedge}(f_1 / f_0)\right|^2 |S_1|^2 < \eta(l_0)|S_0|^2$$
$$|S_1|^2 < \eta(l_1)\left|\psi^{\wedge}(f_0 / f_1)\right|^2 |S_0|^2$$

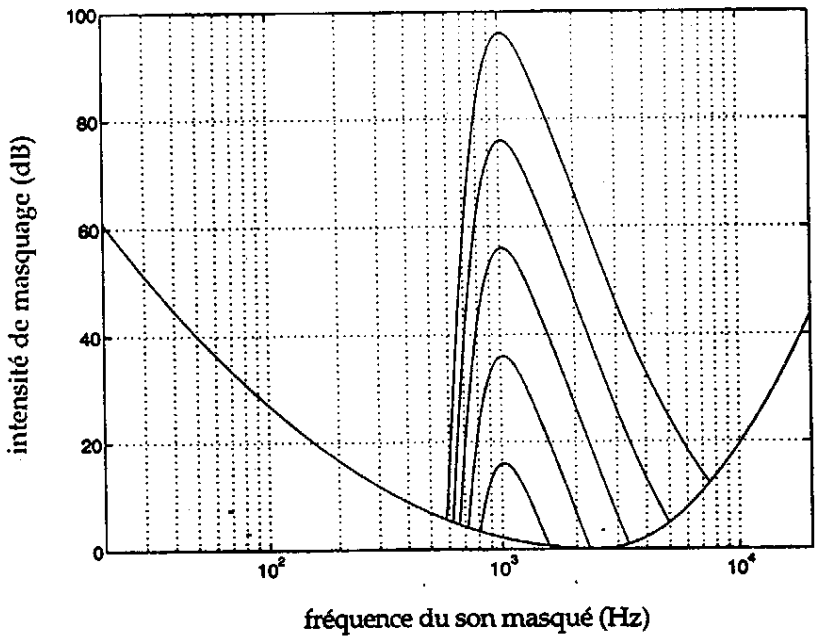
En fait, dès que la seconde inégalité est vérifiée, la première l'est automatiquement tant que les différences entre $\eta(l_0)$ et $\eta(l_1)$ ne sont pas trop importantes, et c'est donc la deuxième inégalité qui donnera les courbes de masquage fréquentiel (exprimées en dB ici)

$$M_{f_0}(f) = 10\log\eta + 20\log\left|\psi^{\wedge}(f_0 / f)\right| \qquad \text{(VII.2)}$$

De la forme des courbes de masquage on peut donc déduire le module de la transformée de Fourier de ψ. Il n'est ainsi pas nécessaire que la fonction $\psi^{\wedge}$ soit dissymétrique autour 1 pour expliquer la dissymétrie des courbes de masquage: cela est dû au simple fait que l'on ne s'intéresse pas exactement à $\psi^{\wedge}$ mais à $\psi^{\wedge}(f_0 / f)$ qui n'a bien sûr plus la symétrie de la fonction initiale. Par exemple, si l'on choisit une gaussienne centrée en 1 pour $\psi^{\wedge}$ on obtiendra des courbes de masquage très fortement dissymétriques. Cependant, notre but étant de nous rapprocher le plus possible des fonctions de masquage publiées par [ZF] il est sans doute préférable d'introduire de la dissymétrie. À titre d'exemple, une fonction $\psi^{\wedge}$ qui donne de bons résultats est la suivante

$$a(\nu) = \frac{1}{2}e^{-\frac{(\nu-0.9)^2}{2\times 0.185^2}} + e^{-\frac{(\nu-1)^2}{2\times 0.14^2}} \qquad \psi^{\wedge}(\nu) = a(\nu) + a(-\nu) \qquad \text{(VII.3)}$$

On a représenté ci-dessous les courbes de masquage correspondantes pour différentes intensités L=20, 40, 60 et 80 dB en insérant un modèle simple de seuil



En fait, on doit comparer ces courbes à celles du masquage d'un son pur par un bruit à bande étroite plutôt que par un autre son pur: cela est dû au fait que nous ne modélisons que la première partie du masquage et non pas les effets non-linéaires impliqués dans la reconnaissance privilégiée des sons purs.

On constate que pour les fortes intensités, elles sous-estiment le masquage. En fait, si l'on avait intégré le fait que η n'est pas constant, et dépend en particulier de $E(\Delta x)$ et que d'autre part, le lieu du masquage fréquentiel de f_1 par f_0 n'est pas nécessairement f_1, mais se trouve en général au-delà de cette valeur (pour $f_1>f_0$), on aurait également pu mettre en évidence une diminution de la pente de la courbe de masquage pour les hautes fréquences et pour des valeurs importantes du son masquant.

D. Techniques existantes de codage numérique de son HiFi

Ces techniques tirent profit du masquage fréquentiel. Par exemple, à un débit initial de 16 bits tous les 32000$^{\text{èmes}}$ de secondes en mono, c'est-à dire 512 kbits/s elles permettent sans aucune perte subjective de descendre à 128 kbits/s, voire d'atteindre 64 kbits/s avec une perte de qualité presque nulle. D'autres fréquences d'échantillonnage sont également utilisées: 44.1 kHz pour l'enregistrement sur disque compact, et 48 kHz pour l'enregistrement en studio, ces fréquences ne changeant que relativement peu le débit de sortie après compression grâce à la plus faible sensibilité de l'oreille aux très hautes fréquences (>16 kHz).

Dans tous les cas, une transformation de type banc de filtres est effectuée sur le signal, puis le niveau du masquage est estimé dans chaque bande, fournissant la base du quantificateur devant y être appliqué. Le calcul de la courbe de masquage s'effectue en réalisant une transformation de type Fourier —fenêtrée— sur une portion du signal d'une durée fixe de l'ordre de plusieurs dizaines de millisecondes. Enfin les contributions en énergie de chaque bande critique à la fonction de masquage sont simplement ajoutées [Mah]. En un peu plus de détails, voici les différents procédés utilisés.

Codeur de Yannick Mahieux [Mah]

Un codeur pour le son HiFi développé au CNET Lannion a fait l'objet d'une commercialisation sous le nom de HiFiScoop. À sa base, se trouve une MDCT (Modified Discrete Cosine Transform) avec 512 (pour un échantillonnage à 32 kHz) ou 1024 (pour un échantillonnage à 48 kHz) sorties. L'intérêt d'utiliser une MDCT plutôt qu'une DCT est d'obtenir une meilleure résolution fréquentielle en perdant l'avantage d'avoir des supports qui ne se chevauchent pas. Par ailleurs un algorithme rapide pour le calcul de cette MDCT existe [DMP] ce qui rend cette transformation aussi attractive qu'une transformée en cosinus.

Avoir autant de bandes de fréquence permet d'estimer directement la fonction de masquage associée et de pouvoir masquer un nombre important de coefficients. Il faut cependant noter que pour les basses fréquences, la résolution spectrale (32 Hz) est nettement supérieure à la résolution fréquentielle de l'oreille (environ 2 Hz) alors que pour les hautes fréquences elle est au contraire bien plus fine (50 Hz à 10 kHz). Elle n'est donc pas adaptée à la sensibilité fréquentielle de l'oreille, mais par contre sur-échantillonne les bandes critiques…

Le problème le plus ennuyeux lié à l'utilisation d'une taille de bloc unique pour estimer la fonction de masquage et pour effectuer la transformation est la perte de certaines évolutions rapides du signal, ce qui se traduit par le phénomène de "préécho": une brusque transition du signal ne sera pas restituée fidèlement après codage et décodage, l'énergie de la transition s'étalant alors sur toute la durée du bloc qui la contient. Une méthode à base de filtrage de Kalman est utilisée ici pour atténuer cet effet de préécho.

Un dernier point , qui est commun à toutes les transformations utilisées ici, est le fort délai nécessaire au traitement des données. Pour le codeur de Mahieux le délai à 32 kHz est de 80 ms et à 48 kHz, de 120 ms, ce qui rend la technique plutôt adaptée à la diffusion qu'à la communication bidirectionnelle.

MPEG Audio [IM]

Il s'agit là d'un procédé qui a été consacré sous la forme d'une norme choisie par l'ISO: une partie en a été développé au CCETT, à l'IRT (Institut für RundfunkTechnik à Munich) et chez Phillips (Hollande), dans le cadre du projet européen Eurêka 147-DAB [DLR]. Il est constitué de trois couches de complexité croissante permettant d'atteindre des qualités de codage variées, les deux premières étant issues du codeur Musicam [DLR].

À la différence du codeur de Mahieux, l'accent est mis ici sur la sélectivité plutôt que sur le nombre de raies spectrales. Le banc de filtres d'analyse utilisé est ainsi constitué de 32 sous-bandes, les filtres correspondants étant déduits par modulation d'un prototype unique de taille 512 ce qui signifie un recouvrement de 16 blocs de 32 échantillons. La couche III ajoute cependant à cette transformation la possibilité de rediviser les sous-bandes en 6 ou 18 sous-sous-bandes à l'aide de MDCT de tailles 12 ou 36. Ces 32 sous-bandes sont à rapprocher des 24 bandes critiques du système auditif. On remarque que, du fait que la transformée est uniforme, la largeur des bandes est nettement supérieure à celle d'une bande critique pour les basses fréquences et nettement inférieure pour les hautes fréquences.

Le calcul de la courbe de masquage est effectué en parallèle à l'aide d'une transformée de Fourier discrète sur 1024 points, ce qui permet de quantifier de façon adéquate les coefficients de chaque ligne spectrale.

Codeur hybride [JB]

Afin d'éviter les problèmes d'inadaptation de la transformée (qui la rendent inefficace dans les hautes fréquences —résolution trop grande induisant un délai important générateur de préécho— et de mauvaise qualité dans les basses fréquences —résolution trop faible—) il a été proposé [JB] de coupler un banc de filtres itéré sur 3 octaves avec des transformations de type Fourier le tout permettant d'obtenir 320 sorties fréquentielles avec une répartition plus proche de l'analyse réalisée par l'oreille. La fonction de masquage est calculée à partir des sorties du banc de filtres donnant ainsi des valeurs toutes les 2.6 ms pour les bandes de fréquences 12-24 kHz, 5.3 ms pour les bandes de fréquences 6-12 kHz, 10.6 ms pour les bandes de fréquence 3-6 kHz et 21.3 ms pour les bandes de fréquences 0-3 kHz. Ceci permet de conserver un certain caractère dynamique à la courbe de masquage, et permet de minimiser les effets de préécho.

En outre, une détection des sons purs est effectuée: en effet l'oreille masque différemment les sons purs des sons aléatoires.

Sinha et Tewfik [ST]

Tournant le dos aux transformations uniformes, [ST] proposent l'itération d'un banc de filtres en octaves à la fois sur le passe-bas et sur le passe-haut —approche de type paquets d'ondelettes—, de façon à obtenir l'équivalent des bandes critiques. Il y a donc peu de sorties fréquentielles (29 sous-bandes) par rapport à [Mah] (512 ou 1024 sous-bandes) et à [JB] (320 sous-bandes). Le prix à payer est un délai plus important dans la mesure où l'on utilise des filtres orthonormés plus longs (par exemple, pour des filtres de longueur 20 —le minimum—, il est nécessaire de prévoir un délai analyse-synthèse de 116 ms).

À nouveau, comme dans [Mah,DLR] la fonction de masquage est estimée à l'aide d'une transformée de Fourier fenêtrée sur 1024 ou 2048 points (suivant la qualité "dynamique" du signal) ce qui induit nécessairement un effet de préécho dû à un mauvais choix du quantificateur (qui dépend de l'ensemble du signal sur une durée de 23 ou 46 ms) pour un signal impulsionnel.

L'intérêt du travail effectué dans [ST] réside dans le fait que les ondelettes utilisées sont optimisées sur chaque trame et que cette optimisation semble apporter des résultats tangibles. L'optimisation est en fait limitée aux ondelettes de Daubechies (à nombre de moments nuls maximal) d'un ordre donné. Cela soulève un certain nombre de questions car les auteurs indiquent avoir minimisé sur d'autres ensembles de filtres —donc en particulier sur d'autres filtres plus sélectifs— sans que cela n'apporte d'amélioration notable. D'un autre côté, observant qu'une augmentation du nombre de moments nuls des ondelettes est sensiblement bénéfique, ils en déduisent qu'il s'agit là d'une conséquence de la meilleure sélectivité des filtres de Daubechies —dont on sait que ce n'est pas la principale qualité!—. Si c'était le cas, on pourrait trouver des filtres plus courts, réguliers (on n'a pas vraiment besoin d'ondelettes 10 fois dérivables!) et bien plus sélectifs, par exemple en utilisant un algorithme comme celui publié par O. Rioul [Ri5].

Enfin, [ST] utilisent un dictionnaire de formes d'ondes pour coder plus efficacement les signaux.

Revenons quelques instants sur l'opposition qui apparaît entre tous ces systèmes au sujet du type de transformée à utiliser, à savoir uniforme ou non uniforme. Les courbes de masquage ont été historiquement mises en évidence pour des signaux de durée infinie et stationnaires [ZF] et c'est dans cette optique que la transformée de Fourier est l'outil le plus adapté. Cependant le caractère non-stationnaire des signaux sonores est bien connu, même si la plupart du temps —c'est le cas en particulier pour la musique— on peut parler de quasi-stationnarité devant la période d'échantillonnage de ces signaux. On utilise donc plutôt la transformée de Fourier fenêtrée glissante.

Il est ainsi clair qu'il s'agit d'un compromis d'autant que, malgré leur caractère stationnaire, les courbes de masquage indiquent bien que l'oreille n'analyse pas toutes les fréquences avec la même précision: en utilisant une transformation uniforme qui se ramène en général à l'équivalent d'une transformée de Fourier, on amalgame donc des caractéristiques dynamiques qui gagneraient au contraire à être dissociées puisque l'oreille elle-même le fait. On peut voir là une cause du préécho qui affecte certains signaux sonores. On peut donc penser qu'une transformation plus proche du mode d'analyse de l'oreille serait "psychoacoustiquement" plus robuste. C'est ce qui est proposé en partie par [JB] et par [ST].

Cependant, il ne suffit pas que l'analyse soit effectuée de façon plus proche de l'analyse en bandes critiques, il faut également donner une forme plus "dynamique" au phénomène de mas-

quage fréquentiel. C'est la raison pour laquelle on a éprouvé la nécessité de reformaliser ce phénomène dans la section précédente, ce qui nous a conduit au modèle de transformée en ondelettes qui donne directement les courbes de masquage (VII.2).

Il faut malgré tout avouer que la conception de transformées non uniformes est considérablement plus complexe que celui de transformations uniformes dans la mesure où pour ces dernières existe déjà la DFT (que l'on calcule rapidement à l'aide de l'algorithme de FFT), alors que la seule méthode fiable actuellement dans le cas non-uniforme est le procédé par itérations: même si les filtres à itérer prennent une forme très "propre" le banc de filtres ainsi engendré peut avoir un comportement catastrophique. Dans le cas uniforme, le banc de filtres est ainsi conçu directement, alors que dans le cas non uniforme c'est par le biais des itérations…

Enfin, les transformations à base de bancs de filtres itérés génèrent en général un important délai: on a vu de manière heuristique au chapitre VI que ce défaut était probablement inhérent à la structure même d'un banc de filtres itéré, et n'est que partiellement lié au fait que les filtres choisis sont paraunitaires. Il en résulte malheureusement une plus grande complexité de conception.

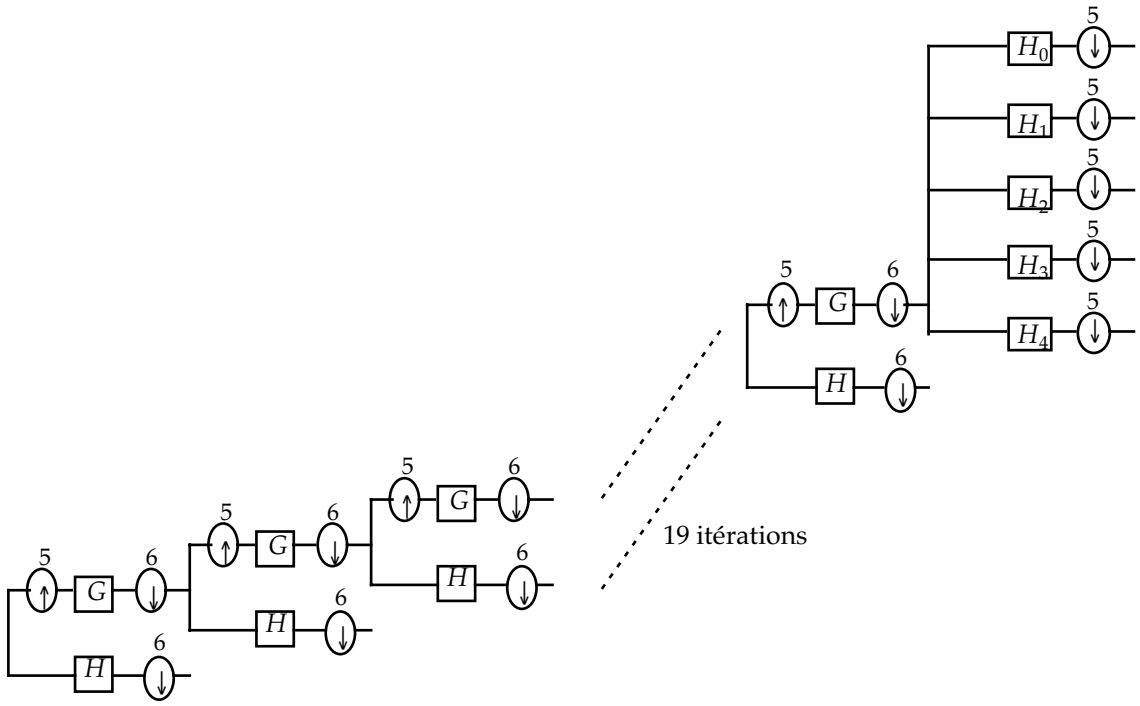
On va maintenant proposer un système de codage de son HiFi à l'aide de bancs de filtres itérés en fraction d'octave.

E. Technique à base de Bancs de filtres rationnels

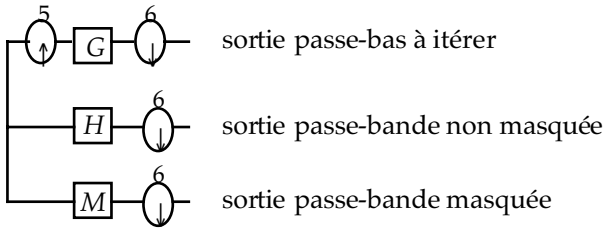
Le fait que le banc de filtres 6/5 itéré fournisse une analyse en tiers d'octave nous permet de réaliser de manière systématique avec un seul couple de filtres et en itérant seulement sur le passe-bas ce qui est effectué par [ST]. Les résultats obtenus lors des chapitres précédents nous conduisent naturellement à des filtres caractérisés par une sélectivité fréquentielle convenable. Il faut noter que dans [ST], aucune optimisation basée sur la sélectivité fréquentielle n'est effectuée: la seule variable correspondant approximativement à ce paramètre étant le nombre de moments nuls de l'ondelette de base. En outre le fait de changer de branche à itérer modifie à nouveau le caractère sélectif de la branche considérée.

1. Implémentation

La solution que je proposerai à 32 kHz sera donc constituée d'un banc de filtres sélectif à reconstruction parfaite itéré 19 fois avec des facteurs p/q=6/5 et dont la dernière itération sera suivie par un banc de filtres uniforme de 5 bandes: ce nombre d'itérations est adapté à des signaux sonores échantillonnés à 32 kHz. Les bandes de fréquences correspondant à cette analyse sont données par leurs extrémités 16000 Hz, 13333 Hz, 11111 Hz, 9259 Hz, 7716 Hz, 6430 Hz, 5358 Hz, 4465 Hz, 3721 Hz, 3101 Hz, 2584 Hz, 2153 Hz, 1795 Hz, 1495 Hz, 1246 Hz, 1038 Hz, 865 Hz, 721 Hz, 601 Hz, 500 Hz, 400 Hz, 300 Hz, 200 Hz, 100 Hz et 0 Hz, c'est-à dire 24 bandes critiques. Un schéma de la transformation est donné ci-après



Afin de tirer profit des caractéristiques de masquage on modifiera le schéma donné plus haut de façon à former non pas une, mais deux sorties passe-bande à chaque itération: à l'itération j on appliquera donc un banc de filtres sur-échantillonné de la forme



où le filtre M est choisi de telle sorte que la sortie correspondante soit très proche de la transformée en ondelettes réalisée par la cochlée, et soit d'autre part synchronisée avec celle du filtre H. Il suffira alors de comparer les niveaux d'énergie entre la sortie H et la sortie M pour déterminer si la première est masquée ou non, et pour déterminer le pas de base de la quantification. Bien sûr, il n'est pas question de transmettre les valeurs de la sortie M: les seules dont on ait besoin sont celles de H et de la transformation uniforme en fin de course.

a. Calcul du filtre *M*

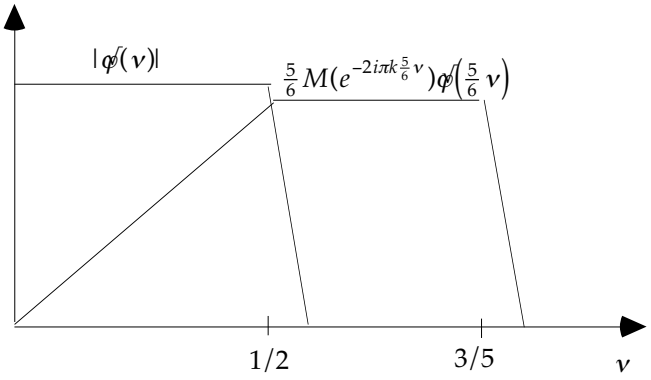
On va supposer que les filtres paraunitaires G et H sont suffisamment sélectifs et correspondent à des pseudo-ondelettes à faible amnésie, elles mêmes sélectives. Cela signifie que le support fréquentiel de ϕ et ψ, les fonctions moyennes associées au passe-bas et au passe-haut, sont proches de $\left[0, \frac{1}{2}\right]$ et $\left[\frac{1}{2}, \frac{3}{5}\right]$ respectivement pour les fréquences positives. Comme on le sait, le filtre M va correspondre à une suite de pseudo-ondelettes —d'autant plus proches d'une fonction unique translatée que l'amnésie des fonctions limites φ_n sera plus faible— dont la fonction moyenne dénotée $\mu(x)$ vérifie

$$\mu(x) = \sum_k m_k \varphi\left(\frac{6}{5}x - k\right)$$

c'est-à dire en fréquence

$$\hat{\mu}(\nu) = \frac{5}{6} M(e^{-2i\pi k\frac{5}{6}\nu})\hat{\varphi}\left(\frac{5}{6}\nu\right)$$

Comme $\hat{\varphi}$ est de module 1 sur son support fréquentiel, on voit qu'il est assez simple de concevoir un filtre M collant avec l'ondelette cochléaire pour les fréquences inférieures à $3/5$ comme c'est indiqué sur le graphique ci-dessous



Ce sont d'ailleurs les fréquences les plus importantes pour le masquage fréquentiel puisqu'elles sont responsables du masquage des fréquences élevées par les fréquences plus basses (voir la formule (VII.2)). Il peut par contre être plus difficile de faire coller les parties fréquentielles supérieures à $3/5$ dans la mesure où cette fois, c'est la fonction passe-bas qui impose la décroissance alors que le filtre passe-haut M est symétrique en valeur absolue par rapport à $1/2$. Un filtre M candidat est donné ci-dessous

n	m_n	n	m_n
0	3.799111854170067e-03	20	-5.748918076161733e-01
1	-2.805027305150366e-03	21	2.941295931644861e-01
2	8.880204224891011e-04	22	-9.255307744582893e-02
3	1.717656915768139e-03	23	3.688448220857387e-03
4	-4.551475509630281e-03	24	2.776154163809237e-02
5	6.960970431848611e-03	25	-3.444729283680766e-02
6	-8.183464049099689e-03	26	2.964318960436592e-02
7	7.463885963473966e-03	27	-2.035276886995649e-02
8	-4.212542128711007e-03	28	1.036050035941482e-02
9	-1.825221412312855e-03	29	-1.825221412312855e-03
10	1.036050035941482e-02	30	-4.212542128711007e-03
11	-2.035276886995649e-02	31	7.463885963473966e-03
12	2.964318960436592e-02	32	-8.183464049099689e-03
13	-3.444729283680766e-02	33	6.960970431848611e-03
14	2.776154163809237e-02	34	-4.551475509630281e-03
15	3.688448220857387e-03	35	1.717656915768139e-03
16	-9.255307744582893e-02	36	8.880204224891011e-04
17	2.941295931644861e-01	37	-2.805027305150366e-03
18	-5.748918076161733e-01	38	3.799111854170067e-03
19	7.188704337600350e-01		

Il ne permet donc de prendre en compte que le masquage des fréquences plus aigües par les fréquences graves. Si l'on voulait en outre inclure le masquage des fréquences graves par les fréquences aiguës on devrait concevoir un autre filtre passe-bas moins sélectif et donc exécuter deux bancs de filtres itérés en parallèle…

On substitue donc pour le masquage auditif une transformée en ondelettes à la transformée de Fourier couramment utilisée. Cela apporte plusieurs avantages:

- le phénomène de masquage devient un simple problème de quantification que l'on peut compliquer à loisir si l'on souhaite se rapprocher des courbes de masquage réelles du système auditif

- on n'a pas besoin de calculer un grand nombre de lignes fréquentielles pour évaluer la courbe de masquage, et qui ne seront pas utiles indépendamment. En outre, si on le souhaite, on peut n'effectuer le calcul que des niveaux des bandes que l'on suppose masquées

- conservation du caractère dynamique du masquage: on évaluera l'énergie dans chaque bande sur un nombre fixe d'échantillons, par exemple 20, ce qui signifiera pour la bande de fréquences la plus élevée une fenêtre d'analyse de 20x6/32≈4 millisecondes alors que pour la bande la plus basse cela signifiera 20/500=40 millisecondes. Une conséquence particulière en sera un meilleur comportement vis-à-vis des signaux à transition forte, et une réduction du phénomène de préécho dont on sait qu'il est essentiellement dû à des parasites haute-fréquence.

Il y a deux types de seuils à insérer dans le schéma présenté: le seuil d'audition absolu et le paramètre du seuil d'audition masqué. Ces nombres dépendent de la bande de fréquence considérée. Dans ce qui suit on notera les sorties H du banc de filtres à reconstruction parfaite et M du banc de filtres "cochléaires" par $y_j^{\psi}[n]$ et $y_j^{\mu}[n]$ respectivement.

b. Seuil d'audition absolu

À partir de la formule (IV.37) (en négligeant l'amnésie) et en définissant T comme période d'échantillonnage du signal d'entrée x

$$y_j^{\psi}[n] = \frac{5^{j/2}}{6^{j/2}} \int \psi\left(5n - \frac{5^j}{6^j} u\right) x(uT) du$$

puisque le filtre passe-bas G est normalisé de telle sorte que $G(1) = \sqrt{6 \times 5}$. Remarquons également que l'ondelette mère donnée par la formule (IV.10) est de module $5/\sqrt{6}$ sur sa bande passante et non pas 1. Si le signal x est un son pur de fréquence f et d'amplitude juste audible $A(f)$, la valeur seuil $y_j^{\psi}[n]$ pour la sortie du banc de filtres sera $S_j(f) = \frac{6^{j/2}}{5^{j/2}} \left|\hat{\psi}(\frac{6^j}{5^j} fT)\right| A(f)$ à la fréquence f considérée. Bien sûr le seuil absolu pour la bande j sera la valeur minimale de ces seuils pour toutes les fréquences de la bande j soit

$$S_j = \frac{6^{(j-1)/2}}{5^{j/2-1}} \sup_{\frac{1}{2} \le \frac{6^j}{5^j} fT \le \frac{p}{2q}} A(f) \qquad \text{(VII.4)}$$

Étant donnée la forme particulière de la courbe $A(f)$ —décroissante jusqu'à 2 kHz puis croissante au-delà— on peut considérer que pour les bandes telles que $\frac{5^j}{2\times 6^j} \le 2000\,\text{Hz} \times T$ on a $S_j = \frac{6^{(j-1)/2}}{5^{j/2-1}} A(\frac{5^{j-1}}{2T\times 6^{j-1}})$ et pour $\frac{5^j}{2\times 6^j} > 2000\,\text{Hz} \times T$ alors $S_j = \frac{6^{(j-1)/2}}{5^{j/2-1}} A(\frac{5^j}{2T\times 6^j})$

c. Seuil d'audition masqué

On doit comparer les énergies moyennes des deux sorties passe-bande ce que l'on peut écrire fréquentiellement à l'aide de la densité spectrale de puissance $D(\nu)$ du signal —supposé stationnaire— audio

$$\left\langle y_j^{\psi}[n]^2 \right\rangle = \frac{1}{T}\int \left|\psi(\frac{6^j}{5^j}\nu)\right|^2 D(\frac{\nu}{T})d\nu \qquad \text{et} \qquad \left\langle y_j^{\mu}[n]^2 \right\rangle = \frac{1}{T}\int \left|\mu(\frac{6^j}{5^j}\nu)\right|^2 D(\frac{\nu}{T})d\nu$$

sortie non-masquée et sortie masquée. Si l'on discrétise la deuxième intégrale avec un pas $\nu_0 T$ on obtient

$$\frac{1}{T}\int \left|\mu(\frac{6^j}{5^j}\nu)\right|^2 D(\frac{\nu}{T})d\nu \cong \sum_n \left|\mu(\frac{6^j}{5^j} n\nu_0 T)\right|^2 D(n\nu_0)\nu_0$$

où l'on peut reconnaître —à une constante multiplicative près— l'expression (15) de [Mah] si l'on identifie la fonction $B(u)$ avec $\left|\mu(\frac{11}{20}\left(\frac{6}{5}\right)^{-u})\right|^2$ et $\nu_0 T$ à $1/N$, en supposant que le bark s'exprime en fonction de la fréquence f sous la forme $u = \log_{\frac{6}{5}} f + \text{Constante}$.

Afin de continuer l'analogie avec [Mah] on va supposer que D est constante sur chaque bande critique ce qui permet de simplifier $\left\langle y_j^{\psi}[n]^2 \right\rangle \cong \frac{5^{j+1}}{6^{j+1}} D(\frac{5^j}{2T\times 6^j})$. Cette hypothèse est cependant mise en défaut quand la bande critique considérée est constituée d'un faible nombre de sons purs puisqu'alors la densité spectrale de puissance a un support bien plus restreint que celui de la bande critique. On verra plus loin comment minimiser ce problème.

Comme la quantité $y^2(m,k)$ de l'équation (2) tirée de [Mah] égale en moyenne $\frac{N}{2} D(k\nu_0)$ on peut affirmer que la sortie y_j^{ψ} sera masquée dès que

$$\left\langle (y_j^{\psi})^2 \right\rangle \le \frac{5^{j-1}}{6^j} \left\langle (y_j^{\mu})^2 \right\rangle \qquad \text{(VII.5a)}$$

(on a supposé N=1024 ici pour une fréquence d'échantillonnage de 32000 Hz). Ce seuil est donné comme simple valeur indicative: dans ce domaine, l'expérimentation est bien plus importante d'autant plus que, comme on va le voir maintenant, il est indispensable de modifier la valeur de ce seuil pour tenir compte de la tonalité de la bande critique.

En effet dans le cas où la densité spectrale de puissance est très concentrée dans la bande j autour d'une fréquence ν, on peut vérifier qu'il est cette fois nécessaire de modifier le coeffi-

cient de masquage puisqu'alors on aura approximativement $\left\langle y_j^{\psi}[n]^2 \right\rangle \cong \frac{5^2}{6} \nu_0 D(\nu)$. La sortie y_j^{ψ} sera alors masquée quand

$$\left\langle (y_j^{\psi})^2 \right\rangle \leq 10^{-3} \left\langle (y_j^{\mu})^2 \right\rangle \qquad \text{(VII.5b)}$$

d. Détection de tonalité

Afin de déterminer le seuil d'audition masqué, il sera nécessaire de disposer d'une méthode permettant d'évaluer si une bande critique est à caractère tonal ou plutôt bruité. Cette détection est ici effectuée de manière très rustique puisque l'on se contente de minimiser le résultat de la convolution du filtre $1\text{-}2az\text{+}z^2$ avec la portion du signal en sous-bande qui nous intéresse: si le rapport entre l'énergie de l'erreur de prédiction et celle du signal est inférieur à une constante donnée (le choix fait ici est de 0.5) alors on considère que le signal est tonal, et en conséquence qu'il faut choisir le seuil de masquage (VII.5b). Sinon on prendra le seuil défini par (VII.5a). Bien sûr, cette méthode a beaucoup de chances d'être prise en défaut dès que plusieurs sons purs sont présents dans la bande critique considérée. Le pari est alors que lorsqu'il y a plus d'une sinusoïde, le système auditif augmente le seuil de masquage. Si ce n'est pas le cas, il faut alors concevoir une méthode plus sophistiquée de détection de tonalité, à l'aide d'un modèle autorégressif contenant plus de pôles.

Afin de tirer profit de la diminution de dynamique du résidu, il est utile de coder celui-ci (avec la même quantification que le signal originel), tandis que le paramètre a du filtre d'autorégression est codé sur 8 bits entre -1 et 1: on peut en effet vérifier que pour une sinusoïde, ce paramètre est le cosinus d'un angle.

e. Tramage

Pour chaque bande critique on déterminera le seuil de masquage et éventuellement le paramètre de tonalité pour un nombre déterminé N_j d'échantillons. Dans la mesure où ce seuil doit être transmis avec précision, on n'a pas intérêt à le transmettre fréquemment ce qui impose N_j relativement grand (on peut cependant s'affranchir en partie de ce problème par une méthode simple de prédiction linéaire de ces seuils). Il ne faut pas non plus qu'il le soit trop, sans cela on risquerait de masquer des sons qui ne le sont pas car trop distants. En fait ce nombre d'échantillons doit être déterminé d'après les valeurs du masquage antérieur (le masquage postérieur étant bien plus grand que le masquage antérieur [ZF]). Comme la valeur maximale de ce masquage temporel est de l'ordre de 4 ms, on en déduit que pour la bande de fréquences la plus élevée (ici à 32 kHz) on doit prendre une taille de fenêtre d'au maximum $4 10^{-3} \times \frac{32000}{6} \cong 20$ échantillons dans cette bande. Bien que cela ne soit pas donné dans [ZF], on peut estimer que cette taille de fenêtre est indépendante de la bande critique considérée. En effet, une conséquence du modèle de l'audition sous forme de transformée en ondelettes est que la taille de la fenêtre temporelle d'analyse croît proportionnellement avec le facteur d'échelle. Exprimée en échantillons, cette taille est invariante dans chaque bande critique de notre banc de filtres (qui sont sous-échantillonnées proportionnellement au facteur d'échelle).

On définit ainsi une trame à l'intérieur de chaque sous-bande analysée. Cependant, à la différence de ce qui se passe avec des sous-échantillonneurs entiers, on ne va pas pouvoir définir naturellement une trame *globale* pour tout le système (comme par exemple dans [ST] 1024 ou

2048 échantillons). En effet, si l'on voulait réunir dans une même trame tous les échantillons des bandes 1 à 19 on serait obligé de recourir à une taille de trame, multiple de… 6^{19} échantillons c'est-à dire un peu moins de 604 ans! Ceci est un bien sûr un problème qu'il faudra résoudre lors de l'implémentation en temps réel d'un algorithme comme celui-ci basé sur des échantillonneurs non-entiers.

f. Quantification

Le seuil de masquage sera finalement le pas de la quantification linéaire associée à la fenêtre de signal analysée (il peut être utile de considérer d'autres formes de quantification pour mieux représenter la sensibilité de l'oreille): c'est essentiellement de cette façon que l'on pourra réduire le débit du signal, bien plus que par le masquage total des coefficients, moins fréquent que dans [Mah] puisque l'on analyse le signal sur des bandes de fréquence plus larges. En échange, les seuils (VII.5a) apparaissent plus élevés.

Le seuil de masquage pour chaque fenêtre devant être transmis il doit être donné avec une précision suffisamment grande pour ne pas être pris en défaut par la sensibilité fréquentielle de l'oreille. Ce seuil A_j s'apparente en fait à la sensibilité à long terme —i-e. pour un signal stationnaire— par exemple celle des taux de modulation juste audibles [ZF]. Comme on l'a vu dans la partie consacrée à la sensibilité de l'oreille, un calcul tiré des données de [ZF,p. 93] montre qu'une quantification non linéaire sur 8 bits est suffisante (sans inclure la moindre prédiction qui devrait, en principe, permettre de limiter davantage le nombre de niveaux de quantification). On peut utiliser la formule empirique suivante (directement interpolée de [ZF]) pour quantifier cette quantité

$$A_j = S_j(1 + 0.14n)^{3.5}$$

où n est un entier positif.

g. Codage

On utilise une technique de codage entropique pour les différentes données à transmettre: pas de quantification, paramètre de tonalité et résidu. On peut transmettre en outre un bit par trame de vingt échantillons indiquant si les données qui suivent sont toute nulles ou non, et un autre bit dans ce dernier cas indiquant la présence ou l'absence de tonalité.

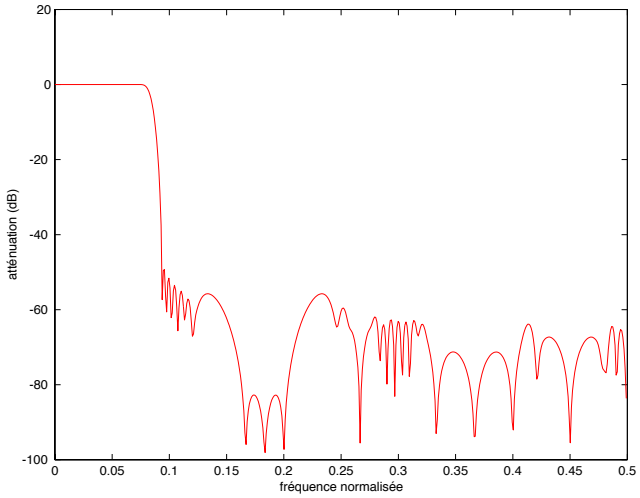
2. Résultats

Afin de vérifier la réduction de débit due au phénomène de masquage on a implémenté un banc de filtres itéré avec un facteur d'échelle de 6/5. Il faut tout de suite reconnaître que le codeur ne représente qu'une ébauche. Ainsi, n'ayant pas à notre disposition une transformation uniforme (nécessaire à l'implémentation des 5 dernières bandes critiques), on a préféré ne pas coder le passe-bas issu de la $19^{\text{ème}}$ itération du banc de filtres: sa contribution au débit final devrait de toutes façon être faible puisque, même si l'on le quantifie sur 16 bits son débit sera de 16 kbits/s, et quand on disposera de la transformation uniforme, on sait qu'il suffira de 8 bits par échantillons (sans intégrer la moindre notion de masquage fréquentiel), soit une contribution de 8 kbits/s, pour le coder.

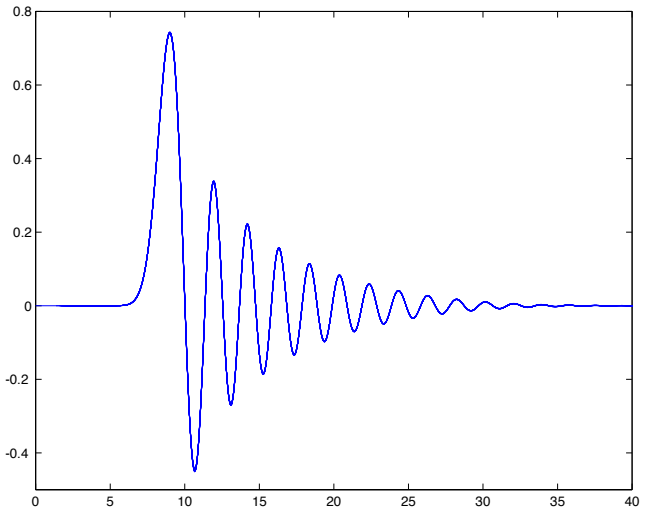
De même, plutôt que d'implémenter un codeur entropique, on a préféré calculer simplement l'entropie statistique de chaque série d'éléments sur toute la durée du signal à traiter.

Le but était plutôt d'appréhender les valeurs optimales des seuils utilisés pour la quantification rendant le débit le plus faible pour la qualité la meilleure. Grâce à l'aimable collaboration d'Alain Le Guyader et Catherine Quinquis du CNET Lannion LAA/TSS/CMC ce but a été en partie atteint (il a en particulier montré la nécessité de détecter des sons purs dans les sous-bandes). Le nombre de tests de paramètres a cependant été bien trop faible pour être déterminant, la plus grande gêne ayant été de ne pas pouvoir entendre moi-même les défauts du traitement que je proposais (sauf sur mon matériel informatique: un MacIntosh Quadra 700 mono, 8 bits par échantillon sonore, loi linéaire, fréquence d'échantillonnage 22000 Hz, soumis en outre au bruit ambiant). En conséquence, il m'a été impossible d'atteindre la transparence.

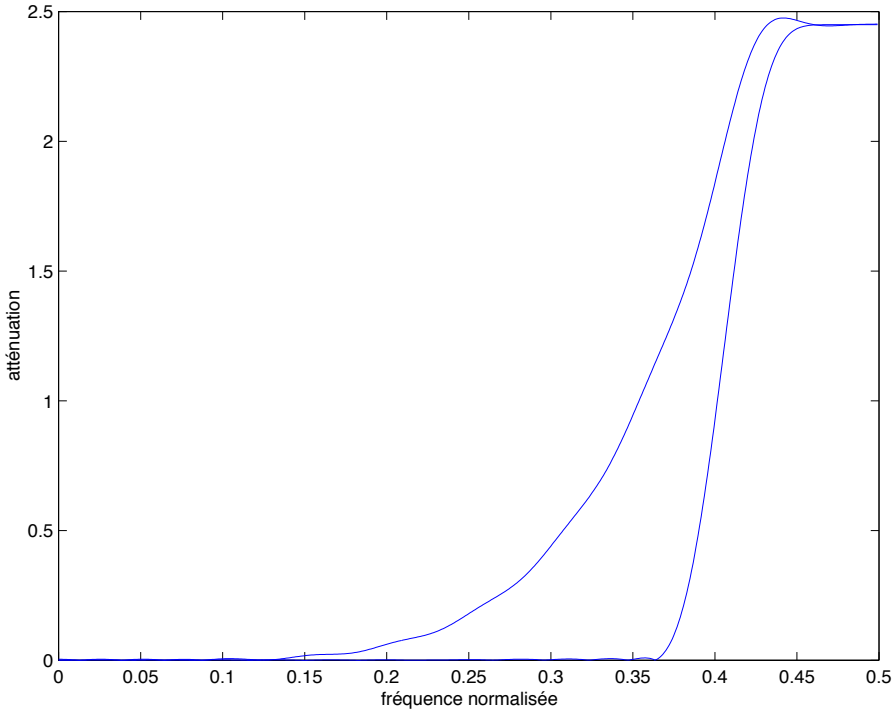
On a utilisé des filtres très longs permettant d'avoir à la fois une bonne atténuation ainsi qu'une amnésie faible (les deux choses étant étroitement liées comme on l'a vu au chapitre V). À l'aide de l'algorithme du chapitre VI, on a choisi un filtre passe-bas de degré 203. Outre son degré, on a fixé le début de sa bande atténuée choisie ici à 0.093 (à comparer à 1/12=0.083) en fréquence normalisée. Le filtre passe-bas est représenté ci-après



Afin de mettre en évidence la faible amnésie des fonctions limites, on a porté sur le même graphique 21 fonctions limites ramenées à l'origine, c'est-à dire $\varphi_n(x+n)$ pour n=0…20



On a également représenté le passe-haut H ainsi que le filtre de masquage M qui lui est associé



Les valeurs exactes des coefficients sont données dans le tableau de la page suivante

n	g_n	n	g_n	n	g_n	n	h_n
0	6.421501841835559e-03	70	-6.305480396601768e-02	140	-2.875202575780823e-03	0	0
1	2.035822326536790e-02	71	-5.493245311880141e-02	141	-2.367980170245853e-03	1	0
2	4.737690370479999e-02	72	-3.245887267299323e-02	142	-1.350496420474851e-03	2	0
3	9.307660714686160e-02	73	-2.786548144995240e-03	143	1.238602490072985e-05	3	2.051129006562199e-04
4	1.620229884496026e-01	74	2.554134578085902e-02	144	1.503456232287635e-03	4	1.110156092710639e-03
5	2.537677342288021e-01	75	4.557483135104463e-02	145	1.852107592899757e-03	5	-6.719606475038887e-04
6	3.636951863277044e-01	76	5.245529537964897e-02	146	1.795509191796484e-03	6	-1.235309698688418e-03
7	4.848339308806263e-01	77	4.491587241228329e-02	147	1.345190394612985e-03	7	3.420055396396006e-03
8	6.041009181616801e-01	78	2.568256869809264e-02	148	5.162648209807550e-04	8	-4.620548192218369e-03
9	7.062614368540692e-01	79	9.061744906606063e-04	149	-6.760815818330118e-04	9	2.795028193922211e-03
10	7.739709456667451e-01	80	-2.231925214983484e-02	150	-1.014851841055215e-03	10	2.526596279325698e-03
11	7.915558581278129e-01	81	-3.838674710764556e-02	151	-1.159575915825041e-03	11	-1.043233264650294e-02
12	7.500056377284322e-01	82	-4.318565782895187e-02	152	-1.036401625562499e-03	12	1.718401624051289e-02
13	6.474875785427708e-01	83	-3.609589754741157e-02	153	-6.031441557998908e-04	13	-1.838436144013547e-02
14	4.907016174995469e-01	84	-1.968446284644281e-02	154	2.389211023522034e-04	14	9.879184594698207e-03
15	2.953621598330397e-01	85	8.352750141514491e-04	155	4.563777934852264e-04	15	8.421083355264258e-03
16	8.598301666557751e-02	86	1.963523375179409e-02	156	6.356741612796858e-04	16	-3.240605585815273e-02
17	-1.086016842566461e-01	87	3.212076028593878e-02	157	6.719744431018508e-04	17	5.264388744299071e-02
18	-2.628633416063915e-01	88	3.509907071436844e-02	158	4.837905598131767e-04	18	-5.864518829670048e-02
19	-3.562112927104615e-01	89	2.848170214639587e-02	159	-6.306930633634076e-05	19	4.159594535200203e-02
20	-3.789491152862369e-01	90	1.454145190053447e-02	160	-1.612800820431013e-04	20	-5.325467946446090e-04
21	-3.312385166237464e-01	91	-2.248875972164744e-03	161	-2.858769702415633e-04	21	-5.653457376321634e-02
22	-2.269050360645323e-01	92	-1.724463058697816e-02	162	-3.686319272059827e-04	22	1.126919353110258e-01
23	-8.912812569903159e-02	93	-2.658682104155344e-02	163	-3.167051877251446e-04	23	-1.473260917699857e-01
24	5.287523385033380e-02	94	-2.815997712034084e-02	164	1.101861176869536e-05	24	1.419273839122830e-01
25	1.712993737122622e-01	95	-2.196652381356903e-02	165	4.257398916755317e-05	25	-8.737448744357858e-02
26	2.456992795004225e-01	96	-1.031733245154886e-02	166	1.010268374984039e-04	26	-1.196956051947099e-02
27	2.621014786465754e-01	97	3.374261390788187e-03	167	1.657953694817184e-04	27	1.380470252950848e-01
28	2.213615431766247e-01	98	1.496566133898619e-02	168	1.740331589463280e-04	28	-2.659384108531052e-01
29	1.358444872696224e-01	99	2.182348666166863e-02	169	-1.185769478503296e-06	29	3.687614375216913e-01
30	2.720978160882842e-02	100	2.212310505915903e-02	170	-7.437953750455363e-06	30	-4.270751830470552e-01
31	-7.896088496780487e-02	101	1.659457640255433e-02	171	-2.666860913368240e-05	31	4.316939506703911e-01
32	-1.592622490323680e-01	102	6.762178146094245e-03	172	-5.859115529391370e-05	32	-3.888145835031687e-01
33	-1.985898138312304e-01	103	-4.049380441299349e-03	173	-7.828806015115288e-05	33	3.144916147404955e-01
34	-1.894285730792098e-01	104	-1.295524396803027e-02	174	7.024018577010878e-08	34	-2.266628079898853e-01
35	-1.366976893775799e-01	105	-1.747774418357101e-02	175	8.004400623849287e-07	35	1.444821940804730e-01
36	-5.568243960372218e-02	106	-1.711846893375239e-02	176	4.659179978269082e-06	36	-7.901859399886974e-02
37	3.266112099433037e-02	107	-1.200875144612225e-02	177	1.546662904929820e-05	37	3.552688159214715e-02
38	1.070745112372469e-01	108	-4.026754477842853e-03	178	2.766688142346521e-05	38	-1.255473962573157e-02
39	1.520492333752546e-01	109	4.533194885954884e-03	179	-1.775944268551618e-09	39	3.314143582122575e-03
40	1.573770789265759e-01	110	1.078457266413189e-02	180	-4.741694486108068e-08	40	-5.790025122894281e-04
41	1.245710779925637e-01	111	1.383422241112020e-02	181	-5.014005221045297e-07	41	6.230971885642698e-05
42	6.275806479123687e-02	112	1.277633178671459e-02	182	-2.702120990310418e-06	42	-3.691140214198587e-06
43	-9.951271872505307e-03	113	8.363623539732189e-03	183	-7.303378642644208e-06	43	9.340457131726568e-08
44	-7.546357737749791e-02	114	1.803511073458547e-03	184	-5.421674135991193e-13		
45	-1.190142692245135e-01	115	-4.370191854652181e-03	185	1.199889235025322e-09		
46	-1.308574916823468e-01	116	-8.901418287770774e-03	186	2.970222882496068e-08		
47	-1.087919293160323e-01	117	-1.057363506087985e-02	187	2.907904150958330e-07		
48	-6.070291685292965e-02	118	-9.270059921517667e-03	188	1.275947895595092e-06		
49	-5.038847657869970e-04	119	-5.371720810570179e-03	189	1.408257187781231e-13		
50	5.584343318241338e-02	120	-4.999030390368307e-04	190	-7.370577816480784e-15		
51	9.574113493324020e-02	121	4.116538714527336e-03	191	-7.515999628275050e-10		
52	1.086879466654266e-01	122	7.065776818764691e-03	192	-1.722599729739582e-08		
53	9.329335404115739e-02	123	7.896673817929890e-03	193	-1.373119225589258e-07		
54	5.468202645345881e-02	124	6.394391556623732e-03	194	-2.078003649132000e-14		
55	4.641369400756087e-03	125	3.303444219031542e-03	195	1.641036048714207e-15		
56	-4.358850893478194e-02	126	-4.421633982177158e-04	196	-1.860692425457434e-15		
57	-7.816115733742274e-02	127	-3.611718782110820e-03	197	4.358951830105655e-10		
58	-9.065139053956177e-02	128	-5.504834598300578e-03	198	8.134157754575101e-09		
59	-7.893146998490252e-02	129	-5.719620716937929e-03	199	-3.233985177374547e-16		
60	-4.732217030729573e-02	130	-4.202414116323206e-03	200	3.570173235767392e-16		
61	-5.377549334705209e-03	131	-1.729654220558808e-03	201	9.847031134821555e-16		
62	3.530711998512841e-02	132	9.271142442816843e-04	202	1.734488278023894e-15		
63	6.482009443190154e-02	133	3.100446698116615e-03	203	-2.058314842376607e-10		
64	7.562718549695772e-02	134	4.267129484618521e-03				
65	6.613924897631566e-02	135	3.830844029248056e-03				
66	3.977929009562152e-02	136	2.512391260681290e-03				
67	4.492532469259769e-03	137	7.015292928781202e-04				
68	-2.967825256771422e-02	138	-1.201983185460466e-03				
69	-5.420877450718561e-02	139	-2.744559758022038e-03				

L'amnésie des fonctions limites est environ de 0.004, c'est-à dire une contribution à l'atténuation de 48 dB (à reporter dans la formule (V.39)).

Les tests ont été effectués sur une série de 15 signaux fournis par Y. Mahieux du CNET Lannion LAA/TSS/CMC. Comme je l'ai indiqué plus haut, les signaux codés sont encore discernables des signaux originaux. Il semble que le problème soit tout simplement dû à un mauvais seuil d'audition qui nécessiterait d'être abaissé de quelques dB. Dans certains cas, le son codé n'était pas discernable du son original, dans d'autres, par exemple sflu3, l'original était plus clair. Dans ce cadre, on ne donne ici que des résultats indicatifs: il semble bien que, dans le but d'améliorer les résultats, outre un seuil d'audition correct, il soit plus efficace d'accorder une forte importance à la détection de tonalité, c'est-à dire tout simplement à une forme de prédiction linéaire.

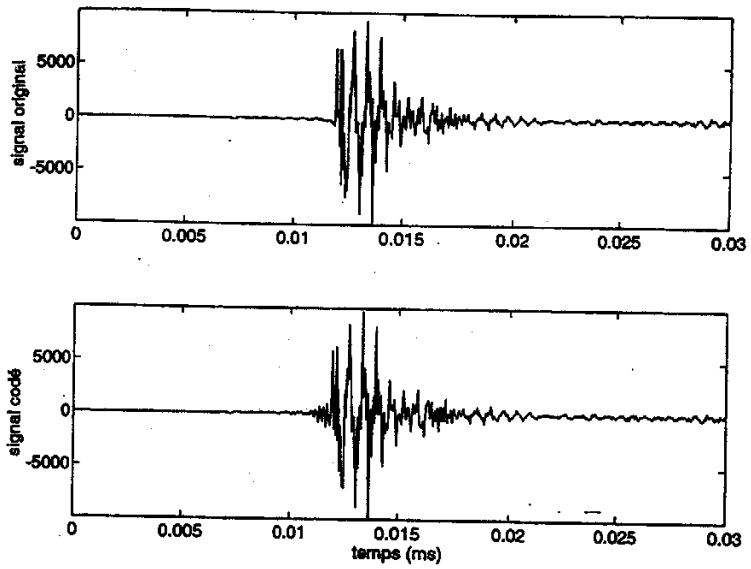
Les résultats sont mis sous la forme du tableau ci-dessous

Séquence sonore et description	Entropie des bandes 1 à 19 (32000 à 1000 Hz)
sbas1: basson	73800 bits/s
scat1: castagnettes	109400 bits/s
sflu3: piccolo	116600 bits/s
sglk1: glockenspiel	27100 bits/s
sobo1: chant + réponse de hautbois + accompagnement d'orchestre	89000 bits/s
spar1: parole, voix de femme (anglais)	76600 bits/s
spia7: piano	84000 bits/s
svib1: vibraphone	73500 bits/s
sbss2: corde frappée	45500 bits/s
sflu1: flûte traversière	62900 bits/s
shar2: harpe	97200 bits/s
sorc1: mouvement d'orchestre	101600 bits/s
spar2: parole, voix d'homme (anglais)	74700 bits/s
spop1: Suzanne Vega *a capella*	118500 bits/s
svio3: violon	114400 bits/s

auxquels il faut rajouter une contribution de 8000 bits/s pour la partie du signal contenue dans la bande de fréquences 0–500 Hz.

Préécho

Ce phénomène est fréquent en présence de brusques variations de dynamique quand on utilise des transformations dont la qualité première est de séparer les zones de fréquence. La transformation que je propose doit donc, à l'instar de ce qui se passe pour des transformations uniformes [Mah] induire un effet de préécho. C'est d'ailleurs ce que l'on peut observer sur cet extrait de son de castagnettes codé



Comme on le voit cependant, l'essentiel du préécho est confiné dans une zone de quelques millisecondes correspondant à la contribution des fréquence les plus élevées. On pourrait cependant s'attendre à avoir un préécho issu des bandes de fréquence plus basses s'étendant sur une durée plus importante puisque c'est justement là l'une des caractéristiques d'une procédure de masquage effectuée sur une période temporelle dépendante de la fréquence analysée: la contribution de chaque bande au préécho ne s'étend pas sur la même durée pour toutes. Dans ce cas précis, il s'agit probablement en partie du fait que le seuil de masquage s'abaisse quand on s'intéresse à des bandes critiques de fréquence plus basse (d'après l'équation (VII.5a)), mais aussi du fait que notre fonction de masquage ne prend pas en compte l'effet des masquage des fréquences basses par les fréquences hautes. Les premières sont donc moins fréquemment masquées.

En détail, deux contributions au préécho doivent être mises en évidence. La première est le support réel du filtre passe-bande de synthèse à l'itération j $H_j(z^{-1})$. On vérifie que les supports réels des filtres $H_j(z^{-1})$ sont de la forme $43 \times (\frac{6}{5})^{j-1}$ ce qui permet de chiffrer la contribution de ces filtres au préécho à autant de symboles: 1.3 ms pour la bande critique la plus aigüe et 35 ms pour la bande correspondant à la 19$^{\text{ème}}$ itération. En fait, on constate expérimentalement (en codant et décodant une impulsion) que ces valeurs doivent être divisées par deux. Les erreurs dues à la quantification se propageront donc sur les périodes correspondantes.

La deuxième contribution est celle de la taille de la fenêtre considérée dans la procédure de masquage, c'est-à dire d'après nos estimations une vingtaine d'échantillons. Cette contribution doit être comprise comme très différente de la première puisqu'elle correspond à une donnée psychoacoustique: elle est dissociée du processus de quantification, et représente plutôt une fenêtre de corrélation. Cette contribution se calcule simplement et vaut $20 \times \frac{6^j}{5^{j-1}}$ échantillons soit près de 6 fois la première contribution: on peut donc affirmer que la contribution "mécanique" du support des filtres est négligeable devant l'effet psychoacoustique et devra donc subir l'effet de masquage antérieur.

Il faut noter combien cette situation est différente de celle de [Mah] où les deux contributions étaient (à 32 kHz) de 46 ms. On voit qu'il ne serait pas correct de penser que le banc de filtres itéré induit un préécho plus important que les transformations uniformes. Ceci est dû au

fait que les filtres itérés, comme on l'a montré au chapitre VI, ne sont pas aussi sélectifs en fréquence que leur longueur le permettrait: leur support réel est en contrepartie plus faible, d'un rapport 6 dans notre cas… Par contre, le délai "mécanique" de la transformation seule (sans inclure la taille de la fenêtre d'analyse nécessaire au masquage) est considérable et s'élève, pour 19 itérations à 200 ms (à 32 kHz): on n'a a priori pas d'espoir de faire baisser cette valeur nettement en dessous de 100 ms dans la mesure où la sélectivité fréquentielle du filtre passe-bas semble être un facteur primordial (en particulier pour obtenir une procédure de masquage raisonnable). Cela semble limiter l'utilisation de cet algorithme aux applications non —ou faiblement— interactives, comme le stockage ou la diffusion.

Points à développer

Bien évidemment, ce que l'on a proposé ici n'est qu'une ébauche de codage de son dont on est encore loin d'avoir fait le tour, en particulier à cause des nombreux paramètres qui devraient permettre de le régler efficacement. Il s'agit maintenant de développer un certain nombre de points en suspens qui permettraient de rendre ce codage opérationnel, et que l'on va exposer ci-dessous

- mettre en œuvre la transformée uniforme à 5 branches destinée aux fréquences 0–500 Hz et les filtres de masquage correspondants. On peut penser qu'il est suffisant de concevoir une MLT (Modulated Lapped Transform) [Malv2] à 5 sorties pour le banc de filtres à reconstruction parfaite, et une autre pour la partie masquage.
- jusqu'à quel point une prédiction linéaire dans chaque sous-bande peut-elle permettre de gagner en débit?
- les filtres passe-bas et passe-haut étant donnés, déterminer les valeurs optimales des paramètres dans chaque bande (seuil de masquage, détection de tonalité et taille de la fenêtre d'analyse), ce qui se décrit en anglais par l'expression "fine tuning". Dans une optique de descente en débit au delà du seuil de transparence, déterminer les modifications de ces paramètres qui limitent au maximum la perte de qualité.
- étudier l'influence de la longueur du filtre passe-bas sur l'efficacité du codage: ceci est extrêmement important dans la mesure où la longueur de ce filtre régit le délai de la transformation totale dans le cas orthonormal, ainsi que le nombre d'opérations de base nécessaires par échantillon. Que penser de filtres IIR pour ce type de problèmes? Peut-on trouver de l'intérêt à concevoir des filtres biorthogonaux, afin de limiter encore une fois le délai de la transformée? On peut penser qu'en outre, à sélectivité fréquentielle identique, les filtres biorthogonaux sont plus courts que leurs homologues orthogonaux.
- choisir un filtre de masquage plus proche de la réalité, de façon à éventuellement mieux refléter les propriétés de masquage temporel. Le filtre donné plus haut est de toutes façons trop long par rapport à sa sélectivité fréquentielle: il est clair qu'il serait possible d'en obtenir un plus court, ce qui limiterait également la complexité de l'algorithme de calcul des seuils de masquage (quoique ce ne soit pas là un problème critique).
- étudier d'autres possibilités de quantification des échantillons dans chaque bande que la quantification linéaire, mieux représentative de la sensibilité de l'oreille,

sachant cependant que ce qui peut alors être gagné en nombre de niveaux de quantification peut être perdu en partie par le codage entropique

- étudier la possibilité d'utiliser d'autres couples d'entiers que 6 et 5: peut-être p/q=5/4 est-il susceptible de donner des résultats acceptables? Ceci aurait l'avantage de diminuer le délai nécessaire au codage-décodage.
- insérer dans le traitement à l'intérieur de chaque bande critique de nouvelles informations issues de la connaissance du fonctionnement des neurones du nerf auditifs: par exemple le fait que pour des faibles intensités, ceux-ci se déchargent à un rythme moins élevé qu'à de plus fortes intensités, réalisant *de facto* un sous-échantillonnage de l'information qu'ils transmettent
- mettre au point une méthode de régulation de débit globale, particulièrement non naturelle dans notre cas, dans la mesure où les échelles de temps dans chaque sous-bandes ne se déduisent pas les unes des autres par un rapport entier.

F. Résumé du chapitre

Après avoir étudié les mécanismes auditifs en détail, on a essayé de convaincre de l'utilité de considérer le processus de masquage comme le résultat d'une opération de filtrage temporel classique —pour chaque bande critique, d'où l'introduction d'une transformation en ondelettes— plutôt que, de façon plus artificielle [Mah,ST,DLR,JB], comme la convolution en échelle de Bark de la densité spectrale de puissance du signal avec un filtre adéquat. Cette interprétation permet de faire intervenir de façon naturelle l'effet de masquage antérieur dont on sera à même de tirer profit en évitant le préécho dont l'apparition est classique dans les techniques de codage actuelles.

Outre cet effet de masquage, on a voulu mettre en évidence la technique de codage réalisée par le système auditif pour traiter l'information sonore: il y a ainsi la partie transformée qui engendre ce phénomène de masquage fréquentiel, mais aussi temporel, suivie par une forme de sous-échantillonnage dont l'originalité est de dépendre de l'intensité du signal —et de valeurs antérieures: propriété d'adaptation des neurones [Del]—. L'utilisation d'une transformée appropriée devrait permettre de tirer profit de ce choix du système auditif.

On a donc proposé une architecture simple réalisant une transformation cochléaire d'une part (partie masquante) et une transformation plus sélective d'autre part dans le but de tirer profit de la mauvaise résolution fréquentielle des filtres cochléaires, qui induit les divers effets de masquage. Cette architecture est basée en grande partie sur un banc de filtres rationnel itéré avec un facteur d'échelle de 6/5. Les chapitres précédents nous ont permis de mieux comprendre les contraintes pesant sur ces bancs de filtres (régularité, amnésie) et par le chapitre VI nous avons été en mesure de concevoir des filtres convenant à notre projet (sélectivité).

L'algorithme de codage a pu être mis au point grâce à la valeur de seuil donnée par [Mah]. Les résultats appliqués à des séquences sonores fournies par Y. Mahieux mériteraient d'être approfondis. Le résultat le plus significatif est l'absence de préécho dans les signaux reconstruits. On peut espérer qu'une fois mis au point, cet algorithme qui paraît proche de l'analyse effectuée par la cochlée démontrera d'autres propriétés intéressantes en terme de robustesse quand on fait baisser le débit…

Conclusion

Cette thèse a tenté de proposer un travail complet sur les bancs de filtres itérés, jusqu'à une possible application. Ces bancs de filtres itérés sont d'une certaine manière une facilité, car ils permettent de réaliser ce que l'on ne sait pas faire autrement à l'heure actuelle: la conception d'un banc de filtres non uniforme. On sait pourtant que les transformations à $\Delta f/f$ constant sont d'un vif intérêt dans bien des domaines, puisqu'ils permettent d'analyser et partant, de coder, plus efficacement les signaux naturels qui présentent fréquemment des spectres en $1/f$ (météorologie, systèmes chaotiques en général, fractals). C'est d'ailleurs une application en géophysique qui avait motivé l'introduction de la transformation en ondelettes en mathématiques et en traitement de signal [GGM].

De ce point de vue, l'application de la transformation en ondelettes au codage de sons apparaît paradoxale. Bien sûr, certains signaux sonores se présentent sous la forme de bruits avec des caractéristiques spectrales en $1/f$, mais la plus grande part des sons qui intéressent l'oreille humaine est éminemment tonale. C'est d'ailleurs cette constatation qui a consacré depuis longtemps la transformée de Fourier glissante en analyse de parole par exemple. Pourtant l'oreille présente dans son analyse du son un comportement qui la rapproche plutôt d'une transformation à $\Delta f/f$ constant, du moins aux fréquences supérieures à 500 Hz. Il s'agit du phénomène de masquage fréquentiel qui s'explique simplement dès lors que l'on suppose que la cochlée, mise en vibration par une onde de pression, réalise une transformation temps-échelle le long de son axe longitudinal proche d'une transformation en ondelettes. C'est seulement après la cochlée, au niveau des cellules ciliées et des fibres du nerf auditif que la nature tonale du son devient importante et que le système auditif déploie des algorithmes non-linéaires plus complexes pour identifier les harmoniques. Le paradoxe évoqué tombe alors puisqu'il devient naturel de vouloir effectuer la même transformation temps-fréquence que la cochlée afin, au minimum, de tirer profit des effets de masquage inhérents à cette transformation, et de façon plus ambitieuse, de modéliser plus avant le comportement postcochléaire du système auditif. Cette nécessité a bien évidemment été ressentie par d'autres que l'auteur, à commencer par son directeur de thèse et par l'équipe de traitement du son large bande du CNET Lannion (A. Le Guyader, Y. Mahieux)! Il y a ainsi le travail intéressant de D. Sinha et A.H. Tewfik [ST] qui implémentent un banc de filtres reproduisant approximativement la décomposition en bandes critiques de l'oreille. Il y a aussi un nombre constant de chercheurs insistant sur la nécessité de disposer d'une transformée "en tiers d'octave".

Il était donc important de pouvoir mettre à la disposition de la communauté de traitement de signal une telle transformation. Là encore, l'auteur n'a bien évidemment rien inventé: les bancs de filtres à échantillonnage fractionnaire existent depuis qu'existent les bancs de filtres (ou presque) [CR]: ils servaient alors à changer le taux d'échantillonnage des signaux d'un rapport non entier. Après l'arrivée des ondelettes et l'unification avec les bancs de filtres [Ma1,Mey1] il était tout naturel de considérer les itérations de tels bancs de filtres. Cependant les choses ne se passaient plus alors simplement. C'est ainsi que J. Kovačević et M. Vetterli qui s'étaient intéressés de façon précise à l'itération de bancs de filtres rationnels [Ko,KV1,KV2, KV3] durent admettre que ces schémas discrets n'engendraient pas, à la différence de ce qui se passe dans le cas dyadique, de fonctions limites.

C'est cette conclusion, qui devait s'avérer erronée, que l'auteur a choisi pour point de départ de ce travail sur les bancs de filtres rationnels itérés conduisant à l'introduction de l'amnésie (ou plutôt à la "shift error" en version anglaise) et à la nécessité d'envisager une série infinie de fonctions limites [Blu1,Blu2]. Bien qu'il ait été clair dès les constatations de J. Kovačević et M. Vetterli que les sorties d'un banc de filtres itéré ne pouvaient constituer une transformation en ondelettes, il est devenu rapidement évident que l'on pouvait rendre l'erreur induite aussi petite que l'on souhaitait en choisissant de façon appropriée les filtres générateurs. En étudiant la régularité des fonctions limites [BR], l'auteur s'est ainsi convaincu que celle-ci joue un rôle direct dans la minimisation de l'amnésie. Cette conclusion hâtive [Blu1] devait être révisée de façon bien plus subtile comme on peut le voir dans le chapitre V de la présente thèse. En fait, les liens principaux de l'amnésie sont plutôt avec la sélectivité, bien qu'au moins un facteur de régularité soit cependant nécessaire pour assurer la continuité des fonctions limites.

La structure même des espaces engendrés par les fonctions limites (toujours emboîtés) imposait d'étendre de façon très naturelle les espaces multirésolution définis par exemple dans [Mey1]. Une façon proche de voir les choses était, comme dans le chapitre I, de considérer des opérateurs d'échantillonnage dépendant du temps, de même que les processus d'interpolation. L'introduction de propriétés d'invariance d'échelle conduit alors aux nouveaux espaces multirésolution. Ceci permet de donner une vision plus géométrique de l'action d'un banc de filtres à reconstruction parfaite. La notion d'échantillonnage et d'interpolation, par ailleurs centrale dans tout le traitement numérique du signal, a d'ailleurs initialement conduit à l'opérateur de base du banc de filtres rationnels. Le chapitre I montre cependant que ce n'est pas le seul à pouvoir réaliser une opération numérique de changement d'échelle non entier, et construit des bancs de filtres généralisés.

C'est véritablement dans les chapitres IV et V que cette thèse étudie de la façon la plus profonde le comportement du banc de filtres dans ces itérations, mettant en évidence le rôle central joué par la régularité et l'amnésie pour, respectivement, stabiliser rapidement le spectre des filtres itérés et obtenir une bonne sélectivité de ceux-ci. Même s'il s'est avéré plus tard que la sélectivité des filtres peut d'une certaine manière se substituer à la régularité pourvu que l'on n'effectue pas trop d'itérations, il n'en reste pas moins que celle-ci reste indispensable dès que l'on n'a pas les moyens d'avoir une sélectivité suffisante. Quant à savoir s'il est important d'avoir des fonctions plus que continues, la réponse, pour les applications que l'on considère ici semble être négative. On n'a cependant pas vraiment pu juger de l'influence de la régularité sur le traitement de son puisque, c'est là un des échecs de cette thèse, il n'a pas été possible de mettre au point un algorithme de conception de filtres orthonormaux aussi réguliers que voulus: l'algorithme du chapitre VI s'arrête à un facteur de régularité pour le cas fractionnaire général, mais cependant peut aller plus loin dans le cas entier. En fait, l'existence même de filtres orthonormaux à coefficients réels de degré supérieur à 1 n'est pas démontrée…

Deux chapitres, indépendants entre eux, ont été dévolus à la conception de filtres. Il s'agit d'abord du chapitre VI qui fournit un algorithme nouveau, pratique, efficace et simple de conception de filtres orthonormés. C'est grâce à lui que ce travail de thèse a pu aller jusqu'à une application plus concrète, et c'est grâce à sa rapidité que l'on a pu sélectionner les filtres

d'amnésie la plus faible. Le chapitre III est plus théorique en ce sens qu'il s'intéresse d'abord à la structure des solutions qui nous intéressent, sans décrire d'algorithme de conception. Les résultats de ce chapitre transcendent d'ailleurs la conception de filtres puisqu'elle s'applique également à l'implémentation d'un banc de filtres en virgule fixe. L'originalité de ce chapitre réside dans la factorisation de *toutes* les matrices polyphases qui sont la base des bancs de filtres à reconstruction parfaite, et non pas seulement de celles qui sont paraunitaires [Vai2]. Il devrait pouvoir déboucher sur un algorithme de conception de filtres biorthogonaux, mais celui-ci rique d'être moins maniable que l'algorithme "direct" du chapitre VI.

Bien sûr, les résultats de ces derniers chapitres n'auraient pas pu être exposés sans ceux du chapitre II qui traitent principalement de la "banalisation" des bancs de filtres rationnels en montrant comment ils peuvent être rendus équivalents à des bancs de filtres uniformes —un résultat déjà connu auparavant [Hsi] et popularisé par [KV1,KV3] sous la forme d'une double transformation—, et des relations d'analyse-synthèse de bancs de deux filtres à échantillonnage fractionnaire. L'analogie avec le cas dyadique des relations (II.9) est alors frappant.

Ainsi qu'on l'a dit, le travail réalisé ici a dû s'efforcer d'être complet, ceci d'autant plus que le domaine était pratiquement vierge malgré l'activité de [KV3], essentiellement dans le domaine de la conception de filtres. Il était donc particulièrement intéressant de pouvoir appliquer tous ces résultats dans un but de codage de son haute fidélité, de pouvoir vérifier pratiquement l'intérêt d'une transformation qui semble poser autant de problèmes, plutôt que de s'en tenir à des techniques plus éprouvées. Il semble que la réponse soit plutôt affirmative, malgré un délai important qui semble prohiber la transformation dans des configurations de communication interactive. Les résultats peuvent être améliorés mais nécessitent, pour cela, un matériel de qualité. Il faut cependant garder à l'esprit que l'apport le plus important du chapitre concernant le codage du son est probablement la modélisation du phénomène de masquage, que l'on espère plus fidèle à la réalité dynamique. En ce sens, on peut imaginer la modification de l'algorithme de [ST] de façon à intégrer la technique de masquage présentée dans cette thèse: le gain immédiat en serait probablement la perte de l'effet de préécho et donc une simplicité accrue de l'algorithme.

Il reste ainsi beaucoup de travail à réaliser touchant aux préoccupations de cette thèse. Sur les bancs de filtres rationnels il reste bien sûr une foule de points non résolus: fonctions régulières orthonormées, conception de filtres biorthogonaux pour ne citer que quelques points. La recherche sur ces différents points serait bien sûr facilitée si de nouveaux sujets d'application étaient soumis à cette technique, qui permet, rappelons-le tout de même, d'analyser les signaux de façon plus fine que les classiques analyses en octave.

Références

[Au] P. Auscher, "Ondelettes Fractales et Applications", Thèse de doctorat, Université Paris IX, 1989

[Bi] G. Bi, "Minimization of Delay Requirements for Rational Sampling Rate Alternating Systems", *Proc. ICASSP* Mai 1991, Vol. 3, pp. 1817-1820, Toronto, Canada

[Blu1] T. Blu, "Iterated Filter Banks with Rational Rate Changes -- Connection with Discrete Wavelet Transforms", *IEEE Trans. SP Special Issue on Wavelets*, Vol. 41 No. 12, pp. 3232-3244, Dec. 1993

[Blu2] T. Blu, "Fractional Octave Filter Banks", *Proceedings du Symposium "Ondelettes et Opérateurs"*, Toulouse, France, 1992

[Blu3] T. Blu, "Lossless Filter Design in Two-Band Rational Filter Banks: A New Algorithm", *Proceedings du GRETSI*, Juan-les-Pins, France, 1993

[BR] T. Blu et O. Rioul, "Wavelet Regularity of Iterated Filter Banks with Rational Sampling Changes", *Proc. ICASSP* Avril 1993, Vol. III, pp. 213-216, Minneapolis, MN

[CD] A. Cohen et I. Daubechies, "Orthonormal bases of compactly supported wavelets III. Better frequency resolution", *Siam J. Math. Analysis*,Vol 24, No 2, pp. 520-527, Mars 1993

[CDF] A. Cohen, I. Daubechies et J.C. Feauveau, "Biorthogonal Basis of Compactly Supported Wavelets", *Comm. Pure and Applied Math.*, Vol 45, No 5, pp. 485-560, 1992

[CEG] A. Croisier, D. Esteban et C. Galand, "Perfect Channel Splitting by Use of Interpolation/Decimation/Tree Decomposition Techniques", *Proc. of Int. Conf. on Inform. Sc. and Sys.*, pp. 443-446, Patras, Août 1976

[CR] R.E. Crochiere et L.R. Rabiner, "Interpolation and Decimation of Digital Signals --- A Tutorial Review", *IEEE Proceedings*, Vol. 69, No. 3, pp. 300-331, Mars 1981

[CV] Z. Cvetković et M. Vetterli "Oversampled Filter Banks", à paraître dans *IEEE Trans. SP*, 1996

[DL] C. D'Alessandro et J.S. Lienard, "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation", *Proc. ICASSP* Avril 1988, pp. 351-354, New York

[Dau1] I. Daubechies , "Orthonormal Bases of Compactly Supported Wavelets", *Comm. on Pure and Applied Math.*, Vol. XLI, pp. 909-996, Novembre 1988

[Dau2] I. Daubechies , "Ten Lectures on Wavelets", *Philadelphia: CBMS-NSF Series in Appl. Math.*, SIAM Publ., 1992

[DauL1] I. Daubechies et J. Lagarias, "Two-Scale Difference Equations I. Existence and Global Regularity of Solutions", *Siam J. Math. Anal.*, Vol. 22, No. 5, pp. 1388--1410, Septembre 1991

[DauL2] I. Daubechies et J. Lagarias, "Two-Scale Difference Equations II. Local Regularity, Infinite Products of Matrices and Fractals", *Siam J. Math. Analysis*, Vol. 23 No. 4, 1031-1079, Juillet 1992

[Del] B. Delgutte, "Codage de la parole dans le nerf optique", Thèse de doctorat, *Univ. Paris VI*, 1984

[DLR] Y.F. Dehery, M. Lever et J.B. Rault, "Une norme de codage sonore de haute qualité pour la diffusion, les télécommunications et les systèmes mulimédias", *L'écho des recherches*, No. 151, pp. 17-28, 1er trimestre 1993

[DVN] Z. Doğanata, P.P. Vaidyanathan et T.Q. Nguyen, "General Synthesis Procedures for FIR Lossless Transfer Matrices, for Perfect-Reconstruction Multirate Filter Bank Applications", *IEEE Trans. ASSP*, Vol.36, No.10, pp.1561-1574, Octobre 1988

[DMP] P. Duhamel, Y. Mahieux et J.P. Petit, "A Fast Algorithm for the Implementation of Filter Banks Based on Time Domain Aliasing Cancellation", *Proc. ICASSP*, Mai 1991, Vol.3, pp.2209-2212, Toronto

[GGM] P. Goupillaud, A. Grossmann et J. Morlet, "Cycle-Octave and Related Transforms in Seismic Signal Analysis", *Geoexploration*, Vol. 23, pp. 85-102, Elsevier Science Publishers, Amsterdam, 1984/1985

[GM] C. Goulaouic et Y. Meyer, "Analyse Fonctionnelle et Calcul Différentiel", *Cours de mathématiques de l'École Polytechnique*, Palaiseau 1984

[HKMT] M. Holschneider, R. Kronland-Martinet, J. Morlet et Ph. Tchamitchian, "A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform", dans *Wavelets,Time-Frequency Methods and Phase Space*, J.M. Combes et al. eds, Springer, pp. 286-297,1989

[Hsi] C.-C. Hsiao, "Polyphase Filter Matrix for Rational Sampling Rate Conversions", *Proc. ICASSP* Avril 1987, pp. 2173-2176, Dallas, TX

[HV] P.Q. Hoang et P.P. Vaidyanathan, "Non-uniform Multirate Filter Banks: Theory and Design", *Proc. IEEE Int. Symp. Circ. Syst.*,pp. 371-374, Portland OR, 1989

[Jo] J. D. Johnston, "A Filter Family Designed for Use in Quadrature Mirror Filter Banks", *Proc. ICASSP* Avril 1980, pp. 291-294

[JB] J. D. Johnston et K. Brandenburg, "Wideband Coding —Perceptual Considerations for Speech and Music", dans *"Advances in Speech Signal Processing"*, S. Furui & M.M. Sondhi ed., New York 1991

[IM] "Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mbit/s",*Norme internationale MPEG1*, ISO No. 11172, 1992

[KB] S. Kadambe et G.F. Boudreaux-Bartels, "A Comparison of Wavelet Functions for Pitch Detection of Speech Signals", *Proc. ICASSP* Mai 1991, pp. 449-452, Toronto, Canada

[Kai] T. Kailath, "Linear Systems" , *Prentice Hall*, Englewood Cliffs, N.J. 1980

[KL] M.R.K. Khansari et A. Leon-Garcia "Subband Decomposition of Signals with Generalized Sampling", *IEEE Trans. SP*, Vol. 41 No. 12, pp. 3365-3376, Décembre1993

[Ko] J. Kovačević, "Filter Banks and Wavelets: Extensions and Applications", PhD Thesis, *Columbia University*, New York, NY, 1991

[KV1] J. Kovačević et M. Vetterli, "Perfect Reconstruction Filter Banks with Rational Sampling Rate Changes", *Proc. ICASSP* Mai 1991, Vol. 3, pp. 1785-1788, Toronto, Canada

[KV2] J. Kovačević et M. Vetterli, "Perfect Reconstruction Filter Banks with Rational Sampling Rate Changes in One and Two Dimensions", *Proc. of SPIE Conf. on Vis. Commun. and Image Proc.* Novembre 1989, pp. 1258-1268, Philadelphia, PA

[KV3] J. Kovačević et M. Vetterli, "Perfect Reconstruction Filter Banks with Rational Sampling Factors", *IEEE Trans. SP*, Vol. 41 No. 6, pp. 2047-2066, Juin 1993

[KMG] R. Kronland-Martinet, J. Morlet et A. Grossmann, "Analysis of Sound Patterns Through Wavelet Transforms", Int. J. Pattern Recognition and Artificial Intelligence, Vol. 1, No. 2, pp. 97-126, 1987

[Ma1] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Decomposition", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 11, No. 7, pp. 674-693, Juillet 1989

[Ma2] S. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models", *IEEE Trans. ASSP*, Vol. 37, No. 12, pp. 2091-2110, Decembre 1989

[Mah] Y. Mahieux, "Codage par transformation à 64 kbit/s pour les signaux audio de haute qualité", *Ann. Téléc.*, Vol. 47 No. 3-4, pp. 95-106, Mars-Avril 1992

[MPC] Y. Mahieux, J.P. Petit et A. Charbonnier, "Codage pour le transport du son de haute qualité sur le réseau des télécommunications", *L'écho des recherches*,No. 146, pp. .25-36, 4ème trimestre 1991

[Malv1] H.S. Malvar, "Lapped Transforms for Efficient Transform/Subband Coding", *IEEE Trans. ASSP*, Vol. 38, pp. 969-978, Juin 1990

[Malv2] H.S. Malvar, "Signal Processing with Lapped Transforms", Artech House, Norwood MA, 1992

[Mey1] Y. Meyer, "Ondelettes", Hermann ed., Paris 1990

[Mey2] Y. Meyer, "Orthonormal Wavelets", dans *Wavelets,Time-Frequency Methods and Phase Space*, J.M. Combes et al. eds, Springer, pp. 21-37, 1989

[NBS1] K. Nayebi, T.P. Barnwell III et M.J.T. Smith, "The Design of Perfect Reconstruction Nonuniform Band Filter Banks", *Proc. ICASSP* Mai 1991, Vol. 3, pp. 1781-1784, Toronto, Canada

[NBS2] K. Nayebi, T.P. Barnwell III et M.J.T. Smith, "Time-Domain Filter Bank Analysis: A New Design Theory", *IEEE Trans. SP*, Vol. 40, No. 6, pp. 1412-1429, Juin 1992

[NV] T.Q. Nguyen et P.P. Vaidyanathan, "Structures for M-Channel Perfect-Reconstruction FIR QMF Banks Which Yield Linear-Phase Analysis Filters", *IEEE Trans. SP*, Vol. 38, No. 3, pp. 433-446, Mars 1990

[PM] T.W. Parks et J.H. McClellan, "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase", *IEEE Trans. Circ. Th.*, Vol.19 No.2, pp.189-194, Mars 1972

[Ri1] O. Rioul, "Simple Regularity Criteria for Subdivision Schemes", *Siam J. Math. Anal.*, Vol. 23, No. 6, Novembre 1992

[Ri2] O. Rioul, "Regular Wavelets: A Discrete-Time Approach", *IEEE Trans. SP*, Vol. 41 No. 12, pp. 3572-3579, Décembre 1993

[Ri3] O. Rioul, "On the Choice of Wavelet Filters for Still Image Compression", *Proc. ICASSP*'93, Minneapolis

[Ri4] O. Rioul, "Ondelettes Régulières: Application à la Compression d'Images Fixes", Thèse de doctorat, *ENST Paris*, 1993

[RB] O. Rioul et T. Blu, "Simple Regularity Criteria for Subdivision Schemes. II. The Rational Case", en préparation

[RD1] O. Rioul et P. Duhamel, "Fast Algorithm for Discrete and Continuous Wavelet Transforms", *IEEE Trans. Inform. Theory*, Vol. 38, No. 2, pp. 569-586, Mars 1992

[RD2] O. Rioul et P. Duhamel, "A Remez Exchange Algorithm for Orthonormal Wavelets", *IEEE Trans. Circ. Syst.—A. and D. SP*, Vol. 41, No. 8, pp. 550-560, Août 1994

[RV] O. Rioul et M. Vetterli, "Wavelets and Signal Processing", *IEEE ASSP Magazine*, Vol. 8, No. 4, pp. 14-38, Octobre 1991

[SR] R.W. Schafer et L.R. Rabiner, "A Digital Signal Processing Approach to Interpolation", *IEEE Proceedings*, Vol. 61, No. 6, pp. 692-702, Juin 1973

[She] M.J. Shensa, "Affine Wavelets: Wedding the "à trous" and Mallat Algorithms", *IEEE Trans. SP*, Octobre 1992

[SB] M. J. T. Smith et T. P. Barnwell, "Exact Reconstruction Techniques for Tree-Structured Subband Coders", *IEEE Trans. ASSP*, Vol. 34, pp. 434-441, Juin 1986

[ST] D. Sinha et A.H. Tewfik, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelets", *IEEE Trans. SP*, Vol. 41, n°12, pp. 3463-3479, Décembre 1993

[SVN] A.K. Soman, P.P. Vaidyanathan et T.Q. Nguyen, "Linear Phase Orthonormal Filter Banks", *Proc. ICASSP*'93, Minneapolis

[UA] M. Unser et A. Aldroubi, "A General Sampling Theory for Nonideal Acquisition Devices", *IEEE Trans. SP*, Vol. 42 No. 11, pp. 2915-2925, Novembre 1994

[Vai1] P.P. Vaidyanathan, "Quadrature Mirror Filter Banks, M-Band Extensions and Perfect-Reconstruction Techniques", *IEEE ASSP Magazine*, Vol. 4, No. 3, pp. 4-20, Juillet 1987

[Vai2] P.P. Vaidyanathan, "Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial", *IEEE Proceedings*, Vol. 78, No. 1, pp. 56-93, Janvier 1990

[VC1] P.P. Vaidyanathan et T. Chen, "Role of Anticausal Inverses in Multirate Filter Banks—Part I: System-Theoretic Fundamentals", *IEEE Trans. SP*, Vol. 43 No. 5, pp. 1090-1102, Mai 1995

[VC2] P.P. Vaidyanathan et T. Chen, "Role of Anticausal Inverses in Multirate Filter Banks—Part II: The FIR Case, Factorizations, and Biorthogonal Lapped Transforms", *IEEE Trans. SP*, Vol. 43 No. 5, pp. 1103-1115, Mai 1995

[VH] P.P. Vaidyanathan et P.Q. Hoang, "Lattice Structures for Optimal Design and Robust Implementation of Two-Channel Perfect Reconstruction QMF Banks", *IEEE Trans. ASSP*, Vol. 36, No. 1, pp. 56-93, Janvier 1988

[VNDS] P.P. Vaidyanathan, T.Q. Nguyen, Z. Doğanata, et T. Saramäki, "Improved Technique for Design of Perfect Reconstruction FIR QMF Banks with Lossless Polyphase Matrices", *IEEE Trans. ASSP*, Vol.37, No.7, pp.1042-1056, Juillet 1989

[Vet] M. Vetterli, "A Theory of Multirate Filter Banks", *IEEE Trans. ASSP*, Vol.35, No.3, pp. 356-372, Mai 1987

[VG] M. Vetterli et D. Le Gall, "Perfect Reconstruction Filter Banks: Some Properties and Factorizations", *IEEE Trans. ASSP*, Vol.37, No.7, pp. 1057-1071, Juillet 1989

[YWS] X. Yang, K.Wang et S. Shamma, "Auditory Representations of Acoustic Signals", *IEEE Trans. Inf. Theory*, Vol. 38, n°2, pp. 824-839, Mars 1992

[Zem] W.R. Zemlin, "Speech and Hearing Science, Anatomy and Physiology", *Prentice Hall*, Englewood Cliffs NJ, 1981

[ZF] E. Zwicker et R. Feldtkeller, "L'oreille récepteur d'information" traduit par C. Sorin, Masson 1981