Codage et Compression des Signaux
Cours de DESS
2003-04, Marseille

Partie II

Bruno Torrésani
Université de Provence
Marseille

Table des matières

Préliminaires

On entend généralement par système de communication une série de traitements faisant intervenir des signaux de différents types (analogiques ou numériques), et dont le but est de transmettre l'information qu'ils contiennent, en optimisant un certain nombre de critères (volume de données, erreur, complexité,...).

On représente en général un système de communication sous la forme d'une série de "boites", affectées à des tâches particulières. Le plus classique de ces diagrammes est donné en FIGURE 1.

La boite `CANAL NUMERIQUE` représente quant à elle les opérations effectuées sur le signal numérisé, à savoir un recodage lui donnant une plus grande robustesse, une transformation en signal "physique" (électrique, ou "Hertzien") qui est le signal transmis, et les opérations inverses. Ces opérations sont synthétisées en FIGURE 2.

Pour compléter ces diagrammes, quelques définitions sont utiles.

- SIGNAL ANALOGIQUE : Signal fonction d'une (ou plusieurs) variable continue (réelle).
- SIGNAL NUMÉRIQUE : Signal indexé par un (ou plusieurs) indice discret.

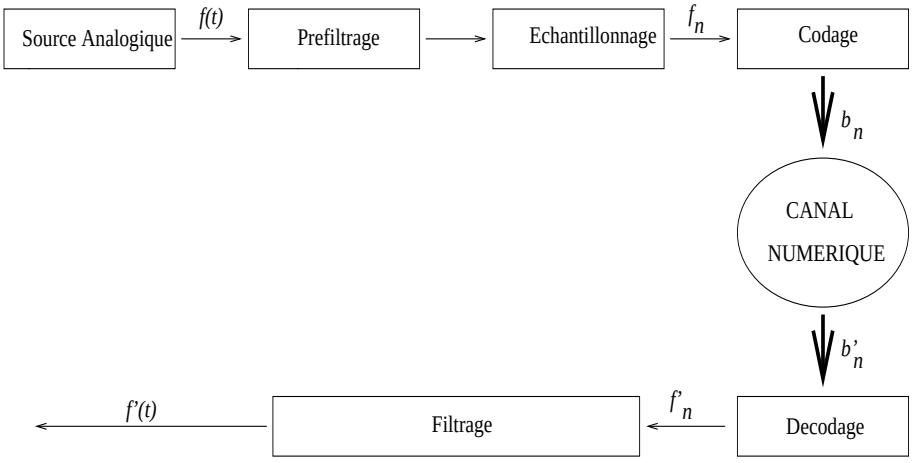

FIG. 1. Diagramme d'un systéme de communication. Les flèches "fines" représentent des signaux analogiques, et les flèches "épaisses" représentent des signaux numériques.

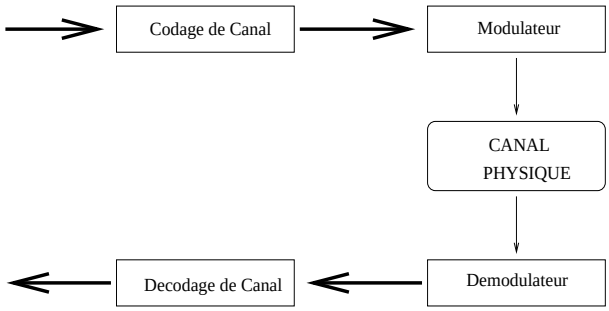

FIG. 2. Diagramme d'un canal numérique.

- FILTRE ANALOGIQUE : Opérateur linéaire continu sur $L^2(\mathbb{R})$, invariant par translation : si $f \in L^2(\mathbb{R})$, et si g est définie par $g(t) = f(t-\tau)$, alors $Tg(t) = Tf(t-\tau)$.
- FILTRE NUMÉRIQUE : Opérateur linéaire continu sur $\ell^2(\mathbb{Z})$, invariant par translation discrète.
- ECHANTILLONNAGE : Opération permettant de passer d'un signal analogique à un signal numérique. La façon la plus usuelle consiste à prendre des valeurs ponctuelles régulièrement espacées $f(n/\eta), n \in \mathbb{Z}$ du signal analogique $t \to f(t)$, quand celles ci sont bien définies. L'échantillonnage est généralement précédé d'un filtrage passe-bas, c'est à dire éliminant les hautes fréquences.
- QUANTIFICATION : passage d'un signal (généralement déjà échantillonné) prenant ses valeurs dans un ensemble continu à un signal prenant ses valeurs dans un ensemble fini.
- CODAGE, OU ALLOCATION DE MÉMOIRE : passage (sans perte d'information) d'un signal échantillonné et quantifié à un signal binaire. On regroupe souvent dans le terme codage la quantification et le codage proprement dit.
- CODAGE DE CANAL : Opération visant à introduire de la redondance dans un signal afin de le rendre plus robuste vis à vis des erreurs de transmission.
- MODULATEUR, DÉMODULATEUR : interface entre signal numérique et signal physique (courant électrique, onde électromagnétique,...).

Un système tel que celui décrit en figure 1 est généralement appelé CODEC (pour codeur-décodeur).

L'objet de ce cours est de dcrire l'ensemble de ces oprations, et de donner quelques exemples de mise en oeuvre de ces ides dans le cadre de schmas classiques de codage de signaux et d'images.

CHAPITRE 1

Introduction ; éléments de théorie du signal

On rappelle dans ce chapitre les aspects mathématiques de la représentation des signaux (signaux analogiques et numériques, signaux déterministes et aléatoires) ainsi que les transformations intégrales les plus couramment utilisées. On décrit également l'une des opérations les plus fondamentales, à savoir le filtrage des signaux.

1. Signaux déterministes

On se contente souvent dans un premier temps de construire des modèles de signaux déterministes, car ils sont plus faciles à manipuler. Les opérations définies sur les signaux déterministes sont généralement étendues ensuite au cadre aléatoire.

On distingue deux classes de signaux : les signaux analogiques, qui sont généralement des signaux "physiques" (ondes acoustiques ou électromagnétiques, courants électriques faibles,...), et les signaux numériques (que ce soit sous forme de suites de nombres ou de suites de bits). Les opérations effectuées dans l'un des deux cadres ont quasiment toujours leur pendant dans l'autre cadre, le monde analogique étant généralement plus complexe que le monde numérique.

1.1. Signaux analogiques ; filtrage analogique. Les signaux analogiques sont des signaux dépendant d'une variable continue (temps ou espace en général). L'un des outils les plus fondamentaux pour la représentation des signaux analogiques est la transformation de Fourier, que nous allons "visiter" dans deux versions ci-dessous.

1.1.1. *Signaux analogiques à support borné.* Un signal analogique à support borné est par définition une fonction $f : t \in [a, b] \to f(t) \in \mathbb{C}$. On notera $T = b - a$, et on identifiera souvent une telle fonction avec la fonction périodique de période $T = b-a$ avec laquelle elle coïncide dans $[a, b]$.

a. Fréquence, spectre,... Il est bien connu que de telles fonctions admettent généralement une représentation en *série de Fourier*

$$f(t) = \sum_{k=-\infty}^{\infty} c_k(f) e^{2i\pi kt/T} \ ,$$

où le coefficient de Fourier $c_k(f)$, défini par l'intégrale

$$c_k(f) = \frac{1}{T} \int_a^b f(t) e^{-2i\pi kt/T} \, dt$$

mesure le "contenu" de f à la fréquence k/T. Cette intégrale converge dès que $f \in L^1([a,b])$; cependant, le cadre L^2 est comme souvent plus "confortable", comme en atteste le résultat suivant :

THÉORÈME 1.1. *Soit $f \in L^2([a,b])$, et considérons la série de Fourier tronquée définie par*

$$f_N(t) = \sum_{k=-N}^{N} c_k(f) e^{2i\pi kt/T} \ . \tag{1.1}$$

Alors on a

$$\lim_{N\to\infty} \|f - f_N\| = 0 \ . \tag{1.2}$$

De plus, on a également la formule de Parseval : $\forall f, g \in L^2([a,b])$

$$\sum_{k=-\infty}^{\infty} c_k(f)\overline{c_k(g)} = \frac{1}{T}\int_a^b f(t)\overline{g(t)}\,dt = \frac{1}{T}\langle f, g\rangle \ . \tag{1.3}$$

En d'autres termes, la transformation linéaire $f \in L^2([a,b]) \to \{c_k(f)/\sqrt{T}, k \in \mathbb{Z}\}$ est une isométrie entre $L^2([a,b])$ et ℓ^2. Les nombres $|c_k(f)|^2$ forment le *spectre* de f.

REMARQUE 1.1. *La transformation $f \in L^2([a,b]) \to f_N$ n'est autre que la projection orthogonale de $L^2([a,b])$ sur le sous-espace $\mathcal{E}_N$ des polynomes trigonométriques de degré inférieurs à N.*

REMARQUE 1.2. *Cette représentation de Fourier se transporte sans difficulté au cas des fonctions de deux variables. Par exemple, pour toute fonction $f \in L^2([a,b]\times[a,b])$, on pourra écrire*

$$f = \lim_{N\to\infty} f_N \ ,$$

où

$$f_N(x,y) = \sum_{k=-N}^{N}\sum_{\ell=-N}^{N} c_{k\ell}(f)\, e^{2i\pi(kx+\ell y)/T} \ ,$$

et

$$c_{k\ell}(f) = \frac{1}{T^2}\int_a^b\int_a^b f(x,y)\, e^{2i\pi(kx+\ell y)/T}\, dxdy \ .$$

b. Filtrage analogique. L'opération de filtrage est l'une des opérations fondamentales du traitement du signal. Avant d'en voir une caractérisation plus simple et intuitive, commençons par en donner la définition. On se limite ici au cas des signaux à support borné dans l'intervalle $[0,1]$, il est facile d'en déduire le cas général. Il est utile d'introduire ici une notation. Pout $t \in \mathbb{R}$, on notera $[t]$ le nombre t modulo 1, en d'autres termes sa partie fractionnaire[1].

DÉFINITION 1.1. *Un filtre analogique sur $L^2([0,1])$ est un opérateur T, continu de $L^2([0,1])$ sur $L^2([0,1])$, qui commute avec les translations, dans le sens suivant : pour tout $f \in L^2([0,1])$ et $s \in [0,1]$,*

$$T\tau_s f = \tau_s T f$$

où τ_s est l'opérateur de translation "périodisé" défini par $\tau_s f(t) = f([t-s])$.

[1]Dans le cas où on travaille avec des fonctions définies dans un intervalle $[a,b]$, on notera $[t]$ le nombre $a+r$, où r est le reste de la division Euclidienne de $t-a$ par $(b-a)$.

Les filtres analogiques admettent la caractérisation simple suivante :

PROPOSITION 1.1. *Soit $T : L^2([0,1]) \to L^2([0,1])$ un filtre analogique. Alors il existe $m \in \ell^\infty(\mathbb{Z})$ tel que pour tout $f \in L^2([0,1])$, on ait*

$$c_k(Tf) = m_k c_k(f) \ , \quad \forall k \in \mathbb{Z} \ . \tag{1.4}$$

la suite m est appelée fonction de transfert *du filtre T.*

On peut également écrire (au sens de la convergence des séries de Fourier dans L^2)

$$Tf(t) = \sum_{k=-\infty}^{\infty} m_k c_k(f) e^{2i\pi kt} \ . \tag{1.5}$$

On donne ainsi une interprêtation tout à fait simple des opérations de filtrage : le filtrage vise essentiellement à modifier le contenu fréquentiel d'un signal, par exemple en atténuant certaines fréquences k (on a alors $|m_k| < 1$ ou en en renforçant d'autres (en prenant des valeurs $|m_k| > 1$).

Les exemples les plus simples (par certains aspects tout du moins) de filtres analogiques sont les filtres dits "idéaux", dont la fonction de transfert est une suite de zéros et de uns. Par exemple, le filtre passe-bas idéal, de fréquence de coupure K, est défini par la série de Fourier tronquée

$$Tf(t) = \sum_{k=-K}^{K} c_k(f) e^{2i\pi kt} \ ,$$

et correspond à $m_k = 1$ si $|k| \leq K$ et zéro sinon.

Une autre classe simple d'exemples est donnée par les filtres de convolution. Etant donnée une fonction $h \in L^2([0,1])$, on définit $K_h : f \to K_h f = h * f = f * h$:

$$K_h f(t) = \int_0^1 h(s) f([t-s])\, ds \ .$$

Un calcul simple montre que dans ce cas, K_h est bel et bien un filtre, de fonction de transfert définie par

$$m_k = c_k(h) \ .$$

La fonction h est appelée *réponse impulsionnelle* du filtre K_h.

REMARQUE 1.3. *Dans le cas des signaux à support borné, les filtres idéaux sont aussi des filtres de convolution ; par exemple le filtre passe-bas idéal ci-dessus est un filtre de convolution, de réponse impulsionnelle*

$$h(t) = \sum_{k=-K}^{K} e^{2i\pi kt} \ ,$$

qui n'est autre que le noyau de Fejér.

1.1.2. *Signaux analogiques à support infini.* Passons maintenant au cas, légèrement plus complexe, des signaux analogiques à support infini.

a. La transformation de Fourier intégrale. On travaillera la plupart du temps dans le cadre de l'espace $L^2(\mathbb{R})$ des fonctions de carré intégrable. La transformation de Fourier est l'opérateur $\mathcal{F}$ défini formellement par

$$[\mathcal{F}f](\nu) := \hat{f}(\nu) = \int_{-\infty}^{\infty} f(t)e^{-2i\pi\nu t}\,dt \tag{1.6}$$

Il est bien connu que cette transformation est bien définie sur l'espace des fonctions intégrables $L^1(\mathbb{R})$. Le résultat essentiel, appelé Théorème de Riemann-Lebesgue est le suivant.

THÉORÈME 1.2 (Riemann-Lebesgue). *Soit $f \in L^1(\mathbb{R})$, et soit $\hat{f}$ sa transformée de Fourier. Alors, $\hat{f}$ est une fonction bornée et continue, et $\hat{f}(\nu) \to 0$ quand $|\nu| \to \infty$.*

La transformation de Fourier est donc continue de $L^1(\mathbb{R})$ dans $L^\infty(\mathbb{R})$. Elle ne possède malheureusement pas d'inverse de $L^\infty(\mathbb{R})$ dans $L^1(\mathbb{R})$. Pour avoir une transformation de Fourier inverse, il faut faire des hypothèses supplémentaires. Introduisons tout d'abord $\overline{\mathcal{F}}$, défini par

$$[\overline{\mathcal{F}}\varphi](t) = \int_{-\infty}^{\infty} \varphi(s)e^{2i\pi ts}ds\ . \tag{1.7}$$

On peut alors montrer, par exemple :

PROPOSITION 1.2 (Dirichlet). *Soit $f \in L^1(\mathbb{R})$, telle que $\hat{f} \in L^1(\mathbb{R})$ également. Si f est continue en $t = t_0$, alors on a*

$$f(t_0) = \int_{-\infty}^{\infty} \hat{f}(\nu)e^{2i\pi\nu t_0}\,d\nu\ . \tag{1.8}$$

REMARQUE 1.4. On peut trouver les hypothèses de ce résultat contraignantes, ou difficiles à vérifier directement. On peut toutefois se contenter de conditions suffisantes : par exemple, si f est telle que $f, f', f'' \in L^1(\mathbb{R})$, alors $\hat{f} \in L^1(\mathbb{R})$, et la proposition s'applique. On va voir ci-dessous un autre cadre dans lequel l'inversion de la transformation de Fourier reste valide, dans un sens différent.

Une propriété remarquable de la transformation de Fourier est con comportement vis à vis de la différentiation : la transformation de Fourier "diagonalise" les opérateurs de différentiation. Plus précisément (dans le cadre $L^1(\mathbb{R})$, des résultats similaires peuvent être dérivés dans des cadres différents).

PROPOSITION 1.3. (1) *Soit $f \in L^1(\mathbb{R})$, telle que les fonction $t \to tf(t), \dots$ $t \to t^m f(t)$ soient elles aussi intégrables. Alors sa transformée de Fourier $\hat{f}$ admet en tout point m dérivées continues, et on a pour tout $k = 0, \dots m$*

$$\frac{d^k\hat{f}}{d\nu^k}(\nu) = \int_{-\infty}^{\infty} (-2i\pi t)^k f(t)e^{-2i\pi\nu t}\,dt \tag{1.9}$$

(2) *Soit $f \in L^1(\mathbb{R})$, et supposons que les dérivées $f^{(k)}, k = 1, \dots m$ de f existent presque partout et sont intégrables. Alors pour tout $k \leq m$, la fonction $f^{(k)}$ a pour transformée de Fourier la fonction $(2i\pi\nu)^k\hat{f}(\nu)$:*

$$[\mathcal{F}f^{(k)}](\nu) = (2i\pi\nu)^k\hat{f}(\nu)\ . \tag{1.10}$$

On travaillera la plupart du temps dans le cadre de l'espace $L^2(\mathbb{R})$ des fonctions de carré intégrable. La transformée de Fourier $\hat{f}$ d'une fonction $f \in L^2(\mathbb{R})$ n'est plus dans ce cas définie point par point, mais via un processus de passage à la limite

$$\hat{f}(\nu) = \lim_{T\to\infty} \int_{-T}^{T} f(t) e^{-2i\pi\nu t} dt$$

utilisant la densité de $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ dans $L^2(\mathbb{R})$. Cependant, $\hat{f}$ est elle même de carré intégrable, et il est possible de donner un sens à la formule d'inversion. On résume ces propriétés dans le résultat suivant

THÉORÈME 1.3. *Soit $f \in L^2(\mathbb{R})$. Alors $\hat{f} \in L^2(\mathbb{R})$, et on a la formule de Plancherel*

$$\|\hat{f}\|^2 = \|f\|^2 \ . \tag{1.11}$$

Plus généralement, si $f, g \in L^2(\mathbb{R})$, on a

$$\langle \hat{f}, \hat{g} \rangle = \langle f, g \rangle \ . \tag{1.12}$$

REMARQUE 1.5. Un corollaire important est que si $f \in L^2(\mathbb{R})$, alors $\hat{f} \in L^2(\mathbb{R})$, et $\overline{\mathcal{F}}\hat{f}$ définit également une fonction de $L^2(\mathbb{R})$, qui coïncide avec f presque partout. Ceci donne un sens à la formule d'inversion de la transformation de Fourier dans $L^2(\mathbb{R})$.

La transformation de Fourier possède un grand nombre de propriétés simples, que l'on trouve dans tous les livres. On donne ici simplement un résultat, qui sera utile par la suite, conne sous le nom d'inégalité de Heisenberg. Cette inégalité montre qu'une fonction et sa transformée de Fourier ne peuvent être simultanément aussi bien localisées que ce que l'on pourrait le vouloir. Il faut pour cela introduire des mesures de localisation. Etant donnée une fonction $f \in L^2(\mathbb{R})$, on introduit sa moyenne μ_f par

$$\mu_f = \frac{1}{\|f\|^2} \int_{-\infty}^{\infty} t|f(t)|^2 dt \ , \tag{1.13}$$

et son écart quadratique moyen σ_f par

$$\sigma_f^2 = \frac{1}{\|f\|^2} \int_{-\infty}^{\infty} (t-\mu_f)^2 |f(t)|^2 dt \ , \tag{1.14}$$

pour peu que ces deux intégrales soient convergentes. On définit de même les quantités analogues dans le domaine fréquentiel : $\mu_{\hat{f}}$ et $\sigma_{\hat{f}}$. Les nombres μ sont des mesures de localisation, alors que les σ sont des mesures de "concentration". On a donc alors :

PROPOSITION 1.4. *Soit $f \in C^1(\mathbb{R})$, telle que les fonctions $f(t), f'(t)$ et $tf(t)$ soient de carré intégrable. Alors on a*

$$\sigma_f \sigma_{\hat{f}} \geq \frac{1}{4\pi} \ . \tag{1.15}$$

Preuve : On peut sans perte de généralité supposer que $\mu_f = \mu_{\hat{f}} = 0$. La preuve est une conséquence de l'inégalité de Cauchy-Schwarz : écrivons

$$\int_{-\infty}^{\infty} t \frac{d}{dt} |f(t)|^2 dt = \left[t|f(t)|^2\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} |f(t)|^2 dt$$

On montre que sous les hypothèses plus haut, $\lim_{t\to\pm\infty} t|f(t)|^2 = 0$. Donc, en prenant la valeur absolue et en développant la dérivée, on a

$$||f||^2 \leq \left| \int_{-\infty}^{\infty} tf(t)\overline{f'}(t)dt \right| + \left| \int_{-\infty}^{\infty} tf'(t)\overline{f}(t)dt \right|$$

En appliquant l'inégalité de Cauchy-Schwarz à ces deux termes, on aboutit à

$$||f||^2 \leq 2\sqrt{\int_{-\infty}^{\infty} t^2|f(t)|^2dt \ \int_{-\infty}^{\infty} |f'(t)|^2dt} \ ,$$

Or, nous savons que la transformée de Fourier de f' n'est autre que la fonction $\nu \to 2i\pi\nu\hat{f}(\nu)$ (voir la PROPOSITION 1.3). En utilisant la formule de Plancherel, nous obtenons

$$\int_{-\infty}^{\infty} |f'(t)|^2dt = 4\pi^2 \int_{-\infty}^{\infty} \nu^2|\hat{f}(\nu)|^2d\nu = 4\pi^2\|f\|^2\sigma_{\hat{f}}^2 \ .$$

On a bien le résultat désiré, dans le cas particulier $\mu_f = \mu_{\hat{f}} = 0$

Pour le cas général, considérons la fonction g définie par

$$g(t) = e^{-2i\pi\mu_{\hat{f}}t}f(t + \mu_f) \ .$$

Un calcul immédiat montre que $\mu_g = \mu_{\hat{g}} = 0$, de sorte que l'on peut appliquer à g le résultat que nous venons de montrer. Or, on a $\|g\| = \|f\|$,

$$\sigma_g^2 = \frac{1}{\|g\|^2} \int_{-\infty}^{\infty} t^2|f(t + \mu_f)|^2dt = \sigma_f^2 \ ,$$

et

$$\sigma_{\hat{g}}^2 = \frac{1}{\|\hat{g}\|^2} \int_{-\infty}^{\infty} \nu^2|\hat{f}(\nu + \mu_{\hat{f}})|^2d\nu = \sigma_{\hat{f}}^2 \ .$$

Ceci conclut la démonstration. ♠

<u>b.</u> *Convolution-Produit ; filtrage analogique.* Les opérations de filtrage sont absolument essentielles en traitement du signal. Commençons par donner la définition mathématique d'un filtre analogique.

DÉFINITION 1.2. *Un filtre analogique est un opérateur T, continu de $L^2(\mathbb{R})$ dans $L^2(\mathbb{R})$, et invariant par translation : si $f \in L^2(\mathbb{R})$, et si g est définie par $g(t) = f(t-\tau)$, alors $Tg(t) = Tf(t-\tau)$. Si pour tout $f \in L^\infty(\mathbb{R})$, $Tf \in L^\infty(\mathbb{R})$, le filtre est dit stable.*

Le résultat suivant donne une caractérisation des filtres analogiques.

THÉORÈME 1.4. *Soit $T : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ un filtre analogique. Alors il existe une fonction $m \in L^\infty(\mathbb{R})$ telle que l'on ait, pour tout $f \in L^2(\mathbb{R})$*

$$Tf(t) = \int_{-\infty}^{\infty} e^{2i\pi\nu t}m(\nu)\hat{f}(\nu)d\nu \ . \tag{1.16}$$

La fonction m est appelée fonction de transfert *du filtre T.*

Le cas le plus simple de filtre analogique est celui des filtres de convolution. Etant donnée une fonction $h \in L^1(\mathbb{R})$, on considère l'opérateur K_h défini par $K_hf = h * f$, autrement dit

$$(K_hf)(t) = \int_{-\infty}^{\infty} h(s)f(t-s)\,ds \ .$$

Il est facile de voir que comme $h \in L^1(\mathbb{R})$, $K_h f \in L^2(\mathbb{R})$ dès que $f \in L^2(\mathbb{R})$; de plus, on a

$$\widehat{K_h f}(\nu) = \hat{h}(\nu)\hat{f}(\nu) \ .$$

REMARQUE 1.6. Evidemment, tous les filtres analogiques ne sont pas des filtres de convolution. Le contre exemple le plus immédiat est l'opérateur identité, qui est évidemment un filtre analogique, mais qui n'est pas un filtre de convolution (sauf à admettre des réponses impulsionnelles qui soient des distributions). D'autres exemples sont fournis par des "systèmes de lignes à retard", de la forme

$$Tf(t) = \sum_{k=0}^{K-1} h_k f(t - t_k) \ ,$$

où $h \in \ell^1(\mathbb{Z})$ est une suite finie, et où les t_k sont des réels quelconques (en général, régulièrement espacés).

DÉFINITION 1.3. *Lorsque la fonction de transfert m d'un filtre analogique est transformée de Fourier d'une fonction intégrable $h \in L^1(\mathbb{R})$, celle-ci est appelée* réponse impulsionnelle *du filtre. Le filtre est dit* réalisable *(ou* causal*) si $h(t) = 0$ pour tout $t \leq 0$. Un filtre stable et causal est dit* dynamique*.*

REMARQUE 1.7. La notion de réalisabilité provient tout droit du traitement du signal analogique. Si h est la réponse impulsionnelle d'un filtre T, on écrit $Tf(t) = \int_{-\infty}^{\infty} h(t-s)f(s)ds$, et on voit immédiatement que si le filtre n'est pas réalisable, le calcul de $Tf(t)$ nécessite la connaissance des valeurs de $f(s)$ pour $s \leq t$, ce qui n'est pas compatible avec une implémentation analogique.

Le filtrage est souvent utilisé pour modifier le contenu fréquentiel des signaux, en particulier pour en diminuer le contenu aux hautes fréquences (on parle alors de filtrage passe-bas). On peut en voir un exemple en FIG. 1, qui représente un signal transitoire (c'est à dire à variations rapides), et deux versions obtenues par filtrage passe-bas. On remarque que dans le premier signal filtré (figure du milieu), les caractéristiques les plus rapidement variables du signal ont disparu, et le signal obtenu est plus "lisse" que l'original. Dans l'autre cas, le filtrage est plus fort (la fonction de transfert a atténué davantage de fréquences), et le signal obtenu est encore bien plus lisse.

Le filtre passe-bas idéal de fréquence de coupure ν_0 a pour fonction de transfert

$$m(\nu) = \chi_{[-\nu_0,\nu_0]}(\nu) \ ,$$

mais il est facile de vérifier qu'un tel filtre n'est ni stable ni réalisable. En fait, on peut associer à ce filtre une réponse impulsionnelle $h : t \to h(t) = 2\nu_0 \mathrm{sinc}(2\pi\nu_0 t)$, mais celle-ci n'est pas intégrable. On a plutôt recours à des filtres plus élaborés, comme par exemple les filtres de "Butterworth"

$$|m(\nu)|^2 = \frac{1}{1 + (\nu/\nu_0)^{2n}}$$

ou les filtres de Chebyshev, montrés en FIG. 2.

$$|m(\nu)|^2 = \frac{1}{1 + \epsilon T_{2n}(2\pi\nu)} \ ,$$

où les T_n sont les polynômes de Chebyshev, définis par

$$T_n(2\pi\nu) = \cos(n \ \mathrm{arc}\cos(2\pi\nu)) \ .$$

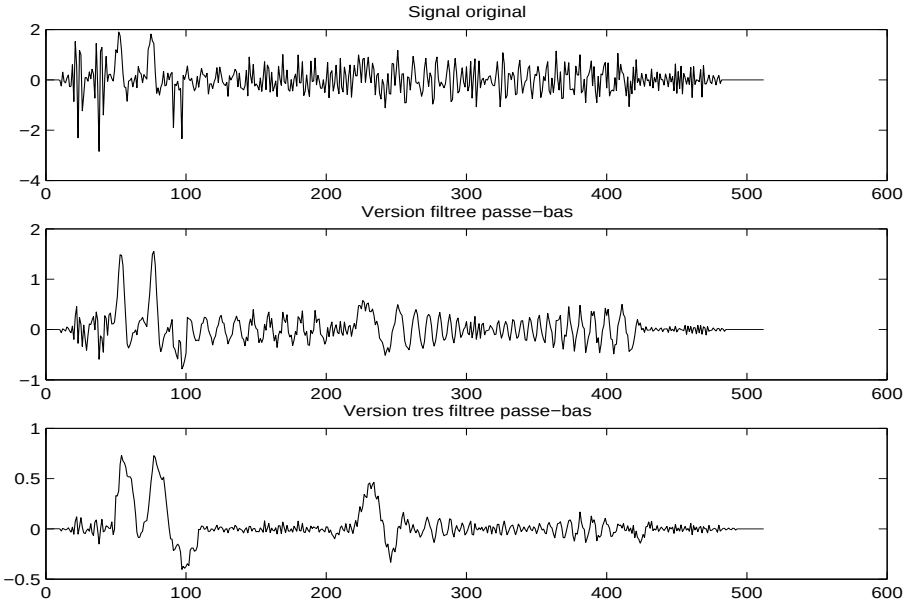

FIG. 1. Exemple de filtrage passe-bas : un signal transitoire, et deux versions filtrées avec des fréquences de coupure différentes.

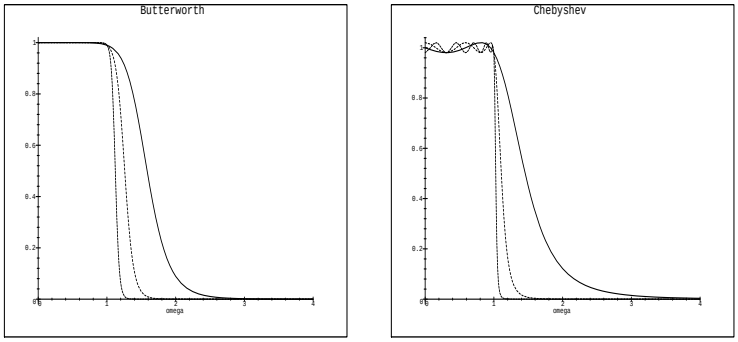

FIG. 2. Fonctions de transfert des filtres de Butterworth (à gauche) et de Chebyshev (à droite)

On montre que les polynômes de Chebyshev de degré pair sont pairs, à coefficients rèels ; les filtres correspondants sont stables et réalisables. Voir plus bas pour plus de détails.

c. Filtres et équations différentielles ; synthèse de filtres analogiques. Une façon simple de construire des filtres réalisables est d'utiliser des circuits électriques. Par exemple, un circuit du type de celui de la FIG. 3.

En notant $v(t) = Q(t)/C$ la tension aux bornes du condensateur, la loi d'Ohm s'écrit

$$Ri(t) + v(t) = u(t) \ , \tag{1.17}$$

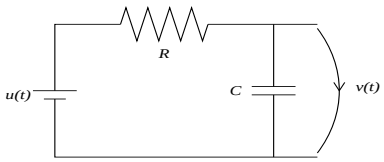

FIG. 3. Le filtre RC

ce qui entraîne, puisque $i(t) = Q'(t) = Cv'(t)$, que la tension $v(t)$ satisfait à l'équation différentielle ordinaire

$$RC\, v'(t) + v(t) = u(t)\ . \tag{1.18}$$

Pour résoudre cette dernière, il est utile d'introduire la fonction $w(t) = v(t)e^{t/RC}$. On a donc

$$w'(t) = \frac{1}{RC} e^{t/RC} u(t) \tag{1.19}$$

et la solution est

$$w(t) = \frac{1}{RC} \int_{-\infty}^{t} e^{s/RC} u(s) ds\ , \tag{1.20}$$

soit

$$v(t) = \frac{1}{RC} \int_{-\infty}^{t} e^{-(t-s)/RC} u(s) ds \tag{1.21}$$

$$= \int_{-\infty}^{\infty} h(t-s) u(s) ds\ . \tag{1.22}$$

où nous avons posé

$$h(t) = \Theta(t) e^{-t/RC}$$

$\Theta(t)$ étant la fonction d'Heaviside, qui vaut 1 pour $t \geq 0$ et 0 pour $t < 0$. Nous sommes bien en présence d'un filtre réalisable et stable.

Ce résultat peut également s'obtenir en utilisant les propriétés remarquables de la transformation de Fourier vis à vis de la dérivation, (PROPOSITION 1.3)

Plus généralement, on peut à partir de circuits analogiques obtenir des systèmes linéaires dont l'entrée $u(t)$ et la sortie $v(t)$ sont liés par une équation différentielle du type

$$a_N v^{(N)} + a_{N-1} v^{(N-1)} + \cdots + a_0 v = b_M u^{(M)} + b_{M-1} u^{(M-1)} + \cdots + b_0 u\ , \tag{1.23}$$

complétée par des conditions initiales adéquates. Un calcul immédiat (basé sur la proposition précédente) montre que la fonction de transfert correspondante $m(\nu)$ est donnée par

$$m(\nu) = \frac{a_N (2i\pi\nu)^N + \cdots + a_0}{b_M (2i\pi\nu)^M + \cdots + b_0} = \frac{N(2i\pi\nu)}{D(2i\pi\nu)} \tag{1.24}$$

Pour que la fonction de transfert soit bornée, on a nécessairement $M \geq N$.

Le résultat suivant donne une première caractérisation de la structure des filtres dynamiques que l'on peut obtenir à base de circuits analogiques.

PROPOSITION 1.5. *Le filtre défini par la fonction de transfert (1.24) est dynamique si et seulement si les racines de $D(z)$ ont une partie réelle négative.*

Preuve : Notons α_k les racines (complexes) de $D(z)$, et soit m_k leur multiplicité :

$$D(z) = C \prod_k (z - \alpha_k)^{m_k} \ .$$

En décomposant $m(\nu)$ en éléments simples, on aboutit à une forme

$$m(\nu) = C' + \sum_k \sum_{\ell=1}^{m_k} \frac{P_\ell(2i\pi\nu)}{(2i\pi\nu - \alpha_k)^\ell} \ ,$$

où P_ℓ est un polynôme de degré inférieur à ℓ. C' est une constante, qui correspond à un multiple de l'identité. On supposera dans la suite que $C' = 0$ (ce qui est le cas dès que $M > N$).

Soit $\rho_0(t) = \Theta(t)e^{\alpha t}$, où $\Re(\alpha) < 0$. Un calcul immédiat donne $\hat{\rho}(\nu) = 1/(2i\pi\nu - \alpha)$. Plus généralement, si

$$\rho_\ell(t) = t^{\ell-1} e^{\alpha t} \Theta(t) \ ,$$

on a

$$\hat{\rho}_\ell(\nu) = \frac{(\ell-1)!}{(2i\pi\nu - \alpha)^\ell} \ .$$

Donc, la transformée de Fourier inverse de $P_\ell(2i\pi\nu)/(2i\pi\nu - \alpha)^\ell$ est

$$P_\ell\left(\frac{d}{dt}\right) \frac{t^{\ell-1}e^{\alpha t}}{(\ell-1)!} \Theta(t) \ .$$

Par conséquent, si $\Re(\alpha_k) > 0$ pour tout k, la réponse impulsionnelle du filtre est de la forme

$$h(t) = \Theta(t) \sum_k Q_k(t) e^{\alpha_k t} \ ,$$

où les Q_k sont des polynômes. Il s'agit bien de filtres stables et réalisables.

Inversement, supposons que pour une certaine racine α de $D(z)$, on ait $\Re(\alpha) > 0$. Un calcul similaire au précédent montre que la transformée de Fourier inverse de $P_\ell(2i\pi\nu)/(2i\pi\nu - \alpha)^\ell$ est proportionnelle à $\Theta(-t)$, ce qui est incompatible avec la causalité. Ceci achève la preuve de la proposition. ♠

On s'intéresse souvent au problème de construire des filtres à partir d'une réponse attendue sur le spectre du signal. Plus précisément, on recherche à construire un filtre de fonction de transfert $m(\nu)$ telle que $|m(\nu)|^2 = M(\nu)$, où M est une fonction donnée, paire à valeurs réelles positives.

LEMME 1.1. *Soit P un polynôme à coefficients réels, pair et non négatif. Alors, il existe un polynôme $\nu \to Q(2i\pi\nu)$ tel que $P(\nu) = |Q(2i\pi\nu)|^2$.*

Preuve : Soit $\gamma \in \mathbb{C}$ une racine de P. D'après la parité de P, $-\gamma$ est également racine de P. Si γ est complexe, $\overline{\gamma}$ et $-\overline{\gamma}$ sont également racines. Si $\gamma \notin \mathbb{R}$ et $\gamma \notin i\mathbb{R}$, $P(\nu)$ contient nécessairement un terme de la forme

$$\begin{aligned}(\nu-\gamma)(\nu-\overline{\gamma})(\nu+\gamma)(\nu+\overline{\gamma}) &= C\,(2i\pi\nu-\alpha)(2i\pi\nu-\overline{\alpha})(2i\pi\nu+\alpha)(2i\pi\nu+\overline{\alpha}) \\ &= C\,|(2i\pi\nu-\alpha)(2i\pi\nu-\overline{\alpha})|^2 \\ &= C\,|(2i\pi\nu+\alpha)(2i\pi\nu+\overline{\alpha})|^2 \ ,\end{aligned}$$

où $C = (2\pi)^{-4}$ est une constante et $\alpha = 2i\pi\gamma$. Ce terme a bien la forme annoncée.

Si $\gamma \in \mathbb{R}$, alors γ est nécessairement de multiplicité paire : en effet, $P(\nu)$ contient obligatoirement un terme en $(\nu^2 - \gamma^2)^\mu$, μ étant la multiplicité de γ, et pour que P soit positif μ doit nécessairement être pair. Donc

$$(\nu^2 - \gamma^2)^\mu = \left|(\nu^2 - \gamma^2)^{\mu'}\right|^2 = C' \left|(2i\pi\nu - \alpha)^{\mu'}(2i\pi\nu + \alpha)^{\mu'}\right|^2 ,$$

avec $\mu' \in \mathbb{Z}^+$ et toujours $\alpha = 2i\pi\gamma$, est lui aussi de la forme annoncée.

Si $\gamma \in i\mathbb{R}$, $P(\nu)$ contient nécessairement un terme en $(\nu^2 - \gamma^2)^\mu$, qui est toujours positif, et de la forme

$$(\nu^2 - \gamma^2)^\mu = C' \left|(2i\pi\nu - \alpha)^\mu\right|^2 .$$

On peut alors construire $Q(2i\pi\nu)$ en ne conservant que les racines de partie réelle négative, et les racines imaginaires avec la moitié de leur multiplicité. ♠

On peut maintenant passer au cas des filtres rationnels. Soit donc M une fonction rationnelle de ν, de la forme

$$M(\nu) = \frac{N(\nu)}{D(\nu)} .$$

On peut alors appliquer le traitement précédent à $N(\nu)$ et $D(\nu)$, et on obtient :

PROPOSITION 1.6. *Soit $M : \nu \to M(\nu) = N(\nu)/D(\nu)$ une fraction rationnelle, où N et D sont deux polynômes réels pairs et strictement positifs. Alors il existe deux polynômes $n(2i\pi\nu)$ et $d(2i\pi\nu)$ tels que*

$$M(\nu) = \frac{N(\nu)}{D(\nu)} = \left|\frac{n(2i\pi\nu)}{d(2i\pi\nu)}\right|^2 . \tag{1.25}$$

De plus, n et d peuvent être choisis de sorte que le filtre dont la fonction de transfert est $\nu \to n(2i\pi\nu)/d(2i\pi\nu)$ soit réalisable.

Preuve : Il suffit d'utiliser le lemme précédent, en sélectionnant la factorisation adaptée du dénominateur. ♠

EXEMPLE 1.1. Les deux familles classiques d'exemples de filtres rationnels approchant des filtres idéaux sont fournies par les filtres de Butterworth et les filtres de Chebyshev. On se limite ici au cas des filtres passe-bas.

Les filtres de Butterworth sont les plus simples, et sont donnés par une fonction de transfert de la forme

$$M_n^B(\nu) = \frac{1}{1 + \left(\frac{\nu}{\nu_c}\right)^{2n}} \tag{1.26}$$

$M_n^B(\nu)$ approche la fonction de transfert (en module carré) d'un filtre passe-bas idéal, de fréquence de coupure ν_c. Les pôles correspondants sont égaux aux racines $2n$-ièmes de -1, multipliées par ν_c. L'avantage des filtres de Butterworth est que la fonction $M_n^B(\nu)$ est "plate" en $\nu \approx 0$. Plus précisément, elle se comporte comme $M_n^B(\nu) \sim 1 + \nu^{2n}$ pour $\nu \approx 0$.

Une alternative est fournie par les filtres de Chebyshev, définis à partir des polynômes de Chebyshev

$$T_n(x) = \cos(n \arccos x) ,$$

par

$$M_n^C(\nu) = \frac{1}{1 + \epsilon\, T_{2n}\left(\frac{\nu}{\nu_c}\right)} \ . \tag{1.27}$$

Le paramètre ν_c contrôle la largeur du filtre. Les filtres de Chebyshev présentent quant à eux l'avantage d'être mieux localisés (plus "étroits"), mais ceci se fait au prix d'oscillations apparaissant pour $\nu \approx 0$. L'amplitude des oscillations est gouvernée par le paramètre ϵ.

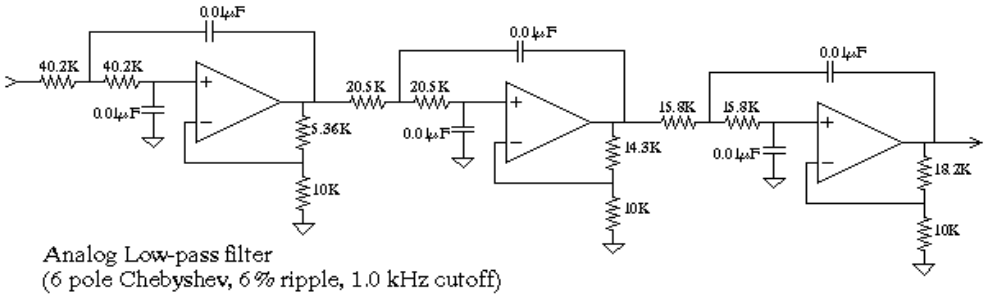

FIG. 4. Circuit électrique correspondant à un filtre de Chebyshev à 6 pôles.

Des exemples de fonctions M pour les cas Chebyshev et Butterworth avec des nombres variables de pôles se trouvent en FIG. 2. Un exemple de circuit électrique produisant un filtre de Chebyshev est montré en FIG. 4 (simplement pour prouver que ça existe pour de vrai...).

d. La transformation de Laplace. La discussion précédente nous a implicitement placés dans un cadre de fonctions d'une variable complexe, puisque nous en sommes arrivés à étudier les zéros et les pôles dans le plan complexe de la fonction de transfert m des filtres considérés. C'est pour cela qu'il est parfois utile d'introduire la *transformation de Laplace*, comme alternative à la transformation de Fourier. C'est en tous cas un langage qu'utilisent souvent les ingénieurs. On utilise deux variantes de la transformation de Laplace, la transformation unilatérale et la transformation bilatérale, adaptées respectivement aux cas des signaux causaux et des signaux plus généraux. On rappelle qu'une fonction f est localement intégrable sur un domaine Ω (ce que l'on note $f \in L^1_{loc}(\Omega)$) si f est intégrable sur tout domaine compact inclus dans Ω.

DÉFINITION 1.4. (1) *Soit $f \in L^1_{loc}(\mathbb{R})$. Sa transformée de Laplace (bilatérale) est la fonction $F = \mathcal{L}f$ de la variable complexe s définie par*

$$F(p) = \int_{-\infty}^{\infty} f(t) e^{-pt}\, dt \ , \tag{1.28}$$

pour tout $s \in \mathbb{C}$ tel que l'intégrale soit convergente.

(2) *Soit $f \in L^1_{loc}(\mathbb{R}^+)$. Sa transformée de Laplace (unilatérale) est la fonction $F = \mathcal{L}f$ de la variable complexe p définie par*

$$F(p) = \int_{0}^{\infty} f(t) e^{-pt}\, dt \ , \tag{1.29}$$

Sans entrer dans les détails (voir par exemple [**14**] pour une discussion plus complète), on peut quand même mentionner les propriétés essentielles suivantes de la transformation de Laplace :

(1) Soit $f \in L^1_{loc}(\mathbb{R})$ ou $L^1_{loc}(\mathbb{R}^+)$. Il existe deux nombres $s_1, s_2 \in \overline{\mathbb{R}}$ (l'axe réel complété par $\pm\infty$) tels que l'intégrale définissant $F(p)$ soit absolument convergente pour tout $p \in \mathbb{C}$ avec $s_1 < \Re(p) < s_2$. Ces nombres sont appelés appelés abscisses d'intégrabilité, et le domaine correspondant dans le plan complexe est la bande d'intégrabilité (qui peut éventuellement être vide si $s1 = s_2$...).

(2) La transformée de Laplace F de f est holomorphe à l'intérieur de la bande d'intégrabilité.

(3) Lorsque f est un signal causal et F est sa transformée de Laplace unilatérale, alors $s_2 = \infty$.

(4) Soit $f \in L^1_{loc}(\mathbb{R})$ ou $L^1_{loc}(\mathbb{R}^+)$, et soit F sa transformée de Laplace (bilatérale ou unilatérale). Si pour tout $\gamma \in \mathbb{R}$ tel que $s_1 < \gamma < s_2$, la fonction $\nu \to F(2i\pi\nu + \gamma)$ appartient à $L^1(\mathbb{R})$, on a la formule d'inversion

$$f(t) = \lim_{u\to\infty} \int_{\gamma-2i\pi u}^{\gamma+2i\pi u} F(p)\, e^{pt}\, dt \ , \tag{1.30}$$

valable pour tout $\gamma \in]s_1, s_2[$, où l'intégrale est prise sur une droite verticale du plan complexe, strictement incluse dans la bande d'intégrabilité. Les calculs de transformée de Laplace inverse font généralement appel à la méthode des résidus (voir par exemple [**13, 14**] pour plus de détails).

Le comportement de la transformation de Laplace vis à vis du filtrage est assez similaire au comportement de la transformation de Fourier. Essentiellement, un filtrage est défini dans le domaine des transformées de Laplace par la multiplication par une fonction holomorphe m, elle aussi appelée fonction de transfert. Le domaine dans lequel est définie la transformée de Laplace est souvent la question essentielle.

Par exemple, supposons $f \in L^1_{loc}(\mathbb{R})$, et soit F sa transformée de Laplace définie pour $\Re(p) \in]s_1, s_2[$. Soit $g \in L^1_{loc}(\mathbb{R})$, de transformée de Laplace G définie dans le domaine $\Re(p) \in]s'_1, s'_2[$. Alors, la transformée de Laplace du produit de convolution $f * g$ est définie dans le domaine

$$\Re(p) \in] \max(s_1, s'_1), \min(s_2, s'_2)[\ ,$$

et est donnée par

$$[\mathcal{L}(f * g)](p) = F(p)\, G(p) \ .$$

1.2. Signaux numériques. Par définition, les signaux numériques sont des suites (signaux indexés par un ensemble discret), finies ou infinies.

1.2.1. *Signaux infinis.* La version de la transformation de Fourier adaptée à cette situation est la transformée de Fourier discrète, qui est fortement liée à la théorie des séries de Fourier. On commence par revenir brièvement sur cette dernière.

a. Coefficients de Fourier. Revenons sur l'espace $L^2_p([a,b])$ des fonctions périodiques de période $b-a$, de carré intégrable sur l'intervalle $[a,b]$ (et donc sur tout intervalle de longueur $b-a$), muni du produit scalaire

$$\langle f,g\rangle = \int_a^b f(t)\overline{g}(t)\,dt \ . \tag{1.31}$$

On considère les fonctions trigonométriques

$$e_n : t \to e_n(t) = \frac{1}{\sqrt{b-a}}\,\exp\left(2i\pi\frac{nt}{b-a}\right) \ . \tag{1.32}$$

On vérifie facilement que la famille de fonctions $\{e_n, n\in\mathbb{Z}\}$ est un système orthonormal dans $L^2_p([a,b])$. Il s'agit en fait d'une base orthonormale, comme le montre le résultat suivant

THÉORÈME 1.5. *La famille des fonctions exponentielles e_n définies en (1.32) est une base orthonormée de $L^2_p([a,b])$. Pour tout $f\in L^2_p([a,b])$, on a*

$$||f-f_N||^2 = \sum_{|n|>N} |c_n(f)|^2 \longrightarrow 0 \ \textit{quand}\ N\to\infty \ . \tag{1.33}$$

De plus, la formule de Parseval s'écrit, pour toutes $f,g\in L^2_p([a,b])$

$$\sum_{-\infty}^{\infty} c_n(f)\overline{c_n}(g) = \frac{1}{b-a}\int_a^b f(t)\overline{g}(t)\,dt \ . \tag{1.34}$$

b. Transformation de Fourier discrète. Les résultats obtenus plus haut (séries de Fourier) se transposent de façon immédiate au cas des signaux numériques. En effet, la théorie L^2 des séries de Fourier permet de construire une isométrie bijective entre $L^2_p([-\pi,\pi])$ et $\ell^2(\mathbb{Z})$. La transformation inverse porte le nom de transformation de Fourier discrète.

DÉFINITION 1.5. *Soit $s=\{s_n\}\in\ell^2(\mathbb{Z})$. Sa transformée de Fourier discrète est la fonction 1-périodique $\nu\to\hat{s}(\nu)$ définie par*

$$\hat{s}(\nu) = \sum_{n=-\infty}^{\infty} s_n e^{-2i\pi\nu n} \ , \tag{1.35}$$

pour tout ν tel que la série soit convergente.

Il résulte de la théorie des séries de Fourier que la TFD d'une suite de $\ell^2(\mathbb{Z})$ est une fonction 1-périodique, de carré intégrable sur $[-1/2,1/2]$, et que la transformation inverse est donnée par le calcul des coefficients de Fourier de $\hat{s}$. Plus précisément, on a

THÉORÈME 1.6. *La transformation de Fourier discrète est multiple d'une isométrie bijective de $\ell^2(\mathbb{Z})$ sur $L^2_p([-1/2,1/2])$: la formule de Parseval*

$$\int_{-1/2}^{1/2} |\hat{s}(\nu)|^2\,d\nu = \sum_{-\infty}^{\infty} |s_n|^2 \tag{1.36}$$

est vérifiée. La transformation inverse est donnée par

$$s_n = \int_{-1/2}^{1/2} \hat{s}(\nu)e^{2i\pi\nu n}\,d\nu \ . \tag{1.37}$$

REMARQUE 1.8. On verra plus loin, au moment de décrire la théorie de l'échantillonnage, l'utilité de cette transformation. Il est souvent nécessaire d'utiliser une variante, définie par

$$\hat{s}(\nu) = \sum_{n=-\infty}^{\infty} s_n e^{-2i\pi\nu n/\eta} \ ,$$

où η est un réel strictement positif fixé (appelé *fréquence d'échantillonnage*). $\hat{s}$ est alors η-périodique, et on a également

$$s_n = \frac{1}{\eta} \int_{-\eta/2}^{\eta/2} \hat{s}(\nu) e^{2i\pi\nu n/\eta} \, d\nu \ .$$

c. Filtrage numérique. Les opérations de filtrage sont les opérations de base du traitement du signal. On se limite pour le moment aux filtres continus de $\ell^2(\mathbb{Z})$ sur $\ell^2(\mathbb{Z})$.

DÉFINITION 1.6. *Un filtre numérique est un opérateur T, continu sur $\ell^2(\mathbb{Z})$, qui commute avec les translations : pour tout $s \in \ell^2(\mathbb{Z})$ et $k \in \mathbb{Z}$, on a*

$$T\tau_k s = \tau_k T s$$

où τ_k est l'opérateur de translation par k : $(\tau_k s)_n = s_{n-k}$.

On a dans le cas numérique la même caractérisation que dans le cas analogique :

PROPOSITION 1.7. *Soit $T : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ un filtre numérique. Alors il existe $m \in L^\infty([-1/2, 1/2])$, appelée* fonction de transfert *de m, telle que pout tout $s \in \ell^2(\mathbb{Z})$ on ait*

$$\widehat{Ts}(\nu) = m(\nu)\, \hat{s}(\nu) \ , \quad \nu \in [-1/2, 1/2] \ . \tag{1.38}$$

Des exemples simples sont fournis par les filtres de convolution : si $h \in \ell^1(\mathbb{Z})$, on définit le filtre K_h par

$$(K_h s)_n = \sum_{k=-\infty}^{\infty} h_m s_{n-m} \ . \tag{1.39}$$

La suite $h = \{h_n, n \in \mathbb{Z}\}$ est appelée *réponse impulsionnelle* du filtre, et la fonction de transfert n'est autre que sa TFD $\hat{h}$.

$$\widehat{K_h s}(\nu) = \sum_{n,m=-\infty}^{\infty} h_m s_{n-m} e^{-2i\pi\nu n} = \sum_{m=-\infty}^{\infty} h_m e^{-2i\pi\nu m} \sum_{k=-\infty}^{\infty} s_k e^{-2i\pi\nu k} = \hat{h}(\nu)\hat{s}(\nu) \ . \tag{1.40}$$

Le filtre est dit causal (ou réalisable) si $h_n = 0$ pour tout $n < 0$.

A l'image des filtres analogiques, les filtres numériques sont essentiellement utilisés pour modifier le contenu fréquentiel des signaux (on en verra des applications par la suite). L'exemple le plus simple est celui du *filtre passe-bas idéal*, qui force à zéro toutes les fréquences supérieures (en valeur absolue) à une certaine *fréquence de coupure* $\nu_0 < 1/2$. Un tel filtre est défini par sa fonction de transfert

$$\hat{h}(\nu) = \chi_{[-\nu_0, \nu_0]}(\nu) \ .$$

Après TFD inverse, il vient

$$h_n = \frac{\sin(2\pi\nu_0 n)}{n} \ .$$

Il est à noter qu'un tel filtre n'est pas causal, et que la réponse impulsionnelle est lentement décroissante (on n'a pas $h \in \ell^1(\mathbb{Z})$). Il est donc difficile à utiliser en pratique (si on doit éviter de passer par une TFD), et on en cherche souvent une approximation.

EXEMPLE 1.2. Les exemples les plus simples de filtres sont les filtres à réponse impulsionnelle finie (filtres *FIR*), c'est à dire tels que la suite h soit de support fini. La fonction de transfert est alors un polynôme trigonométrique.

EXEMPLE 1.3. Les filtres FIR ne sont souvent pas suffisants, et il est nécessaire de recourir à des filtres à réponse impulsionnelle infinie (filtres *IIR*). Les modèles les plus simples sont les *filtres récursifs*, dans lesquels les signaux d'entrée s et de sortie s' du filtre sont reliés par une relation du type

$$\sum_{m=0}^{N} \alpha_m s'_{n-m} = \sum_{m=0}^{M} \beta_m s_{n-m} \ . \tag{1.41}$$

L'équation précédente admet une solution si et seulement si la matrice A définie par $A_{mn} = \alpha_{m-n}$ est inversible. Dans ce cas, la valeur s'_n est calculée à partir des valeurs précédentes du signal d'entrée s, ainsi que de ses propres valeurs précédentes : en supposant $\alpha_0 \neq 0$,

$$s'_n = \frac{1}{\alpha_0} \left(\sum_{m=0}^{M} \beta_m s_{n-m} - \sum_{m=1}^{N} \alpha_m s'_{n-m} \right) \ .$$

Il est facile de voir qu'après transformation de Fourier discrète, on aboutit à une relation du type

$$\left(\sum_{m=0}^{N} \alpha_m e^{-2i\pi\nu m} \right) \hat{s}'(\nu) = \left(\sum_{m=0}^{M} \beta_m e^{-2i\pi\nu m} \right) \hat{s}(\nu) \ , \tag{1.42}$$

de sorte que la fonction de transfert m du filtre correspondant prend la forme d'une fraction rationnelle de deux polynômes trigonométriques

$$m(\nu) = \frac{\sum_{m=0}^{M} \beta_m e^{-2i\pi\nu m}}{\sum_{m=0}^{N} \alpha_m e^{-2i\pi\nu m}} \ . \tag{1.43}$$

Il est évident que le propriétés du filtre (à commencer par son existence en tant qu'opérateur borné sur $\ell^2(\mathbb{Z})$) dépendent fortement des zéros et des pôles de la fonction de transfert m. Il est pour cela utile d'utiliser un outil voisin de la TFD, à savoir la transformation en z.

d. Transformation en z. Etant donné un signal numérique $\{s_n, n \in \mathbb{Z}\}$, il existe des cas où sa transformée de Fourier discrète n'est pas définie au sens classique. On a parfois recours à une alternative, la transformée en z, dont on décrit ci-dessous les propriétés essentielles, sans entrer dans les détails.

DÉFINITION 1.7. *Soit $s = \{s_n, n \in \mathbb{Z}\}$ un signal numérique. Sa transformée en z est la série de Laurent*

$$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \ , \tag{1.44}$$

définie dans la couronne de convergence (éventuellement vide) $r_1 < |z| < r_2$.

On sait d'après des résultats généraux sur les séries de Laurent que S est holomorphe dans sa couronne de convergence. Inversement, étant donnée une fonction S holomorphe dans une couronne $r_1 < |z| < r_2$, elle admet un unique développement en série de Laurent. De plus, on a le lemme classique suivant :

LEMME 1.2. *Le rayon de convergence ρ de la série entière $z \to \sum_0^\infty a_n z^n$ est donné par*

$$\frac{1}{\rho} = \limsup_{n\to\infty} |a_n|^{1/n} \ .$$

On en déduit immédiatement la couronne de convergence de la transformée en z d'un signal numérique :

COROLLAIRE 1.1. *Soit S la transformée en z de la série s. Les bornes de la couronne de convergence de S sont données par*

$$r_1 = \limsup_{n\to\infty} |s_n|^{1/n} \ , \quad \frac{1}{r_2} = \limsup_{n\to\infty} |s_{-n}|^{1/n} \ . \tag{1.45}$$

EXEMPLE 1.4. On dit qu'un signal numérique s est *causal* si $s_n = 0$ pour tout $n < 0$. Inversement, s est dit *anticausal* si $s_n = 0$ pour tout $n \geq 0$. Supposons que s soit causal. Alors il est évident que $r_2^{-1} = 0$, de sorte que la transformée en z de s est bien définie dans le domaine $|z| > r_1$, c'est à dire à l'extérieur d'un cercle de rayon r_1. De même, si s est anticausal, $r_1 = 0$, et $S(z)$ est bien défini à l'intérieur du cercle de rayon r_2.

Inversion de la transformation en z Il existe plusieurs techniques permettant d'inverser une transformation en z. La plus simple consiste à expliciter un développement en série de Laurent de la fonction S considérée. Le développement en série de Laurent étant unique, ceci fournit directement une transformée inverse.

EXEMPLE 1.5. Prenons l'exemple de la fonction

$$S(z) = \frac{z}{z - z_0} \ , \quad |z| < |z_0| \ ;$$

on peut alors écrire, pour $|z| < |z_0|$,

$$S(z) = \frac{z}{z - z_0} = -\frac{z}{z_0} \frac{1}{1 - z/z_0} = \frac{z}{z_0} \sum_{n=0}^{\infty} \left(\frac{z}{z_0}\right)^n = \sum_{n=-\infty}^{-1} z_0^n \, z^{-n} \ ,$$

ce qui, conjugué à l'unicité du développement en série de Laurent, fournit

$$s_n = \begin{cases} z_0^n & \text{pour } n < 0 \\ 0 & \text{sinon .} \end{cases}$$

Une alternative consiste à utiliser la TFD. Soit S la transformée en z d'un signal s, et soit r un nombre tel que $r_1 < r < r_2$. Calculons

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} S\left(re^{i\theta}\right) e^{in\theta} \, d\theta = \sum_m s_m r^{-m} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(n-m)\theta} \, d\theta = r^{-n} s_n \ .$$

On peut donc écrire

$$s_n = \frac{r^n}{2\pi} \int_{-\pi}^{\pi} S\left(re^{i\theta}\right) e^{in\theta} \, d\theta \ .$$

Par un changement de variables complexes $z = re^{i\theta}$, on obtient donc

PROPOSITION 1.8. *Soit s un signal numérique, et soit S sa transformée en z, définie dans la couronne de convergence $r_1 < |z| < r_2$. Les coefficients de s sont donnés par*

$$s_n = \frac{1}{2i\pi} \oint_C S(z) z^n \frac{dz}{z} \ , \tag{1.46}$$

où C est un cercle centré sur l'origine du plan complexe, de rayon $r \in]r_1, r_2[$.

On a généralement recours à la méthode des résidus pour calculer de telles intégrales. Transformation en z et filtrage numérique L'un des intérêts de la transformation en z est son comportement vis à vis des transformations simples, et en particulier des translations. Etant donnée une suite $\{s_n, n \in \mathbb{Z}\}$, et une suite filtres $\{s'_n, n \in \mathbb{Z}\}$ donnée par $s'_n = s_{n-k}$, on voit immédiatement que leurs transformées en z sont reliées par $S'(z) = z^k S(z)$. Le corollaire immédiat est le comportement de la transformation en z vis à vis du filtrage numérique. numériques. Etant donné un signal numérique s et une filtre numérique de réponse impulsionnelle h, alors pour tout z à l'intérieur de l'intersection des couronnes de convergence des transformées en z S et H de s et h respectivement, on a

$$S'(z) = \sum_n s'_n z^{-n} = \sum_n \sum_k h_k z^{-k} s_{n-k} z^{-(n-k)} \ ,$$

de sorte que l'on a

$$S'(z) = H(z) S(z) \ . \tag{1.47}$$

La fonction H est elle aussi appelée *fonction de transfert* du filtre.

En particulier, dans le cas d'un filtre récursif comme précédemment, on a

$$S'(z) = \frac{\sum_{m=0}^{M} \beta_m z^{-m}}{\sum_{m=0}^{N} \alpha_m z^{-m}} S(z) \ ,$$

c'est à dire que la transformée en z de h prend la forme d'une fraction rationnelle. Cette expression est à rapprocher de l'expression (1.43) obtenue avec la TFD.

e. Factorisation des filtres causaux d'ordre fini réels. On dit qu'un filtre causal K_h est d'ordre fini si K_h peut être réalisé comme un filtre récursif comme en (1.41). On impose que le filtre K_h soit causal (donc la couronne de convergence de $H(z)$ est de la forme $|z| > r_1$) et stable (donc le cercle unité est inclus dans la couronne de convergence. Par conséquent, les pôles de H se trouvent à l'intérieur du cercle unité.

On se limite ici aux filtres réels, c'est à dire tels que les coefficients α_k et β_k sont réels. Dans ce cas, il est facile de voir que

$$\overline{\hat{h}}(\nu) = \hat{h}(-\nu) \ ,$$

de sorte que le spectre prend la forme

$$|\hat{h}(\nu)|^2 = H(z) H(z^{-1})|_{z=e^{2i\pi\nu}} \ .$$

Notons z_k les pôles de H (les zéros du dénominateur de H) et ζ_ℓ les zéros de H. Il est immédiat que $\hat{h}(\nu)$ est caractérisé par des facteurs de la forme

$$(z - z_k)(z^{-1} - z_k) = 1 + z_k^2 - (z + z^{-1}) \ , \quad \text{et} \quad (z - \zeta_k)(z^{-1} - \zeta_k) = 1 + \zeta_k^2 - (z + z^{-1}) \ ,$$

(avec $z = e^{2i\pi\nu}$), et est donc une fonction (positive rationnelle) de

$$w = \frac{1}{2}(z + z^{-1}) = \cos(2\pi\nu) \ .$$

Inversement, soit $\nu \to W(\cos(2\pi\nu)) = N(\cos(2\pi\nu))/D(\cos(2\pi\nu))$ une fonction rationnelle positive. Notons $w = \cos(2\pi\nu)$, et w_k les zéros (dans le plan complexe) de D (le numérateur N se traite de façon identique). On voit facilement que l'équation en z

$$w_k = \frac{1}{2}\,(z + z^{-1})$$

possède deux solutions inverses l'une de l'autre, notées z_k et ${z_k}^{-1}$. Par convention, on choisit $|z_k| < 1$. On peut alors poser

$$d(z) = \prod_k (z - z_k)\ .$$

De même, en notant v_k les zéros de N, et ζ_k et ${\zeta_k}^{-1}$ les solutions et ζ de

$$v_k = \frac{1}{2}\,(\zeta + \zeta^{-1})$$

(sans nécessairement imposer $|\zeta_k| < 1$), on est naturellement conduit à introduire

$$n(z) = \prod_k (z - \zeta_k)\ .$$

Il résulte de cette analyse que la fonction

$$z \to \frac{n(z)}{d(z)}$$

est la fonction de transfert d'un filtre causal stable d'ordre fini.

PROPOSITION 1.9. *Soit K_h un filtre causal stable d'ordre fini. Alors son spectre $A^2 = |\hat{h}|^2$ est une fonction rationnelle non-négative de* $\cos(2\pi\nu)$. *Inversement, étant donnée une fonction rationnelle non-négative de* $\cos(2\pi\nu)$, *il existe un filtre causal stable d'ordre fini K_h dont le spectre coïncide avec A^2.*

REMARQUE 1.9. Le filtre K_h n'est pas unique, car il reste la liberté de choisir les zéros ζ_k à l'intérieur ou à l'extérieur du disque unité pour former la fonction d. Choisir tous les zéros à l'intérieur du disque unité conduit aux filtres dits *à phase minimale*.

1.2.2. *En pratique : transformation de Fourier finie (TFF).* Les suites de longueur finie se prêtent à la même analyse. On peut également leur associer une transformée de Fourier (qui est elle aussi une suite de longueur finie), et la transformation correspondante est de nouveau une isométrie (à une constante près). Plus précisément, à la suite finie $u = \{u_n, n = 0 \dots N-1\}$ on associe la suite $\hat{u} = \{\hat{u}_k, k = 0, \dots N-1\}$, définie par

$$\hat{u}_k = \sum_{n=0}^{N-1} u_n e^{-2i\pi kn/N}\ . \tag{1.48}$$

C'est alors un jeu d'enfant que de montrer des propriétés analogues aux propriétés que nous avons déjà vues : formule de Plancherel et inversion. De fait, on a

$$\sum_{k=0}^{N-1} |\hat{u}_k|^2 = N \sum_{n=0}^{N-1} |u_n|^2\ , \tag{1.49}$$

et

$$u_n = \frac{1}{N} \sum_{k=0}^{N-1} \hat{u}_k e^{2i\pi \frac{kn}{N}} \ . \tag{1.50}$$

REMARQUE 1.10. Dans le langage que nous avons utilisé plus haut, ceci est équivalent à dire que la famille des vecteurs $e_k \in \mathbb{C}^N$ définis par leurs composantes

$$e_k = \left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}} e^{2i\pi k/N}, \dots \frac{1}{\sqrt{N}} e^{2i\pi k(N-1)/N} \right) \tag{1.51}$$

est une base orthonormée de $\mathbb{C}^N$, et un a posé $\hat{u}_k = \langle u, e_k \rangle \sqrt{N}$.

De nouveau, on verra plus loin l'utilité de cette autre version, ainsi qu'un algorithme rapide de calcul.

1.3. Echantillonnage.

1.3.1. *Le cas des signaux à support fini.* Le traitement du signal numérique s'adresse aux signaux échantillonnés, c'est à dire à des suites (finies) de nombres. L'échantillonnage est une opération qui consiste à générer de telles suites finies à partir de signaux analogiques, c'est à dire de fonctions d'une variable continue. On s'intéresse ici au cas des fonctions à support borné f, dont on considère des échantillons, c'est à dire des valeurs régulièrement espacées $f_n = f(n/\eta)$, où η est un nombre fixé, appelé *fréquence d'échantillonnage*.

Supposons donc que $f \in L^2([0,T])$, et notons $c_n(f)$ ses coefficients de Fourier. Supposons que $c_n(f) = 0$ pour tout $n \in \mathbb{Z}$ tel que $|n| > N$, pour un $N \in \mathbb{Z}^+$ fixé. On peut alors montrer que la fonction f est continue, et écrire pour tout $t \in [0,T]$

$$f(t) = \sum_{n=-N}^{N} c_n(f) \exp\left\{ 2i\pi \frac{nt}{T} \right\} \ .$$

Soit maintenant $M \geq N$ un entier, et considérons $2M+1$ valeurs t_k de t, régulièrement espacées

$$t_k = \frac{kT}{2M+1} \ , \quad k = 0, \dots 2M \ .$$

On peut alors écrire (en gardant en mémoire le fait que $c_n(f) = 0$ si $|n| > N$)

$$f(t_k) = \sum_{n=-M}^{M} c_n(f) \exp\left\{ 2i\pi \frac{nk}{2M+1} \right\} \ .$$

Un calcul simple (somme d'une série géométrique) montre que

$$\sum_{k=0}^{2M} \exp\left\{ 2i\pi \frac{nk}{2M+1} \right\} = (2M+1)\delta_{n,0} \ ,$$

où $\delta_{n,0}$ est le symbole de Kronecker qui vaut 1 si $n = 0$ et 0 sinon. On obtient ainsi

$$c_n(f) = \frac{1}{2M+1} \sum_{k=0}^{2M} f(t_k) \exp\left\{ -2i\pi \frac{nk}{2M+1} \right\} \ ,$$

pour tout $n = -M, 1-M, \dots M$. Par conséquent, on obtient, pour tout $t \in [0,T]$

$$f(t) = \frac{1}{2M+1} \sum_{k=0}^{2M} f(t_k) \sum_{n=-M}^{M} \exp\left\{ -2i\pi \frac{nk}{2M+1} \right\} \exp\left\{ 2i\pi \frac{nt}{T} \right\} \ .$$

Pour effectuer la somme sur n, il faut utiliser l'identité

$$\sum_{n=-N}^{N} e^{in\alpha} = \frac{\sin\left(\left(N+\frac{1}{2}\right)\alpha\right)}{\sin\left(\frac{\alpha}{2}\right)} ,$$

et on obtient finalement le résultat suivant

$$f(t) = \frac{1}{2N+1}\sum_{k=0}^{2N} f\left(\frac{kT}{2N+1}\right) \frac{\sin\left(\frac{2\pi}{T}\left(N+\frac{1}{2}\right)\left(t-\frac{kT}{2N+1}\right)\right)}{\sin\left(\frac{\pi}{T}\left(t-\frac{kT}{2N+1}\right)\right)} ,$$

qui montre que f est complètement caractérisée par ses valeurs sur les points régulièrement espacés $kT/(2N+1)$. L'expression ci-dessus est une formule d'interpolation pour f à partir des échantillons.

Plus généralement, soit M un entier supérieur ou égal à N. En posant $c'_n = c_n(f)$ si $|n| \leq N$, et $c'_n = 0$ si $|n| = N+1, \dots M$, on peut écrire

$$f(t) = \sum_{n=-M}^{M} c'_n \exp\left\{2i\pi\frac{nt}{T}\right\} ,$$

et reproduire le calcul ci-dessus en utilisant cette fois les $2M+1$ échantillons $f(kT/(2M+1)$, $k = 0, \dots 2M$. On aboutit ainsi au résultat suivant :

PROPOSITION 1.10. *Soit $f \in L^2([0,T])$, telle que $c_n(f) = 0$ si $|n| > N$, Alors pour tout entier $M \geq N$, la fonction f est complètement caractérisée par les $2M+1$ échantillons régulièrement espacés $f(kT/(2M+1))$, et on a la formule d'interpolation*

$$f(t) = \frac{1}{2M+1}\sum_{k=0}^{2M} f\left(\frac{kT}{2M+1}\right) \frac{\sin\left(\frac{2\pi}{T}\left(M+\frac{1}{2}\right)\left(t-\frac{kT}{2M+1}\right)\right)}{\sin\left(\frac{\pi}{T}\left(t-\frac{kT}{2M+1}\right)\right)} . \tag{1.52}$$

On aboutit donc à la règle empirique suivante : la règle est donc d'avoir au moins autant d'échantillons que ce que l'on a de coefficients de Fourier non nuls. Inversement, si le nombre d'échantillons considérés est inférieur au nombre de coefficients de Fourier, le problème de caractérisation de f par les échantillons devient mal posé, et il n'existe plus de formule d'interpolation.

1.3.2. *Le cas des signaux à support infini.* Le théorème d'échantillonnage se perd dans la nuit des temps. Il est généralement attribué à Shannon et Kotelnikov, qui en ont proposé une preuve vers 1945, peu après Nyquist. En fait, il avait été démontré bien avant par Whittaker (1936), et probablement par Cauchy encore plus avant.

Le cadre naturel du théorème d'échantillonnage est l'espace des signaux à bande limitée, ou espace de Paley-Wiener

$$PW_{\nu_0} = \left\{ f \in L^2(\mathbb{R}), \hat{f}(\nu) = 0 \text{ pour tout } \nu \notin [-\nu_0, \nu_0] \right\} \tag{1.53}$$

Il est facile de voir que PW_{ν_0} est un espace de fonctions continues, de sorte que les valeurs ponctuelles des fonctions de PW_{ν_0} ont un sens. On peut alors introduire l'opérateur d'échantillonnage E, associé à la fréquence d'échantillonnage η : si $f \in PW_{\nu_0}$,

$$(Ef)_n = f\left(\frac{n}{\eta}\right) , \qquad n \in \mathbb{Z} . \tag{1.54}$$

THÉORÈME 1.7. *Soit $f \in PW_{\nu_0}$, et soit $\eta > 0$ la fréquence d'échantillonnage. On considère la suite des échantillons définie en (1.54).*

(1) *Si $\eta < 2\nu_0$, la suite des échantillons $(Ef)_n$ ne permet pas de déterminer la fonction f sans hypothèse supplémentaire.*

(2) *Si $\eta > 2\nu_0$, alors il existe une infinité de formules de reconstruction de f à partir des échantillons. Soit φ telle que $\hat{\varphi} \in C^\infty$, $\hat{\varphi}(\nu) = 0$ pour tout $\nu \notin [-\eta/2, \eta/2]$ et $\hat{\varphi}(\nu) = 1$ pour tout $\nu \in [-\nu_0, \nu_0]$. Alors on a*

$$f(t) = \sum_{n=-\infty}^{\infty} f(n/\eta)\varphi(t - n/\eta) \ . \tag{1.55}$$

(3) *Si $\eta = 2\nu_0$, alors il n'existe plus qu'une seule formule de reconstruction de f à partir des échantillons :*

$$f(t) = \sum_{n=-\infty}^{\infty} f(n/\eta)\frac{\sin(\pi\eta(t - n/\eta))}{\pi\eta(t - n/\eta)} \ . \tag{1.56}$$

Preuve : Commençons par considérer la fonction périodique

$$\Gamma(\nu) = \sum_{k=-\infty}^{\infty} \hat{f}(\nu + k\eta) \ . \tag{1.57}$$

Il est immédiat que $\Gamma \in L^1_p([-\eta/2, \eta/2])$, et on peut donc s'intéresser à ses coefficients de Fourier. Un calcul simple montre que

$$c_n(\Gamma) = \frac{1}{\eta}\int_{-\eta/2}^{\eta/2} \Gamma(\nu)e^{-2i\pi\frac{\nu n}{\eta}}d\nu = \frac{1}{\eta}\int_{-\infty}^{\infty} \hat{f}(\nu)e^{-2i\pi\frac{\nu n}{\eta}}d\nu = \frac{1}{\eta}f\left(\frac{-n}{\eta}\right) \ .$$

En fait, on dit que la fonction Γ est la TFD (transformée de Fourier discrète, étudiée un peu plus loin) de la série d'échantillons $\{f_n\}$, et le problème de retrouver f à partir des échantillons est équivalent au problème de retrouver $\hat{f}$ à partir de Γ. Or, Γ n'est autre (à une constante multiplicative près) qu'une version "périodisée" de $\hat{f}$, de période η. On peut donc considérer les trois cas de figure.

- Supposons que $\eta > 2\nu_0$. Alors, il est clair que l'on peut toujours trouver une fonction φ, dont la transformée de Fourier $\hat{\varphi}$ est C^∞, à support compact dans l'intervalle $[-\nu_0, \nu_0]$, et vaut uniformément 1 dans $[-\eta/2, \eta/2]$. On a donc $\hat{f}(\nu) = \Gamma(\nu)\hat{\varphi}(\nu)$, ce qui se traduit, après transformation de Fourier inverse, par la relation (1.55).
- Dans le cas critique, le raisonnement est similaire, à ceci près que la fonction $\hat{\varphi}$ ne peut plus être choisie continue, et est nécessairement de la forme $\varphi(\nu) = \chi_{[-\nu_0,\nu_0]}(\nu)$. La transformée de Fourier inverse de cette dernière étant le sinus cardinal, on obtient (1.56).
- Si $\eta < 2\nu_0$, le "truc" précédent ne fonctionne plus : la périodisation "mélange des morceaux de $\hat{f}$ congrus modulo η, de sorte que l'on ne peut plus inverser le processus. C'est le phénomène de *repliement de spectre*.

Ceci conclut la démonstration. ♠

En fait, l'étude du cas critique $\nu_0 = \eta/2$ nous montre :

COROLLAIRE 1.2. *La famille des sinus cardinaux normalisés ϕ_n définis par*

$$\phi_n(t) = \sqrt{\eta}\, sinc(\pi\eta(t - n/\eta)) = \frac{sin(\pi\eta(t - n/\eta))}{\pi\sqrt{\eta}(t - n/\eta)} \ , \quad n \in \mathbb{Z} \tag{1.58}$$

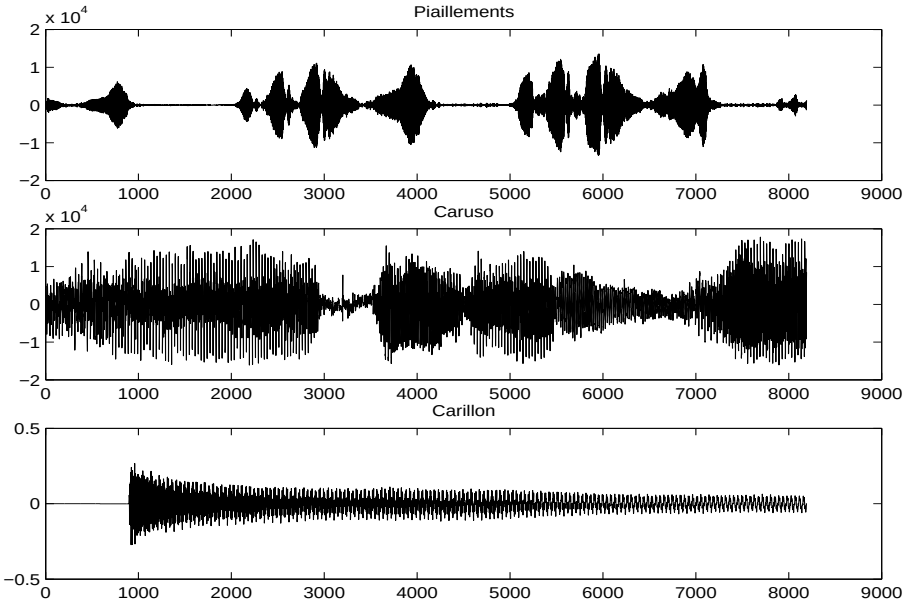

FIG. 5. Trois exemples de signaux audiophoniques : piaillements d'oiseaux, un vieil enregistrement de Caruso, et un son de carillon.

est une base orthonormée de l'espace $PW_{\eta/2}$.

REMARQUE 1.11. Dans le cas favorable $\eta > 2\nu_0$, la famille de fonctions $t \to \varphi(t - n/\eta), n \in \mathbb{Z}$ considérée n'est plus une base orthonormée, car elle est redondante. Elle forme alors ce que l'on appelle un repère, comme on le verra au chapitre suivant.

REMARQUE 1.12. En pratique, l'échantillonnage est souvent (toujours) précédé d'un filtrage passe-bas, dont le but est de réduire la largeur de bande pour l'adapter à la fréquence d'échantillonnage prévue. Les filtres passe-bas idéaux n'étant pas réalisables, on se rabat plutôt sur des filtres rationnels, comme par exemple un des filtres de Chebyshev ou de Butterworth que nous avons vus plus haut.

2. Signaux aléatoires

On a souvent recours à des modèles de signaux faisant intervenir des quantités aléatoires. On peut trouver à cela deux justifications essentielles :

- La nécessité de modéliser des classes relativement larges de signaux, regroupés par certaines propriétés génériques : par exemple, des signaux audiophoniques (voir trois exemples en FIG. 5), le signal de parole, des images...
- Le besoin de modéliser divers types de "bruits" (bruits de mesure par exemple), généralement difficilement contrôlables.

Le cadre mathématique adapté à cette situation est celui des processus aléatoires. L'objectif de cette section est d'aboutir à la représentation spectrale des processus stationnaires (puis par la suite à la représentation de Karhunen-Loève dans un cadre plus général, comme on le verra dans le chapitre suivant), qui fournit une version de l'analyse de Fourier adaptée au cadre des signaux aléatoires.

2.1. Définitions, propriétés simples. Dans cette section on désignera par $(\mathcal{A}, \mathcal{F}, \mathbb{P})$ un espace probabilisé. On note par $\mathcal{L}^0(\mathcal{A}) = \mathcal{L}^0(\mathcal{A}, \mathbb{P})$ l'espace des variables aléatoires sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, à valeurs réelles ou complexes. Etant données deux variables aléatoires $X, Y \in \mathcal{L}^0(\mathcal{A})$, on dit que $X \sim Y$ si $X = Y$ presque surement. Ceci définit une relation d'équivalence, et on note

$$L^0(\mathcal{A}) = \mathcal{L}^0(\mathcal{A})/\sim$$

l'espace quotient, c'est à dire l'espace des variables aléatoires différentes presque surement. Etant donnée une variable aléatoire $X \in L^0$, on en notera $\mathbb{E}\{X\}$ l'espérance

$$\mathbb{E}\{X\} = \int x \, d\mathbb{P}_X(x) \ .$$

2.1.1. *Premières définitions.*

DÉFINITION 1.8. *Soit $T \subset \mathbb{R}$ une partie (continue ou discrète) de $\mathbb{R}$. On appelle processus stochastique indexé par T à valeurs réelles ou complexes une application*

$$t \in T \to X_t \in L^0(\mathcal{A}) \ . \tag{1.59}$$

Etant donné $a \in \mathcal{A}$, l'application $t \to X_t(a)$ est appelée trajectoire du processus.

Un processus aléatoire sera aussi appelé signal aléatoire, ou série chronologique. On introduit de même des signaux aléatoires multidimensionnels (pour lesquels T est une partie de $\mathbb{R}^n$), mais on se limitera ici au cas unidimensionnel.

DÉFINITION 1.9. *Etant donnés un processus aléatoire $\{X_t, t \in T\}$, et n valeurs $(t_1, t_2, \dots t_n) \in T^n$. $(X_{t_1}, \dots X_{t_n})$ est une variable aléatoire vectorielle. L'ensemble des distributions de toutes ces variables forme le système de lois marginales du processus.*

Un théorème célèbre de Kolmogorov (le théorème d'extension de Kolmogorov, voir par exemple [**2**]) montre que la connaissance du système de lois marginales est suffisante pour caractériser la distribution du processus.

2.1.2. *Exemples.*

EXEMPLE 1.6. L'exemple le plus simple est celui d'un bruit blanc discret. On considère pour cela une famille $\{W_0, \dots W_{N-1}\}$ de variables aléatoires indépendantes, identiquement distribuées, par exemple $\mathcal{N}(0, \sigma^2)$. Il s'agit d'un processus aléatoire indexé par $\{0, 1, \dots N-1\}$, que l'on appelle bruit blanc discret Gaussien.

EXEMPLE 1.7. Partant de l'exemple précédent, et d'une suite finie $\{h_0, \dots h_{N-1}\}$, on peut former la suite $\{X_0, \dots X_{N-1}\}$ définie par le produit de convolution circulaire

$$X_k = \sum_{n=0}^{N-1} h_n W_{(k-n)[\mathrm{mod} N]}$$

On a alors par exemple $\mathbb{E}\{X_n\} = 0$ pour tout n, et aussi

$$\mathbb{E}\{X_k X_\ell\} = \sum_n h_n h_{((k-\ell)+n)[\mathrm{mod} N]}$$

EXEMPLE 1.8. On s'intéressera souvent à des processus à temps continu, c'est à dire au cas où T n'est pas dénombrable. Prenons par exemple $T = \mathbb{R}^+$, et introduisons les temps $t_0 = 0 < t_1 < t_2 < \dots$. Soient $Z_0, Z_1, \dots$ une suite de variables

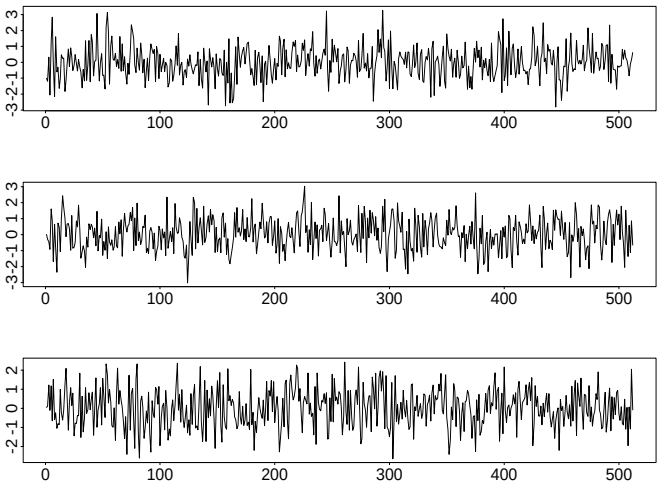

FIG. 6. 3 trajectoires de bruit blanc.

aléatoires sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$; on peut alors introduire le processus "à sauts" X défini par

$$X_t = \sum_{n=0}^{\infty} Z_n \chi_{[t_n, t_{n+1}]}(t) \ . \tag{1.60}$$

X est bien un processus aléatoire sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$; ses trajectoires sont des fonctions constantes par morceaux, généralement discontinues (voir la notion de continuité presque sûre des trajectoires plus bas).

EXEMPLE 1.9. Un processus harmonique est un processus défini sur $\mathbb{R}^+$, de la forme

$$X_t = Ae^{-t/\tau} \cos(2\pi\nu t + \varphi) \ , \tag{1.61}$$

où A, ν et τ sont des constantes, et où φ est une variable aléatoire uniformément distribuée sur $[0, 2\pi]$. X est aussi un processus aléatoire sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, indexé par $\mathbb{R}^+$.

En fait, on peut mettre l'accent sur deux classes de processus particulièrement intéressantes, car basées sur des hypothèses simplificatrices relativement réalistes dans de nombreux cas pratiques.

(1) Processus à accroissements indépendants : ce sont les processus tels que pour tous temps $t_1 < t_2 < \cdots < t_M$, la famille de variables aléatoires $\{X_{t_1}, X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots X_{t_M} - X_{t_{M-1}}\}$ soit une famille de variables aléatoires indépendantes. On verra plus loin un exemple avec le processus de Wiener.

(2) Processus Gaussiens : toutes les mesures de probabilités du système de lois marginales sont Gaussiennes.

Notons que ces deux hypothèses ne sont pas exclusives (voir l'exemple du processus de Wiener). L'hypothèse de "Gaussianité" est particulièrement utile, car elle permet de caractériser les distributions de probabilités par leurs moments d'ordre 1 et 2.

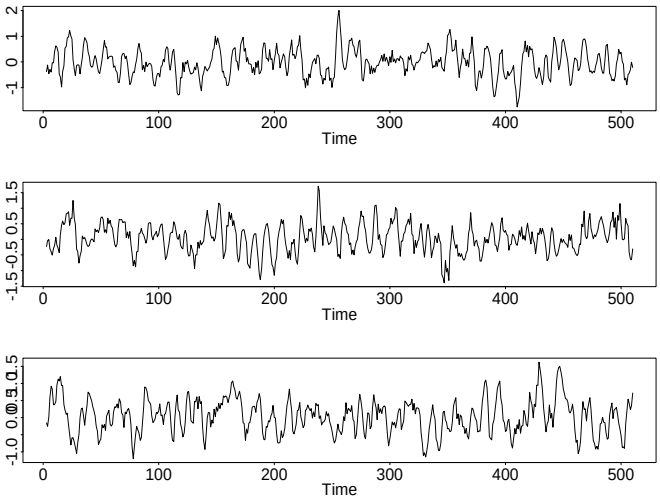

FIG. 7. 3 trajectoires de bruit blanc filtré (filtrage passe-bas).

2.1.3. *Processus du second ordre.* On se limitera dans ce cours au cas des processus du second ordre c'est à dire des processus tels que leur covariance est bien définie.

DÉFINITION 1.10. *Un processus aléatoire* $\{X_t, t \in T\}$ *est dit du second ordre si pour tout* $t \in T$*, on a* $\mathbb{E}\left\{|X_t|^2\right\} < \infty$*. Lorsque* T *est un espace continu, un processus du second ordre est dit continu en moyenne d'ordre 2 si pour tout* $t \in T$*,* $\mathbb{E}\left\{|X_{t+\delta} - X_t|^2\right\} \to 0$ *quand* $\delta \to 0$*.*

Notons que grâce à l'inégalité

$$\mathbb{E}\left\{|X|\right\} \leq 1 + \mathbb{E}\left\{|X|^2\right\} ,$$

on peut introduire la moyenne du processus

$$m_t = \mathbb{E}\left\{X_t\right\} . \tag{1.62}$$

On introduit également la covariance du processus

$$C_X(t,s) = \mathbb{E}\left\{(X_t - m_t)(\overline{X_s} - \overline{m_s})\right\} = R_X(t,s) - m_t\overline{m_s} , \tag{1.63}$$

où

$$R_X(t,s) = \mathbb{E}\left\{X_t\overline{X_s}\right\} \tag{1.64}$$

est l'autocorrélation. Ces deux fonctions vérifient la propriété suivante

PROPOSITION 1.11. *Les fonctions* C_X *et* R_X *sont définies positives.*

On rappelle qu'une fonction de deux variables F est définie positive si pour tous $t_1, \dots t_n \in \mathbb{R}$ et $\alpha_1, \dots \alpha_n \in \mathbb{C}$, on a

$$\sum_{k,\ell=1}^{n} \alpha_k\overline{\alpha}_\ell F(t_k, t_\ell) \geq 0 . \tag{1.65}$$

Preuve : Il suffit de le prouver pour R_X (la preuve pour C_X est identique). On a

$$\sum_{k,\ell=1}^{n} \alpha_k \overline{\alpha_\ell} R_X(t_k, t_\ell) = \sum_{k,\ell=1}^{n} \alpha_k \overline{\alpha_\ell} \mathbb{E}\left\{X_{t_k} \overline{X_{t_\ell}}\right\} = \mathbb{E}\left\{ |\sum_{k=1}^{n} \alpha_k X_{t_k}|^2 \right\} \geq 0 \ .$$

♠

Les variables aléatoires X sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$ telles que $\mathbb{E}\{X\} = 0$ et $\mathbb{E}\left\{|X|^2\right\} < \infty$ engendrent un espace linéaire, noté $\mathcal{L}^2(\mathcal{A}, \mathbb{P})$., ou plus généralement $\mathcal{L}^2(\mathcal{A})$. Soit $L^2(\mathcal{A})$ l'espace quotient de $\mathcal{L}^2(\mathcal{A})$ dans lequel on a identifié les variables aléatoires égales presque sûrement. $L^2(\mathcal{A})$ est naturellement muni d'un produit scalaire défini par

$$(X|Y) = \mathbb{E}\left\{X\overline{Y}\right\} \ , \tag{1.66}$$

qui en fait un espace de Hilbert. Etant donné un processus du second ordre $\{X_t, t \in T\}$, on notera $\mathcal{M}_X$ le sous espace fermé de $L^2(\mathcal{A})$ engendré par les variables aléatoires $X_t, t \in T$.

2.2. Signaux aléatoires numériques. On se limitera ici au cas des processus du second ordre, indexés par $\mathbb{Z}$. Soit donc $X = \{X_n, n \in \mathbb{Z}\}$ un processus du second ordre sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, de moyenne $m_X(n) = \mathbb{E}\{X_n\}$ et de fonction de corrélation R_X.

Définition 1.11. *X est dit stationnaire en moyenne d'ordre deux si ses statistiques d'ordre un et deux sont invariantes par translation, c'est á dire si*

$$m_X(n) = m_X(0) := m_X \ , \quad \forall n \in \mathbb{Z} \tag{1.67}$$

$$R_X(n+\tau, m+\tau) = R_X(n,m) := R_X(n-m) \ , \quad \forall n, m, \tau \in \mathbb{Z} \tag{1.68}$$

Il est facile de vérifier que si X est stationnaire en moyenne d'ordre deux, on a aussi

$$C_X(n+\tau, m+\tau) = C_X(n,m) := C_X(n-m) \ , \quad \forall n, m, \tau \in \mathbb{Z}$$

2.2.1. *Filtrage.* Les signaux aléatoires du second ordre restent du second ordre par filtrage numérique. En effet, soit $h \in \ell^1(\mathbb{Z})$, et soit X un processus du second ordre, stationnaire en moyenne d'ordre deux. Soit Y défini par

$$Y_n = (K_h X)_n = \sum_k h_k X_{n-k} \ .$$

Alors Y est du second ordre :

$$\mathbb{E}\left\{|Y_n|^2\right\} = \sum_{k,\ell} h_k \overline{h}_\ell C_X(n-k, n-\ell) \leq C_X(0) \, \|h\|_1^2 \ .$$

De plus, on a

$$m_Y(n) = \sum_k h_k m_X(n-k) = (h * m_X)(n) \ ,$$

$$C_Y(n,m) = \sum_{k,\ell} h_k \overline{h}_\ell C_X(n-k, m-\ell) \ .$$

Si de plus X est stationnaire en m.o.d., alors on a de plus

$$m_Y(n) = m_X \sum_k h_k = m_Y(0) \ ,$$

$$C_Y(n,m) = \sum_{k,\ell} h_k \overline{h}_\ell C_X(n-k-m+\ell) = C_Y(n-m) \ .$$

2.2.2. *Mesure spectrale et densité spectrale pour les processus stationnaires en m.o.d.* Soit donc X un processus stationnaire en m.o.d., que l'on suppose centré pour simplifier. Si tel n'est pas le cas, on peut toujours écrire $X = Y + m_X$ et travailler sur le signal aléatoire centré Y. En corollaire de ce qui précède, la covariance C_X est une suite définie positive : pour tous $n_1, \dots n_N \in \mathbb{R}$ et $\alpha_1, \dots \alpha_N \in \mathbb{C}$, on a

$$\sum_{k,\ell=1}^{n} \alpha_k \overline{\alpha}_\ell F(n_k - n_\ell) \geq 0 \ . \tag{1.69}$$

Un résultat général d'analyse fonctionnelle (voir par exemple [**8**] pour la démonstration) permet d'introduire la mesure spectrale de X :

THÉORÈME 1.8 (Herglotz). *Soit ϕ une suite définie positive. Alors il existe une unique mesure non-négative sur $[-1/2, 1/2]$ telle que pour tout n, on ait*

$$\phi(n) = \int_{-1/2}^{1/2} e^{2i\pi\nu n}\, d\mu(\, nu) \ . \tag{1.70}$$

Preuve : Commençons par calculer la quantité suivante (qui est toujours positive ou nulle), pour $\nu \in [-1/2, 1/2]$

$$\sum_{j,k=1}^{N} \phi(j-k) e^{-2i\pi\nu(j-k)} = \sum_{n=1-N}^{N-1} \phi(n)\, e^{-2i\pi\nu n}\, (N - |n|) \ ,$$

et posons

$$\gamma_N(\nu) = \sum_{n=1-N}^{N-1} \phi(n)\, e^{-2i\pi\nu n} \left(1 - \frac{|n|}{N}\right) \ .$$

Il est clair que $\gamma_N(\nu) \geq 0$ pour tout ν, et que

$$\int_{-1/2}^{1/2} \gamma_N(\nu)\, d\nu = \phi(0) \ .$$

Soient $d\mu_N$ les mesures définies par

$$d\mu_N(\nu) = \gamma_N(\nu)\, d\nu \ .$$

Il s'agit de mesures bornées, définies sur un domaine compact. Par conséquent, il est possible d'extraire une sous-suite $d\mu_{N'}$ qui converge faiblement vers une limite $d\mu$ (c'est le théorème de Helly). De plus, pour tout m tel que $|m| \leq N'$, on a

$$\int_{-1/2}^{1/2} e^{2i\pi\nu m} d\mu_{N'}(\nu) = \left(1 - \frac{|m|}{N}\right) \phi(m) \longrightarrow \phi(m) \text{ pour } N' \to \infty \ .$$

Par définition de la convergence faible, on en déduit que

$$\int_{1/2}^{1/2} e^{2i\pi\nu m} d\mu_{N'}(\nu) \longrightarrow \int_{1/2}^{1/2} e^{2i\pi\nu m} d\mu(\nu) \text{ pour } N' \to \infty \ ,$$

ce qui prouve l'existence de μ.

Pour ce qui est de l'unicité : soient μ et μ' deux limites ; soit $g \in C([-1/2, 1/2])$; on sait que toute fonction continue est arbitrairement bien approximée par les polynômes trigonométriques ; μ et μ' coïncidant sur les polynômes trigonométriques, on a bien

$$\int_{1/2}^{1/2} g(\nu)\, d\mu(\nu) = \int_{1/2}^{1/2} g(\nu)\, d\mu'(\nu) \ ,$$

pour tout $g \in C([-1/2, 1/2]$, ce qui prouve que $\mu' = \mu$, et donc l'unicité. ♠

COROLLAIRE 1.3. *Soit X un signal numérique aléatoire du second ordre, centré et stationnaire en moyenne d'ordre deux. Il existe une unique mesure μ_X, appelée* mesure spectrale *de X, telle que pour tout n on ait*

$$C_X(n) = \int_{-1/2}^{1/2} e^{2i\pi\nu n}\, d\mu_X(\nu) \ . \tag{1.71}$$

La mesure spectrale μ_X n'est pas nécessairement absolument continue par rapport à la mesure de Lebesgue. Si c'est le cas, on peut alors écrire

$$d\mu_X(\nu) = \mathcal{S}_X(\nu)\, d\nu \ ,$$

où $\mathcal{S}_X \in L^\infty([-1/2, 1/2])$ est appelé *densité spectrale* (ou *spectre*) de X.

2.2.3. *Quelques exemples.*

(1) L'exemple le plus simple est celui du bruit blanc Gaussien : les variables aléatoires X_n sont des variables aléatoires Gaussiennes indépendantes et identiquement distribuées $\mathcal{N}(0, \sigma)$. On a alors $m_X = 0$ et $C_X(m, n) = \sigma^2 \delta_{mn}$, et X est stationnaire en m.o.d. L'unicité de la mesure spectrale montre facilement que

$$d\mu_X(\nu) = \sigma^2\, d\nu \ ,$$

d'où X admet une densité spectrale $\mathcal{S}_X$ constante.

(2) Bruit blanc filtré (signal MA) : si X est le bruit blanc précédent, et si $h \in \ell^1(\mathbb{Z})$, on a déjà vu que $Y = K_h X$ est toujours un processus du second ordre, stationnaire en m.o.d. Un calcul simple montre que Y admet une densité spectrale $\mathcal{S}_Y$ de la forme

$$\mathcal{S}_Y(\nu) = |\hat{h}(\nu)|^2\, \mathcal{S}_X(\nu) \ .$$

Plus généralement, si X est un signal aléatoire du second ordre centré, stationnaire en moyenne d'ordre deux, de densité spectrale $\mathcal{S}_X$, alors le même calcul montre que Y est lui aussi du second ordre, centré et stationnaire en m.o.d., et tel que

$$\mathcal{S}_Y(\nu) = |\hat{h}(\nu)|^2\, \mathcal{S}_X(\nu) \ .$$

Ainsi, comme dans le cas des signaux déterministes, un filtrage de convolution revient à "modeler" le contenu en fréquences d'un signal.

(3) Signal AR : Soit X un bruit blanc Gaussien comme ci-dessus, et soient $\alpha_0, \ldots \alpha_N$ des nombres complexes. Si il existe un processus Y solution de

$$\sum_{k=0}^{N} \alpha_k Y_{n-k} = X_n \ ,$$

alors Y est stationnaire en moyenne d'ordre deux, et admet une densité spectrale de la forme

$$\mathcal{S}_Y(\nu) = \frac{1}{\left|\sum_k \alpha_k e^{-2i\pi k\nu}\right|^2} \ .$$

(4) Signal harmonique : On considère une variable aléatoire uniforme ϕ sur l'intervalle $[-1/2, 1/2]$ (donc, de densité de probabilités $\rho_\phi(\alpha) = \chi_{[-1/2,1/2]}(\alpha)$), et on lui associe le signal aléatoire X défini par

$$X_n = Ae^{2i\pi(n\nu_0+\phi)} \ ,$$

où $A \in \mathbb{C}$ et $\nu_0 \in [-1/2, 1/2]$ sont deux constantes. Il est facile de vérifier que X est du second ordre ($\mathbb{E}\left\{|X_n|^2\right\} = |A|^2$ pour tout n) et centré. De plus,

$$\mathbb{E}\left\{X_n\overline{X_m}\right\} = |A|^2\, e^{2i\pi\nu_0(n-m)} \ ,$$

de sorte que X est stationnaire en moyenne d'ordre deux. Finalement, on a

$$C_X(n) = |A|^2\, e^{2i\pi n\nu_0} = \int_{-1/2}^{1/2} e^{2i\pi\nu_0 n}\, d\mu_X(\nu) \ ,$$

d'où on déduit que la mesure spectrale de X n'est autre que la mesure de Dirac δ_{ν_0} en ν_0, à un facteur $|A|^2$ près. Les signaux harmoniques fournissent l'exemple le plus simple de signaux aléatoires stationnaires en moyenne d'ordre deux ne possédant pas de densité spectrale.

2.2.4. *Représentation spectrale pour les processus stationnaires en m.o.d.* Les sections précédentes nous ont donné une représentation spectrale (i.e. de type "Fourier") pour la covariance d'un signal numérique aléatoire stationnaire. La covariance est un objet déterministe. Nous allons maintenant obtenir une représentation spectrale pour le processus lui même.

On note $\mathcal{M}_X \subset L^2(\mathcal{A})$ le sous-espace de $L^2(\mathcal{A})$ engendré par les variables aléatoires X_k. Soit ψ l'application linéaire de $\mathcal{M}_X$ dans $L^2([-1/2, 1/2], d\mu_X)$ définie par

$$\psi(X_k) = \epsilon_k : \nu \to e^{2i\pi k\nu} \ .$$

Il est clair que $\epsilon_k \in L^2([-1/2, 1/2], d\mu_X)$. De plus, on a

$$\langle \epsilon_k, \epsilon_\ell \rangle = \int_{-1/2}^{1/2} e^{2i\pi(k-\ell)\nu}\, d\mu_X(\nu) = C_X(k-\ell) = (X_k|X_\ell) \ .$$

Ainsi, ψ s'étend à une isométrie de $\mathcal{M}_X$ sur $L^2(d\mu_X)$.

Soit maintenant $A \subset [-1/2, 1/2]$ un Borélien, et soit χ_A l'indicatrice de A. A χ_A correspond une variable aléatoire $Z_A \in \mathcal{M}_X$, telle que

$$\mathbb{E}\left\{Z_A\overline{Z}_B\right\} = (Z_a|Z_B) = \mu_X(A \cap B) \ .$$

Ceci s'étend par linéarité aux fonctions simples de la forme $\sum_{k=1}^K \alpha_k^K \chi_{A_k^K}$ où les A_k sont des Boréliens de $[-\pi, \pi]$. On a $\psi^{-1}(\sum_{k=1}^K \alpha_k^K \chi_{A_K^k}) = \sum_{k=1}^K \alpha_k^K Z_{A_k^K}$. Finalement, on sait que toute toute fonction $\varphi \in L^2(d\mu_X)$ s'écrit comme limite de telles fonctions simples. Le résultat suivant montre que cette limite a également un sens dans $\mathcal{M}_X$.

THÉORÈME 1.9 (Cramèr). *Soit $\varphi \in L^2(d\mu_X)$, et considérons une suite de fonctions simples de la forme $\sum_{k=1}^K \alpha_k^K \chi_{A_k^K}$, telle que $\lim_{K\to\infty} \|\varphi - \sum_{k=1}^K \alpha_k^K \chi_{A_k^K}\| = 0$. La suite de variables aléatoires $\sum_{k=1}^K \alpha_k^K Z_{A_k^K}$ converge presque sûrement vers une limite notée*

$$\lim_{K\to\infty} \sum_{k=1}^K \alpha_k^K Z_{A_k^K} = \int_{-1/2}^{1/2} \varphi(\nu)\, dZ(\nu) \ , \tag{1.72}$$

et la limite est indépendante de la suite approximante.

En appliquant ce résultat au cas particulier $\varphi = \epsilon_k$, on obtient la représentation spectrale

COROLLAIRE 1.4. *On a pour tout k*

$$X_k = \int_{-1/2}^{1/2} e^{2i\pi k\nu}\, dZ(\nu) \ . \tag{1.73}$$

2.3. Le cas fini ; application à la simulation de processus stationnaires en m.o.d. La théorème de Wiener-Khinchin fournit une *représentation spectrale* pour la fonction d'autocovariance des signaux aléatoires stationnaires en moyenne d'ordre deux. La question suivante est : pouvons nous obtenir une représentation similaire (de type "Fourier") pour les signaux aléatoires eux mêmes ?

Nous allons voir dans la section suivante un tel théorème de représentation spectrale pour les signaux numériques infinis. Il est utile, pour motiver cette discussion, de faire une parenthèse avec le cas des signaux numériques aléatoires de longueur finie. Il est tout d'abord nécessaire d'adapter la définition de stationnarité à cette situation. Il faut pour cela tenir compte des conditions aux bords, que l'on suppose ici périodiques.

DÉFINITION 1.12. *Soit $X = \{X_n, n = 0, \dots N-1\}$ un signal numérique aléatoire du second ordre de longueur N. X est stationnaire en moyenne d'ordre deux si*

$$m_X(n) = m_X(0) := m_X \ , \quad \forall n = 0, \dots N-1 \tag{1.74}$$

et

$$R_X((n+\tau)\ mod N, (m+\tau)\ mod N) = R_X(n,m) := R_X((n-m)\ mod N) \ , \quad \forall n, m, \tau \tag{1.75}$$

Dans ce cas, on a aussi

$$C_X((n+\tau)\ mod N, (m+\tau)\ mod N) = C_X(n,m) := C_X((n-m)\ mod N) \ , \quad \forall n, m, \tau \ .$$

Soit X un tel signal aléatoire, que l'on suppose de plus centré. Soit C_X son autocovariance, et soit $\mathcal{S}_X$ le vecteur défini par

$$\mathcal{S}_X(k) = \sum_{n=0}^{N-1} C_X(n)\, e^{-2i\pi kn/N} \ . \tag{1.76}$$

On considère la transformée de Fourier finie de X : le vecteur aléatoire $\{\hat{X}_0, \dots \hat{X}_{N-1}\}$, défini par

$$\hat{X}_k = \sum_{n=0}^{N-1} X_n\, e^{-2i\pi kn/N} \ . \tag{1.77}$$

Soit C_X son autocovariance Un calcul simple montre que

$$\mathbb{E}\left\{\hat{X}_k\right\} = 0 \ , \quad \forall k = 0, \dots N-1 \ ,$$

et que

$$\mathbb{E}\left\{\hat{X}_k \overline{\hat{X}}_\ell\right\} = N\, \mathcal{S}_X(k)\, \delta_{k\ell} \ .$$

Les composantes de $\hat{X}$ sont donc décorrélées. Utilisant la transformée de Fourier finie inverse, on peut alors écrire

$$X_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} e^{2i\pi kn/N}\, Y_k \ , \tag{1.78}$$

où les variables aléatoires

$$Y_k = \frac{1}{\sqrt{N}}\, \hat{X}_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X_n e^{-2i\pi kn/N} \tag{1.79}$$

sont décorrélées :

$$\mathbb{E}\left\{Y_k \overline{Y}_\ell\right\} = \mathcal{S}_X(k)\, \delta_{k\ell} \ . \tag{1.80}$$

La représentation (1.78) porte parfois le nom de *représentation de Cramèr* en dimension finie.

Application à la simulation numérique de signaux aléatoires stationnaires : Lorsque l'on souhaite sinuler numériquement un signal aléatoire, on se place *de facto* dans une situation de dimension finie. Les ordinateurs proposent généralement des générateurs de nombres pseudo-aléatoires (par exemple, les fonctions de type `rand` sour `UNIX`), capables de fournir des séquences de nombres aléatoires aussi proches que possible de vecteurs identiquement distribués et décorrélés. Dans ces conditions, si on souhaite générer une réalisation d'un signal stationnaire en moyenne d'ordre deux, de spectre $\mathcal{S}_X$ donné, on peut procéder comme suit : on génère tout d'abord une séquence de nombres pseudo-aléatoires $\{W_0, W_1, \dots W_{N-1}\}$, qui sont tels que

$$\mathbb{E}\left\{X_k\right\} = 0 \ , \qquad \mathbb{E}\left\{W_k \overline{W}_\ell\right\} = \delta_{k\ell} \ ,$$

puis on exploite la représentation de Cramèr en formant

$$X_n = \frac{1}{N} \sum_{k=0}^{N-1} e^{2i\pi kn/N} \sqrt{\mathcal{S}_X(k)}\, W_k \ ;$$

il est alors facile de vérifier que $\mathbb{E}\left\{X_n\right\} = 0$ pour tout n, et que

$$C_X(n-m) = \mathbb{E}\left\{X_n \overline{X}_m\right\} = \frac{1}{N} \sum_{k=0}^{N-1} \mathcal{S}_X(k)\, e^{2i\pi k(n-m)/N} \ ,$$

ce qui est bien le résultat recherché.

2.4. Signaux aléatoires analogiques. Les signaux aléatoires analogiques sont des signaux aléatoires indexés par un ensemble continu. La majorité des opérations que nous avons effectuées dans le cas des signaux aléatoires numériques restent valables. Cependant, la continuité introduit des difficultés supplémentaires, comme par exemple des notions de régularité.

2.4.1. *Diverses notions de continuité.* Compte tenu de la structure supplémentaire apportée par l'aléa, il existe plusieurs notions différentes de continuité que l'on peut imposer à un processus.

DÉFINITION 1.13. *Soit X un processus aléatoire sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, indexé par T.*

(1) *On dit que X est continu en moyenne d'ordre p si pour tout t, on a*

$$\lim_{t' \to t} \left[\mathbb{E}\left\{|X_{t'} - X_t|^p\right\}\right]^{1/p} = 0 \ . \tag{1.81}$$

(2) *X est continu en probabilité si pour tout t et pour tout $\epsilon > 0$,*

$$\lim_{t' \to t} \mathbb{P}\{|X_{t'} - X_t| \geq \epsilon\} = 0 \ . \tag{1.82}$$

(3) *X est continu presque sûrement si pour tout t,*

$$\lim_{t' \to t} X_{t'} = X_t \quad \mathbb{P} \text{ presque surement} \ . \tag{1.83}$$

(4) *X a presque sûrement ses trajectoires continues si, en notant*

$$\mathcal{A}_0 = \{a \in \mathcal{A}, X(a) \text{ est discontinue}\} \ ,$$

on a

$$\mathbb{P}\{\mathcal{A}_0\} = 0 \ . \tag{1.84}$$

Il existe une hiérarchie (partielle) entre ces notions de continuité. Par exemple, la continuité en probabilités est conséquence de la continuité en moyenne d'ordre p. Ceci résulte de l'inégalité de Chebyshev

LEMME 1.3. *Si X est une variable aléatoire sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, et si $f : x \to f(x)$ est une fonction monotone non décroissante, alors pour tout $\epsilon > 0$, on a*

$$\mathbb{P}\{|X| \geq \epsilon\} \leq \frac{1}{f(\epsilon)} \mathbb{E}\{f(|X|)\} \ . \tag{1.85}$$

Supposons que X soit un processus continu en moyenne d'ordre p, et prenons $f(x) = x^p$. Alors, pour tout t fixé, on a

$$\mathbb{P}\{|X_{t'} - X_t|^p \geq \epsilon\} \leq \epsilon^{-p} \mathbb{E}\{|X_{t'} - X_t|^p\} \ ,$$

et donc

$$\lim_{t' \to t} \mathbb{P}\{|X_{t'} - X_t|^p \geq \epsilon\} \leq \lim_{t' \to t} \frac{1}{\epsilon} \mathbb{E}\{|X_{t'} - X_t|^p\} = 0 \ .$$

On peut également montrer que la continuité presque sûre implique la continuité en probabilité. Par contre, la continuité presque sûre n'implique pas nécessairement la continuité presque sûre des trajectoires, comme le montre l'exemple suivant.

EXEMPLE 1.10. Prenons $T = [0, 1]$, $\mathcal{A} = [0, 1]$, muni de la mesure de Lebesgue $\mathbb{P} = m$, et définissons X par

$$X_t(\omega) = \begin{cases} 0 & \text{si } t < \omega \\ 1 & \text{sinon} \ . \end{cases}$$

X est presque sûrement continu : pour tout t, $\lim_{t' \to t} X_{t'} = X_t$, $\mathbb{P}$ presque sûrement. Par contre, les trajectoires de X sont presque sûrement discontinues.

En revanche, le processus est continu en moyenne d'ordre p : pour cela, pour tous $0 < t < t' < 1$, écrivons

$$\begin{aligned} \mathbb{E}\{|X_t - X_{t'}|^p\} &= \int_0^1 |X_t(\omega) - X_{t'}(\omega)|^p \, d\mathbb{P}(\omega) \\ &= \int_0^t |X_t(\omega) - X_{t'}(\omega)|^p \, d\mathbb{P}(\omega) + \int_t^{t'} |X_t(\omega) - X_{t'}(\omega)|^p \, d\mathbb{P}(\omega) \\ &\qquad + \int_{t'}^1 |X_t(\omega) - X_{t'}(\omega)|^p \, d\mathbb{P}(\omega) \\ &= t' - t \to 0 \text{ quand } t' \to t. \end{aligned}$$

Le résultat suivant, donné sans démonstration, donne une condition suffisante pour qu'un processus ait presque sûrement ses trajectoires continues.

LEMME 1.4 (Totoki). *Soit T un ouvert quelconque de $\mathbb{R}$. Si pour tout compact $K \subset T$, il existe trois constantes α, β, γ_K telles que pour tous $t, t' \in K$, on ait*

$$\mathbb{E}\{|X_t - X_{t'}|^\alpha\} \leq \gamma_K |t - t'|^{\beta+1} , \tag{1.86}$$

alors X a presque sûrement ses trajectoires continues.

2.4.2. *Processus continus en moyenne d'ordre 2.*

DÉFINITION 1.14. *Un processus du second ordre est dit continu en moyenne d'ordre 2 si pour tout $t \in T$, $\mathbb{E}\left\{|X_{t+\delta} - X_t|^2\right\} \to 0$ quand $\delta \to 0$.*

La proposition suivante donne une caractérisation simple de la continuité en moyenne d'ordre 2.

PROPOSITION 1.12. *Le processus X est continu en moyenne d'ordre 2 si la fonction d'autocorrélation*

$$t, t' \to R_X(t, t')$$

est continue sur la diagonale (c'est à dire dans la limite $t \to t'$).

La preuve utilise le lemme suivant :

LEMME 1.5. *Soit H un espace de Hilbert. Une suite $\{v_0, v_1, \ldots\}$ d'éléments de H converge vers $v \in H$ tel que $\|v\|^2 = x > 0$ si et seulement si*

$$\lim_{m,n\to\infty} \langle v_n, v_m \rangle = x \ . \tag{1.87}$$

Preuve du lemme : Si $\lim\langle v_n, v_m\rangle = x$, alors $\|v_n - v_m\|^2 = \|v_n\|^2 + \|v_m\|^2 - 2\Re\langle v_n, v_m\rangle \to 0$ quand $m, n \to \infty$. Inversement, si la suite $\{v_n\}$ converge, alors $\langle v_n, v_m\rangle \to \|v\|^2 = x$. ♠

Preuve de la proposition : Il suffit d'appliquer le lemme 1.5 à l'espace $H = L^2(\mathcal{A})$; pour un t fixé, on s'intéresse à la convergence de $X_{t'}$ vers X_t. D'après le lemme, celle-ci est équivalente à

$$\lim_{m,n\to\infty} R_X(t_m, t_n) = T_X(t, t) \ ,$$

ce qui permet de conclure. ♠

Remarquons que les processus continus en moyenne d'ordre 2 n'ont en général pas leurs trajectoires continues.

2.4.3. *Processus stationnaires en moyenne d'ordre 2.* Après quelques définitions et propriétés générales, on se focalisera plus particulièrement sur le cas des processus indexés sur $\mathbb{R}$.

DÉFINITION 1.15.
(1) *Un processus $X = \{X_t, t \in T\}$ est stationnaire si pour tout τ, les processus $\{X_t, t \in T\}$ et $\{X_{t+\tau}, t \in T\}$ ont le même système de lois marginales.*
(2) *Le processus X est stationnaire en moyenne d'ordre deux si $\mu_X(t) = \mu_X(0) := \mu_X$ pour tout t, et si $R_X(t+\tau, \tau) = R_X(t, 0) := R_X(t)$ pour tout t, τ.*

La stationnarité au sens strict est généralement une propriété trop contraignante. C'est pourquoi on se borne généralement à supposer la stationnarité à l'ordre deux. Cette hypothèse est déjà extrêmement puissante, comme on va le voir, tout d'abord dans le cas des processus sur $\mathbb{R}$.

PROPOSITION 1.13 (Wiener-Khintchin). *Soit $\{X_t, t \in \mathbb{R}\}$ un processus du second ordre, continu et stationnaire en moyenne d'ordre deux. Alors il existe une mesure non-négative bornée μ_X sur $\mathbb{R}$, telle que*

$$R_X(\tau) = \int_{-\infty}^{\infty} e^{2i\pi\nu\tau}\, d\mu_X(\nu) \ . \tag{1.88}$$

μ_X est appelée mesure spectrale du processus.

La proposition est une conséquence immédiate du fait que l'autocorélation R_X est définie positive(voir proposition 1.11) et du théorème de Bochner (voir par exemple [**8**] pour la démonstration) :

THÉORÈME 1.10 (Bochner). *Soit $\Phi : \mathbb{R} \to \mathbb{R}$ une fonction définie positive et continue. Alors il existe une mesure non-négative bornée μ sur $\mathbb{R}$, telle que*

$$\Phi(t) = \int_{-\infty}^{\infty} e^{2i\pi\nu t}\, d\mu(\nu) \ . \tag{1.89}$$

On peut alors faire une décomposition de Lebesgue de la mesure μ_X :

$$\mu_X = (\mu_X)_{ac} + (\mu_X)_s \ ,$$

où $(\mu_X)_{ac}$ est absolument continue par rapport à la mesure de Lebesgue, et $(\mu_X)_s$ est singulière. On peut alors écrire

$$d(\mu_X)_{ac}(\nu) = \mathcal{S}_X(\nu) d\nu \ . \tag{1.90}$$

La fonction $\mathcal{S}_X$ est appelée densité spectrale de puissance du processus, et on a

COROLLAIRE 1.5. *La densité spectrale $\mathcal{S}_X$ est non-négative, et intégrable.*

REMARQUE 1.13. Il n'existe pas toujours de densité spectrale : par exemple, dans le cas d'un processus harmonique, on a $R_X(t) = \sigma^2 \cos(2\pi\nu_0 t)/2$; le processus est bien du second ordre, continu en moyenne d'ordre 2 (car R_X est continue), mais n'admet pas de densité spectrale.

REMARQUE 1.14. Si $m_X \neq 0$, le processus n'admet pas de densité spectrale.

EXEMPLE 1.11. *Bruit blanc, ou bruit blanc filtré :*Un exemple "pathologique" est donné par le cas $\mathcal{S}_X = \sigma^2$, qui échappe à notre analyse car $\mathcal{S}_X \notin L^1(\mathbb{R})$. C'est ce que l'on appelle le bruit blanc, pour lequel il est nécessaire de développer une approche adaptée (son autocorrélation n'existe pas en tant que fonction). Cependant, il est utile (et réaliste en pratique) de considérer le bruit blanc filtré, défini par

$$\mathcal{S}_X(\nu) = \sigma^2\, \chi_{[-\nu_0,\nu_0]}(\nu) \ .$$

Il est clair que $\mathcal{S}_X \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. Par transformation de Fourier inverse, on aboutit à la fonction d'autocorrélation suivante :

$$R_X(\tau) = \int_{-\infty}^{\infty} \mathcal{S}_X(\nu)\, e^{2i\pi\nu\tau}\, d\nu = 2\nu_0\sigma^2\, \mathrm{sinc}(2\pi\nu_0\tau) \ .$$

On obtient un tel processus par filtrage passe-bas (voir plus loin) d'un bruit blanc.

2.4.4. *Filtrage linéaire.* On peut effectuer des opérations de filtrage linéaire sur des signaux aléatoires, comme dans le cas des signaux déterministes. Pour simplifier, on se limitera ici au cas des filtres de convolution. On considère pour cela une fonction $h \in L^1_{loc}(\mathbb{R})$. Etant donné un processus aléatoire $X = \{X_t, t \in \mathbb{R}\}$ sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, on s'intéresse au processus $Y = \{Y_t, t \in \mathbb{R}\}$ défini sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$ également, par

$$Y_t = (K_h X)_t = \int_{-\infty}^{\infty} h(t-s) X_s \, ds \ . \tag{1.91}$$

K_h est un filtre linéaire de convolution, de réponse impulsionnelle h. Quand $\hat{h} = m$ existe, m est appelé *fonction de transfert* du filtre, comme dans le cas déterministe.

PROPOSITION 1.14. *Si $h \in L^1(\mathbb{R})$, et si X est un processus du second ordre, continu et stationnaire en moyenne d'ordre deux, alors Y est également un processus du second ordre, continu et stationnaire en moyenne d'ordre deux, et on a*

$$m_Y = \hat{h}(0) m_X \tag{1.92}$$

$$R_Y(\tau) = \int_{-\infty}^{\infty} h(s)\overline{h}(t) R_X(\tau + (t-s)) \, dt \, ds \tag{1.93}$$

$$d\mu_Y(\nu) = \left|\hat{h}(\nu)\right|^2 d\mu_X(\nu) \ . \tag{1.94}$$

Preuve : Pour tout compact $K \subset \mathbb{R}$, posons $Y_t^K = \int_K h(t-s) X_s ds$. Puisque $X_s \in L^2(\mathcal{A})$ pour tout s, on a

$$\mathbb{E}\left\{|Y_t^K|^2\right\} \leq \sup_{s \in K} \mathbb{E}\left\{|X_s|^2\right\} \int_K |h(t-s)| ds \leq \sup_{s \in K} \mathbb{E}\left\{|X_s|^2\right\} \|h\|_1 < \infty \ ,$$

d'où Y est du second ordre (la borne est uniforme en K). Calculons maintenant

$$\mathbb{E}\left\{Y_t^K \overline{Y}_{t'}^{K'}\right\} = \int_{K \times K'} h(t-s)\overline{h}(t'-s') R_X(s,s') \, ds \, ds' \ .$$

La limite $K, K' \to \mathbb{R}$ de cette expression est bien définie car $R_X, h \in L^1(\mathbb{R})$; donc Y est du second ordre. Calculons

$$m_Y = \int_{-\infty}^{\infty} h(t-s) \mathbb{E}\left\{X_s\right\} ds = m_X \int_{-\infty}^{\infty} h(t-s) = m_X \, \hat{h}(0) \ .$$

D'autre part,

$$\begin{aligned} R_Y(t,t') = \mathbb{E}\left\{Y_t, Y_{t'}\right\} &= \int_{-\infty}^{\infty} h(t-s)\overline{h}(t'-s') R_X(s-s') \, ds \, ds' \\ &= \int_{-\infty}^{\infty} h(u)\overline{h}(v) R_X(t-t'-(u-v)) \, du \, dv \end{aligned}$$

est bien une fonction de $t - t'$, d'où la stationnarité. Finalement, la propriété (1.94) s'obtient par une transformation de Fourier. ♠

EXEMPLE 1.12. *Filtrage passe-bas idéal* :Supposons que le signal aléatoire du second ordre, continu et stationnaire en moyenne d'ordre deux, X ait une densité de spectrale $\mathcal{S}_X$:

$$R_X(t) = \int_{-\infty}^{\infty} \mathcal{S}_X(\nu) e^{2i\pi\nu t} \, d\nu \ ,$$

et soit T le filtre passe bas idéal, défini par

$$Tf(t) = \int_{-\nu_0}^{\nu_0} \hat{f}(\nu) e^{2i\pi\nu t}\, d\nu \ .$$

T est bien un filtre linéaire, continu sur $L^2(\mathbb{R})$ et stable. Si on introduit le processus filtré $Y = TX$, on voit donc que Y est lui aussi du second ordre, continu et stationnaire en moyenne d'ordre 2, et admet une densité spectrale $\mathcal{S}_Y$ de la forme

$$\mathcal{S}_Y(\nu) = |\hat{h}(\nu)|^2\, \mathcal{S}_X(\nu) \ .$$

Ce type de filtrage passe-bas possède le même statut que le filtrage passe bas des signaux déterministes. Il est soumis aux mêmes contraintes (à savoir que le filtre parfait n'est pas causal, mais peut être approximé par un filtre causal).

DÉFINITION 1.16. *Un signal aléatoire du socond ordre, continu et stationnaire en moyenne d'ordre 2, est dit à bande limitée si sa mesure spectrale a un support borné.*

Comme dans le cas déterministe, les signaux aléatoires à bande limitée offrent un cadre adéquat pour développer une théorie d'échantillonnage.

2.4.5. *La représentation de Cramèr.* Le théorème de Wiener-Khintchin énoncé ci-dessus est la clé de voûte de la théorie spectrale des processus stationnaires. En fait, il est possible de montrer que tout processus stationnaire peut s'écrire sous la forme d'une intégrale de Fourier, c'est ire une superposition d'exponentielles complexes, avec des poids aléatoires. C'est ce que l'on appelle la représentation de Cramèr, que l'on obtient (très succintement) comme suit. Etant donné un processus du second ordre X_t, continu et stationnaire en moyenne d'ordre deux, on établit une correspondance entre les variables aléatoires engendrées par le processus et des fonctions de la variable ν comme suit : $\forall t \in \mathbb{R}$, on établit la correspondance

$$X_t \longleftrightarrow e^{2i\pi\nu t} \ , \tag{1.95}$$

que l'on prolonge par linéarité à l'espace fermé $\mathcal{M}_X$ engendré par les variables aléatoires $X_t, t \in \mathbb{R}$. On déduit de (1.88) l'identité

$$\langle X_t, X_s \rangle_{L^2(\mathcal{A})} = \langle e^{2i\pi\nu t}, e^{2i\pi\nu s} \rangle_{L^2(\mathbb{R}, d\mu)} \tag{1.96}$$

Par conséquent, nous avons obtenu une isométrie entre le sous-espace $\mathcal{M}_X \subset L^2(\mathcal{A})$ et $L^2(\mathbb{R}, d\mu)$. Pour toutes $f, g \in L^2(\mathbb{R}, d\nu)$ notons $F, G \in \mathcal{M}_X$ respectivement les variables aléatoires associées :

$$F \longleftrightarrow f(\nu) \ , \qquad G \longleftrightarrow g(\nu) \tag{1.97}$$

Nous avons ainsi

$$\mathbb{E}\left\{F\overline{G}\right\} = \int f(\nu)\overline{g}(\nu) d\mu(\nu) \tag{1.98}$$

Considèrons maintenant un Borèlien $A \subset \mathbb{R}$ et associons-lui son indicatrice $\chi_A(\nu)$. On lui associe donc une variable aléatoire $Z(A) \in \mathcal{M}_X$, telle que pour toute paire de Borèliens A, B

$$\mathbb{E}\left\{Z(A)\overline{Z(B)}\right\} = \mu(A \cap B) \tag{1.99}$$

Toute fonction $f(\nu)$ peut être obtenue comme limite de fonctions de la forme $\sum_k \alpha_k \chi_{A_k}(\nu)$ pour des Borèliens A_k et des coefficients α_k convenablement choisis ;

à ces dernières on fait correspondre la variable aléatoire $\sum \alpha_k Z(A_k)$. Par passage à la limite, on fait donc correspondre à $f(\nu)$ la variable aléatoire

$$\int f(\nu) Z(d\nu) \longleftarrow f(\nu) \tag{1.100}$$

Du cas particulier $f(\nu) = e^{2i\pi\nu t}$ on déduit

$$X_t = \int e^{2i\pi\nu t} Z(d\nu) \ , \tag{1.101}$$

ce qui est appelé la représentation de Cramèr du processus. Il résulte des discussions précédentes que

$$\mathbb{E}\left\{ Z(d\nu)\overline{Z(d\nu')} \right\} = \mathcal{S}(\nu)\delta(\nu - \nu')d\nu \tag{1.102}$$

La représentation de Cramèr du processus est parfois mise sous la forme

$$X_t = \int_{-\infty}^{\infty} e^{2i\pi\nu t} \sqrt{\mathcal{S}(\nu)} dW_\nu \ , \tag{1.103}$$

où dW_ν est une mesure aléatoire (Gaussienne) telle que

$$\mathbb{E}\left\{ dW_\nu \overline{dW_{\nu'}} \right\} = \delta(\nu - \nu')d\nu \tag{1.104}$$

2.5. Echantillonnage. Nous avions vu une version "déterministe" du théorème d'échantillonnage. Il est possible de montrer un résultat analogue dans le cas des signaux aléatoires. Le résultat est essentiellement le même : si X est un processus stationnaire admettant une densité spectrale à support borné, alors X est caractérisé par une version "échantillonnée", pour peu que la fréquence d'échantillonnage soit choisie assez grande. Plus précisément :

THÉORÈME 1.11. *Soit X un processus du second ordre centré sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, continu et stationnaire en moyenne d'ordre deux, admettant une densité spectrale $\mathcal{S}_X$ à support borné inclus dans l'intervalle $[-\nu_0, \nu_0]$. Alors X est caractérisé par le processus discret $\{X_{n/\eta}, n \in \mathbb{Z}\}$ si et seulement si $\eta \geq 2\nu_0$. On a dans ce cas la formule d'interpolation : pour tout $t \in \mathbb{R}$,*

$$X_t = \sum_{n=-\infty}^{\infty} X_{n/\eta} \, \frac{\sin(\pi\eta(t - n/\eta))}{\pi\eta(t - n/\eta)} \tag{1.105}$$

au sens de $L^2(\mathcal{A})$.

En d'autres termes, ceci signifie que pour tout t,

$$\lim_{N\to\infty} \left\| X_t - \sum_{n=-N}^{N} X_{n/\eta} \right\| = 0 \ .$$

CHAPITRE 2

Quantification ; PCM et DPCM

Le PCM est le schéma le plus simple de codage des signaux, et est utilisé par exemple pour le codage des signaux audio sur les CDs. Il est essentiellement basé sur un échantillonnage, suivi d'une quantification des échantillons.

1. Quantification scalaire ; le codeur PCM

PCM est le sigle désignant le *Pulse Code Modulation*, un standard (ou plutôt une famille de standards) adopté de façon assez universelle. Le système PCM est connu pour offrir des performances assez moyennes en termes de compression, mais aussi pour sa grande robustesse (notamment par rapport aux erreurs de transmission) et sa faible complexité (algorithme peu gourmand en mémoire et CPU). On décrit ici le PCM de façon assez sommaire, dans le but d'introduire quelques idées simples, notamment en ce qui concerne la quantification scalaire.

Le PCM consiste essentiellement en un échantillonneur, suivi d'un quantificateur (uniforme) appliqué aux échantillons, et enfin d'un système simple de codage. La phase d'échantillonnage a déjà été décrite dans le chapitre précédent. Le codage est basé sur le principe d'une attribution "démocratique" des bits : chaque valeur quantifiée sera codée sur un nombre de bits constant. On parle de code de longueur constante. On se concentre maintenant sur la quantification.

1.1. Quantification scalaire. Considérons donc des échantillons f_n d'un signal f. La base du PCM est de modéliser chacun de ces échantillons comme une variable aléatoire, sans se préoccuper des corrélations entre échantillons. Supposons que la variable aléatoire X soit une variable aléatoire continue du second ordre, de densité de probabilités ρ_X.

On considère donc un quantificateur

$$Q : \mathbb{R} \to E_M = \{y_-, y_0, \dots y_{M-1}, y_+\} \ ,$$

et la variable aléatoire discrète $Y = Q(X)$, à valeurs dans l'ensemble fini E_M. Q est défini à partir d'intervalles de la forme $[x_k, x_{k+1}[$ par

$$Q(x) = \begin{cases} y_- & \text{si } x \leq x_0 \\ y_k & \text{si } x \in [x_k, x_{k+1}[\ , \quad k = 0, \dots M-1 \\ y_+ & \text{si } x > x_M \end{cases} \tag{2.1}$$

On s'intéresse particulièrement à l'erreur de quantification, c'est à dire à la variable aléatoire

$$Z = X - Y = X - Q(X) \ , \tag{2.2}$$

que l'on va chercher à évaluer. On doit donc se donner une façon de mesurer cette quantité. La quantité d'intérêt la plus simple est ici la variance de l'erreur de quantification

$$\sigma_Z^2 = \mathbb{E}\left\{Z^2\right\} = \int (x - Q(x))^2 \rho_X(x) dx \tag{2.3}$$

$$= \sum_{k=0}^{M-1} \left(\int_{x_k}^{x_{k+1}} (x - y_k)^2 \rho_X(x) dx \right) + \int_{-\infty}^{x_0} (x - y_-)^2 \rho_X(x) dx + \int_{x_M}^{\infty} (x - y_+)^2 \rho_X(x) dx \ , \tag{2.4}$$

et c'est celle-ci que nous allons évaluer dans certaines situations simples. Les deux derniers termes forment le *bruit de saturation*, alors que les autres forment le *bruit granulaire*. Dans le cas d'un signal borné (c'est à dire quand ρ_X a un support borné), le bruit de saturation est généralement évité [1].

DÉFINITION 2.1. *On considère un quantificateur comme décrit ci-dessus. Le* facteur de performance *du quantificateur est le quotient*

$$\epsilon^2 = \frac{\sigma_X^2}{\sigma_Z^2} \ , \tag{2.5}$$

où σ_X^2 et σ_Z^2 sont les variances respectives du signal X et du bruit de quantification Z. Le Rapport Signal à Bruit de Quantification *est quant à lui défini par*

$$SNR_Q = 10 \log_{10}(\epsilon^2) = 10 \log_{10} \left(\frac{\sigma_X^2}{\sigma_Z^2} \right) \ . \tag{2.6}$$

Il est bien évident que l'objectif que l'on se fixe en développant un quantificateur est de maximiser le rapport signal à bruit, pour un débit R fixé. Ou, plus ambitieusement, on cherche à construire une *théorie Débit-Distorsion*, qui décrive l'évolution de la distorsion en fonction de R. On va voir que ceci est possible au prix d'approximations simplificatrices. On commencera par étudier le cas le plus simple, à savoir le cas de la quantification uniforme.

1.2. Quantification uniforme. On s'intéresse maintenant au cas le plus simple, à savoir le cas de la quantification uniforme. L'effet de la quantification uniforme sur un signal est décrit en FIG. 1.

Pour simplifier (en évitant d'avoir à considérer le bruit de saturation), on suppose pour cela que la variable aléatoire X est bornée, et prend ses valeurs dans un intervalle I. Une quantification uniforme consiste à découper I en $M = 2^R$ sous-intervalles de taille constante, notée Δ. Plus précisément :

DÉFINITION 2.2. *Soit X une variable aléatoire dont la densité ρ_X est à support borné dans un intervalle $I = [x_{min}, x_{max}]$. Soit R un entier positif, et soit $M = 2^R$. Le quantificateur uniforme de débit R est donné par le choix*

$$x_0 = x_{min} \ , \quad x_m = x_0 + m\Delta \ , \tag{2.7}$$

avec

$$\Delta = |I|/M = |I| 2^{-R} \ , \tag{2.8}$$

[1] Encore que ceci ne soit pas obligatoire ; on peut parfois se permettre une certaine quantité de bruit de saturation.

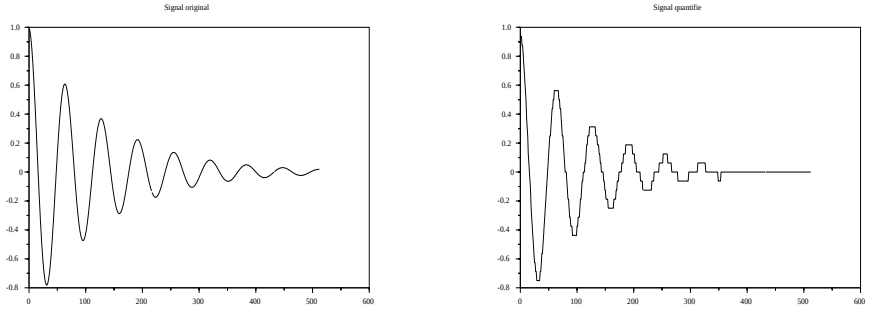

FIG. 1. Exemple de quantification : un signal simple (à gauche) et le même signal après quantification de chaque échantillon sur 5 bits (32 niveaux de quantification).

et

$$y_m = \frac{x_m + x_{m+1}}{2} \ . \tag{2.9}$$

Supposons que le quantificateur soit un quantificateur "haute résolution", c'est à dire qu'à l'intérieur d'un intervalle $[x_k, x_{k+1}]$, la densité de probabilités $x \to \rho_X(x)$ soit lentement variable, et puisse être approximée par la valeur $\rho_k = \rho(y_k)$. Sous ces conditions, on montre facilement que

$$\mathbb{E}\left\{X - Q(X)\right\} \approx 0 \ , \tag{2.10}$$

et on écrit alors

$$\int_{x_k}^{x_{k+1}} (x - y_k)^2 \rho_X(x) dx \approx \frac{\rho_k}{3} \left((x_{k+1} - y_k)^3 - (x_k - y_k)^3 \right) \ . \tag{2.11}$$

Notons que pour un ρ_k donné, cette dernière quantité atteint son minimum (par rapport à y_k en $y_k = (x_k + x_{k+1})/2$, c'est à dire la valeur donnée en hypothèse, de sorte que $x_{k+1} - y_k = y_k - x_k = \Delta/2$. Par conséquent, on obtient

$$\int_{x_k}^{x_{k+1}} (x - y_k)^2 \rho_X(x) dx \approx \rho_k \frac{\Delta^3}{12} \ . \tag{2.12}$$

Par ailleurs, on écrit également $1 = \int \rho_X(x) dx \approx \Delta \sum_k \rho_k$, d'où

$$\sum_k \rho_k \approx \frac{1}{\Delta} \ . \tag{2.13}$$

Finalement, en faisant le bilan, on aboutit à

$$\sigma_Z^2 \approx \frac{\Delta^3}{12} \sum_0^{M-1} \rho_k \approx \frac{\Delta^2}{12} \tag{2.14}$$

REMARQUE 2.1. Notons que d'après ces estimations, on obtient une estimation de la courbe débit-distortion fournie par la quantification uniforme :

$$D = \sigma_Z^2 = C^{ste}\, 2^{-2R} \ .$$

Plus précisément, on montre que

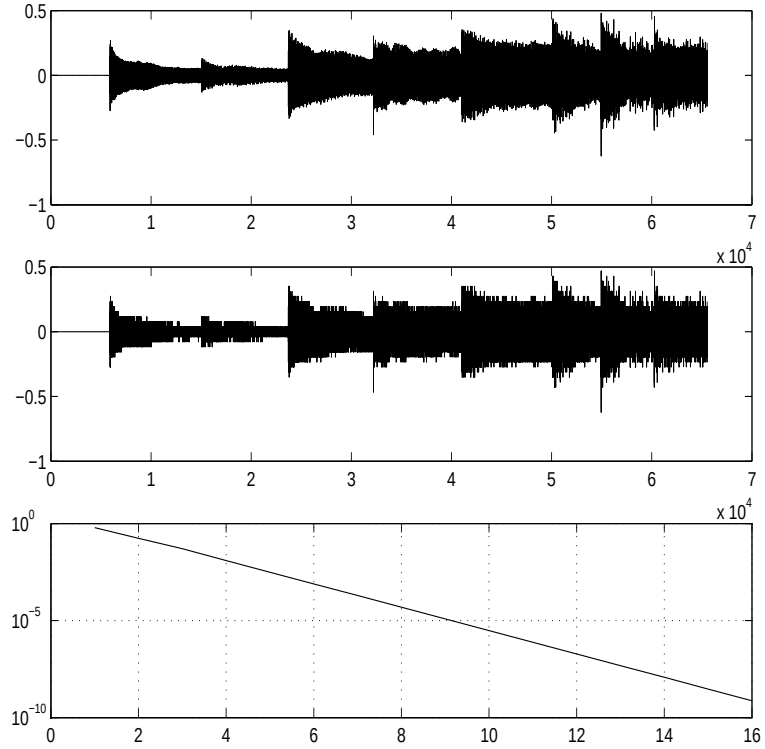

FIG. 2. Quantification d'un signal audio : un signal "test" (le "carillon" test des codeurs MPEG audio, en haut), une version quantifiée sur 4 bits (milieu), et le logarithme de la distorsion en fonction du débit (en bas).

LEMME 2.1. *Soit X une variable aléatoire bornée dans I. Supposons en outre que $\rho_X \in C^1(\mathbb{R})$. Soit Q un quantificateur uniforme sur R bits par échantillon. Alors, on a*

$$\sigma_Z^2 = \frac{\Delta^2}{12} + r \ , \tag{2.15}$$

avec

$$|r| \leq C^{ste}\, 2^{-3R} \sup_x |\rho'_X(x)| \ . \tag{2.16}$$

Preuve : Il suffit de donner un sens plus précis à l'approximation (2.14). Par accroissements finis, on obtient

$$\int_{x_k}^{x_{k+1}} (x - y_k)^2 \rho_X(x) dx = \rho_k \frac{\Delta^3}{12} + \int_{x_k}^{x_{k+1}} (x - y_k)^3 \rho'_X(y) dx = \rho_k \frac{\Delta^3}{12} + r_k \ ,$$

pour un certain $y = y(x) \in [x_k, x_{k+1}]$. On a donc

$$|r_k| \leq \sup_{y \in [x_k, x_{k+1}]} |\rho'_X(y)| \int_{x_k}^{x_{k+1}} |x - y_k|^3 \, dx \ .$$

Cette derniére intégrale vaut

$$2\int_0^{\Delta/2} u^3 du = \frac{\Delta^4}{32} .$$

Donc,

$$|r| \leq \sum_1^M |r_k| \leq M \sup|\rho'_X| \frac{\Delta^4}{32} = \frac{|I|^4}{32} 2^{-3R} \sup|\rho'_X| .$$

L'estimation (2.13) se fait de façon similaire. Ceci conclut la démonstration ♠

Ces approximations permettent d'obtenir une première estimation pour l'évolution du SNR_Q en fonction du taux R. En effet, nous avons

$$SNR_Q = 10\log_{10}\left(\frac{\sigma_X^2}{\sigma_Z^2}\right) = 20R\log_{10}(2) - 10\log_{10}\left(\frac{|I|^2}{12\sigma_X^2}\right) \approx 6,02\,R + C^{ste}$$

où la constante dépend de la loi de X. Ainsi, on aboutit à la règle empirique suivante :

Ajouter un bit de quantification revient à
augmenter le rapport signal à bruit de quantification de $6dB$ *environ.*

Ceci est très bien illustré par la FIG. 2, qui représente un signal audio (un son de carillon, utilisé comme signal test par le consortium `MPEG`). La figure du haut représente le signal original, et la figure du milieu représente une version quantifiée sur 4 bits. Les distorsions sont assez visibles. La figure du bas représente quant à elle le rapport signal à bruit SNR_Q en fonction du débit R, pour un R variant de 1 à 16. On voit que comme attendu, le comportement est remarquablement proche d'un comportement linéaire.

EXEMPLE 2.1. Prenons par exemple une variable aléatoire X, avec une loi uniforme sur un intervalle $I = [-x_0/2, x_0/2]$. On a alors

$$\sigma_X^2 = \frac{1}{x_0}\int_{-x_0/2}^{x_0/2} x^2 dx = \frac{x_0^2}{12}$$

de sorte que l'on obtient pour le rapport signal à bruit de quantification, exprimé en décibels (dB) :

$$SNR_Q \approx 6,02\,R .$$

C'est le résultat standard que l'on obtient pour le codage des images par PCM.

1.3. Compensation logarithmique. Une limitation de la quantification uniforme que nous avons vue plus haut est que, tant que l'on reste dans le cadre de validité des approximations que nous avons faites, la variance σ_Z^2 du bruit de quantification (qui vaut $\Delta^2/12$) ne dépend pas de la variance du signal. Donc, le rapport signal à bruit de quantification décroît quand la variance du signal décroît. Or, celle-ci est souvent inconnue à l'avance, et peut aussi avoir tendance à varier (lentement) au cours du temps. Il peut donc être avantageux de "renforcer" les faibles valeurs du signal de façon "autoritaire". Pour ce faire, une technique classique consiste à effectuer sur les coefficients une transformation, généralement non-linéaire, afin de rendre la densité de probabilités plus "plate", plus prôche d'une densité de variable aléatoire uniforme. Il s'agit généralement d'une transformation logarithmique, modifiée à l'origine pour éviter la singularité.

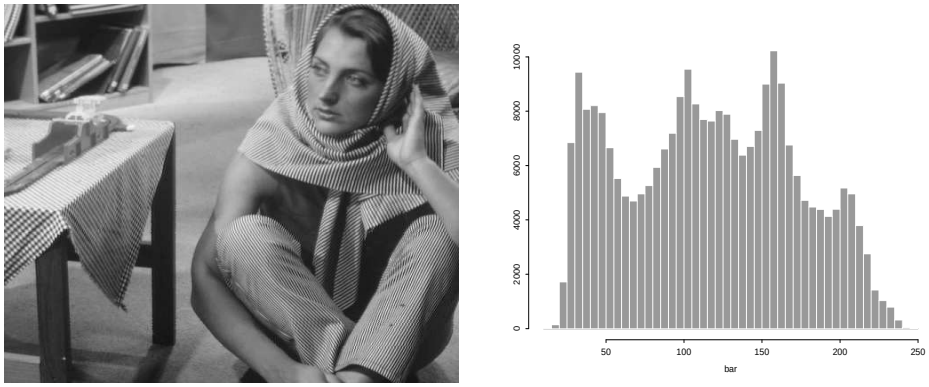

FIG. 3. Image, et densité de probabilités empirique des valeurs de pixel.

Deux exemples de telles transformations sont couramment utilisées, en téléphonie notamment : la *loi A*, qui correspond au standard européen, et la *loi* μ (standard nord-américain).

1.3.1. *La loi A.* Le premier exemple est la loi A, qui consiste en une modification linéaire pour les faibles valeurs du signal, et d'une compensation logarithmique pour les plus grandes valeurs. Plus précisément, la fonction de compensation est donnée par

$$c(x) = \begin{cases} \frac{A|x|}{1+\log A}\,\mathrm{sgn}(x) & \text{si } |x| \leq \frac{x_{max}}{A} \\ x_{max}\,\frac{1+\log(A|x|/x_{max})}{1+\log A}\,\mathrm{sgn}(x) & \text{si } |x| > \frac{x_{max}}{A} \end{cases} \tag{2.17}$$

On peut alors montrer que le rapport signal à bruit de quantification devient

$$SNR \approx \begin{cases} 6,02\,R + 4,77 - 20\log_{10}(1+\log A) & \text{pour les grandes valeurs du signal} \\ SNR_Q + 20\log_{10}\left(\frac{A}{1+\log A}\right) & \text{pour les faibles valeurs du signal .} \end{cases} \tag{2.18}$$

EXEMPLE 2.2. Valeur typique du paramètre, pour le codage du signal de parole : $A = 87.56$ (standard PCM européen). Le SNR correspondant, exprimé en décibels (dB), vaut

$$SNR_A \approx 6,02\,R - 9,99\ .$$

1.3.2. *La loi* μ. La fonction de compensation est dans ce cas donnée par

$$c(x) = x_{max}\,\frac{\log(\mu|x|/x_{max})}{\log(1+\mu)}\,\mathrm{sgn}(x)\ . \tag{2.19}$$

On montre alors que le rapport signal à bruit de parole est approximativement donné par

$$SNR \approx \begin{cases} 6,02\,R + 4,77 - 20\log_{10}(\log(1+\mu)) & \text{pour les grandes valeurs du signal} \\ SNR_Q + 20\log_{10}\left(\frac{\mu}{\log(1+\mu)}\right) & \text{pour les faibles valeurs du signal .} \end{cases} \tag{2.20}$$

EXEMPLE 2.3. Valeur typique du paramètre : pour le signal de parole : $\mu = 255$ (standard PCM US). Le SNR correspondant, exprimé en décibels (dB), est de la forme

$$SNR_\mu \approx 6,02\,R - 10,1\ .$$

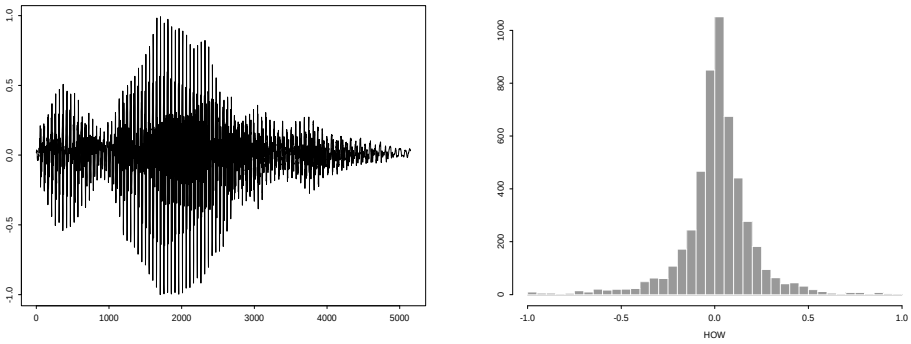

FIG. 4. Signal de parole, et densité de probabilités empirique des valeurs du signal.

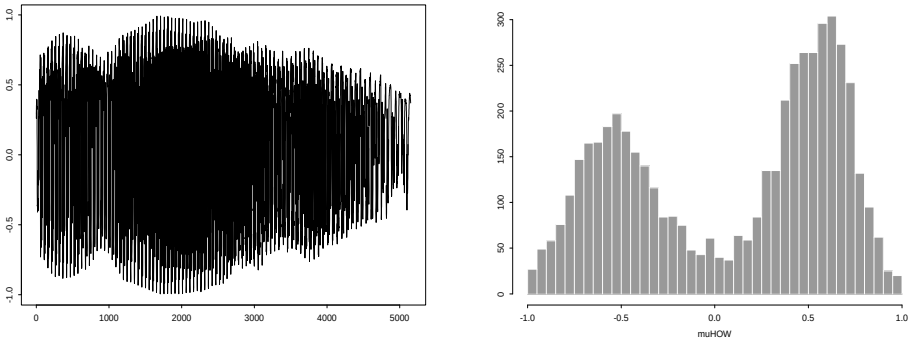

FIG. 5. Signal de parole, et densité de probabilités empirique des valeurs du signal, après compensation logarithmique par loi μ.

La figure 5 représente le signal de parole de la figure 4 après compensation logarithmique, et la densité de probabilités correspondante, qui est beaucoup plus régulière que celle de la figure 4. La figure 6 reprend l'exemple de la figure 2, et montre l'effet de la compensation logarithmique sur la courbe débit-distorsion : la nouvelle courbe est toujours essentiellement linéaire, mais se situe en dessous de la précédente. La compensation logarithmique a donc bien amélioré les performances du quantificateur.

1.4. Quantification scalaire optimale. Par définition, un quantificateur scalaire optimal est un quantificateur qui, pour un nombre de niveaux de quantification N fixé, minimise la distorsion. Il existe un algorithme, appelé *algorithme de Lloyd-Max*, qui permet d'atteindre l'optimum connaissant la densité de probabilités ρ_X de la variable aléatoire X à quantifier.

Supposons par exemple que nous ayions à quantifier une variable aléatoire X de densité ρ_X à support non borné. Pour obtenir le résultat, on commence par

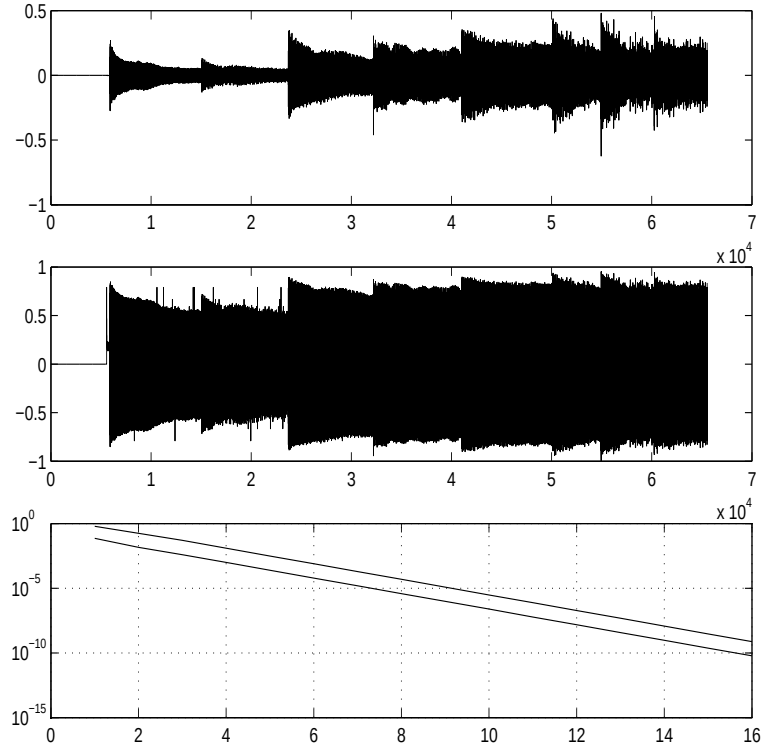

FIG. 6. Quantification logarithmique d'un signal audio : le signal test "carillon" (en haut), une version corrigée par loi μ (milieu), et le logarithme de la distorsion en fonction du débit (en bas) pour le signal original et le signal compensé logarithmiquement (en bas).

calculer

(2.21)
$$\sigma_Z^2 = \sum_{k=0}^{M-1} \left(\int_{x_k}^{x_{k+1}} (x - y_k)^2 \rho_X(x) dx \right) + \int_{-\infty}^{x_0} (x - y_-)^2 \rho_X(x) dx + \int_{x_M}^{\infty} (x - y_+)^2 \rho_X(x) dx \ ,$$

qu'il s'agit de minimiser par rapport aux variables x_k, y_k et y_+, y_-. Il s'agit d'un problème classique de minimisation de forme quadratique, dont la solution est fournie par les équations d'Euler. En égalant à zéro la dérivée par rapport à x_k, on obtient facilement les expressions suivantes

$$x_k = \frac{1}{2}(y_k + y_{k-1}) \ , \quad k = 1, \dots M-1 \ , \tag{2.22}$$

$$x_0 = \frac{1}{2}(y_0 + y_-) \tag{2.23}$$

$$x_M = \frac{1}{2}(y_{M-1} + y_+) \tag{2.24}$$

De même, en égalant à zéro les dérivées par rapport aux y_k, à y_- et y_+, on obtient

$$y_k = \frac{\int_{x_k}^{x_{k+1}} x\rho_X(x)\,dx}{\int_{x_k}^{x_{k+1}} \rho_X(x)\,dx} , \tag{2.25}$$

$$y_- = \frac{\int_{-\infty}^{x_0} x\rho_X(x)\,dx}{\int_{-\infty}^{x_0} \rho_X(x)\,dx} , \tag{2.26}$$

$$y_+ = \frac{\int_{x_M}^{\infty} x\rho_X(x)\,dx}{\int_{X_M}^{\infty} \rho_X(x)\,dx} . \tag{2.27}$$

On obtient facilement des expressions similaires dans des situations où ρ_X est bornée. Ceci conduit naturellement à un algorithme itératif, dans lequel les variables x et y sont mises à jour récursivement. Naturellement, comme il s'agit d'un algorithme de type "algorithme de descente", rien ne garantit qu'il converge obligatoirement vers le minimum *global* de la distorsion. Lorsque tel n'est pas le cas (ce qui est en fait la situation la plus générale), on doit recourir à des méthodes plus sophistiquées.

2. Quantification vectorielle

Le défaut essentiel du codeur PCM est son incapacité à prendre en compte les corrélations existant dans le signal codé. La quantification utilisée dans le PCM traite en effet chaque échantillon individuellement, et ne tient aucun compte de la "cohérence" du signal. Pour tenir compte de celle-ci, il faudrait a priori effectuer une quantification non plus sur des échantillons individuels, mais sur des familles, ou vecteurs, d'échantillons. C'est le principe de la quantification vectorielle. Cependant, alors que l'idée de la quantification vectorielle est somme toute assez simple, sa mise en œuvre effective peut être extrêmement complexe.

Considérons un signal $\{x_0, x_1, \ldots\}$, et commençons par le "découper" en segments de longueur fixée N. Chaque segment représente donc un vecteur $X \in \mathbb{R}^N$, et on se pose donc le problème de construire un quantificateur :

$$Q : \mathbb{R}^N \to E_M = \{y_0, \ldots y_{M-1}\} ,$$

en essayant de minimiser l'erreur de quantification

$$D = \mathbb{E}\left\{|X - Q(X)|^2\right\} ,$$

où on a noté $|X|$ la norme Euclidienne de $X \in \mathbb{R}^N$.

Définition 2.3. *L'ensemble des "vecteurs quantifiés" est appelé le* dictionnaire *du quantificateur (*codebook *en anglais).*

Le problème est le même que celui que nous avons vu dans le cas de la recherche du quantificateur scalaire optimal, mais la situation est cette fois bien plus complexe, à cause de la dimension supérieure dans laquelle on se place. En effet, dans le cas d'un quantificateur scalaire, le problème est essentiellement de "découper" un sous-domaine de l'axe réel, ou l'axe réel lui même, en un nombre fini de boîtes de quantification (c'est à dire d'intervalles). En dimension supérieure, il s'agit maintenant d'effectuer une partition d'une partie de $\mathbb{R}^N$ en sous-domaines. On doit pour ce faire effectuer de multiples choix, notamment en ce qui concerne la forme des frontières : dans le cas de frontières planes, on parlera de quantificateur régulier, dans le cas contraire de quantificateur irrégulier (voir la Fig. 7. La problèmatique de

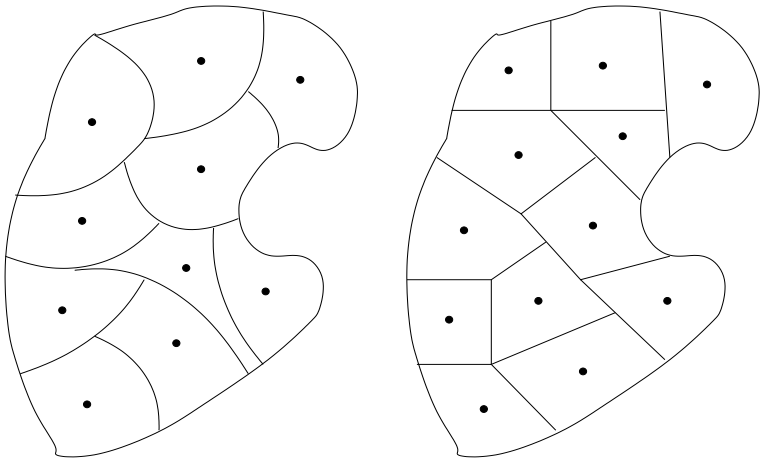

FIG. 7. Quantification vectorielle : deux partitions d'un domaine de $\mathbb{R}^2$: quantifications régulière (à droite) et irrégulière (à gauche) : boîtes de quantification et leurs centroïdes.

la quantification vectorielle présente des similarités certaines avec la problématique du "clustering".

Il existe de multiples façons différentes de construire un quantificateur vectoriel. On peut par exemple se référer à [**5**] pour une description relativement complète de l'état de l'art. Le problème de la recherche du quantificateur vectoriel optimal peut se formuler suivant les lignes ébauchées dans la section 1.4 : à partir du moment où on s'est donné le nombre de sous-domaines recherchés pour la partition, il reste à optimiser la distorsion globale

$$D = \sum_{k=0}^{M-1} \int_{\Omega_k} (x - y_k)^2 \rho_X(x)\, dx \ ,$$

où on a noté $\Omega_0, \ldots \Omega_{M-1}$ les sous-domaines considérés, et $y_0, \ldots y_{M-1}$ les valeurs de quantification correspondantes. L'optimisation par rapport à y_k reste assez simple, et fournit la *condition de centroïde*

$$y_k = \frac{\int_{S_k} x \rho_X(x)\, dx}{\int_{S_k} \rho_X(x)\, dx} \ ,$$

mais l'optimisation par rapport aux sous-domaines S_k, généralement effectuée numériquement, peut s'avérer extrêmement complexe suivant les hypothèses que l'on fait sur la forme des sous-domaines.

3. PCM différentiel (DPCM)

3.1. Principes généraux. Le DPCM (*differential PCM*) a pour but d'exploiter les redondances souvent présentes dans un signal pour en améliorer le codage, via un schéma plus simple que la quantification vectorielle. L'idée essentielle est que si un signal présente une certaine redondance, il doit être prédictible dans un certain sens, et à un certain niveau de précision. L'idée est d'introduire un modèle de signal, et de ne coder que les écarts au modèle. On peut ainsi espérer dimiunuer le domaine de valeurs (en d'autre termes, la variance) du signal avant quantification,

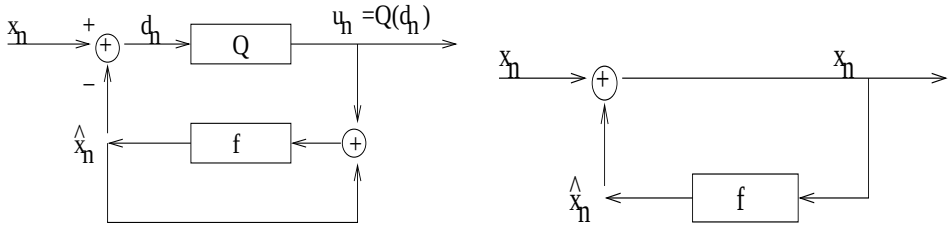

FIG. 8. Le codeur DPCM : décomposition et reconstruction.

et par là même diminuer le bruit de quantification (à nombre de bits par échantillon fixé).

Prenons l'exemple le plus simple, celui des signaux numériques (une version analogique peut être développée aussi), plus simple à formaliser. Soit donc $x = \{x_n, n \in \mathbb{Z}\}$ un signal numérique, et supposons que l'on puisse écrire un modèle de la forme

$$x_n = f(x_{n-1}, x_{n-2}, x_{n-3}, \dots x_{n-N}) + w_n \ , \tag{2.28}$$

où f est une fonction de prédiction, et w est l'erreur de prédiction. Il suffirait alors de coder les w_n (et les N premières valeurs non nulles de x_n), via les formules :

$$\begin{aligned} \hat{x}_n &= f(x_{n-1}, x_{n-2}, x_{n-3}, \dots x_{n-N}) \\ w_n &= x_n - \hat{x}_n \ . \end{aligned}$$

En ce qui concerne le décodage, connaissant un coefficient w_n et les valeurs précédentes $x_{n-1}, \dots x_{n-N}$, on reconstitue tout d'abord la prédiction $\hat{x}_n$ puis la vraie valeur

$$x_n = w_n + f(x_{n-1}, x_{n-2}, \dots x_{n-N}) \ .$$

En pratique, ce type d'approche présente le risque d'induire une propagation d'erreurs : en effet, entre le codage et le décodage, les coefficients w_n sont quantifiés, de sorte que les coefficients x_n sont reconstitués avec une erreur. Comme ces coefficients sont de plus utilisés pour calculer la prédiction des suivants (via la fonction de prédiction f), les erreurs introduites par la quantification se propagent bel et bien. On dit qu'un tel schéma n'est pas stable.

Pour pallier cet inconvénient, il est plus efficace d'utiliser pour la prédiction les valeurs $\overline{x}_n$ qui seront disponibles à partir du signal décodé :

$$\hat{x}_n = f(\overline{x}_{n-1}, \overline{x}_{n-2}, \dots \overline{x}_{n-N}) \ . \tag{2.29}$$

On note alors

$$d_n = x_n - \hat{x}_n \tag{2.30}$$

l'erreur de prédiction, et

$$u_n = Q(d_n) \tag{2.31}$$

l'erreur de prédiction quantifiée. La valeur $\overline{x}_n$ est quant à elle donnée par

$$\overline{x}_n = \hat{x}_n + u_n \ , \tag{2.32}$$

et est également la valeur obtenue au décodage.

L'erreur commise au décodage est de la forme

$$z_n = \overline{x}_n - x_n = Q(d_n) - d_n \ , \tag{2.33}$$

c'est à dire coïncide avec l'erreur de quantification. Par conséquent, il n'y a effectivement pas de propagation d'erreur dans un tel schéma.

Comment évaluer l'amélioration apportée par le DPCM par rapport au PCM ? Si on introduit le *gain de prédiction*

$$G_p = \frac{\sigma_d^2}{\sigma_x^2} \ , \tag{2.34}$$

qui évalue la diminution relative de variance du signal apportée par la prédiction, on obtient

$$SNR_{DPCM} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_z^2} = 10 \log_{10} \frac{\sigma_d^2}{\sigma_z^2} - 10 \log_{10} G_p \ .$$

Or le premier terme n'est autre que la valeur de rapport signal à bruit que l'on aurait obtenue par codage PCM : on aboutit donc à

$$SNR_{DPCM} \approx SNR_{PCM} - 10 \log_{10} G_p \ . \tag{2.35}$$

On obtient bien ainsi l'effet recherché : une diminution de la variance du signal à quantifier, et donc une amélioration du rapport signal à bruit de quantification. A titre d'exemple, dans le cas du signal de parole, on peut ainsi obtenir une amélioration de l'ordre de $8dB$ en utilisant une prédiction d'ordre $K = 3$.

Reste à comprendre comment obtenir la fonction de prédiction f.

3.2. Prédiction optimale, prédiction linéaire. Il est intéressant de formuler le problème dans un cadre plus large, qui permettra des généralisations à d'autres problèmes.

On considère deux vecteurs aléatoires $\underline{X} = (X_1, \dots X_K)^t$ et $\underline{Y} = (Y_1, \dots Y_N)^t$ à valeurs dans $\mathbb{R}^K$ et $\mathbb{R}^N$ respectivement (considérés comme des "vecteurs colonne"), et on se pose le problème de prédire les valeurs de $\underline{Y}$ connaissant les valeurs de $\underline{X}$. Plus précisément, on recherche le prédicteur $\underline{\hat{Y}}$ de $\underline{Y}$ qui est optimal dans le sens suivant : il minimise l'erreur en moyenne quadratique

$$\mathbb{E} \left\{ \| \underline{\hat{Y}} - \underline{Y} \|^2 \right\} \ .$$

Ce problème, formulé comme un problème d'optimisation, admet la solution classique suivante :

THÉORÈME 2.1. *Etant donnés deux vecteurs aléatoires du second ordre $\underline{X}$ et $\underline{Y}$ sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, le prédicteur optimal $\underline{\hat{Y}}^*$ au sens de la moyenne quadratique de $\underline{Y}$ connaissant $\underline{X}$ est donné par*

$$\underline{\hat{Y}}^* = \mathbb{E} \{ \underline{Y} | \underline{X} \} \ . \tag{2.36}$$

Ce résultat est cependant difficile à exploiter dans un contexte de traitement du signal, car le calcul de l'espérance conditionnelle de $\underline{Y}$ demande de connaitre avec précision la distribution conditionnelle de $\underline{Y}$ sachant $\underline{X}$, ce qui est très difficile. On doit donc se limiter à une sous-classe de prédicteurs, que l'on appelle les prédicteurs linéaires.

Ces derniers sont de la forme

$$\underline{\hat{Y}} = \underline{\underline{A}} \, \underline{X} \ , \tag{2.37}$$

et trouver le meilleur prédicteur linéaire au sens des moindres carrés revient à rechercher la matrice $N \times K$ $\underline{\underline{A}}$ qui minimise l'erreur en moyenne quadratique. En notant génériquement $A_{k\ell}$ les éléments de la matrice $\underline{\underline{A}}$, l'équation normale

$$\frac{\partial}{\partial A_{k\ell}} \mathbb{E}\left\{\|AX - Y\|^2\right\} = 0$$

conduit à l'équation

$$\sum_{\ell'=1}^{K} A_{k\ell'} (\underline{\underline{R}}_X)_{\ell\ell'} = \mathbb{E}\left\{Y_k X_\ell\right\} , \tag{2.38}$$

où on a introduit la matrice de corrélation $\underline{\underline{R}}_X$ de $\underline{X}$, définie par

$$(\underline{\underline{R}}_X)_{k\ell} = \mathbb{E}\left\{X_k X_\ell\right\}$$

En utilisant une notation matricielle, on aboutit ainsi au résultat suivant

THÉORÈME 2.2. *Etant donnés deux vecteurs aléatoires du second ordre $\underline{X}$ et $\underline{Y}$ sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, le prédicteur linéaire optimal $\hat{\underline{Y}}^*$ au sens de la moyenne quadratique de $\underline{Y}$ connaissant $\underline{X}$ est donné par la matrice A solution de l'équation*

$$\underline{\underline{A}}\,\underline{\underline{R}}_X = \mathbb{E}\left\{\underline{Y}\,\underline{X}^t\right\} . \tag{2.39}$$

Si de plus la matrice de covariance $\underline{\underline{R}}_X$ de $\underline{X}$ est inversible, on a la solution

$$\underline{\underline{A}} = \mathbb{E}\left\{\underline{Y}\,\underline{X}^t\right\} {\underline{\underline{R}}_X}^{-1} . \tag{2.40}$$

Cette dernière équation est généralement résolue numériquement.

3.3. Application à la prédiction de signaux. Revenons au cas de la prédiction de signaux unidimensionnels. Dans ce cas, il est nécessaire de supposer (au moins dans un premier temps) que le prédicteur reste valable pour tout le signal. Ceci est assuré dès que l'on suppose le signal stationnaire en moyenne d'ordre deux. On suppose aussi pour simplifier que le signal est centré. Alors, dans les notations précédentes, on a $\underline{Y} = X_n$ (donc $N = 1$), et $\underline{X} = (X_{n-1}, \dots X_{n-K})^t$. $\underline{\underline{A}}$ est donc une matrice $1 \times K$, c'est à dire un vecteur ligne, que l'on notera $\underline{h}^t = (h_1, \dots h_K)$. Le prédicteur linéaire est de la forme

$$\hat{X}_n = \sum_{k=1}^{K} h_k X_{n-k} . \tag{2.41}$$

Les coefficients h_k du prédicteur linéaire de rang K optimal sont solutions de l'équation

$$\underline{\underline{R}}_X\, \underline{h} = \underline{r} , \tag{2.42}$$

où on a noté $\underline{r}$ le vecteur de coordonnées

$$r_k = R_X(k) = \mathbb{E}\left\{X_n X_{n-k}\right\} .$$

Cette équation est appelée *Equation de Yulle-Baxter*, ou *équation de Wiener-Hopf*.

Dans les cas les plus simples, la solution est explicite, et fait appel aux coefficients de corrélation

$$\rho_k = \frac{\mathbb{E}\left\{X_n X_{n-k}\right\}}{\mathbb{E}\left\{X_n^2\right\}} . \tag{2.43}$$

3.3.1. *Le cas* $K = 1$. Dans ce cas, $\underline{X}$ est lui aussi unidimensionnel, et on ne manipule que des nombres. $\underline{\underline{R}}_X = \sigma_X^2$, et $\underline{h}$ est un nombre

$$h = h_1 = \frac{\mathbb{E}\{X_n X_{n-1}\}}{R_X} = \rho_1 \ .$$

3.3.2. *Le cas* $K = 2$. Dans ce cas, $\underline{\underline{R}}_X$ est une matrice 2×2, de la forme

$$\underline{\underline{R}}_X = \sigma_X^2 \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \ .$$

De plus, on a

$$\mathbb{E}\left\{\underline{Y} X^t\right\} = \sigma_X^2(\rho_1, \rho_2) \ ; \quad \underline{\underline{R}}_X{}^{-1} = \frac{1}{\sigma_X^2} \frac{1}{1-\rho_1^2} \begin{pmatrix} 1 & -\rho_1 \\ -\rho_1 & 1 \end{pmatrix} \ ,$$

d'où on déduit

$$\underline{h}^t = \frac{1}{1-\rho_1^2} \left(\rho_1 - \rho_1\rho_2^2, \rho_2 - \rho_1^2\right) \ .$$

3.3.3. *Le cas général.* Dans les cas où K est plus grand, il est nécessaire de revenir à l'équation de Yulle-Baxter (2.42), qui doit être résolue numériquement. Il existe des algorithmes classiques pour résoudre ce problème, comme par exemple l'algorithme de Levinson-Durbin. On pourra se rapporter à [**7**] pour plus de détails sur ces algorithmes. Il existe de nombreuses implémentations disponibles, dans des logiciels libres comme dans des produits commerciaux.

3.4. Quelques remarques. Il est utile de conclure cette section par quelques remarques "pratiques".

(1) Les systèmes de type DPCM sont très utiles dans certains cas bien précis, notamment pour le codage du signal de parole. Il a en effet été montré que l'introduction de prédiction linéaire se traduit par un gain substantiel en termes de SNR. Ceci est particulièrement vrai pour des prédicteurs d'ordre faible ($K = 1, 2, 3$) pour lesquels la prédiction peut faire gagner jusqu'à 8 ou 10 decibels. Des prédicteurs d'ordre plus élevé continuent d'améliorer encore les performances, mais le gain devient moins spectaculaire. De plus, augmenter l'ordre de la prédiction améliore certes le SNR, mais augmente aussi la complexité du codeur (et dans une moindre mesure, du décodeur). Il y a donc (comme d'habitude) un compromis à trouver, en fonction de l'application visée.

(2) Il est clair que les performances du prédicteur seront d'autant meilleures que le signal présentera de fortes redondances. Donc, des signaux échantillonnés à une cadence plus élevée (qui sont donc a priori plus redondants) sont plus facilement prédictibles, et les performances du DPCM sont d'autant meilleures. Ceci suggère donc qu'augmenter la fréquence d'échantillonnage doit permettre de diminuer le nombre de bits alloué par coefficient. C'est effectivement ce qui est observé en pratique, et qui a conduit au dernier standard SACD (Super Audio CD) de Sony, qui est basé sur un codage DPCM à **1 bit par coefficient**, avec des taux d'échantillonnage très élevés. Les performances du SACD semblent comparables à celles obtenues avec les codeurs basés sur les idées plus "modernes" de codage par transformation que nous allons voir dans le chapitre suivant.

CHAPITRE 3

Représentation des signaux ; Codage par transformation

On a vu dans le chapitre précédent deux premiers exemples de systèmes de codage de signaux, à savoir les systèmes PCM et DPCM. Le premier d'entre eux est très simple, et présente en particulier le défaut de ne pas tenir compte de la redondance présente dans les signaux. Le second (DPCM) prend en compte cette redondance via l'incorporation d'une procédure de prédiction ; on a vu que ceci permet de diminuer le rapport signal à bruit de quantification, et donc d'obtenir une meilleure compression avec un taux R réduit.

On va voir dans ce chapitre une alternative qui perment, non pas de réduire le nombre de bits par échantillon R, mais plutôt le nombre d'échantillons significatifs, toujours en tenant compte des redondances. Le principe est de représenter un signal, non plus par des échantillons, mais par les coefficients de son développement par rapport à une base bien choisie. Le but est d'obtenir une famille de coefficients telle que peu d'entre eux soient numériquement significatifs, de sorte que le nombre de coefficients à coder soit faible.

On a déjà vu un premier exemple de représentation "analytique" des signaux, avec le théorème d'échantillonnage : un signal à bande limitée peut être reconstruit à partir de ses échantillons, pour peu que ceux-ci soient choisis avec une fréquence suffisamment élevée. Nous avons vu également que dans ce cas, le développement correspondant du signal n'est autre qu'un développement par rapport à une base particulière, à savoir la base des sinus cardinaux. Dans ce chapitre, on approfondit cette notion, et on examine quelques alternatives pour ce choix de base.

Il est utile à ce point de rappeler les propriétés essentielles des espaces de Hilbert et des bases orthonormées. Etant donné un espace de Hilbert, muni d'une norme $\| \ \|$ et d'un produit scalaire $\langle ., . \rangle$, une famille de vecteurs $\{e_\lambda, \lambda \in \Lambda\}$ (où Λ est un index fini ou infini dénombrable) est dite orthonormée si pour tous $\lambda, \lambda' \in \Lambda$, on a

$$\langle e_\lambda, e_{\lambda'} \rangle = \delta_{\lambda\lambda'} \ .$$

Le résultat suivant donne des conditions nécessaires et suffisantes pour qu'une famille orthonormée soit une base orthonormée.

THÉORÈME 3.1. *Soit $\{e_\lambda, \lambda \in \Lambda\}$ un système orthonormé dans un espace de Hilbert H. Les assertions suivantes sont équivalentes :*

(1) *$\{e_\lambda, \lambda \in \Lambda\}$ est une base orthonormée de H.*

(2) *La famille $\{e_\lambda, \lambda \in \Lambda\}$ est complète dans H.*

(3) *Pour tout $x \in H$, on a la formule de Parseval :*

$$||x||^2 = \sum_{\lambda \in \Lambda} |\langle x, e_\lambda \rangle|^2 \ . \tag{3.1}$$

(4) *Pour tous $x, y \in H$, on a*

$$\langle x, y \rangle = \sum_{\lambda \in \Lambda} \langle x, e_\lambda \rangle \langle e_\lambda, y \rangle \ . \tag{3.2}$$

L'objet central que nous considèrerons dans ce cadre est l'application "coefficients"

$$\mathcal{L} : x \in H \to c \in \ell^2(\Lambda) \ : \quad c(\lambda) = \langle x, e_\lambda \rangle \ .$$

Une relecture du résultat ci-dessus indique que $\mathcal{L}$ est une isométrie bijective entre H et $\ell^2(\mathbb{Z})$.

On va voir ci-dessous un certain nombre de choix possibles de bases orthonormées dans un contexte de traitement du signal. Ces bases seront étudiées sous l'angle de leur utilisation potentielle pour la compression des signaux. On se placera toujours dans le cadre des espaces de signaux d'énergie finie, que ce soit dans le cas de signaux analogiques (espaces de fonctions de module carré intégrable) ou de signaux numériques (espaces de suites de module carré sommable). Il faut cependant remarquer que les espaces de type L^2 ne sont pas toujours les mieux adaptés pour la modélisation de classes de signaux (déterministes). On a maintenant de plus en plus recours à des modélisations faisant intervenir des espaces (de Banach) plus sophistiqués, comme les espaces de Besov, les espaces de modulation ou les espaces de fonctions à variation bornée. Le problème de la construction de bases dans ces espaces devient alors bien plus complexe, et on ne le considèrera pas ici.

A coté des bases dans les espaces de Hilbert séparables, on évoquera également une alternative, à savoir le choix de repères de préférence aux bases. La différence entre une base et un repère est qu'un repère est généralement une famille $\{f_\lambda, \lambda \in \Lambda\}$ *surcomplète* d'élements d'un espace de Hilbert (ou de Banach). Dans ce cas de figure, l'application coefficients

$$\mathcal{L} : x \in H \to c \in \ell^2(\Lambda) \ : \quad c(\lambda) = \langle x, f_\lambda \rangle$$

n'est plus surjective (c'est à dire, il existe une certaine redondance entre les coefficients $c(\lambda)$), mais on peut quand même lui associer un "pseudo-inverse". Dans un contexte de codage par transformation, les repères prennent tout leur intérêt lorsque l'on s'attend à ce que la transmission des coefficients soit "bruitée". Il est en effet possible de montrer que la présence de redondance améliore la stabilité de la reconstruction en présence de bruit.

1. Bases classiques

1.1. Bases trigonométriques pour les signaux analogiques. On commence par les signaux analogiques, définis sur un intervalle, disons $[0, a]$. Le premier résultat classique est donné par les séries de Fourier. Dans ce qui suit, on notera $L^2_p([a, b])$ l'espace des fonctions périodiques[1] de période $b - a$, de carré intégrable

[1]Dans la suite, on identifiera généralement $L^2_p([a, b])$ à $L^2([a, b])$.

dans l'intervalle $[a,b]$ (et donc dans tout intervalle de longueur $b-a$). $L^2_p([a,b])$ est un espace de Hilbert, grâce au produit scalaire défini par

$$\langle f,g\rangle = \int_a^b f(t)\overline{g}(t)\,dt\ , \quad f,g\in L^2_p([a,b])\ . \tag{3.3}$$

Rappelons que d'après le Théorème 1.5, si on introduit les fonctions définies en (1.32)

$$e_n(t) = \sqrt{\frac{1}{b-a}}\,\exp\left\{2i\pi\frac{nt}{b-a}\right\}\ ,$$

la collection des fonctions $\{e_n, n\in\mathbb{Z}\}$ est une base orthonormée de $L^2_p([a,b])$. Pour toute fonction $f\in L^2_p([a,b])$, on a

$$f = \sum c'_n e_n\ , \qquad \text{où} \tag{3.4}$$

$$c'_n = \langle f,e_n\rangle = \sqrt{\frac{1}{b-a}}\int_a^b f(t)e^{-2i\pi\frac{nt}{b-a}}\,dt\ . \tag{3.5}$$

Plus conventionnellement, on utilise les coefficients de Fourier

$$c_n = c_n(f) = \frac{1}{b-a}\int_a^b f(t)e^{-2i\pi\frac{nt}{b-a}}\,dt\ , \tag{3.6}$$

ce qui donne

$$f(t) = \sum_n c_n(f)e^{2i\pi\frac{nt}{b-a}}\ , \tag{3.7}$$

et

$$\sum_n |c_n(f)|^2 = \frac{1}{b-a}\|f\|^2\ . \tag{3.8}$$

Remarque 3.1. *Une conséquence directe de ce résultat est la réinterprétation suivante du théorème d'échantillonnage. Partant de l'espace de Paley-Wiener PW_{ν_0}, il est bien clair que la transformation de Fourier intégrale $\mathcal{F}$ constitue une isométrie entre PW_{ν_0} et $L^2([-\nu_0,\nu_0])$. Or, la famille des fonctions e_n formant une base orthonormée de $L^2([-\nu_0,\nu_0])$, on en déduit que la famille des fonctions φ_k définies par*

$$\varphi_k(t) = \int_{-\infty}^{\infty} e_k(\nu)e^{2i\pi k\nu/2\nu_0}\,d\nu = \frac{1}{\pi\sqrt{2\nu_0}}\frac{\sin\left(2\pi k\nu_0(t-k/2\nu_0)\right)}{t-k/2\nu_0}$$

est une base orthonormée de PW_{ν_0}. Ceci permet donc de réinterpréter le théorème 1.7 (dans le cas critique $\eta = 2\nu_0$) en termes de base orthonormée : les échantillons $f_k = f(k/2\nu_0)$ ne sont autres que les coefficients du développement de f sur la base des φ_k :

$$\langle f,\varphi_k\rangle = \frac{1}{2\nu_0}f\left(\frac{-k}{2\nu_0}\right)\ .$$

Il est bien connu que les séries de Fourier sont également utilisées pour décrire les fonctions définies sur un intervalle. Comme toute fonction à support dans un intervalle, disons $[a,b]$ peut être vue comme le produit d'une fonction périodique par la fonction caractéristique $\chi_{[a,b]}$ de l'intervalle en question, on en déduit directement que les fonctions

$$t\to e_n(t)\chi_{[a,b]}(t)$$

forment une base orthonormée de $L^2([a,b])$. Par contre, on peut se poser la question de l'adéquation de telles bases aux problèmes de compression des signaux. Le résultat suivant nous donne quelques pistes.

LEMME 3.1. *Soit f une fonction périodique, k fois continûment différentiable. Alors, il existe une constante $C > 0$ telle que*

$$|c_n(f)| \leq \frac{C}{|n|^k} \ . \tag{3.9}$$

Preuve : Prenons pour simplifier $a = -b = -\pi$. Il est clair que f étant continue, elle appartient à $L^1([-\pi,\pi])$, de même que toutes ses dérivées jusqu'à l'ordre k, dont les coefficients de Fourier sont donc bien définis. On a, par intégrations par parties successives,

$$c_n(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-int}dt = \left(\frac{1}{in}\right)^k c_n(f^{(k)}) \ ,$$

et le fait que $c_n(f)$ soit borné montre le résultat. ♠

Ce résultat est encourageant, car il montre que plus le signal étudié est régulier, plus la décroissance de ses coefficients de Fourier est rapide, et plus l'erreur commise en remplaçant la série de Fourier par une somme finie sera faible.

Ceci étant, l'opération visant à ne considérer que la restriction d'un signal à un intervalle $[a,b]$ donné n'est pas innocente. En effet, si $f \in C^\infty$ est telle que $f(b) \neq f(a)$, alors le lemme précédent ne nous donne aucune indication intéressante quant à la décroissance des coefficients de Fourier. On peut même montrer que dans ce cas, ils ne peuvent pas décroître plus vite que $C/|n|$...

Tout ceci suggère d'essayer de modifier la base trigonométrique pour limiter ce problème. Une solution est apportée par le résultat suivant, qui est basée non plus sur une périodisation de f, mais tout d'abord sur une symétrisation par rapport à l'origine, suivie d'une périodisation de période $2a$. On aboutit alors à la collection de fonctions

$$u_0(t) = \sqrt{\frac{1}{b-a}}\chi_{[a,b]}(t) \ , \tag{3.10}$$

$$u_n(t) = \sqrt{\frac{2}{b-a}}\cos\left(\pi n\frac{t-a}{b-a}\right)\chi_{[a,b]}(t) \ , \tag{3.11}$$

et on a le résultat suivant :

COROLLAIRE 3.1. *La collection des fonctions $\{u_n, n = 0,1,2,\ldots\}$ est une base orthonormée de $L^2([a,b])$. Pour toute fonction $f \in L^2([a,b])$, on a*

$$f = \sum_{n=0}^{\infty} d'_n u_n \ , \qquad \text{où} \tag{3.12}$$

$$d'_0 = \langle f, u_0\rangle = \sqrt{\frac{1}{b-a}}\int_a^b f(t)\,dt \ , \tag{3.13}$$

$$d'_n = \langle f, u_n\rangle = \sqrt{\frac{2}{b-a}}\int_a^b f(t)\cos\left(\pi n\frac{t-b}{b-a}\right)dt \ . \tag{3.14}$$

On a de plus la relation de Parseval

$$\frac{1}{b-a}\sum_n |\langle f, u_n\rangle|^2 = \|f\|^2 \ . \tag{3.15}$$

Preuve : On note f_p la fonction définie sur $[-a,a]$ par $f_p(t) = f(t)$ si $t \geq 0$, et $f_p(t) = f(-t)$ si $t \leq 0$. Il suffit alors d'appliquer le premier théorème, qui donne

$$f_p(t) = \sum_{-\infty}^{\infty} c_n(f_p) e^{i\frac{nt}{2a}} \ ,$$

avec

$$c_n(f_p) = \frac{1}{2a} \int_{-a}^{a} f_p(t) e^{-i\frac{nt}{2a}}\, dt = c_{-n}(f_p) \ .$$

Il suffit alors de poser

$$d_n(f) = 2c_n(f_p) \ ,$$

de remarquer que d_n ne dépend que de f, et le tour est joué. Les coefficients d'_n s'obtiennent par une renormalisation appropriée. ♠

Le codeur d'images `JPEG` utilise des bases de type "bases trigonométriques locales" qui sont de simples extensions des bases précédentes : étant donné un intervalle, disons $[A,B]$, il suffit de le "découper en morceaux" et de considérer une base trigonométrique dans chacun des segments. Le choix de bases trigonométriques apparaitra plus naturel par la suite.

Corollaire 3.2 (la base JPEG). *Etant donné un intervalle $[A,B] \subset \mathbb{R}$, on considère une partition de celui-ci en sous-intervalles $[a_n, a_{n+1}]$, $n = 0, \dots N-1$, où $a_0 = A, a_N = B$ et $a_n < a_{n+1}$ pour tout n. On pose $\ell_k = a_{k+1} - a_k$. Alors la famille de fonctions $v_{k\nu}, k = 0, \dots N-1, \nu = 0, \dots \infty$ définies par*

$$v_{k\nu}(t) = \begin{cases} \sqrt{\frac{1}{\ell_k}} \chi_{[a_k, a_{k+1}]}(t) & \text{si } \nu = 0 \\ \sqrt{\frac{2}{\ell_k}} \chi_{[a_k, a_{k+1}]}(t) \cos\left(\nu\pi \frac{t-a_k}{\ell_k}\right) & \text{pour } \nu = 1, 2, \dots \end{cases} \tag{3.16}$$

est une base orthonormée de $L^2([A,B])$.

Remarque 3.2. Comme on le verra, il est possible de diminuer encore les problèmes aux bords en introduisant de nouvelles bases, obtenues en remplaçant les fonctions caractéristiques par des fenêtres plus régulières.

Nous sommes maintenant en position de décrire un premier prototype de codeur par transformation basé sur la DCT :

Codage par transformation DCT

(1) Transformation : à $f \in L^2([A,B])$ on associe la famille des coefficients $c_{k\nu} = \langle f, v_{k\nu} \rangle$, où $k = 0, \dots K-1$ et $\nu = 0, 1, \dots$. Cette famille caractérise f via

$$f = \sum_{k=0}^{K-1} \sum_{\nu=0}^{\infty} c_{k\nu} v_{k\nu} \ .$$

(2) Approximation : (comparer au filtrage passe-bas) : les coefficients $c_{k\nu}$ correspondant aux grandes valeurs de ν sont ''effacés''. Ceci produit l'approximation suivante de f :

$$f_N = \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} c_{k\nu} v_{k\nu} \ ,$$

et une erreur (distorsion fonctionnelle) estimée en norme L^2 comme

$$\|f - f_N\|^2 = \sum_{k=0}^{K-1} \sum_{\nu=N}^{\infty} |c_{k\nu}|^2 \ .$$

(3) Quantification : Les coefficients $c_{k\nu}$ restants sont quantifiés, ce qui produit l'approximation

$$\tilde{f}_N = \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} Q(c_{k\nu}) v_{k\nu} \ ,$$

et une erreur (distorsion de quantification) estimée en norme L^2 comme

$$\|f_N - \tilde{f}_N\|^2 = \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} |c_{k\nu} - Q(c_{k\nu})|^2 \ .$$

Au final, la distorsion totale est estimée comme

$$\|f - \tilde{f}_N\|^2 = \sum_{k=0}^{K-1} \sum_{\nu=N}^{\infty} |c_{k\nu}|^2 + \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} |c_{k\nu} - Q(c_{k\nu})|^2 \ .$$

Cette dernière équation provient du fait que

$$\langle f - f_N, f_N - \tilde{f}_N \rangle = 0 \ ,$$

ces deux fonctions appartenant à deux sous-espaces différents de l'espace de départ.

REMARQUE 3.3. *Comme on le verra par la suite, de nombreuses variantes sont possibles. Mentionnons toutefois la variante la plus immédiate, qui consiste à utiliser des quantificateurs Q_ν différents pour différentes fréquences. Ceci permet d'optimiser la distorsion globale.*

1.2. Bases trigonométriques pour les signaux numériques. En pratique, les signaux à représenter sont souvent déjà discrétisés, et il faut alors disposer de bases adaptées à la situation. La base discrète qui généralise au cas discret les bases trigonométriques est naturellement la base trigonométrique discrète. Par exemple, dans le cas de signaux de longueur M, on utilisera la TFF, et les suites

$$e_m(k) = \frac{1}{\sqrt{M}} \ e^{2i\pi \frac{km}{M}} \ . \tag{3.17}$$

Cependant, la discussion précédente concernant la périodisation implicite qui est effectuée quand on utilise une telle base est toujours d'actualité. Il est donc nécessaire de disposer d'une base généralisant la base de cosinus précédente. Les résultats suivants sont facilement vérifiés.

THÉORÈME 3.2 (Cosinus II). *La famille des vecteurs $\{e_k, k = 0, \dots M-1\}$ définis par*

$$e_0(m) = \sqrt{\frac{1}{M}} \tag{3.18}$$

$$e_k(m) = \sqrt{\frac{2}{M}} \cos\left(\frac{k\pi}{M}(m+1/2)\right) \ , \quad k = 1, \dots M-1 \tag{3.19}$$

est une base orthonormée de $\mathbb{C}^M$, appelée base de cosinus II.

Une variante est donnée par le résultat suivant

THÉORÈME 3.3 (Cosinus IV). *La famille des vecteurs* $\{f_k, k = 0, \dots M-1\}$ *définis par*

$$f_k(m) = \sqrt{\frac{2}{M}} \cos\Big(\frac{\pi}{M}(k+1/2)(m+1/2)\Big) \ , \quad k = 0, \dots M-1 \tag{3.20}$$

est une base orthonormée de $\mathbb{C}^M$*, appelée* base de cosinus IV.

REMARQUE 3.4. *Le prototype de codeur par transformation DCT que nous avons vu plus haut dans le cas analogique se transpose de façon quasiment immédiate au cas numérique. Il suffit pour cela de remplacer dans l'algorithme précédent les sommes infinies sur* ν *par des sommes finies sur* k *variant de 0 à* $M-1$*, qui sont ensuite tronquées à des suites de longueur* $N < M$*.*

Ces bases sont très utilisées en pratique. Par exemple, le codeur d'images JPEG est basé sur un premier découpage de l'image en blocs de 8 pixels sur 8 pixels. Puis, chaque bloc est décomposé sur une base de $\mathbb{C}^8 \times \mathbb{C}^8$, obtenue par produit tensoriel de deux bases de cosinus IV de $\mathbb{C}^8$. Les coefficients ainsi obtenus sont ensuite quantifiés puis codés.

1.3. Bases trigonométriques locales "adoucies". Nous reprenons ici les notations du COROLLAIRE 3.2. Le problème induit par la segmentation "brutale" précédente est qu'elle introduit potentiellement une singularité à chaque noeud a_k. On se pose donc le problème suivant : *peut-on remplacer les fonctions caractéristiques dans les bases trigonométriques locales par des fenêtres plus régulières ?*

Considérons donc, dans un intervalle $[A, B]$, une suite de nombres

$$A = a_0 < a_1 < a_2 < \cdots < a_K = B \ ,$$

et pour $k = 1, \dots K-2$ une série de nombres $\eta_k > 0$, tels que

$$a_k + \eta_k < a_{k+1} - \eta_{k+1} \ . \tag{3.21}$$

On considère également un ensemble de fonctions w_k telles que

$$\left\{ \begin{array}{rcl} w_k(t) & = & 1 \text{ si } t \in [a_k + \eta_k, a_{k+1} - \eta_{k+1}] \ , \\ w_k(t) & = & 0 \text{ si } t \notin [a_k - \eta_k, a_{k+1} + \eta_{k+1}] \ , \end{array} \right. \tag{3.22}$$

destinées à remplacer les fonctions caractéristiques des intervalles $[a_k, a_{k+1}]$ utilisées précédemment. Par convention, on prend $\eta_0 = \eta_{K-1} = 0$. Posons maintenant

$$u_{k\nu}(t) = \sqrt{\frac{2}{\ell_k}}\, w_k(t) \cos\left[\frac{\pi}{\ell_k}\left(\nu + \frac{1}{2}\right)(t - a_k)\right] \ , \tag{3.23}$$

où $\ell_k = a_{k+1} - a_k$. Remarquons que (3.21) s'écrit alors

$$\eta_k + \eta_{k+1} \leq \ell_k$$

On a alors

THÉORÈME 3.4. *Si les fenêtres* w_k *sont telles que* $\forall \tau$ *tel que* $0 \leq |\tau| \leq \eta_k$*,*

$$\left\{ \begin{array}{rcl} w_{k-1}(a_k + \tau) & = & w_k(a_k - \tau) \ , \\ w_k(a_k + \tau)^2 & + & w_{k-1}(a_k + \tau)^2 = 1 \ , \end{array} \right. \tag{3.24}$$

alors les fonctions $u_{k\nu}$ *,* $k, \nu \in \mathbb{Z}$ *forment une base orthonormée de* $L^2([A, B])$*.*

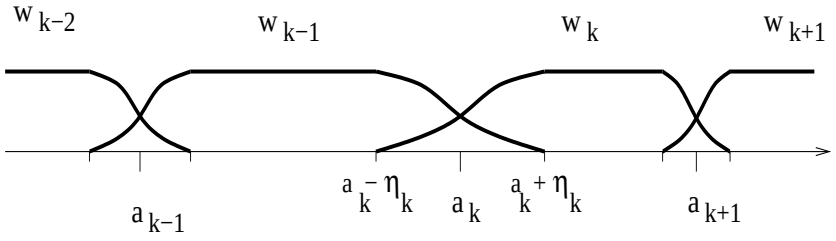

FIG. 1. Fenêtres adoucies

Un exemple de telles fenêtres adoucies se trouve en FIGURE 1.
Preuve : Considérons deux fonctions $u_{k\nu}$ et $u_{k'\nu'}$, et commençons par le premier cas simple $k' \neq k$. Si $|k'-k| \geq 2$, on a $\langle u_{k\nu}, u_{k'\nu'}\rangle = 0\ \forall k, \nu$ pour des raisons de support. Prenons donc $k' = k-1$

$$\begin{aligned}
\langle u_{k\nu}, u_{k-1\nu'}\rangle &\sim \int_{a_k-\eta_k}^{a_k+\eta_k} w_k(t) w_{k-1}(t) \cos\left[\frac{\pi\,(\nu+1/2)}{\ell_k}(t-a_k)\right] \\
&\qquad \cos\left[\frac{\pi\,(\nu'+1/2)}{\ell_{k-1}}(t-a_{k-1})\right] dt \\
&\sim \int_{-\eta_k}^{\eta_k} w_k(a_k+\tau) w_{k-1}(a_k+\tau) \cos\left[\frac{\pi\,(\nu+1/2)}{\ell_k}\tau\right] \\
&\qquad \cos\left[\frac{\pi\,(\nu'+1/2)}{\ell_{k-1}}(\tau+\ell_{k-1})\right] d\tau \\
&= 0
\end{aligned}$$

pour des raisons de symétrie (intégrand impair).

Le cas $k' = k$ se traite avec des arguments de trigonométrie élémentaire. Reste donc à montrer la complétude de la base. Soit donc $f(t) \in L^2(\mathbb{R})$, et considérons

$$F(t) = \sum_{k,\nu} \langle f, u_{k\nu}\rangle u_{k\nu}(t)\ .$$

Nous avons donc

$$\begin{aligned}
F(t) &= \sum_{k,\nu} \frac{1}{\ell_k} \int w_k(t) w_k(s) f(s) \left\{ \cos\left[\frac{\pi\,(\nu+1/2)}{\ell_k}(t-s)\right] \right. \\
&\qquad \left. + \cos\left[\frac{\pi\,(\nu+1/2)}{\ell_k}(t+s-2a_k)\right]\right\} ds\ .
\end{aligned} \tag{3.25}$$

Mais nous pouvons écrire, comme conséquence de la formule de Poisson

$$\sum_0^\infty \cos\left[\left(\nu+\frac{1}{2}\right)\alpha\right] = \pi e^{i\alpha/2} \sum_{k=-\infty}^{\infty} \delta(\alpha - 2\pi k)$$

Ceci nous donne

$$\begin{aligned} F(t) &= \sum_{k,n} \{w_k(t)w_k(2a_k+2n\ell_k-t)f(2a_k+2n\ell_k-t) \\ &\qquad +w_k(t+2n\ell_k)w_k(t)f(t+2n\ell_k)\}\, e^{in\pi} \\ &= \sum_{k} \big\{w_k(t)^2 f(t) + w_k(t)w_k(2a_k-t)f(2a_k-t) \\ &\qquad -w_k(t)w_k(2a_{k+1}-t)f(2a_{k+1}-t)\big\} \ . \end{aligned}$$

Reste maintenant à examiner cas par cas. Clairement, sur l'intervalle $[a_k+\eta_k, a_{k+1}-\eta_{k+1}]$, nous avons $F(t)=f(t)$. Supposons maintenant par exemple $t=a_k+\tau,\ \tau\le\eta_k$. On a alors

$$\begin{aligned} F(t) &= \left[w_k(t)^2+w_{k-1}(t)^2\right] f(t) \quad + \ w_k(a_k+\tau)w_k(a_k-\tau)f(a_k-\tau) \\ &\qquad -w_{k-1}(a_k+\tau)w_{k-1}(a_k-\tau)f(a_k-\tau) \\ &= f(t)\ . \end{aligned}$$

Ceci conclut la preuve du théorème. ♠

Notons que la différence de signe provient du terme $1/2$ présent dans la définition des fonctions $u_{k\nu}(t)$.

On en déduit de façon quasi-automatique la forme du codeur par transformation correspondant, simple transposition du codeur DCT précédent :

CODAGE PAR TRANSFORMATION MDCT

(1) Transformation : à $f\in L^2([A,B])$ on associe la famille des coefficients $c_{k\nu}=\langle f,u_{k\nu}\rangle$, où $k=0,\dots K-1$ et $\nu=0,1,\dots$. Cette famille caractérise f via

$$f=\sum_{k=0}^{K-1}\sum_{\nu=0}^{\infty} c_{k\nu}u_{k\nu}\ .$$

(2) Approximation : (comparer au filtrage passe-bas) : les coefficients $c_{k\nu}$ correspondant aux grandes valeurs de ν sont ''effacés''. Ceci produit l'approximation suivante de f :

$$f_N=\sum_{k=0}^{K-1}\sum_{\nu=0}^{N-1} c_{k\nu}u_{k\nu}\ ,$$

et la distorsion fonctionnelle estimée en norme L^2 comme

$$\|f-f_N\|^2=\sum_{k=0}^{K-1}\sum_{\nu=N}^{\infty}|c_{k\nu}|^2\ .$$

(3) Quantification : Les coefficients $c_{k\nu}$ restants sont quantifiés, ce qui produit l'approximation

$$\tilde f_N=\sum_{k=0}^{K-1}\sum_{\nu=0}^{N-1} Q(c_{k\nu})u_{k\nu}\ ,$$

et la distorsion de quantification estimée en norme L^2 comme

$$\|f_N - \tilde{f}_N\|^2 = \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} |c_{k\nu} - Q(c_{k\nu})|^2 \ .$$

La distorsion totale est estimée comme

$$\|f - \tilde{f}_N\|^2 = \sum_{k=0}^{K-1} \sum_{\nu=N}^{\infty} |c_{k\nu}|^2 + \sum_{k=0}^{K-1} \sum_{\nu=0}^{N-1} |c_{k\nu} - Q(c_{k\nu})|^2 \ .$$

1.4. Le cas multidimensionnel. Jusqu'à présent, on ne s'est intéressé qu'au cas des signaux unidimensionnels. Les bases trigonométriques s'étendent facilement au cas des signaux bidimensionnels. Pour prendre l'exemple le plus simple, étant donnée une fonction de deux variables f, de module carré intégrable sur le carré $[0,1] \times [0,1]$:

$$\int_0^1 \int_0^1 |f(x,y)|^2 \, dx \, dy < \infty \ ,$$

on lui associe ses coefficients de Fourier

$$c_{m,n}(f) = \int_0^1 \int_0^1 f(x,y) e^{-2i\pi(mx+ny)} \, dx \, dy \ .$$

Le développement de f en série de Fourier s'écrit sous la forme :

$$\lim_{M,N\to\infty} \|f - f_{M,N}\| = 0 \ ,$$

où on a posé

$$f_{M,N}(x,y) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} c_{m,n}(f) \, e^{2i\pi(mx+ny)} \ .$$

Toutes les variations autour de ce thème (bases de cosinus, bases trigonométriques adoucies,...) que nous avons vues dans le cas unidimensionnel peuvent être développées dans le cas bidimensionnel également.

L'application au codage des signaux analogiques bidimensionnels (images analogiques) est immédiate. Etant donnée une image, modélisée comme une fonction de deux variables, on calcule tout d'abord ses coefficients de Fourier $c_{m,n}(f)$ pour $|m| \leq M$, $|n| \leq N$, et ces coefficients de Fourier sont ensuite quantifiés et codés. Il y a donc deux sources d'erreur dans ce schéma de codage : une première erreur est commise en tronquant le développement en série de Fourier (on ne considère qu'un nombre fini de coefficients de Fourier), puis une deuxième est commise lors de la quantification.

En ce qui concerne le codage d'images déjà numérisées (ce qui est le cas le plus courant), on utilise les versions discrètes des bases trigonométriques. Le principe est strictement le même.

2. Ondelettes et codage en sous bandes

Les bases trigonométriques vues à la section précédente sont généralement bien adaptées à des signaux qui présentent des oscillations régulières, comme par exemple la parole, ou les signaux audio. Par contre, elles introduisent une échelle de référence dans le problème, qui est la taille des fenêtres utilisées. L'introduction de ces fenêtres peut se traduire par l'apparition d'effets de "bloc", comme on peut en voir sur les

images comprimées avec JPEG avec de forts taux de compression. Ces effets sont atténués lorsque l'on utilise des bases "adoucies" comme on vient de le voir, mais rarement supprimés. Le codage en sous bandes que nous allons maintenant voir (et son pendant, la transformation en ondelettes) constitue une alternative bien adaptée au codage des images, qui présente l'avantage de ne pas introduire d'effet de bloc.

2.1. Bases d'ondelettes et multitésolution. Plaçons nous tout d'abord dans le contexte des signaux analogique. Le résultat fondamental de la théorie est le suivant :

THÉORÈME 3.5 (Meyer, Mallat). (1) *Il existe une fonction $\psi \in L^2(\mathbb{R})$, appelée* ondelette mère *telle que si l'on introduit les fonctions (les ondelettes) ψ_{jk} définies par*

$$\psi_{jk}(t) = 2^{-j/2}\psi\big(2^{-j}t - k\big) \ , \quad j,k \in \mathbb{Z} \tag{3.26}$$

la famille $\{\psi_{jk}, j,k \in \mathbb{Z}\}$ est une base orthonormée de $L^2(\mathbb{R})$.

(2) *Par conséquent, pour tout $f \in L^2(\mathbb{R})$, il existe une unique décomposition*

$$f = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_k^j \psi_{jk} \ , \tag{3.27}$$

où l'égalité est à prendre au sens de la convergence dans $L^2(\mathbb{R})$, et où les coefficients d_k^j sont donnés par

$$d_k^j = \langle f, \psi_{jk}\rangle = 2^{-j/2} \int_{-\infty}^{\infty} f(t)\overline{\psi}\left(2^{-j}t - k\right)\, dt \tag{3.28}$$

La preuve complète de ce résultat est assez longue et complexe, et repose sur la théorie des *Analyses multirésolution*. Sans entrer dans les détails, on peut en tracer les grandes lignes, au moins dans le cas unidimensionnel

DÉFINITION 3.1. *Une analyse multirésolution de $L^2(\mathbb{R})$ consiste en une suite emboitée de sous espaces fermés $\mathcal{V}_j$ de $L^2(\mathbb{R})$*

$$\cdots \subset \mathcal{V}_1 \subset \mathcal{V}_0 \subset \mathcal{V}_{-1} \subset \cdots \ , \tag{3.29}$$

tels que

(1) $\overline{\bigcup_{j=-\infty}^{\infty} \mathcal{V}_j} = L^2(\mathbb{R})$ *et* $\bigcap_{j=-\infty}^{\infty} \mathcal{V}_j = \emptyset$.

(2) *Pour tout $f \in \mathcal{V}_0$ et $k \in \mathbb{Z}$, on a $f_k \in \mathcal{V}_0$ où on a posé $f_k(t) = f(t-k)$. Pour tout $f \in \mathcal{V}_j$ et $k \in \mathbb{Z}$, on a $g \in \mathcal{V}_{j+1}$, où on a posé $g(t) = f(t/2)$.*

(3) *Il existe une fonction $\phi \in \mathcal{V}_0$ telle que la collection $\{\phi_k, k \in \mathbb{Z}\}$ de ses translatées entières ($\phi_k(t) = \phi(t-k)$) soit une base orthonormée de $\mathcal{V}_0$.*

Les espaces $\mathcal{V}_j$ sont appelés "espaces d'approximation". Un corollaire immédiat de la définition est que si l'on pose

$$\phi_{jk}(t) = 2^{-j/2}\phi\big(2^{-j}t - k\big) \ , \quad j,k \in \mathbb{Z} \tag{3.30}$$

la famille $\{\phi_{jk}, k \in \mathbb{Z}\}$ est une base orthonormée de $\mathcal{V}_j$. Etant donnée une fonction $f \in L^2(\mathbb{R})$, sa projection orthogonale sur $\mathcal{V}_j$, de la forme

$$P_j f = \sum_k \langle f, \phi_{jk}\rangle\, \phi jk$$

représente une approximation, ou version "lissée" de f à l'échelle 2^j.

Il est aussi facile de montrer que la fonction ϕ (appelée "fonction d'échelle) possède des propriétés particulières :

PROPOSITION 3.1. *Etant donnée une analyse multirésolution, et une fonction d'échelle associée ϕ, cette dernière vérifie*

$$\sum_{k=-\infty}^{\infty} \left|\hat{\phi}(\nu+k)\right|^2 = 1 \ p.p. \ . \tag{3.31}$$

ϕ satisfait de plus une relation de raffinement *du type*

$$\phi(t) = \sqrt{2} \sum_k \overline{h}_k \phi(2t+k) \ , \tag{3.32}$$

pour une certaine suite $\{h_k, k \in \mathbb{Z}\} \in \ell^1(\mathbb{Z})$.

Cette dernière propriété est une conséquence immédiate de l'inclusion $\mathcal{V}_0 \subset \mathcal{V}_{-1}$.

Etant donnée une analyse multirésolution et une fonction d'échelle associée ϕ, on en déduit alors une base d'ondelettes via une procédure très simple. On introduit une seconde suite $\{g_k, k \in \mathbb{Z}\}$, définie précisément par la relation :

$$g_k = (-1)^k \overline{h}_{1-k} \ ,$$

et on peut alors introduire l'ondelette ψ, définie elle aussi par une relation de raffinement :

$$\psi(t) = \sqrt{2} \sum_k \overline{g}_k \phi(2t+k) \ , \tag{3.33}$$

et on montre que la collection des ondelettes $\{\psi_{jk}, j, k \in \mathbb{Z}\}$ est effectivement une base orthonormée de $L^2(\mathbb{R})$.

On verra plus loin le lien entre les bases d'ondelettes et le codage en sous bandes. Pour le moment, notons que l'analyse multirésolution détermine complètement les suites h et g. La réciproque est "à peu près" vraie : par transformation de Fourier intégrale, l'équation (3.32) devient

$$\hat{\phi}(\nu) = \frac{1}{\sqrt{2}} \overline{\hat{h}}\left(\frac{\nu}{2}\right) \hat{\phi}\left(\frac{\nu}{2}\right) = m_0\left(\frac{\nu}{2}\right) \hat{\phi}\left(\frac{\nu}{2}\right) \ ,$$

où on a posé

$$m_0(\nu) = \frac{1}{\sqrt{2}} \sum_k h_k e^{2i\pi\nu k} = \frac{1}{\sqrt{2}} \overline{\hat{h}}(\nu) \ .$$

En itérant cette équation, on obtient

$$\hat{\phi}(\nu) = m_0\left(\frac{\nu}{2}\right) \hat{\phi}\left(\frac{\nu}{2}\right) = m_0\left(\frac{\nu}{2}\right) m_0\left(\frac{\nu}{4}\right) \hat{\phi}\left(\frac{\nu}{4}\right) = \cdots = \prod_{j=1}^{\infty} m_0\left(2^{-j}\nu\right) \tag{3.34}$$

Ainsi, la fonction ϕ est complètement caractérisée par m_0, donc $\hat{h}$, et donc le filtre h, à condition bien sûr que le produit infini converge vers une fonction de $L^2(\mathbb{R})$. Il est possible d'obtenir des conditions suffisantes assurant la convergence de ce produit infini, en particulier dans le cas de filtres à réponse impulsionnelle finie. On en verra des exemples dans la section suivante.

De même, l'équation (3.33) devient

$$\hat{\psi}(\nu) = \frac{1}{\sqrt{2}} \overline{\hat{g}}\left(\frac{\nu}{2}\right) \hat{\phi}\left(\frac{\nu}{2}\right) = m_1\left(\frac{\nu}{2}\right) \hat{\phi}\left(\frac{\nu}{2}\right) \ ,$$

où on a posé

$$m_1(\nu) = \frac{1}{\sqrt{2}} \sum_k g_k e^{2i\pi\nu k} = \frac{1}{\sqrt{2}}\, \overline{\hat{g}}(\nu) \ .$$

Ainsi, l'ondelette ψ est elle aussi complètement caractérisée par le filtre h.

2.2. Codage en sous bandes. Le codage en sous bandes consiste essentiellement en une variation autour du thème de l'échantillonnage et du sous-échantillonnage. On se place dès le départ dans le cadre des signaux échantillonnés, c'est à dire dans $\ell^2(\mathbb{Z})$. On considère deux suites absolument sommables $\{h_n, n \in \mathbb{Z}\}$ et $\{g_n, n \in \mathbb{Z}\}$, similaires à la suite $\{h_n, n \in \mathbb{Z}\}$ précédente. Plus précisément, considérant les restrictions de leurs transformées de Fourier à $[-1/2, 1/2]$, on impose que celle de $\{h_n, n \in \mathbb{Z}\}$ soit concentrée au voisinage de $[-1/4, 1/4]$, et que celle de $\{g_n, n \in \mathbb{Z}\}$ soit concentrée au voisinage de $[-1/2, -1/4] \cup [1/4, 1/2]$. On va alors procéder de manière similaire au chapitre précédent : la convolution de $\{s_n, n \in \mathbb{Z}\}$ par $\{h_n, n \in \mathbb{Z}\}$ ou $\{g_n, n \in \mathbb{Z}\}$ produit une suite dont la bande a été réduite d'un facteur 2 (à une erreur provenant d'un mauvais échantillonnage près). On sous-échantillonne alors chacun de ces deux produits de convolution d'un facteur 2. Ce faisant, on commet dans les deux cas une erreur (aliasing), et on va rechercher dans quels cas ces erreurs peuvent se compenser.

Pour cela, considérons en détail les opérations effectuées lors du schéma décomposition-reconstruction, et ce sur une seule étape tout d'abord.

Sous-échantillonnage. Commençons par évaluer l'effet du sous-échantillonnage. Soit $s = \{s_n, n \in \mathbb{Z}\}$ une suite de carré sommable, et considérons la suite ρ définie par

$$\rho_n = \begin{cases} s_n & \text{si } n \text{ est pair} \\ 0 & \text{sinon.} \end{cases}$$

Un calcul immédiat montre que

$$\hat{\rho}(\nu) = \frac{1}{2}\left(\hat{s}(\nu) + \hat{s}(\nu + 1/2)\right) \ .$$

On introduit maintenant la nouvelle suite sous-échantillonnée σ, définie par

$$\sigma_n = \rho_{2n} = s_{2n} \ .$$

On a alors

$$\hat{\sigma}(\nu) = \hat{\rho}\left(\frac{\nu}{2}\right) = \frac{1}{2}\left(\hat{s}\left(\frac{\nu}{2}\right) + \hat{s}\left(\frac{\nu}{2} + 1/2\right)\right) \ . \tag{3.35}$$

Filtres conjugués en quadrature. On considère maintenant les deux filtres $h, g \in \ell^1(\mathbb{Z})$, et les opérateurs H et $G : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$, appelés *opérateurs de décomposition*, définis par

$$(Hf)_n = \textstyle\sum_k h_{2n-k} f_k = \displaystyle\sum_k h_k f_{2n-k} \ ; \tag{3.36}$$

$$(Gf)_n = \textstyle\sum_k g_{2n-k} f_k = \displaystyle\sum_k g_k f_{2n-k} \ . \tag{3.37}$$

Ces deux opérations sont essentiellement des convolutions, suivies d'un sous-échantillonnage. Par conséquent, il vient, en posant $s = Hf$ et $d = Gf$:

$$\hat{s}(\nu) = \frac{1}{2}\left[\hat{h}\left(\frac{\nu}{2}\right)\hat{f}\left(\frac{\nu}{2}\right) + \hat{h}\left(\frac{\nu}{2} + 1/2\right)\hat{f}\left(\frac{\nu}{2} + 1/2\right)\right] \tag{3.38}$$

$$\hat{d}(\nu) = \frac{1}{2}\left[\hat{g}\left(\frac{\nu}{2}\right)\hat{f}\left(\frac{\nu}{2}\right) + \hat{g}\left(\frac{\nu}{2} + 1/2\right)\hat{f}\left(\frac{\nu}{2} + 1/2\right)\right] \ . \tag{3.39}$$

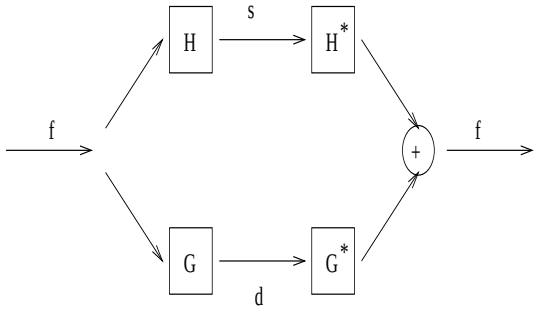

FIG. 2. Un schéma "analyse-synthèse".

On considère maintenant les opérateurs adjoints H^* et G^* de H et G, à savoir

$$(H^*f)_k = \sum_n \overline{h}_{2n-k} f_n \ ; \tag{3.40}$$

$$(G^*f)_k = \sum_n \overline{g}_{2n-k} f_n \ . \tag{3.41}$$

H^* et G^* sont appelés *opérateurs de reconstruction*, ou de *synthèse*. Un calcul direct montre que

$$\widehat{H^*f}(\nu) = \overline{\hat{h}}(\nu)\hat{f}(2\nu) \ ; \quad \widehat{G^*f}(\nu) = \overline{\hat{g}}(\nu)\hat{f}(2\nu) \ . \tag{3.42}$$

Considérons maintenant un système tel que décrit dans la figure 2. On impose que ce système soit à reconstruction parfaite, c'est à dire que

$$H^*H + G^*G = 1 \ . \tag{3.43}$$

En imposant une telle contrainte dans l'espace de Fourier, on aboutit à

$$\begin{aligned} \hat{f}(\nu) = \tfrac{1}{2} \Big[& \overline{\hat{h}}(\nu) \left(\hat{h}(\nu)\hat{f}(\nu) + \hat{h}(\nu+1/2)\hat{f}(\nu+1/2) \right) \\ & + \overline{\hat{g}}(\nu) \left(\hat{g}(\nu)\hat{f}(\nu) + \hat{g}(\nu+1/2)\hat{f}(\nu+1/2) \right) \Big] \ , \end{aligned} \tag{3.44}$$

ceci pour tout $f \in L^2([-1/2, 1/2])$. On reconnait là les deux termes importants : le terme qui nous intéresse, proportionnel à $\hat{f}(\nu)$, et le terme de "repliement de spectre", proportionnel à $\hat{f}(\nu + 1/2)$, que l'on souhaite annuler. La condition de reconstruction parfaite se met donc sous la forme suivante : pour tout ν,

$$|\hat{h}(\nu)|^2 + |\hat{g}(\nu)|^2 = 2 \tag{3.45}$$

$$\hat{h}(\nu)\overline{\hat{h}}(\nu+1/2) + \hat{g}(\nu)\overline{\hat{g}}(\nu+1/2) = 0 \ , \tag{3.46}$$

ou encore, sous forme matricielle

$$\begin{pmatrix} \hat{h}(\nu) & \hat{g}(\nu) \\ \hat{h}(\nu+1/2) & \hat{g}(\nu+1/2) \end{pmatrix} \begin{pmatrix} \overline{\hat{h}}(\nu) \\ \overline{\hat{g}}(\nu) \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{3.47}$$

Soit $\Delta(\nu)$ le déterminant de cette matrice, que l'on suppose non nul pour presque tout ν. On a alors la caractérisation suivante des filtres permettant une reconstruction parfaite :

PROPOSITION 3.2 (Smith-Barnwell,Vaidyanathan). (1) *Les filtres à reconstruction parfaite doivent satisfaire les équations de compatibilité*

$$\begin{pmatrix} \overline{\hat{h}}(\nu) \\ \overline{\hat{g}}(\nu) \end{pmatrix} = \frac{2}{\Delta(\nu)} \begin{pmatrix} \hat{g}(\nu+1/2) \\ -\hat{h}(\nu+1/2) \end{pmatrix} \tag{3.48}$$

et

$$|\hat{h}(\nu)|^2 + |\hat{h}(\nu+1/2)|^2 = 2 \ . \tag{3.49}$$

(2) *En supposant que les filtres h et g soient des suites finies, alors il existe $\varphi \in \mathbb{R}$ et $\ell \in \mathbb{Z}$ tels que*

$$\hat{g}(\nu) = e^{i\varphi} e^{2i\pi(2\ell+1)(\nu+1/2)} \, \overline{\hat{h}}(\nu+1/2) \ . \tag{3.50}$$

Preuve : Remarquons tout d'abord qu'en itérant cette équation, on a

$$\overline{\Delta}(\nu)\Delta(\nu+1/2) = -4 \ .$$

La première partie est alors une conséquence immédiate de (3.47) (si $\Delta(\nu) \neq 0$) pour la première équation, et de cette dernière égalité pour la seconde. Pour la seconde partie du résultat, la périodicité de $\hat{h}$ et $\hat{g}$ fait que $\Delta(\nu+1/2) = -\Delta(\nu)$, d'où $|\Delta(\nu)| = 2$. Si h et g sont des suites finies, alors $\hat{h}$ et $\hat{g}$ sont des polynômes trigonométriques. Donc Δ est également un polynôme trigonométrique, de même que Δ^{-1}. Par conséquent, $\Delta(\nu)$ est nécessairement de la forme

$$\Delta(\nu) = 2e^{-i\varphi} e^{2i\pi(2\ell+1)\nu} \ ,$$

ce qui complète la preuve. ♠

DÉFINITION 3.2. *Des suites $h = \{h_n, n \in \mathbb{Z}\}$ et $g = \{g_n, n \in \mathbb{Z}\}$ telles que les conditions (3.49) et (3.50) soient satisfaites sont appelés* filtres miroirs conjugués.

On note que la relation (3.50) se traduit, par TFD inverse, par

$$g_k = e^{i\varphi}(-1)^k \overline{h}_{-k-2\ell-1} \ . \tag{3.51}$$

Algorithme récursif. L'algorithme de codage en sous bandes repose sur l'utilisation récursive du schéma d'analyse-synthèse que nous venons de voir. Etant donnée une suite $f = \{f_n, n \in \mathbb{Z}\} \in \ell^2(\mathbb{Z})$, on pose

$$s^0 = f \ , \tag{3.52}$$

$$s^n = Hs^{n-1} \ , \tag{3.53}$$

$$d^n = Gs^{n-1} \ . \tag{3.54}$$

Ce schéma de décomposition est représenté en FIG. 3 (image de gauche).

On a alors de façon évidente

$$\begin{aligned} f &= H^*s^1 + G^*d^1 \\ &= H^*(H^*s^2 + G^*d^2) + G^*d^1 \\ &= H^*(H^*(H^*s^3 + G^*d^3) + G^*d^2) + G^*d^1 \\ &= \dots \ , \end{aligned}$$

ce qui conduit à un schéma simple de reconstruction de f à partir des suites $d^1, d^2, \dots, d^L, s^L$, où L est un nombre entier fixé. Ce schéma est représenté en FIG. 3, image de droite.

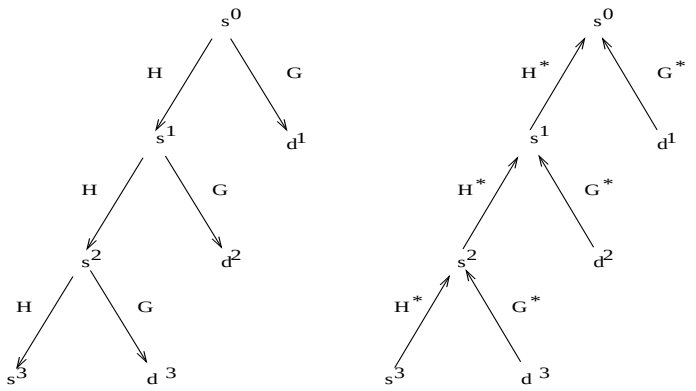

FIG. 3. Schéma de décomposition en sous bandes (à gauche) et de reconstruction à partir des sous bandes (à droite).

REMARQUE 3.5. Dans ce que nous avons vu, le codage en sous bandes s'effectue en divisant de façon récursive la bande de fréquences en deux, et en sous-échantillonnant en conséquence (d'un facteur 2). Bien que ce soit la solution la plus simple, rien n'oblige en fait à se limiter à un facteur 2. On peut construire des schémas de codage en sous bandes associés à des sous-échantillonnages presque arbitraires. En pratique, le facteur 2 est souvent préféré.

REMARQUE 3.6. On généralise facilement le codage en sous bandes au contexte des images. Il suffit pour cela de considérer deux schémas de codage en sous bandes (généralement identiques) unidimensionnels, et de les appliquer aux deux directions des signaux 2D.

2.3. Lien entre les bases d'ondelettes et le codage en sous bandes. Etant donnée une analyse multirésolution, il est facile de vérifier la relation avec le codage en sous bandes. Pour cela, soit $f \in L^2(\mathbb{R})$, et posons

$$d_k^j = \langle f, \psi_{jk} \rangle \; ; \qquad s_k^j = \langle f, \phi_{jk} \rangle \; . \tag{3.55}$$

On a alors, avec les notations précédentes,

$$\begin{aligned} s_k^j &= 2^{-j/2} \int_{-\infty}^{\infty} f(t) \overline{\phi} \left(2^{-j} t - k\right) \, dt \\ &= \sqrt{2} \sum_{\ell} h_\ell \, 2^{-j/2} \int_{-\infty}^{\infty} f(t) \overline{\phi} \left(2 \left(2^{-j} t - k\right) - \ell\right) \, dt \\ &= \sum_{\ell} h_\ell s_{2k-\ell}^{j-1} \\ &= (H s^{j-1})_k \; . \end{aligned}$$

De même, on montre facilement que

$$d_k^j = (G s^{j-1})_k \; .$$

Ceci permet de faire le lien entre les bases d'ondelettes et le codage en sous bandes : une base d'ondelettes "multirésolution" est naturellement associée à un schéma de

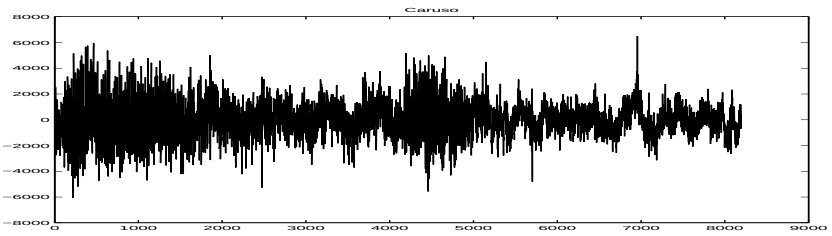

FIG. 4. Un vieil enregistrement de Caruso

codage en sous bandes : pour tout $f \in L^2(\mathbb{R})$, on a les décompositions

$$f = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \langle f, \psi_{jk} \rangle \, \psi_{jk} \tag{3.56}$$

$$= \sum_{k=-\infty}^{\infty} \langle f, \phi_{j_0 k} \rangle \, \phi_{j_0 k} + \sum_{j=-\infty}^{j_0} \sum_{k=-\infty}^{\infty} \langle f, \psi_{jk} \rangle \, \psi_{jk} \ , \tag{3.57}$$

et les coefficients $d_k^j = \langle f, \psi_{jk} \rangle$ et $s_k^j = \langle f, \phi_{jk} \rangle$ peuvent être calculés via un algorithme de codage en sous bandes.

2.4. Application au codage des signaux. Pour coder un signal, on peut donc se contenter de coder ses coefficients d (ainsi que les coefficients s à une échelle grossière). Un exemple est présenté en FIGURES 4, 5 et 6. Le signal original (un vieil enregistrement de Caruso) est montré en FIGURE 4, et ses coefficients d se trouvent en FIGURE 5 (les petites échelles sont vers le bas, et les grandes échelles vers le haut). Compte tenu du sous-échantillonnage inhérent au codage par sous bandes, il est difficile de comparer les sous bandes entre elles. C'est pour cela qu'on s'intéresse aussi à la représentation multirésolution montrée en FIGURE 6, dans laquelle on représente des reconstruction partielles obtenues à partir de coefficients d_{jk} avec une échelle j fixée. On voit apparaitre très nettement sur cette figure les différentes bandes de fréquence considérées : les courbes du haut de la figure varient bien plus lentement que celles du bas. Ceci confirme l'interprétation de la décomposition multirésolution comme un "banc de filtres" passe-bande, correspondant à des bandes de fréquence variables.

Ceci conduit au schéma simple de codage par transformation en ondelettes pour les signaux analogiques. On utilise généralement des bases d'ondelettes de $L^2([A, B])$, obtenues soit par de légères modifications de la construction présentée ci-dessous, soit via une opération de périodisation sur laquelle nous n'insisterons pas ici. Au final, on construit une famille de fonctions $\{\psi_{jk}, j = 0, \dots \infty, k = 0, \dots 2^j - 1\}$ qui forme une base orthonormée de $L^2([A, B])$.

CODAGE PAR TRANSFORMATION EN ONDELETTES

(1) Transformation : à $f \in L^2([A, B])$ on associe la famille des coefficients $d_k^j = \langle f, \psi_{jk} \rangle$, où $j = 0, , \dots$ et $k = 0, \dots 2^k - 1$. Cette famille caractérise f via

$$f = \sum_{j=-\infty}^{0} \sum_{k=0}^{2^{-j}-1} d_k^j \psi_{jk} \ .$$

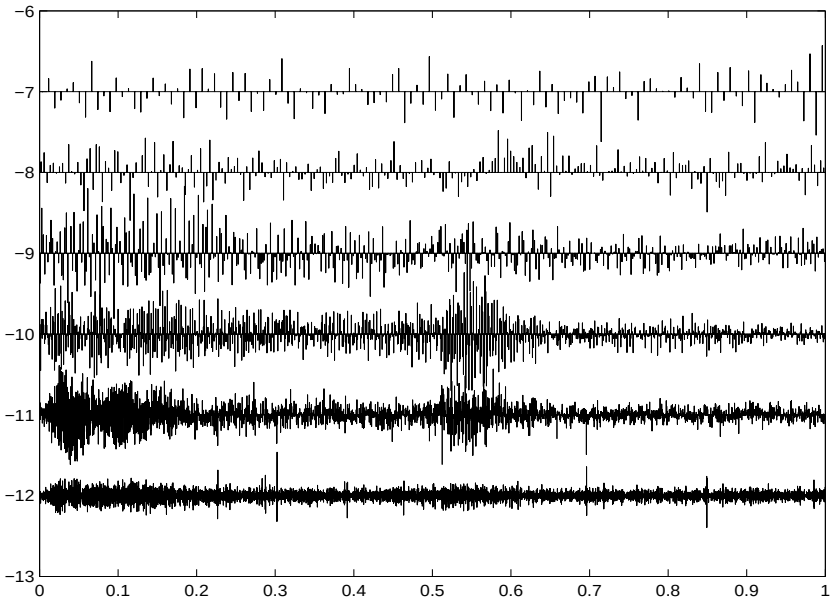

FIG. 5. Un vieil enregistrement de Caruso : coefficients d'ondelettes

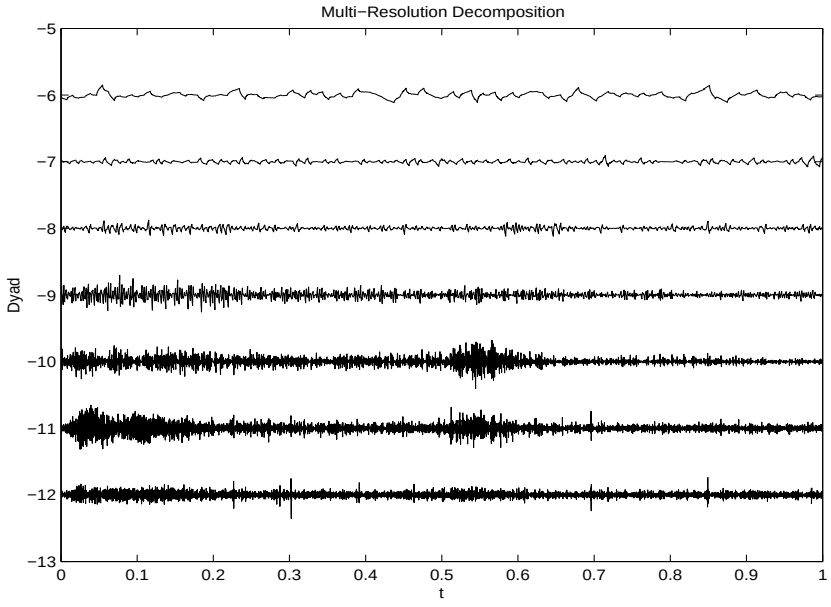

FIG. 6. Un vieil enregistrement de Caruso : décomposition multirésolution

(2) Approximation : (comparer au filtrage passe-bas) : les coefficients d_k^j correspondant aux plus petites valeurs de j sont ‘‘effacés’’. Ceci produit l'approximation

suivante de f :

$$f_J = \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} d_k^j \psi_{jk} \ ,$$

et la distorsion fonctionnelle estimée en norme L^2 comme

$$\|f - f_J\|^2 = \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} |d_k^j|^2$$

(3) Quantification : Les coefficients d_k^j restants sont quantifiés, ce qui produit l'approximation

$$\tilde{f}_J = \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} Q(d_k^j) \psi_{jk} \ ,$$

et la distorsion de quantification estimée en norme L^2 comme

$$\|f_J - \tilde{f}_J\|^2 = \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} |d_k^j - Q(d_k^j)|^2 \ .$$

La distorsion totale est finalement estimée comme

$$\|f - \tilde{f}_J\|^2 = \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} |d_k^j|^2 + \sum_{j=-\infty}^{J} \sum_{k=0}^{2^{-j}-1} |d_k^j - Q(d_k^j)|^2 \ .$$

REMARQUE 3.7. *Les schémas que nous avons décrits plus haut sont basés sur une approximation linéaire des signaux par des combinaisons linéaires de fonctions de base. En fait, il s'avère qu'après codage en sous bandes, les coefficients d obtenus sont souvent assez bien décorrélés, en tous cas bien plus que les échantillons. Ceci implique que la redondance ayant été réduite, beaucoup de ces coefficients sont très faibles numériquement. On peut en voir un exemple sur la* FIGURE *7, où on a représenté un petit échantillon de signal audio, un histogramme de ses échantillons, et un histogramme des coefficients d d'un développement en sous bandes.*

On voit très nettement que les coefficients d sont très majoritairement proches de 0, et pourront donc être négligés. Ceci suggère d'employer un schéma de codage dans lequel seuls les coefficients significatifs seront codés. Ceci dit, les coefficients qui sont petits et seront négligés ne sont pas connus à l'avance, et les valeurs correspondantes de j et k (ou k et ν dans le cas de la DCT ou la MDCT), c'est à dire leurs adresses, dépendent du signal. On parle alors d'approximation non-linéaire, nous y reviendrons plus loin. Ceci contraint alors à coder l'adresse des coefficients significatifs, ce qui pose d'autres problèmes, qui seront abordés dans le chapitre suivant.

En termes de résultats, le codage par sous bandes s'avère être le principe de codage le plus efficace à l'heure actuelle pour ce qui est du codage des images. Il y a des raisons à cela, qui tiennent à la nature même des images. On verra dans la section suivante quelques éléments d'explication.

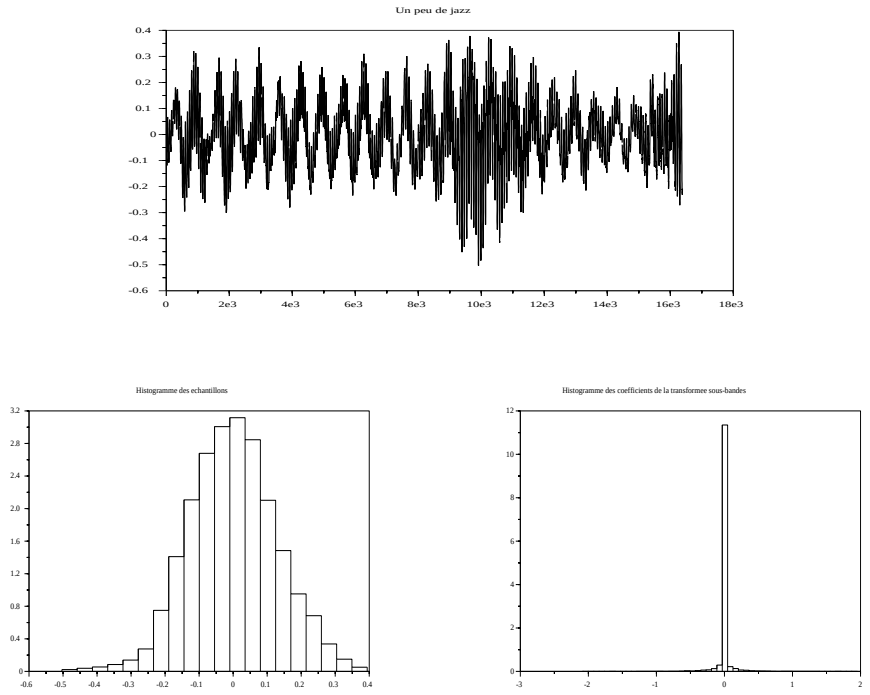

FIG. 7. Signal audio (jazz), en haut ; densité de probabilités empirique des échantillons (en bas à gauche), en des coefficients sous-bande (en bas à droite).

2.5. Les filtres à support borné de Daubechies. La connection entre les bases d'ondelettes et le codage en sous bandes a été étudiée en détails par I. Daubechies [**1**], qui a construit une famille de filtres h et g de réponse impulsionnelle finie, associés à des bases orthonormées d'ondelettes. Les filtres h sont de longueur paire $N = 2L$, par défaut à support entre 0 et $N-1$, et les filtres g correspondants en sont déduits par une relation de type (3.51) :

$$g_k = (-1)^k \overline{h}_{N-1-k} \ . \tag{3.58}$$

Des tables de coefficients peuvent être trouvées dans [**1**]. On montre que ces filtres sont effectivement associés à des bases d'ondelettes. Les fonctions ϕ et ψ correspondantes sont à support borné, et la taille du filtre N contrôle le support de ψ et ϕ, ainsi que son nombre de moments nuls : ψ est telle que

$$\int_{-\infty}^{\infty} t^m \psi(t)\, dt = 0 \ , \quad \forall m = 0, \dots L-1 \ . \tag{3.59}$$

REMARQUE 3.8. Cette dernière propriété a une grande importance en pratique, dans une perspective de compression de signaux : en effet, si ψ possède L moments nuls comme ci-dessus, et si un signal f se comporte comme un polynôme de degré inférieur ou égal à $L-1$ sur le support d'une ondelette ψ_{jk}, alors on a automatiquement $d_k^j = \langle f, \psi_{jk} \rangle = 0$. De telles ondelettes sont "aveugles" aux polynômes de

degré inférieur à L, ce qui se traduit en pratique par le fait que beaucoup de coefficients d_k^j sont nuls. Dans ce cas, un codage par transformation sera extrêmement efficace. C'est en particulier le cas pour le codage des images : les images présentent des zones où elles varient extrêmement peu, et sont donc bien approximées par des polynômes de bas degré. Ces zones là seront caractérisées par un faible nombre de coefficients d_k^j significatifs.

Il existe également de nombreuses variantes, notamment basées sur des fonctions "splines".

3. Comment choisir une base ?

3.1. Position du problème. On a vu dans les section précédentes des exemples de bases orthonormées pouvant être utilisées pour représenter un signal avant quantification. Le problème posé est maintenant le suivant : pour un signal donné, ou une classe donnée de signaux, quelle est la base qui sera la plus adaptée. Répondre à cette question suppose que l'on se soit donné un critère de choix. Pour cela, deux critères s'imposent naturellement :

- *Décorrélation :* On a jusque là traité les échantillons (ou les coefficients du développement par rapport à une base donnée) un à un, c'est à dire en se désintéressant de leur corrélation. C'est particulièrement vrai dans le cas du PCM, qui code échantillon par échantillon (moins pour ce qui est du DPCM, dans lequel le codage PCM est précédé d'une procédure visant à prendre en compte les corrélation). Or, les signaux sont généralement très corrélés, c'est à dire qu'il existe une forte redondance entre les divers coefficients. Ne pas utiliser cette redondance pénalise les performances du codeur. Un exemple extrême est celui d'un signal constant (c'est à dire que la corrélation entre coefficients est toujours égale à 1). Dans ce cas, coder tous les coefficients du signal serait un non sens. On peut donc se fixer comme objectif de trouver une base telle que les coefficients du signal soient aussi décorrélés que possible.
- *Concentration :* On peut également se donner comme objectif de trouver directement une base par rapport à laquelle le développement des signaux considérés sera le plus "court" possible, c'est à dire pour lequel un maximum de coefficients seront nuls ou très petits ; il sera alors inutile de coder et transmettre ces coefficients ; ou alors, on pourra se permettre de les coder avec une précision moindre, c'est à dire en leur affectant peu de bits.

Il faut signaler que ces deux objectifs ne sont nullement incompatibles, comme on le verra dans la section suivante, où il sera montré que la recherche d'une base qui décorrèle optimalement s'accompagne automatiquement d'une sélection des coefficients significatifs. Il faut également noter que cette "philosophie" ne se justifie véritablement que si la décomposition du signal est suivie d'un codage par code de longueur variable.

3.2. Approximation linéaire et approximation non-linéaire. Etant donnée une base orthonormée $\{e_n, n \in \mathbb{Z}^+\}$ d'un espace de Hilbert $\mathcal{H}$, on peut concevoir deux types d'approximation d'un signal $x \in \mathcal{H}$ par un sous ensemble fini de N vecteurs de base.

La première consiste à choisir à l'avance N vecteurs, disons $\{e_1, \dots e_N\}$, et projeter x sur le sous-espace $\mathcal{H}_N$ de $\mathcal{H}$ engendré par les N vecteurs :

$$x_{(N)} = \sum_{n=1}^{N} \langle x, e_n \rangle e_n \ . \tag{3.60}$$

Il s'agit d'une opération linéaire (projection orthogonale) et on parle d'*approximation linéaire*.

La seconde est une alternative non-linéaire : partant de $x \in \mathcal{H}$, on commence par calculer tous les coefficients $\langle x, e_n \rangle$, puis on retient les N plus grands (en valeur absolue), disons $\langle x, e_{n_1} \rangle, \dots \langle x, e_{n_N} \rangle$, à partir desquels on forme l'approximation

$$\tilde{x}_{(N)} = \sum_{i=1}^{N} \langle x, e_{n_i} \rangle e_{n_i} \ . \tag{3.61}$$

On parle alors d'*approximation non-linéaire*. Il est facile de montrer (en utilisant la formule de Parseval) le résultat suivant

PROPOSITION 3.3. *Pour tout $x \in \mathcal{H}$ et pour tout N fixé, on a*

$$\|x - \tilde{x}_{(N)}\| \leq \|x - x_{(N)}\| \ . \tag{3.62}$$

Ainsi, l'approximation non-linéaire est toujours meilleure que l'approximation linéaire. Cependant, dans un contexte de codage par transformation, elle a néanmoins un prix, en termes de codage. En effet, autant un signal approximé par approximation linéaire n'entraîne pas de surcoût en termes de codage (le sous-espace $\mathcal{H}_N$ est fixé une bonne fois pour toutes, et est connu à l'avance par le décodeur), autant dans le cas non-linéaire, il est nécessaire de transmettre, en même temps que les coefficients $\langle x, e_{n_i} \rangle$, les "adresses" $n_1, \dots n_N$ des coefficients retenus.

Le passage à l'approximation non-linéaire permet ainsi de gagner en précision, tout en perdant en débit. On examine plus en détails ce point ci-dessous.

3.3. Codage par transformation non-linéaire : carte de signifiance. Le résultat d'une "bonne" transformation est généralement que beaucoup des coefficients obtenus sont nuls, ou en tous cas très petits, de sorte qu'on peut penser qu'il n'est plus nécessaire de les encoder, se plaçant donc dans un schéma d'approximation non-linéaire. On peut en voir un exemple dans la FIGURE 8, qui représente les distributions des coefficients d'un signal audiophonique dans trois bases différentes : en haut, une base de sinus cardinaux (les coefficients sont alors simplement des échantillons du signal) ; au milieu une base d'ondelettes (algorithme de codage en sous bandes) ; en bas, une base trigonométrique locale adoucie (avec une fenêtre dont la longueur a été ajustée de façon à pouvoir "raisonnablement" considérer que le signal est stationnaire à l'intérieur de la fenêtre). On voit en particulier que la représentation trigonométrique locale est "redoutablement efficace".

Cependant, si tous les coefficients ne sont plus considérés, il devient nécessaire comme on l'a vu de prendre en compte les adresses des coefficients conservés, ce qui n'était pas nécessaire auparavant. On doit donc coder, en plus des coefficients significatifs, une *carte de signifiance*, qui précise les positions de ces derniers. Une carte de signifiance est donc une suite de 0 et 1, indiquant la conservation ou la non conservation d'un coefficient : par exemple

0001111100111111000011111101110000001100000000011111

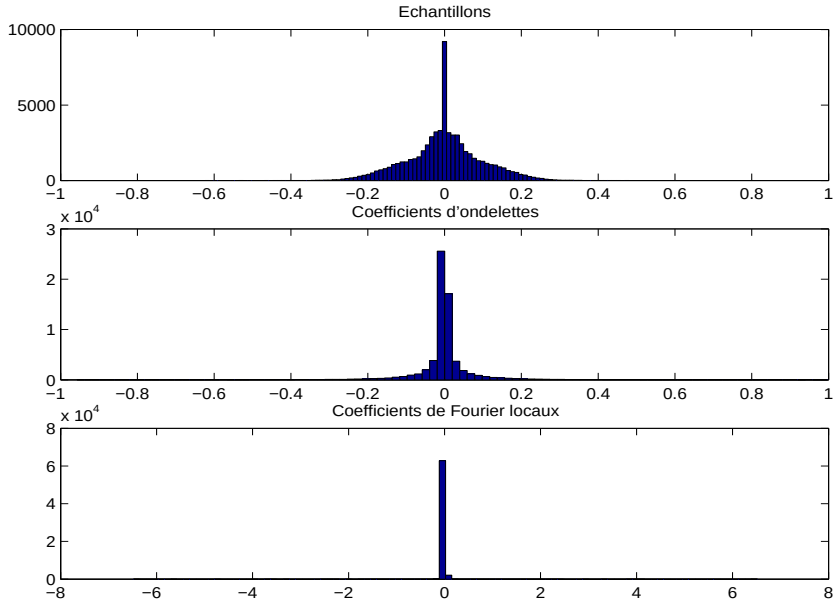

FIG. 8. Codage d'un signal audio : histogrammes des échantillons (en haut), des coefficients dans une base d'ondelettes (milieu) et des coefficients dans une base trigonométrique locale (bas).

Ceci consomme donc 1 bit supplémentaire par coefficient, qu'il soit ou pas conservé : dans notre exemple, 50 bits sont nécessaires.

Or, il arrive rarement que l'on ait des suites complètement désordonnées de 0 et de 1 : on observe souvent des plages de zéros et des plages de uns, en alternance. Si tel est le cas, on a intérêt à tirer avantage de cette "structure", et coder différemment la carte de signifiance. Si on décide par exemple de coder uniquement les longueurs des plages, l'exemple précédent devient

3 5 2 6 2 8 1 3 6 2 8 4

On a donc 12 plages, de longueur inférieure à 8. Les longueurs de plages peuvent donc être codées sur 3 bits, ce qui induit un coût de 36 bits pour coder la carte de signifiance.

Là encore 'comme pour le codage des coefficients), nous avons utilisé un code de longueur constante pour coder les longueurs de plages. On verra par la suite qu'il est en général plus utile d'utiliser un code de longueur variable (code entropique).

Le codage de la carte de signifiance est en fait un cas particulier du problème de codage d'une suite de bits, ou plus généralement d'un système à deux niveaux. L'exemple le plus classique est le codage des facsimilés : chaque ligne d'un fax est en fait une suite de pixels noirs et blancs (1728 pixels dans la norme standard). Le standard de codage des fax inclut un codage de plages, dans lequel les longueurs des plages noires et blanches utilisent un code de Huffman (code entropique).

4. Codage par transformation linéaire : le choix d'une base pour les signaux aléatoires

On considère tout d'abord le cadre du codage par transformation "linéaire", abordé sous l'angle de la recherche de "bases qui décorrèlent". Pour mettre en oeuvre cette approche, il il est nécessaire de se donner une modélisation du signal. On considère généralement des modèles de signaux donnés par des processus aléatoires possédant certaines caractéristiques relativement bien définies. Dans cette section on désignera par $(\mathcal{A}, \mathcal{F}, \mathbb{P})$ un espace probabilisé. Les notations sont les mêmes que dans le chapitre 1.

On considère un signal aléatoire, modélisé comme processus du second ordre sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$, pas nécessairement stationnaire en moyenne d'ordre deux. On limitera la discussion au cas des signaux analogiques (le cas des signaux numériques se traite de façon similaire).

On introduit la moyenne du signal

$$m_X(t) = \mathbb{E}\{X_t\} , \tag{3.63}$$

et sa covariance

$$C_X(t,s) = \mathbb{E}\left\{(X_t - \mu_t)(\overline{X_s} - \overline{\mu_s})\right\} = R_X(t,s) - \mu_t\overline{\mu_s} , \tag{3.64}$$

où

$$R_X(t,s) = \mathbb{E}\left\{X_t\overline{X_s}\right\} \tag{3.65}$$

est l'autocorrélation. On a déjà vu que les deux fonctions C_X et R_X sont définies positives : pour tous $t_1, \dots t_n \in \mathbb{R}$ et $\alpha_1, \dots \alpha_n \in \mathbb{C}$, on a

$$\sum_{k,\ell=1}^{n} \alpha_k\overline{\alpha_\ell}C_X(t_k,t_\ell) \geq 0 , \tag{3.66}$$

et de même pour R_X.

On se limitera pour simplifier au cas des processus centrés ($m_X(t) = 0$ pour tout t). On considère l'espace $\mathcal{L}^2(\mathcal{A})$ des variables aléatoires centrées X sur $(\mathcal{A}, \mathcal{F}, \mathbb{P})$ telles que $\mathbb{E}\left\{|X|^2\right\} < \infty$, et l'espace $L^2(\mathcal{A})$, quotient de $\mathcal{L}^2(\mathcal{A})$ dans lequel on a identifié les variables aléatoires égales presque sûrement. Comme on l'a déjà vu, $L^2(\mathcal{A})$ est naturellement muni d'un produit scalaire défini par

$$(X|Y)_{L^2(\mathcal{A})} = (X|Y) = \mathbb{E}\left\{X\overline{Y}\right\} . \tag{3.67}$$

Etant donné un processus du second ordre $\{X_t, t \in T\}$, on considère le sous-espace $\mathcal{M}_X$ de $L^2(\mathcal{A})$ engendré par les variables aléatoires $X_t, t \in T$.

4.1. La base de Karhunen-Loève. Dans cette section, on se limite au cas où T est une partie compacte de $\mathbb{R}^2$. On considère un processus du second ordre $\{X_t, t \in T\}$, centré, et continu en moyenne d'ordre 2, et sa fonction d'autocorrélation $R_X(t,s)$. Celle-ci est donc une fonction continue (et bornée) de t et s. T étant compact, on en déduit que $R_X \in L^2(T^2)$:

$$\|R_X\|^2 = \int_{T\times T} |R_X(t,s)|^2 \, dt \, ds < \infty$$

[2]Le cas où T est non compact donne lieu à une analyse et des résultats similaires, mais est plus complexe du point de vue "technique".

LEMME 3.2. *Soit $X = \{X_t, t \in T\}$ un processus comme ci-dessus, défini sur un domaine compact T. Pour toute fonction continue $\varphi \in C(T)$, la variable aléatoire*

$$Z_\varphi = \int_T X_t \overline{\varphi}(t)\, dt \tag{3.68}$$

est du second ordre, et on a

$$\mathbb{E}\left\{|Z_\varphi|^2\right\} = \int_{T\times T} R_X(t,s)\varphi(s)\overline{\varphi}(t)\, ds\, dt\ . \tag{3.69}$$

Preuve : Si on pose pour tout $t \in T$, $Y_t = X_t\overline{\varphi}(t)$, f étant continue sur T, Y_t est du second ordre, et $R_Y(t,s) = \mathbb{E}\left\{Y_t\overline{Y}_s\right\} = R_X(t,s)\overline{\varphi}(t)\varphi(s)$ pour tous $t, s \in T$. Il résulte de la continuité de R_X et φ que R_Y est continue sur $T \times T$. On en déduit que la variable aléatoire Z_φ est du second ordre, et

$$\mathbb{E}\left\{|Z_\varphi|^2\right\} = \int_{T\times T} R_Y(t,s)\, dt\, ds < \infty\ ,$$

ce qui conclut la preuve. ♠

Ceci permet d'introduire l'opérateur de corrélation $\mathcal{C}$, défini par ses élements de matrice

$$\langle \mathcal{C}f, g\rangle = \mathbb{E}\left\{\langle X, g\rangle\langle f, X\rangle\right\}\ . \tag{3.70}$$

On a donc

$$\langle \mathcal{C}f, g\rangle = \int_{T\times T} R_X(t,s) f(s)\overline{g}(t)\, dt\, ds\ . \tag{3.71}$$

Il résulte de l'analyse précédente que $\mathcal{C}$ est un opérateur linéaire, auto-adjoint, continu sur $L^2(T)$, de $Hilbert - Schmidt$:

$$\|\mathcal{C}\|^2_{HS} = \int_{T\times T} |R_X(t,s)|^2\, dt\, ds < \infty\ , \tag{3.72}$$

et donc compact. On peut par conséquent utiliser des résultats classiques pour le développer sur ses fonctions propres. Le résultat suivant est une conséquence directe du théorème de Mercer :

PROPOSITION 3.4 (Mercer). (1) *Le spectre de $\mathcal{C}$ est dénombrable, et forme une suite décroissante de valeurs propres positives, notées*

$$\lambda_1 \geq \lambda_2 \geq \cdots \to 0\ , \tag{3.73}$$

chacune étant de multiplicité finie.

(2) *On a la formule de Parseval*

$$\|\mathcal{C}\|^2_{HS} = \sum_{n=1}^{\infty} \lambda_n^2 < \infty\ . \tag{3.74}$$

(3) *Il existe une famille de vecteurs propres correspondants $\{\varphi_n, n = 1, 2, \ldots\}$:*

$$\mathcal{C}\varphi_n = \lambda_n\varphi_n\ , \tag{3.75}$$

qui forment une base orthonormée de $L^2(T)$. De plus, on a $\varphi_n \in C(T)$ pour tout n.

(4) *Pour tous $t, s \in T$, on a*

$$R_X(t,s) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(t) \overline{\varphi}_n(s) \ , \tag{3.76}$$

où la série est uniformément convergente.

La conséquence de ce résultat est le théorème suivant, qui explicite la base de Karhunen-Loève :

THÉORÈME 3.6 (Karhunen-Loève). *Soit $\{X_t, t \in T\}$ un processus du second ordre, défini sur un domaine compact $T \subset \mathbb{R}$, continu en moyenne d'ordre 2. Pour tout $n = 1, 2 \ldots$ on définit la variable aléatoire*

$$Z_n = \langle X, \varphi_n \rangle = \int_T X_t \overline{\varphi}_n(t)\, dt \ . \tag{3.77}$$

Alors les Z_n sont des variables aléatoires du second ordre, orthogonales dans $L^2(\mathcal{A})$:

$$\mathbb{E}\left\{ Z_m \overline{Z_n} \right\} = \lambda_m \delta_{mn} \ , \tag{3.78}$$

et on a pour tout $t \in T$

$$X_t = \sum_{n=1}^{\infty} Z_n \varphi_n(t) \ , \tag{3.79}$$

au sens de la convergence dans $L^2(\mathcal{A})$.

Preuve : Commençons par évaluer

$$\mathbb{E}\left\{ X_t \overline{Z}_m \right\} = \int_T \mathbb{E}\left\{ X_t \overline{X}_s \right\} \varphi_m(s)\, ds = \int_T R(t,s) \varphi_m(s)\, ds = \lambda_m \varphi_m(t) \ . \tag{3.80}$$

On a également

$$\begin{aligned} \mathbb{E}\left\{ Z_m \overline{Z_n} \right\} &= \int_{T\times T} \mathbb{E}\left\{ X_t \overline{X}_s \right\} \varphi_n(s) \overline{\varphi}_m(t)\, dt\, ds \\ &= \int_{T\times T} R(t,s) \varphi_n(s) \overline{\varphi}_m(t)\, dt\, ds \\ &= \langle \mathcal{C}\varphi_n, \varphi_m \rangle = \lambda_m \delta_{mn} \ , \end{aligned}$$

Finalement, étant donné un $M \geq 1$, calculons, pour t quelconque

$$\begin{aligned} \mathbb{E}\left\{ \left| X_t - \sum_{m=1}^{M} Z_m \varphi_m(t) \right|^2 \right\} &= R(t,t) - 2 \sum_{m=1}^{M} \Re(\mathbb{E}\left\{ Z_m \overline{X_t} \right\} \varphi_m(t)) \\ &\quad + \sum_{m,n=1}^{M} \mathbb{E}\left\{ Z_m \overline{Z}_n \right\} \varphi_m(t) \overline{\varphi}_n(t) \\ &= R(t,t) - \sum_{m=1}^{M} \lambda_m \varphi_m(t) \overline{\varphi}_m(t) \ . \end{aligned}$$

De là, la proposition 3.4 permet de conclure que

$$\lim_{M\to\infty} \mathbb{E}\left\{ \left| X_t - \sum_{m=1}^{M} Z_m \varphi_m(t) \right|^2 \right\} = 0 \ , \tag{3.81}$$

ce qui conclut la preuve du théorème. ♠

Dans un cadre de codage par transformation *linéaire*, ce résultat est exploité de la façon suivante. Supposons que l'on cherche à coder une classe de signaux, modélisée par un signal aléatoire X sur $[0,T]$, du second ordre, centré, et continu en m.o.d., dont la corrélation R_X (et donc la base de Karhunen-Loève correspondante) est connue.

Si on décide d'approximer X par sa projection sur l'espace engendré par les φ_n correspondant aux N plus grandes valeurs propres, c'est à dire par sa $\varphi_1, \dots \varphi_N$

$$X_{(N)} = \sum_{n=1}^{N} Z_n \varphi_n \ ,$$

on aura directement une estimation de l'erreur moyenne commise

$$\mathbb{E}\left\{\|X - X_{(N)}\|^2\right\} = \sum_{N+1}^{\infty} \lambda_n \ .$$

En d'autres termes, pour une erreur fixée à l'avance, on disposera d'une estimation du nombre de coefficients Z_n à conserver. Ces coefficients sont ensuite quantifiés et codés.

4.2. Le cas des processus stationnaires en m.o.d. Il est intéressant de considérer le cas particulier des processus stationnaires en moyenne d'ordre 2.

DÉFINITION 3.3. *Un processus aléatoire du second ordre $\{X_t, t \in T\}$ est dit stationnaire en moyenne d'ordre deux si*

$$\mathbb{E}\{X_t\} = \mathbb{E}\{X_0\} \ , \quad \forall t \in T \ , \quad \textit{et} \tag{3.82}$$

$$R(t+\tau, s+\tau) = R(t,s) = R(t-s,0) \ , \quad \textit{pour tous } t,s,\tau \ . \tag{3.83}$$

Dans la mesure où nous avons supposé que le domaine T considéré est borné, il faut être plus précis, notamment en ce qui concerne les conditions aux bords. On supposera donc que T est un intervalle dans $\mathbb{R}$, et que les équations ci-dessus sont considérées "modulo la longueur de l'intervalle". En notant

$$L = |T|$$

cette dernière, on écrira pour tous $s,t \in T$

$$R((t+\tau) \bmod L, (s+\tau) \bmod L) = R(t,s) \ , \quad \text{pour tout } \tau \ .$$

On a donc

$$R_X(t,s) = R_X((t-s) \bmod L, 0) := R_X(t-s) \tag{3.84}$$

Dans ce cas, $\mathcal{C}$ est un opérateur de convolution dans $L^2(T)$, et il est bien connu qu'un tel opérateur est diagonal dans la base trigonométrique. Ainsi, dans ce cas, la base de Karhunen-Loève est donnée par

$$\varphi_n(t) = \frac{1}{\sqrt{L}} e^{2i\pi \frac{t}{L}} \ , \quad n \in \mathbb{Z} \ . \tag{3.85}$$

REMARQUE 3.9. Ceci nous permet de comprendre le pourquoi du choix de bases trigonométriques dans les codeurs de signaux audiophoniques. L'hypothèse de stationnarité semble dans ce cas assez réaliste, à condition de se limiter à des segments de signal (les domaines T ci-dessus) suffisamment courts (de l'ordre de 20 millisecondes en pratique). On peut donc dans ce cas utiliser des bases trigonométriques locales et aboutir à des codecs efficaces.

REMARQUE 3.10. Dans le cas des standards de type JPEG pour la compression des images, on commence par considérer des blocs de 8 pixels par 8 pixels. Supposer que l'image est stationnaire dans un bloc donné a longtemps été considéré (à tort ou à raison...) comme une hypothèse raisonnable. Par conséquent, il était légitime de se tourner vers des bases trigonométriques. Le choix de bases de cosinus de préférence à la base trigonométrique standard a déjà été justifié plus haut.

REMARQUE 3.11. Comme on l'a vu, on utilise plus volontiers en pratique des bases de cosinus plutôt que des bases d'exonentielles. La différence entre les deux tient essentiellement aux conditions aux bords que l'on impose. Dans le calcul ci-dessus, le choix des exponentielles était lié à des conditions aux bords périodiques.

REMARQUE 3.12. Des développements plus récents conduisent à penser que de telles hypothèses ne sont pas vraiment réalistes, et que par conséquent les bases trigonométriques ne sont pas les mieux adaptées. On montre que des hypothèses d'invariance par changement d'échelle conduisent naturellement à choisir des techniques de type "codage en sous bandes", c'est à dire des bases orthonormées d'ondelettes.

5. Une alternative aux bases : repères

Il existe des situations dans lesquelles on a intérêt, plutôt que d'utiliser des bases orthonormées, à utiliser des familles de fonctions qui sont complètes mais pas libres. On parle alors de familles surcomplètes, ou de repères. La famille sur laquelle l'on décompose le signal n'étant pas libre, les coefficients de la décomposition sont redondants, et la représentation n'est donc pas "économique". Cependant, on verra que cette représentation présente l'avantage d'être plus robuste, c'est à dire moins sensible aux perturbations, qu'une décomposition par rapport à une base.

5.1. Définitions.

DÉFINITION 3.4. *Une famille* $\{f_\lambda, \lambda \in \Lambda\}$ *de vecteurs d'un espace de Hilbert* H *est un repère de cet espace si il existe deux constantes réelles* $0 < A < B < \infty$ *telles que pour tout* $f \in H$*, on ait*

$$A\|f\|^2 \leq \sum_{\lambda\in\Lambda} |\langle f, f_\lambda\rangle|^2 \leq B\|f\|^2 \ . \tag{3.86}$$

Les constantes A *et* B *sont appelées* Bornes *du repère. Si* $A = B$*, le repère est dit strict.*

EXEMPLE 3.1. Considérons le plan $\mathbb{R}^2$, muni d'une base orthonormée $\{e_1, e_2\}$. Alors, en posant $f_1 = e_1, f_2 = (-e_1 + e_2\sqrt{3})/2$ et $f_3 = (-e_1 - e_2\sqrt{3})/2$, il est immédiat que $\{f_1, f_2, f_3\}$ est un repère strict de $\mathbb{R}^2$: pour tout $f \in \mathbb{R}^2$, on a $\sum_{n=1}^3 |\langle f, f_n\rangle|^2 = 3||f||^2/2$. Plus généralement, toute famille finie de vecteurs est un repère de l'espace qu'elle engendre.

On associe naturellement à un repère les deux opérateurs suivants : l'opérateur $U : H \to \ell^2(\Lambda)$, défini par

$$(Uf)_\lambda = \langle f, f_\lambda\rangle \ , \tag{3.87}$$

et $\mathcal{R} = U^*U : H \to H$, défini par

$$\mathcal{R}f = \sum_{\lambda\in\Lambda} \langle f, f_\lambda\rangle \, f_\lambda \ . \tag{3.88}$$

Le résultat suivant est vérifié sans difficulté.

LEMME 3.3. *$\mathcal{R}$ est borné, inversible et à inverse borné.*

L'opérateur $\mathcal{R}$ est évidemment auto-adjoint par construction. Donc, il résulte de (3.86) que son spectre est inclus dans l'intervalle $[A, B]$, ce que l'on écrit aussi

$$A \leq \mathcal{R} \leq B \ .$$

$\mathcal{R}$ est inversible, de sorte que l'on peut écrire, pour tout $f \in \mathcal{H}$

$$f = \sum_{\lambda \in \Lambda} \langle f, f_\lambda \rangle \, \tilde{f}_\lambda \ , \tag{3.89}$$

oú on a posé

$$\tilde{f}_\lambda = \mathcal{R}^{-1} f_\lambda \ . \tag{3.90}$$

On peut également vérifier facilement le résultat suivant :

PROPOSITION 3.5. *La famille $\{\tilde{f}_\lambda, \lambda \in \Lambda\}$ est un repère de $\mathcal{H}$, de bornes B^{-1} et A^{-1}, appelé* repère dual *du repère $\{f_\lambda, \lambda \in \Lambda\}$.*

On a donc aussi l'égalité, pour tout $f \in \mathcal{H}$

$$f = \sum_{\lambda \in \Lambda} \langle f, \tilde{f}_\lambda \rangle \, f_\lambda \ , \tag{3.91}$$

EXEMPLE 3.2. Des exemples utiles de repères sont donnés par les repères trigonométriques. On sait que le système trigonométrique, formé des fonctions

$$e_n(t) = e^{2int} \ , \quad n \in \mathbb{Z} \tag{3.92}$$

est une base de $L^2([0, \pi])$. Considérons le système de fonctions

$$f_n(t) = e^{int} \ , \quad n \in \mathbb{Z} \ . \tag{3.93}$$

Soit $f \in L^2([0, \pi])$. Alors

$$\langle f, f_{2n} \rangle = \langle f, e_n \rangle = \pi c_n(f) \ , \quad \text{et}$$

$$\langle f, f_{2n+1} \rangle = \langle g, e_n \rangle = \pi c_n(g) \ ,$$

où g est définie par

$$g(t) = f(t) e^{-it} \ .$$

L'égalité de Parseval donne alors

$$\sum_n |\langle f, f_n \rangle|^2 = \pi^2 \sum_n (|c_n(f)|^2 + |c_n(g)|^2) = \pi ||f||^2 \ . \tag{3.94}$$

Donc, la famille $\{f_n, n \in \mathbb{Z}\}$ est un repère strict de $L^2([0, \pi])$, de borne $A = B = \pi$.

De tels repères, ou plutôt leurs analogues finis, sont utilisés par exemple en restauration d'images, c'est à dire pour reconstituer des images dont certains pixels sont manquants.

Dans le cas d'une famille $\{e_\lambda, \lambda \in \Lambda\}$ qui est une base orthonormée de $\mathcal{H}$, le théorème de Riesz-Fisher établit une correspondance bijective entre $\mathcal{H}$ et $\ell^2(\Lambda)$. Dans le cas d'un repère, la suite des coefficients $\langle f, f_\lambda \rangle$ d'un $f \in \mathcal{H}$ est bien dans $\ell^2(\Lambda)$, mais la correspondance n'est plus surjective. Le résultat suivant donne une description plus "géométrique" de la situation.

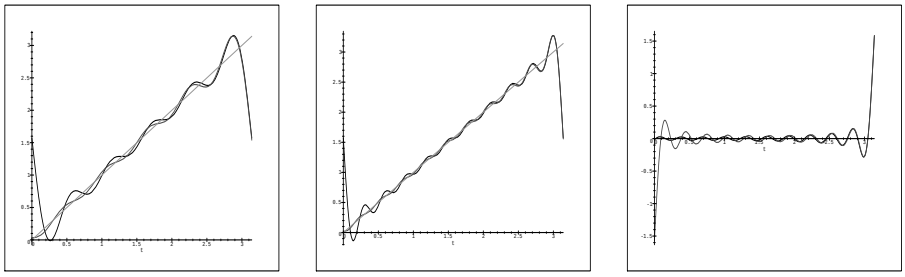

FIG. 9. Comparaison d'une série de Fourier usuelle et d'une décomposition redondante : le cas d'une fonction linéaire. A gauche, la fonction, sa reconstruction à partir de 11 modes de Fourier e_n et 21 fonctions f_n ; au centre, même chose, avec 21 fonctions e_n et 41 fonctions f_n ; à droite, les erreurs de reconstruction.

PROPOSITION 3.6. *Etant donné un repère comme ci-dessus, l'opérateur U possède une infinité d'inverses à gauche. L'inverse à gauche de norme minimale est donné par*

$$\tilde{U}^{-1} = (U^*U)^{-1}U^* \ , \tag{3.95}$$

et s'annulle sur $(U\mathcal{H})^\perp$. De plus,

$$||\tilde{U}^{-1}|| \leq 1/\sqrt{A} \ . \tag{3.96}$$

COROLLAIRE 3.3. *$U\tilde{U}^{-1}$ est le projecteur orthogonal sur l'image de U. De plus, pour tout $x \in \ell^2(\Lambda)$, on a*

$$(U\tilde{U}^{-1}x)_\mu = \sum_\lambda x_\lambda \langle \tilde{f}_\lambda, f_\mu \rangle \ . \tag{3.97}$$

REMARQUE 3.13. *Utilité des décompositions redondantes :* Les décompositions redondantes apportent une stabilité supplémentaire aux décompositions. En effet, soit $f \in \mathcal{H}$ un vecteur fixé, et soit $x = Uf$. Supposons qu'une erreur ϵ soit commise sur x : soient

$$y = x + \epsilon \quad \text{et} \quad \tilde{f} = \tilde{U}^{-1}y = f + \tilde{U}^{-1}\epsilon \ .$$

En notant ϵ_1 la projection de ϵ sur $U\mathcal{H}$, et ϵ_2 sa projection sur $(U\mathcal{H})^\perp$, on voit immédiatement que $\tilde{U}^{-1}\epsilon_2 = 0$, et que donc

$$||f - \tilde{f}|| \leq ||\epsilon_1||/\sqrt{A} \ .$$

Ainsi, la composante ϵ_2 de l'erreur disparait lors de l'inversion de la décomposition. On voit donc que dans des situations où on se doute à l'avance qu'une erreur importante va être commise sur les coefficients, lors d'une étape de transmission par exemple, on a intérêt à utiliser des décompositions par rapport à des repères de préférence à des décompositions sur des bases, car une partie de l'erreur disparaitra lors de la resynthèse.

EXEMPLE 3.3. *L'algorithme de Gershberg-Papoulis :* On se pose le problème de restaurer des valeurs manquantes de signaux à bande limitée. Soit $f = \{f_0, \dots f_{N-1}\} \in$

$\mathbb{C}^N$, tel que

$$\hat{f}_k = \sum_0^{N-1} f_n e^{-2i\pi\frac{kn}{N}} = 0 \text{ si } k \notin [k_1, k_2] \ .$$

On suppose que les valeurs f_n sont connues à l'exception des $f_n, n \in [n_1, n_2]$, et on cherche à restaurer ces valeurs.

(1) Soit $E = \left\{ g \in \mathbb{C}^N, g_n = f_n \ \forall n \notin [n_1, n_2] \right\}$. On montre facilement que E est un espace convexe, c'est à dire que pour tous $g, g' \in E$, $\epsilon g + (1-\epsilon) g' \in E$, pour tout $\epsilon \in [0, 1]$. Soit $P_E : \mathbb{C}^N \to E$ l'opérateur défini par

$$[P_E g]_n = \begin{cases} g_n & \text{si } n \in [n_1, n_2] \\ f_n & \text{sinon} \end{cases}$$

P_E est en fait un projecteur sur E.

(2) Soit $F = \left\{ g \in \mathbb{C}^N, \ \hat{g}_k = 0 \ \forall k \in [k_1, k_2] \right\}$. F est un espace vectoriel, et on note $P_F : \mathbb{C}^N \to F$, défini par

$$\widehat{P_F g}_k = \begin{cases} \hat{g}_k & \text{si } k \notin [k_1, k_2] \\ 0 & \text{sinon} \end{cases}$$

Un opérateur de projection sur F (P_F peut être implémenté simplement au prix de deux `FFTs`).

(3) Etant donnée une suite $f = f^{(0)} \in \mathbb{C}^N$, on pose pour $j = 1, 2, \ldots$:

$$f^{(j+1)} = P_E P_F f^{(j)}$$

Des résultats généraux d'analyse fonctionnelle montrent que si $E \cap F$ n'est pas vide, alors l'algorithme dit *de projection alternée* défini par l'itération précédente converge dans $E \cap F$ quand $j \to \infty$.

5.2. Inversion. La question qui se pose en pratique est la suivante : étant donnés les coefficients de $f \in H$ par rapport à un repère $\{f_n, n \in \Lambda\}$, comment retrouver f à partir de ces coefficients ?

Considérons l'opérateur de repère $\mathcal{R}$. Comme $A \leq \mathcal{R} \leq B$, on a aussi

$$\frac{2A}{A+B} \leq \frac{2}{A+B}\,\mathcal{R} \leq \frac{2B}{A+B} \ .$$

Posons

$$T = 1 - \frac{2}{A+B}\,\mathcal{R} \ .$$

Un calcul immédiat montre que

$$||T|| \leq \frac{B-A}{A+B} < 1 \ . \tag{3.98}$$

Donc, l'opérateur $1 - T$ est inversible, et la série de Neumann correspondante

$$(1-T)^{-1} = 1 + T + T^2 + T^3 + \ldots$$

est convergente. On peut donc écrire

$$\mathcal{R}^{-1} = \frac{2}{A+B}\left(1 + T + T^2 + T^3 + \ldots\right) \ . \tag{3.99}$$

Ceci conduit à l'algorithme suivant : si $\alpha_n = \langle f, f_n \rangle$, on commence par évaluer $f^{(1)} = \frac{2}{A+B} \sum_n \alpha_n f_n$. On sait alors que

$$f - f^{(1)} = \frac{2}{A+B} \left(T + T^2 + T^3 + \dots\right) \left(\sum_n \alpha_n f_n\right) = Tf \ ,$$

de sorte que

$$||f - f^{(1)}|| \leq ||T||\,||f|| \leq \frac{B-A}{A+B}\,||f|| \ .$$

Si la précision est suffisante, c'est à dire si la constante $(B-A)/(A+B)$ est assez faible, on se contentera de $f^{(1)}$ comme approximation de f. Si tel n'est pas le cas, il faut pousser plus loin le développement, et considérer

$$f^{(2)} = \frac{2}{A+B}(1+T) \left(\sum_n \alpha_n f_n\right) \ .$$

On a alors évidemment un ordre d'approximation supplémentaire :

$$||f - f^{(2)}|| \leq \left(\frac{B-A}{A+B}\right)^2 ||f|| \ .$$

REMARQUE 3.14. L'algorithme d'inversion qu'on a vu ci-dessus a l'avantage d'être simple, mais n'est pas optimal. En pratique, il est souvent plus avantageux d'utiliser des méthodes classiques d'inversion, telles que des méthodes de gradient conjugué par exemple.

REMARQUE 3.15. Il est possible de montrer que les bases orthonormées que nous avons vues plus haut peuvent être remplacées par des repères construits de la même manière. C'est en particulier le cas des bases trigonométriques locales (on construit facilement des repères trigonométriques locaux), et des ondelettes, pour lesquelles il est même plus facile de construire des repères ue des bases.

CHAPITRE 4

Quantification et Codage entropique

1. Généralités ; allocation de bits

La dernière étape du codage d'un signal, postérieure à la quantification, est le codage proprement dit. On désigne par le vocable codage un ensemble de techniques visant à associer des suites binaires (c'est à dire des suites de zéros et de uns) à un dictionnaire de mots. Dans le cas qui nous préoccupe, ces mots sont les valeurs (discrètes) de coefficients d'un signal dans une base donnée, après quantification. Cependant, les techniques de codage s'appliquent également à d'autres situations, telles que le codage de fichiers informatiques par exemple.

On a vu dans le cas des codeurs PCM et DPCM des exemples utilisant un code de longueur constante : tous les coefficients étaient codés sur un nombre R constant de bits, le code étant fixé une fois pour toutes. Cependant, il s'avère souvent plus efficace d'utiliser des codes de longueur variable, c'est à dire des codes dans lesquels on n'affecte pas nécessairement un code de la même longueur à différents symboles (dans notre cas, des coefficients quantifiés). On pourra ainsi affecter un code très court à des symboles très fréquent, et un code plus long à des symboles moins fréquents, afin d'optimiser le débit total. On rappelle que dans le cas d'un code de longueur constante, chaque symbole est codé sur R bits, ce qui correspond à $M = 2^R$ symboles.

Les codes de longueur variable posent des problèmes différents, qu'on va étudier ci dessous.

2. Codes de longueur variable

2.1. Introduction. Pour fixer les idées, on considère une suite de variables aléatoires X_n, prenant leurs valeurs dans un alphabet fini $A = \{a_0, a_1, \dots a_{M-1}\}$, avec des probabilités $\Pr X_n = a = p(a)$. Un codeur associera à chaque symbole $a \in A$ un mot binaire $\alpha(a)$, de longueur $\ell(a)$ (mesurée en bits). Le meilleur codeur sera celui qui minimise la longueur moyenne

$$\overline{\ell} = \sum_{a \in A} p(a)\ell(a) \ . \tag{4.1}$$

On rappelle que dans le cas d'un code de longueur constante, chaque symbole a_k est codé sur $\ell(a_k) = R$ bits, ce qui correspond à $M = 2^R$ symboles. Par conséquent, on a

$$\sum_{a \in A} 2^{-\ell(a)} = 1 \ . \tag{4.2}$$

On verra par la suite la signification de cette identité simple.

DÉFINITION 4.1. *Un code scalaire sans perte de longueur variable consiste en*

(1) *Un codeur : une application α qui associe à tout symbole d'entrée $a \in A$ une suite binaire $\alpha(a)$, de longueur $\ell(a)$.*
(2) *Un décodeur : une application β qui associe à toute suite binaire u un symbole $a = \beta(u) \in A$, tel que pour tout $a \in A$, $\beta(\alpha(a)) = a$.*

Le codeur est ensuite étendu aux suites de symboles (les mots) par concaténation : la suite binaire associée au mot $x_1x_2 \dots x_n$ est la concaténation

$$\alpha(x_1x_2 \dots x_n) = \alpha(x_1)\alpha(x_2) \dots \alpha(x_n) \ .$$

La concaténation pose des problèmes si le code est un code de longueur variable. Dans le cas général, on ne sait pas où commencent et où s'arrètent les mots. On introduit donc une sous-classe de codes, appelés codes uniquement décodables, définis comme suit.

DÉFINITION 4.2. *Le code est dit uniquement décodable si le codeur α est injectif : c'est à dire si toute suite binaire $\alpha(x_1x_2 \dots x_n)$ n'est l'image que d'un et un seul mot, à savoir $x_1x_2 \dots x_n$.*

EXEMPLE 4.1. On considère les deux codeurs α_1 et α_2 et α_3 définis par les tableaux donnés dans la TABLE 1. On vérifie immédiatement que le code donné dans le tableau de gauche n'est pas uniquement décodable, alors que les deux autres le sont. Par contre, le code du tableau du milieu n'est pas *instantané*, au sens où il faut attendre le début du mot suivant pour savoir si un mot est terminé : 11 peut être suivi de 1, auquel cas il s'agit d'un a_4, ou d'un 0, auquel cas il s'agit d'un a_3. Le code présenté dans le tableau de droite est quant à lui un code uniquement décodable instantané.

input	code
a_0	0
a_1	10
a_2	101
a_3	0101

input	code
a_0	0
a_1	01
a_2	011
a_3	111

input	code
a_0	0
a_1	10
a_2	110
a_3	111

TAB. 1. Trois exemples de codeurs : celui de gauche n'est pas uniquement décodable, les deux autres le sont. Le codeur du centre n'est pas instantané.

2.2. Codes à préfixe, et codes associés à un arbre binaire. Il existe une façon simple d'assurer qu'un code est uniquement décodable et instantané. Il suffit d'imposer une condition, appelée *condition de préfixe*.

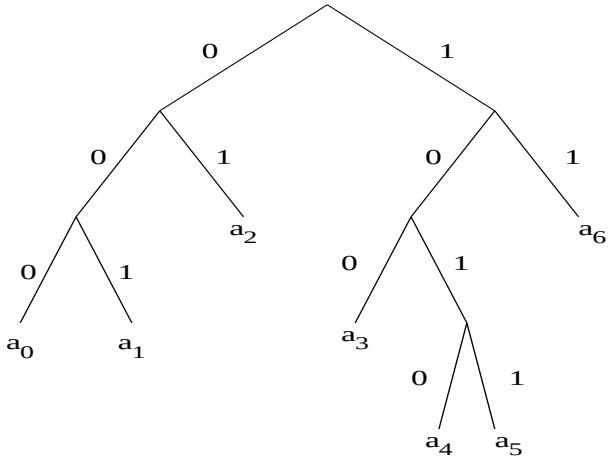

FIG. 1. Exemple d'arbre binaire associé à un alphabet.

input	a_0	a_1	a_2	a_3	a_4	a_5	a_6
code	000	001	01	100	1010	1011	11

TAB. 2. Code associé à l'arbre binaire

DÉFINITION 4.3. *Un codeur α satisfait la condition de préfixe si aucun mot binaire $\alpha(a), a \in A$ n'est préfixe d'un autre mot $\alpha(b), b \in A$. Un code satisfaisant la condition de préfixe est appelé* code à préfixe.

Il existe une construction générique de codes à préfixes. On peut associer un tel code à tout arbre binaire. Plus précisément, étant donné un alphabet $A = \{a_0, \dots a_{M-1}\}$ à M symboles, on considère un arbre binaire à M feuilles. On associe à chaque symbole $a \in A$ une des feuilles, et on numérote les branches de l'arbre par des bits : par exemple, les branches partant vers la gauche sont notées "0", et les branches partant vers la droite sont notées "1". Le code associé à chaque symbole est alors la concaténation des étiquettes des branches consécutives menant de la racine de l'arbre à la feuille considérée.

3. Codes entropiques

3.1. L'inégalité de Kraft. On s'intéresse donc maintenant à des codes de longueur variable. On a vu que dans le cas de codes de longueur constante, si l'alphabet a $M = 2^R$ symboles, chaque mot $\alpha(a)$ est de longueur $\ell(a) = R$, de sorte que

$$\sum_{a \in A} 2^{-\ell(a)} = \sum_{a \in A} 2^{-R} = 1 \ .$$

On va voir que cette relation est une borne pour les codes de longueur variable uniquement décodables.

THÉORÈME 4.1. *Soit $a = \{a_0, \dots a_{M-1}\}$ un alphabet à M symboles. Le codeur α, qui associe à chaque $a \in A$ le mot $\alpha(a)$ de longueur $\ell(a)$, est uniquement décodable seulement si l'inégalité suivante, appelée* inégalité de Kraft*, est vérifiée :*

$$\sum_{a\in A} 2^{-\ell(a)} \leq 1 \ . \tag{4.3}$$

En fait, l'inégalité de Kraft montre que pour une distribution de probabilités de symboles donnée, les longueurs de mots binaires associés à ces symboles doivent, *dans une certaine moyenne*, être supérieures à une valeurs seuil. *Preuve :* Soit K un entier positif fixé. On considère une suite de K symboles $b = \{b_0, \dots, b_{K-1}\}$; la longueur totale de $\alpha(b_0 \dots b_{K-1})$ vaut

$$\ell(b) = \ell(b_0) + \cdots + \ell(b_{K-1}) \ .$$

On notera $N(L)$ le nombre de suites de K symboles ayant une longueur totale égale à L. On pose $\ell_{max} = \max_{a\in A} \ell(a)$. On a alors

$$\ell(b) \leq K\ell_{max} := L_{max} \ .$$

Le code étant uniquement décodable, les $N(L)$ suites ont un code différent. Comme il y a au plus 2^L codes de longueur L différents, on a

$$N(L) \leq 2^L \ .$$

Calculons alors

$$\begin{aligned}
\left[\sum_{m=0}^{M-1} 2^{-\ell(a_m)}\right]^K &= \left[\sum_{a\in A} 2^{-\ell(a)}\right]^K \\
&= \sum_{b_0\in A}\sum_{b_1\in A}\cdots\sum_{b_{K-1}\in A} 2^{-\ell(b_0)-\cdots-\ell(b_{K-1})} \\
&= \sum_{L=1}^{L_{max}} N(L)2^{-L} \\
&\leq L_{max} = K\ell_{max} \ .
\end{aligned}$$

Par conséquent, on a

$$\sum_{m=0}^{M-1} 2^{-\ell(a_m)} \leq (K\ell_{max})^{1/K} \ .$$

Ceci étant vrai pour tout K, on obtient bien l'inégalité de Kraft par passage à la limite $K \to \infty$. Ceci conclut la preuve du théorème. ♠

Ce premier théorème fournit donc une condition nécessaire pour l'unique décodabilité d'un code. Cette condition est en fait également suffisante dans un certain sens, comme l'exprime le théorème suivant.

THÉORÈME 4.2. *Soit $a = \{a_0, \dots a_{M-1}\}$ un alphabet à M symboles, et soit $\{\ell_0, \dots, \ell_{M-1}\}$ une suite d'entiers positifs satisfaisant l'inégalité de Kraft (4.3). Alors il existe un codeur α uniquement décodable, tel que pour tout $k = 0, \dots M-1$, $\ell(a_k) = \ell_k$.*

3.2. L'entropie de Shannon. Considérons maintenant la longueur moyenne $\overline{\ell}(\alpha)$ des mots codés par un codeur α,

$$\overline{\ell}(\alpha) = \sum_{a\in A} p(a)\ell(a) \ . \tag{4.4}$$

et supposons que l'inégalité de Kraft soit vérifiée. On a alors

$$\begin{aligned}
\overline{\ell}(\alpha) &= -\sum_{a\in A} p(a)\log_2\left(2^{-\ell(a)}\right) \\
&\geq -\sum_{a\in A} p(a)\log_2\left(\frac{2^{-\ell(a)}}{\sum_{b\in A}2^{-\ell(b)}}\right) \\
&\geq -\sum_{a\in A} p(a)\log_2(q(a)) \ ,
\end{aligned}$$

où $q = \{q_0, \dots q_{M-1}\}$ est une autre distribution de probabilités, définie par

$$q_k = q(a_k) = \frac{2^{-\ell(a_k)}}{\sum_{b\in A}2^{-\ell(b)}} \ . \tag{4.5}$$

Notons que cette inégalité devient une égalité si l'inégalité de Kraft est une égalité, c'est à dire lorsque pour tout $a \in A$, $q(a) = 2^{-\ell(a)}$.

On peut introduire l'entropie de la distribution de probabilités p :

$$H(p) = -\sum_{a\in A} p(a)\log_2(p(a)) \ . \tag{4.6}$$

L'entropie associée à une distribution de probabilités représente la quantité d'information reçue lorsqu'un évènement $a \in \mathcal{A}$ est observé. L'entropie possède nombre de propriétés simples. En particulier, on a

LEMME 4.1.

$$0 \leq H(p) \leq \log_2\left(Card(A)\right) \ .$$

Preuve : L'entropie est positive par construction. Pour montrer la borne supérieure, il suffit de trouver la distribution de probabilités sur A qui maximise H (avec évidemment la contrainte $\sum_{a\in A} p(a) = 1$). En introduisant un multiplicateur de Lagrange pour imposer la contrainte, on se ramène à maximiser

$$-\sum_{a\in A} p(a)\log_2(p(a)) + \lambda\left(\sum_{a\in A} p(a) - 1\right)$$

par rapport aux nombres $p(a)$ et au multiplicateur λ. On voit facilement que la solution est donnée par $p(a) = 1/\mathrm{Card}(A)$, d'où le résultat. ♠

Ainsi, les faibles valeurs de l'entropie correspondent à des situations dans lesquelles un petit nombre de symboles sont très probables (donc pour lesquels $\log p(a)$ est faible) et un grand nombre de symboles sont très probables (donc pour lesquels $p(a)$ est faible).

On a aussi le résultat important suivant

LEMME 4.2. *Etant données deux distributions de probabilités p et q indexées par l'alphabet A, on a*

$$-\sum_{a\in A} p(a)\log_2(q(a)) \geq H(p) \ . \tag{4.7}$$

avec égalité si et seulement si les deux distributions sont identiques.

Preuve : On va utiliser le fait que $\log(x) \leq x - 1$, avec égalité si et seulement si $x = 1$. Calculons

$$\begin{aligned} \sum_{a\in A} p(a)\log_2\left(\frac{q(a)}{p(a)}\right) &= \frac{1}{\log(2)}\sum_{a\in A} p(a)\log\left(\frac{q(a)}{p(a)}\right) \\ &\leq \frac{1}{\log(2)}\sum_{a\in A} p(a)\left(\frac{q(a)}{p(a)} - 1\right) \\ &= 0\ , \end{aligned}$$

avec égalité si et seulement si $p(a) = q(a)$ pour tout $a \in A$. Ceci prouve le lemme. ♠

En combinant ce résultat avec le précédent, on a donc la première partie du théorème suivant :

THÉORÈME 4.3. *Soit $A = \{a_0, \ldots, a_{M-1}\}$ un alphabet, et soit p la distribution de probabilités de ses symboles.*

(1) *Soit α un codeur dont les longueurs de mots satisfont à l'inégalité de Kraft. Alors on a*

$$\overline{\ell}(\alpha) \geq H(p)\ , \tag{4.8}$$

avec égalité si et seulement si $p(a) = 2^{-\ell(a)}$ pour tout $a \in A$.

(2) *Il existe un code uniquement décodable tel que*

$$\overline{\ell}(\alpha) < H(p) + 1\ . \tag{4.9}$$

Preuve : La première partie est une conséquence directe du lemme :

$$\overline{\ell}(\alpha) \geq -\sum_{a\in A} p(a)\log_2 q(a) \geq H(p)\ ,$$

et on a égalité (i.e. les deux inégalités intervenant dans celle-ci sont des égalités) si et seulement si $p(a) = q(a) = 2^{-\ell(a)}$ pour tout a. Pour la seconde partie, étant donnée la distribution p, soient les nombres ℓ_k définis par

$$2^{-\ell_k} \leq p(a_k) < 2^{1-\ell_k}\ .$$

Il est évident que $\sum_k 2^{-\ell_k} \leq \sum_k p(a_k) = 1$, donc l'inégalité de Kraft est satisfaite. Il existe donc un code uniquement décodable α associé aux longueurs de mots ℓ_k. De plus, on a

$$\ell_k < 1 - \log_2(p(a_k))\ ,$$

et donc

$$\overline{\ell}(\alpha) = \sum_k p(a_k)\ell_k < \sum_k p(a_k)(1 - \log_2(p(a_k))) = 1 + H(p)\ .$$

Ceci conclut la démonstration. ♠

On donne maintenant sans démonstration quelques résultats importants. Une démonstration peut être trouvée dans [**7, 9**].

THÉORÈME 4.4. *Soit (α, β) un code de longueur variable, dont on note $\ell_k = \ell(a_k)$ les longueurs de mots. Il existe un code à préfixe qui a les mêmes longueurs de mots, et en particulier la même longueur moyenne $\overline{\ell}(\alpha)$.*

DÉFINITION 4.4. *Etant donné un alphabet A et une distribution de probabilités pour les symboles de A, on dit qu'un code α est optimal si il minimise la longueur moyenne des mots $\overline{\ell}(\alpha)$.*

THÉORÈME 4.5. *Soit α un code à préfixe optimal.*

(1) *Si les symboles a et b sont tels que $p(a) > p(b)$, alors $\ell(a) \leq \ell(b)$.*
(2) *Les deux symboles les moins probables sont associés à des mots de même longueur, qui ne diffèrent que par leur bit le moins significatif.*

3.3. Le code de Huffman. Le code de Huffman est un algorithme récursif qui permet d'atteindre les performances optimales. Il est basé sur la règle d'agrégation suivante.

Partant d'un alphabet $A = \{a_0, a_1, \dots a_{M-1}\}$ dont les symboles ont probabilités $(p(a_0), p(a_1), \dots p(a_{M-1}))$, on suppose (sans perte de généralité)que les symboles sont ordonnés par probabilités décroissantes : $p(a_0) \geq p(a_1) \geq \dots$ (si nécessaire, on peut les ré-ordonner). On construit un nouvel alphabet de longueur $M-1$ en agrégeant les symboles a_{M-1} et a_{M-2} en un nouveau symbole $\tilde{a}$ auquel on affecte la probabilité

$$p(\tilde{a}) = p(a_{M-1}) + p(a_{M-2}) \ .$$

THÉORÈME 4.6. *Un arbre de préfixe optimal pour l'alphabet à M symboles ordonnés $A = \{a_0, a_1, \dots a_{M-1}\}$ s'obtient à partir d'un arbre de préfixe optimal pour $\{a_0, a_1, \dots a_{M-3,\tilde{a}}\}$, en adjoignant à ce dernier deux branches liant $\tilde{a}$ aux symboles a_{M-2} et a_{M-1}.*

L'algorithme de Huffman exploite ce résultat de la façon suivante : a_{M-1} et a_{M-2} sont les feuilles extrêmes de l'arbre. Les deux branches correspondantes sont caractérisées par un bit.

On est donc en présence d'un nouvel alphabet de $M-1$ symboles

$$\{a_0, a_1, \dots a_{M-3}, \tilde{a}\} \ .$$

Pour poursuivre l'algorithme, il suffit alors de réordonner ce nouvel alphabet par ordre de probabilités décroissantes, et d'itérer la procédure : prendre les deux (nouveaux) symboles de probabilités minimales, et leur associer deux nouvelles branches.

On en déduit

COROLLAIRE 4.1. *Le code de Huffman est le code à préfixe optimal, et satisfait en particulier la borne (4.9).*

EXEMPLE 4.2. Prenons l'exemple suivant d'un alphabet de 8 symboles, de probabilités données dans la table 3.

L'arbre de préfixes correspondant se trouve en FIGURE 2. On peut à partir de cet exemple estimer la longueur moyenne des mots :

$$\overline{\ell}(\alpha) = 0.25{\times}2{+}0.2{\times}2{+}0.2{\times}2{+}0.18{\times}3{+}0.09{\times}4{+}0.05{\times}5{+}0.02{\times}6{+}0.01{\times}6 = 2.63 \ ,$$

ce qui représente un gain par rapport à un codeur uniforme qui aurait nécessité 3 bits par symbole.

REMARQUE 4.1. On a vu au chapitre précédent l'exemple du codage de plages, et de son application au codage des facsimilés. Un fax consiste essentiellement en des plages de pixels noirs et des plages de pixels blancs. La norme pour le codage des fax considère des plages de longueur inférieure à 64 pixels seulement, de sorte qu'il est possible d'avoir à coder plusieurs plages blanches (ou noires) consécutives. Les distributions de probabilités des longueurs de plages blanches et noires ont été estimées expérimentalement, et ont conduit à un code de Huffman "standard" pour les longueurs de plages blanches et noires. Pour plus de détails, on pourra se référer à [**7**].

symbole	probabilité	code
a	0.01	110000
b	0.02	110001
c	0.05	11001
d	0.09	1101
e	0.18	111
f	0.2	00
g	0.2	01
h	0.25	10

TAB. 3. Exemple de code de Huffman.

REMARQUE 4.2. Il arrive souvent que le code optimal ne soit pas unique : pour une distribution de probabilités donnée, l'algorithme de Huffman peut laisser une certaine liberté dans les mots binaires à associer aux symboles (typiquement, lorsque deux probabilités sont égales). Il est alors facile de voir que bien que les codes soient différents, la longueur moyenne des mots $\overline{\ell}$ est la même, ce qui est la seule chose réellement importante.

REMARQUE 4.3. Il existe des alternatives au codage de Huffman. On peut en particulier mentionner le *codage arithmétique*, qui permet d'associer, dans un certain sens, des "nombre de bits non entiers" aux symboles de A. On montre que ceci permet d'améliorer légèrement les performances d'un codeur.

3.4. Codage par blocs. On a vu que le codage entropique (par exemple le code de Huffman) permet d'atteindre une longueur moyenne de mots presque optimale (essentiellement, comprise entre l'entropie et l'entropie +1). Cependant, lorsque l'entropie est faible, on peut ne plus s'en satisfaire. Le codage par blocs permet d'affiner le résultat.

Supposons que l'on ait à coder une suite de variables aléatoires i.i.d. (typiquement, des coefficients issus d'un codeur par transformation, quantifiés). L'idée est alors de ne plus coder les variablea aléatoires individuellement, mais plutôt de les regrouper par blocs de N variables aléatoires (formant donc des vecteurs aléatoires, dont les composantes sont i.i.d.). On peut facilement montrer le résultat suivant :

LEMME 4.3. *Soient $X_1, \dots X_N$ des variables aléatoires i.i.d., dont on note p la distribution de probabilités. Soit P la distribution de probabilités jointe de $X_1, \dots X_N$ (la loi produit). Alors on a*

$$H(P) = NH(p) \ . \tag{4.10}$$

On peut alors adopter la stratégie simple suivante. Etant donné un signal donc on connait des coefficients quantifiés $X_1, X_2, \dots$, on commence par les regrouper en

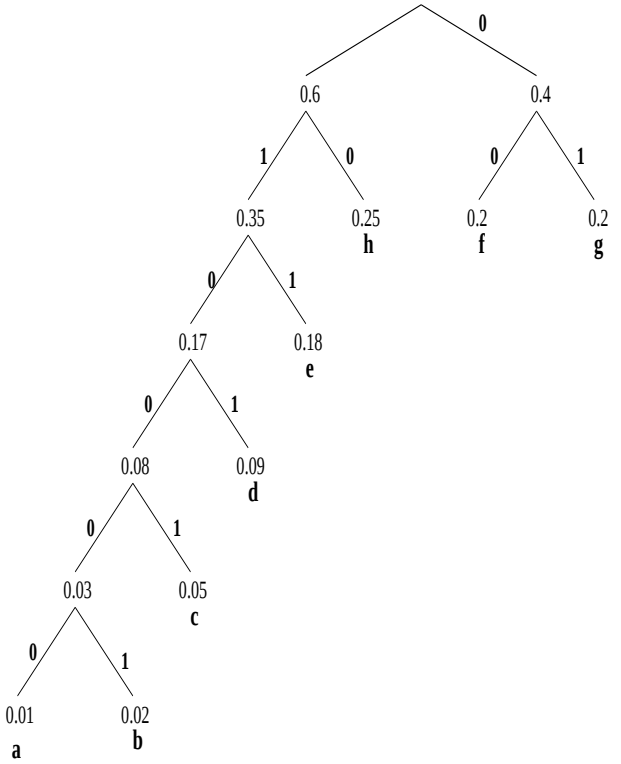

FIG. 2. Arbre de préfixes pour les symboles de la Table 3

blocs de N coefficients. A partir de la loi de ces blocs, on peut facilement construire un code entropique (le code de Huffman par exemple), qui fournit une longueur de mots moyenne telle que

$$\overline{\ell}_{\text{blocs}} < H(P) + 1 \ .$$

De là, si les blocs sont effectivement constitués de coefficients indépendants identiquement distribués, on en déduit que la longueur moyenne des mots associés aux coefficients (et non plus aux blocs) est telle que

$$\overline{\ell}_{\text{coeffs}} < H(p) + \frac{1}{N} \ .$$

Ainsi, en considérant des blocs de coefficients plutôt que des coefficients individuels, on peut s'approcher de la borne optimale. Par contre, il est clair que les algorithmes de codage et de décodage correspondants sont plus complexes.

4. Quantification non-uniforme et codage entropique

Dans les chapitres précédents, on n'avait abordé la quantification que dans une perspective de code de longueur constante. Il est maintenant utile de revenir sur ce problème de quantification à la lumière de ce que nous avons vu. On se limitera aux codes "haute résolution", ou "haut débit".

Supposons donc (pour simplifier) qu'on ait à quantifier une variable aléatoire de densité ρ_X à support borné (pour éviter d'avoir à considérer le bruit de saturation).

On considère donc comme précédemment les valeurs de quantification $y_0 < y_1 < \cdots < y_{M-1}$ et les bornes $x_0, \dots x_M$, telles que

$$Q(x) = y_k \quad \text{si } x \in [x_k, x_{k+1}[\ .$$

Après quantification, les "intervalles de quantification" $[x_k, x_{k+1}[$ ont probabilité

$$p_k = \int_{x_k}^{x_{k+1}} \rho_X(x)\, dx \ .$$

La distorsion s'écrit alors

$$D = \sum_{k=0}^{M-1} \int_{x_k}^{x_{k+1}} (x - y_k)^2\, \rho_X(x)\, dx \ .$$

Si on suppose comme précédemment que ρ_X varie peu dans chacun des intervalles $[x_k, x_{k+1}[$ (hypothèse de haute résolution), et si on pose

$$\Delta_k = x_{k+1} - x_k \ ,$$

et si on choisit $y_k \in [x_k, x_{k+1}[$ (par exemple $y_k = \frac{x_k + x_{k+1}}{2}$), on peut approximer, au premier ordre

$$p_k \approx \rho_X(y_k)\Delta_k \ ,$$

et

$$D \approx \sum_{k=0}^{M-1} \rho_X(y_k)\frac{\Delta_k^3}{12} \ .$$

Si on s'intéresse maintenant à l'entropie

$$\begin{aligned} H = -\sum_k p_k \log p_k &\approx -\sum_k \rho_X(y_k)\Delta_k \log \rho_X(y_k)\Delta_k \\ &\approx -\int \rho_X(x) \log \rho_X(x)\, dx - \frac{1}{2}\sum_k p_k \log \Delta_k^2 \ . \end{aligned}$$

Le premier terme est une caractéristique de la distribution de probabilités initiale de la variable aléatoire X, est appelé "entropie différentielle" de la source :

$$h(X) = -\int \rho_X(x) \log \rho_X(x)\, dx \ , \tag{4.11}$$

et caractérise la "concentration" de la variable aléatoire X.

Pour ce qui est du second terme, on peut utiliser la concavité du logarithme, qui donne l'inégalité de Jensen

$$\sum_k p_k \log \Delta_k^2 \le \log\left(\sum_k p_k \Delta_k^2\right) \ ,$$

et approximer

$$\frac{1}{2}\sum_k p_k \log \Delta_k^2 \le \frac{1}{2}\log(12\, D) \ .$$

Mettant ensembles ces estimations, on aboutit à l'inégalité

$$H \ge h(X) - \frac{1}{2}\log(12\, D) \ ,$$

ou en d'autres termes

$$D \ge \frac{1}{12}\, 2^{-2[H - h(X)]} \ . \tag{4.12}$$

Ceci nous fournit une estimation de la distorsion minimale atteignable par le codeur considéré. On voit donc que pour améliorer les performances (optimales) d'un codeur (en termes de distorsion), on a intérêt à maximiser la différence entre l'entropie H et l'entropie différentielle de la source. En particulier, le résultat sera d'autant plus favorable que l'entropie différentielle $h(X)$ (qui caractérise la variabilité intrinsèque de la source) sera faible.

Cette expression est à comparer à l'expression analogue que nous avions obtenue dans le cas de la quantification uniforme :

$$D \approx C2^{-2R} ,$$

avec laquelle elle coïncide dans le cas d'un code de longueur constante (toujours sous des hypothèses "haute résolution").

Références Bibliographiques

[1] I. Daubechies (1992) : *Ten Lectures on Wavelets.* Vol. 61, CBMS-NFS Regional Series in Applied Mathematics.

[2] J.I. Doob (1953) : *Stochastic Processes*, Wiley, New York.

[3] C. Gasquet et P. Witomski (1990) : *Analyse de Fourier et Applications*, Editions Masson, Paris.

[4] I.M. Gelfand et G.E. Shilov : *Les distributions*

[5] A. Gersho et D. Gray (1992) : Vector quantization

[6] B.B. Hubbard (1996) : *Ondelettes : la saga d'un outil mathématique*, Editions *Pour la Science.*

[7] N.S. Jayant et P. Noll (1984) : *The digital coding of waveforms*, Prentice Hall.

[8] J. Lamperti (1977) : *Stochastic Processes : a survey of the Mathematical Theory.* Applied Mathematical Sciences **23**, Springer Verlag.

[9] S. Mallat (1998) : *A Wavelet Tour of Signal Processing.* Academic Press, New York, N.Y.

[10] A. Papoulis *Signal Processing.* McGraw&Hill, New York.

[11] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T Wetterling (1986) : *Numerical Recipes.* Cambridge Univ. Press, Cambridge, UK.

[12] F. Riesz et B. Nagy (1955) : *Leçons d'Analyse Fonctionnelle*, Gauthier-Villars.

[13] W. Rudin : *Analyse rélle et complexe.*, McGraw et Hill.

[14] C. Soize (1993) : *Méthodes mathématiques en traitement du signal.* Masson, Paris.

[15] M. Vetterli and J. Kovacevic (1996) : *Wavelets and SubBand Coding*, Prentice Hall, Englewood Cliffs, NJ.

[16] M.V. Wickerhauser (1994) : *Adapted Wavelet Analysis, from Theory to Software.* A.K. Peters Publ.