FILTRES NUMÉRIQUES

G. Baudoin
École Supérieure d'Ingénieurs en Électronique et Électrotechnique

---

Octobre 2006 – version 2.0

Ce polycopié constitue le support du cours de filtres numériques de l'ESIEE. Il correspond en partie à une refonte des anciens documents dont certains étaient en word (chapitres 2 à 4), d'autres en latex (chapitre 1) et à une nouvelle rédaction pour le dernier chapitre (chapitre 5).
Il s'agit donc d'une version largement améliorable. Toutes les remarques de forme (typographie, présentation, orthographe) et de fond (clarté, enchaînements, rappels nécessaires, etc) seront bienvenues.

G. Baudoin,
Octobre 2006

Chapitre I

Table des matières

Chapitre I

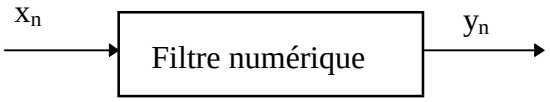
INTRODUCTION AU FILTRAGE NUMÉRIQUE

Ce chapitre définit la notion de filtrage numérique et présente les propriétés générales des filtres numériques. Il étudie par ailleurs les cellules de filtrage élémentaires d'ordre 1 et 2.

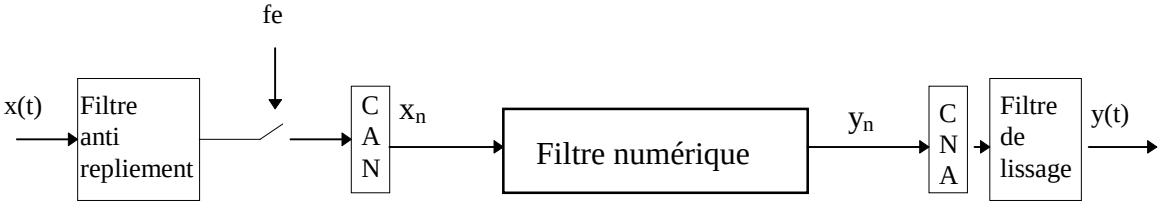
On appelle filtre numérique un système discret linéaire et invariant en temps, opérant sur des signaux discrets.

Un signal discret (x_n) est une suite prenant ses valeurs dans R, le plus souvent, parfois dans C.

Les premiers filtres numériques ont servi à simuler des filtres analogiques sur ordinateurs. Les suites d'entrée et de sortie du filtre sont alors dépourvues de tout support physique.



Puis on a réalisé des systèmes discrets travaillant sur des signaux physiques échantillonnés et numérisés. Dans ce cas, en notant T_e la période d'échantillonnage, les valeurs de la suite (x_n) valent : $x_n = x(nT_e)$.



Il faut alors que les performances attendues du filtre numérique soient cohérentes avec la précision de la conversion analogique numérique.

Si f_e est peu supérieure à $2f_{max}$ (f_{max} étant la fréquence maximale du signal x(t)), le filtre antirepliement de spectre peut être assez difficile à réaliser. Il pourra alors être intéressant de suréchantillonner l'entrée analogique, ce qui simplifie le filtre antirepliement analogique en élargissant sa bande de transition. On sous-échantillonnera ensuite le signal numérique en lui appliquant un filtre antirepliement numérique.

Les principaux avantages des filtres numériques sont les suivants :

- Ils sont reproductibles sans réglages,
- Ils sont programmables,
- Ils ne dérivent pas, ni en temps ni en température,
- On peut les rendre facilement adaptatifs (dans ce cas là, le système n'est plus invariant en temps),
- Ils permettent de réaliser des filtres à phase parfaitement linéaire,

Leurs principaux inconvénients sont :

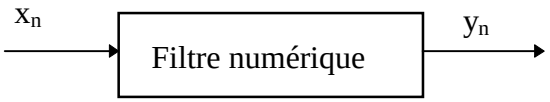
- Leur consommation (par comparaison aux circuits analogiques passifs),
- Leur limitation en fréquence : ils sont limités par la fréquence des convertisseurs analogiques numériques et par la vitesse des opérateurs de calcul numérique comme les multiplieurs.
- Leur coût (mais ce dernier point n'est pas toujours vrai).

Les fonctions des filtres numériques sont analogues à celles des filtres analogiques. On les utilise en général dans le but d'atténuer une ou plusieurs bandes de fréquences. On parle de :

- filtres passe-bas quand on atténue les hautes fréquences,
- filtres passe-haut quand on atténue les basses fréquences,
- filtres coupe-bande quand on atténue une bande de fréquences,
- filtres passe-bande quand on favorise une bande de fréquences.

Un autre rôle des filtres est de corriger la fonction de transfert d'un canal qui introduit de la distortion. Dans ce cas le canal est dit dispersif et le filtre est appelé égaliseur.

1 Systèmes linéaires discrets invariants en temps



1.1 Définition

Un filtre numérique est un système discret linéaire invariant en temps. Il associe à la suite d'entrée (x_n) une suite de sortie (y_n).

1.1.1 Linéarité

Soit 2 suites $x_1(n)$ et $x_2(n)$ avec les sorties correspondantes $y_1(n)$ et $y_2(n)$. Dire que le système est linéaire signifie que :

$$\begin{cases} \forall \lambda_1 \in R \\ \forall \lambda_2 \in R \end{cases} \qquad \lambda_1 x_1(n) + \lambda_2 x_2(n) \rightarrow \lambda_1 y_1(n) + \lambda_2 y_2(n)$$

1.1.2 Invariance en temps

Soit la suite $x(n)$ et la sortie correspondante $y(n)$, dire que le système est invariant en temps signifie qu'à la suite $x(n - n_0)$ correspond la sortie $y(n - n_0)$, et ceci quelque soit n_0.

1.2 Réponse impulsionnelle

La réponse impulsionnelle d'un système discret linéaire invariant en temps est la réponse du filtre à une entrée impulsion u_n. Cette réponse impulsionnelle est généralement notée h_n.

La suite impulsion u_n est définie par :

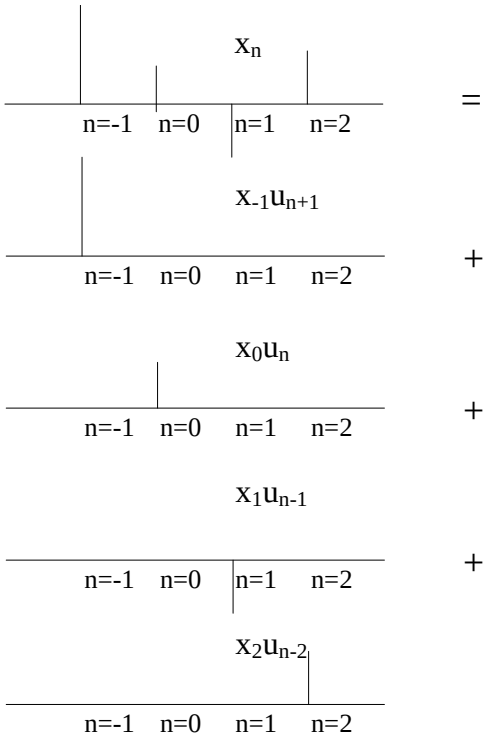
$$\begin{cases} \forall n < 0 \quad u_n = 0 \\ \qquad u_0 = 1 \\ \forall n > 0 \quad u_n = 0 \end{cases}$$

1.3 Relation entrée-sortie, convolution discrète

Soit une entrée quelconque x_n, on va montrer que la sortie correspondante y_n peut s'exprimer comme une convolution discrète de l'entrée x_n et de la réponse impulsionnelle h_n.

Toute suite x_n, peut s'écrire à l'aide de suites impulsionnelles u_{n-k} sous la forme :

$$x_n = \sum_{k=-\infty}^{+\infty} x_k u_{n-k}$$



D'après l'invariance en temps du filtre, la sortie correspondant à une entrée u_{n-k} (notée sortie(u_{n-k})) est la suite h_{n-k}.

D'après la linéarité du filtre, la sortie du filtre y_n pour une entrée x_n vaut :

$$y_n = \sum_{k=-\infty}^{+\infty} x_k \text{sortie}(u_{n-k}) = \sum_{k=-\infty}^{+\infty} x_k h_{n-k}$$

D'où :

$$\boxed{y_n = \sum_{k=-\infty}^{+\infty} x_k h_{n-k} = \sum_{k=-\infty}^{+\infty} h_k x_{n-k}}$$

La relation liant h_n et x_n est appelée convolution discrète de la suite x_n avec la suite h_n.

La sortie d'un filtre est donc le produit de convolution de l'entrée du filtre avec la réponse impulsionnelle du filtre.

En analogique, la convolution n'avait pas d'utilité pratique pour la réalisation d'un filtre. En numérique, cette relation sera, dans certains cas, mise en œuvre explicitement pour la réalisation du filtre.

1.4 Réponse en fréquence

1.5 Réponse à une entrée fréquence pure

Soit une entrée x_n ne contenant qu'une fréquence f_0 (soit une seule pulsation $\omega_0 = 2\pi f_0$).

$$x_n = e^{j\omega_0 n T_e}$$

On a supposé ici que x_n était la suite obtenue par échantillonnage à la période T_e du signal analogique $x(t) = e^{j\omega_0 t}$, mais on pourrait raisonner de même sur une suite numérique $x_n = e^{j\omega_0 n}$.

La sortie y_n correspondante se calcule par la relation de convolution :

$$y_n = \sum_{k=-\infty}^{+\infty} x_k h_{n-k} = \sum_{k=-\infty}^{+\infty} h_k x_{n-k} = \sum_{k=-\infty}^{+\infty} h_k e^{j\omega_0 (n-k) T_e} = e^{j\omega_0 n T_e} \sum_{k=-\infty}^{+\infty} h_k e^{-j\omega_0 k T_e}$$

$$y_n = x_n H(\omega_0)$$

Avec :

$$\boxed{H(\omega_0) = \sum_{k=-\infty}^{+\infty} h_k e^{-j\omega_0 k T_e}}$$

Pour une entrée ne contenant qu'une fréquence, la sortie est proportionnelle à l'entrée, c'est à dire que l'on retrouve la même fréquence à la sortie du filtre. Le coefficient de proportionnalité $H(\omega_0)$ est complexe. Son module est le coefficient qui multiplie l'amplitude réelle de l'entrée, et son argument est l'angle de déphasage de l'entrée.

$$H(\omega_0) = |H(\omega_0)|\, e^{j \arg(H(\omega_0))} = A(\omega_0) e^{j\Phi(\omega_0)}$$

Et pour $x_n = e^{j\omega_0 n T_e}$, la sortie s'écrit $y_n = A(\omega_0) e^{j(\omega_0 n T_e + \Phi(\omega_0))}$.

Les suites $x_n = e^{j\omega_0 n T_e}$ sont les fonctions propres des filtres numériques.

La fonction $H(\omega)$ est la transformée de Fourier à temps discret de la suite h_n. Elle est appelée la fonction de transfert en fréquence du filtre.

$H(\omega)$ est une fonction périodique de période $2\pi/T_e$. On notera $f_e = 1/T_e$.

L'inverse d'une transformée de Fourier à temps discret H() est donnée par :

$$\boxed{h_n = \frac{1}{2\pi f_e} \int_{-\pi f_e}^{\pi f_e} H(\omega) e^{j\omega n T_e} d\omega}$$

1.5.1 Relation entre les transformées de Fourier de l'entrée et de la sortie

Soit une suite d'entrée x_n quelconque et sa transformée de Fourier $X(\omega)$. La sortie y_n a pour transformée de Fourier $Y(\omega)$.

$$y_n = \sum_{k=-\infty}^{+\infty} h_k x_{n-k}$$

$$X(\omega) = \sum_{n=-\infty}^{+\infty} x_n e^{-j\omega n T_e} \quad Y(\omega) = \sum_{n=-\infty}^{+\infty} y_n e^{-j\omega n T_e}$$

$$Y(\omega) = \sum_{n=-\infty}^{+\infty} y_n e^{-j\omega n T_e} = \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} h_k x_{n-k} e^{-j\omega n T_e} = \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} h_k x_{n-k} e^{-j(n-k)T_e} e^{-j\omega k T_e}$$

$$Y(\omega) = \sum_{k=-\infty}^{+\infty} h_k e^{-j\omega k T_e} \left(\sum_{n-k=-\infty}^{+\infty} x_{n-k} e^{-j\omega(n-k)T_e} \right) = H(\omega) X(\omega)$$

$$\boxed{Y(\omega) = H(\omega) X(\omega)}$$

On peut donc caractériser un filtre numérique soit par sa réponse impulsionnelle h_n soit par sa fonction de transfert en fréquence $H(\omega)$, qui est la transformée de Fourier de la suite h_n.

En pratique on s'intéressera souvent au module de la fonction de transfert H, ce module caractérisant l'atténuation ou le gain du filtre pour chaque fréquence.

Le déphasage introduit par le filtre pour chaque fréquence sera noté :

$$\Phi(\omega) = \arg\left(H\left(\omega\right)\right)$$

Lorsque ce déphasage est proportionnel à ω : $\Phi(\omega) = k\omega T_e$ alors le filtre ne déforme pas les signaux dans la bande passante.

En effet soit par exemple :

$$x_n = a_1 \cos(2\pi f_1 n T_e + \Phi_1) + a_2 \cos(2\pi f_2 n T_e + \Phi_2)$$

En supposant que f_1 et f_2 soient dans la bande passante avec un gain unité pour la fonction de transfert, on peut écrire :

$$\begin{aligned} y_n &= a_1 \cos(2\pi f_1 n T_e + \Phi_1 + \Phi(f_1)) + a_2 \cos(2\pi f_2 n T_e + \Phi_2 + \Phi(f_2)) \\ y_n &= a_1 \cos(2\pi f_1 n T_e + \Phi_1 + 2k\pi f_1 T_e) + a_2 \cos(2\pi f_2 n T_e + \Phi_2 + 2k\pi f_2 T_e) \\ y_n &= a_1 \cos(2\pi f_1 T_e (n+k) + \Phi_1) + a_2 \cos(2\pi f_2 T_e (n+k) + \Phi_2) = x(n+k) \end{aligned}$$

La sortie est donc égale à l'entrée retardée de -k échantillons.

On appelle temps de propagation de groupe cette valeur $-kT_e$.

Dans le cas général, on définit le **retard de groupe** $\tau(\omega)$ par :

$$\tau(\omega) = -\frac{\partial \Phi(\omega)}{\partial \omega}$$

Le retard de groupe représente un retard introduit par le filtre pour chaque fréquence.
Lorsque la phase est linéaire, le retard de groupe est constant.

Dans de nombreuses applications, en particulier en transmission de données il est important de ne pas déformer le signal utile et on utilisera des filtres à phase linéaire.

Dans d'autres cas, on cherchera à obtenir un retard de groupe aussi petit que possible.

1.6 Fonction de transfert en z

1.6.1 Définition

On étudie les propriétés d'un filtre analogique à l'aide de H(p), sa fonction de transfert en p, qui est la transformée de Laplace de sa réponse impulsionnelle.

De même, on étudie les propriétés d'un filtre numérique, à l'aide de H(z), sa fonction de transfert en z, qui est la transformée en z de sa réponse impulsionnelle h_n.

$$H(z) = \sum_{n=-\infty}^{+\infty} h_n z^{-n}$$

On remarquera que $H(\omega)$ est la restriction de H(z) au cercle unité (|z|=1) :

$$H(\omega) = \sum_{n=-\infty}^{+\infty} h_n e^{-j\omega n T_e} = H(z)|_{z=e^{j\omega T_e}}$$

1.6.2 Relation entre les transformées en z de l'entrée et de la sortie d'un filtre

On rappelle que la trasnformée en z d'une suite x_n est définie comme la limite d'une série de Laurent lorsque cette série converge.

On notera respectivement X(z), Y(z) et H(z) les transformées en z de l'entrée x_n, de la sortie y_n et de la réponse impulsionnelle h_n.

$$X(z) = \sum_{n=-\infty}^{+\infty} x_n z^{-n} \quad H(z) = \sum_{n=-\infty}^{+\infty} h_n z^{-n} \quad Y(z) = \sum_{n=-\infty}^{+\infty} y_n z^{-n}$$

$$y_n = \sum_{k=-\infty}^{+\infty} h_k x_{n-k}$$

$$\begin{aligned} Y(z) &= \sum_{n=-\infty}^{+\infty}\left(\sum_{k=-\infty}^{+\infty} h_k x_{n-k}\right) z^{-n} = \sum_{n=-\infty}^{+\infty}\left(\sum_{k=-\infty}^{+\infty} h_k x_{n-k}\right) z^{-k} z^{-(n-k)} \\ Y(z) &= \sum_{k=-\infty}^{+\infty} h_k z^{-k}\left(\sum_{m=n-k=-\infty}^{+\infty} x_m z^{-m}\right) = X(z)H(z) \\ Y(z) &= X(z)H(z) \end{aligned}$$

En résumé pour un filtre numérique, on peut écrire :

$$\boxed{\begin{array}{l} y_n = \sum\limits_{k=-\infty}^{k=+\infty} x_k h_{n-k} = \sum\limits_{k=-\infty}^{k=+\infty} h_k x_{n-k} \\ \\ Y(\omega) = H(\omega)X(\omega) \\ \\ Y(z) = X(z)H(z) \end{array}}$$

2 Quelques rappels sur la transformée en z

2.1 Domaine de convergence

Le critère de Cauchy permet d'étudier l'existence de la transformée en z :

$$\left|\sum_{n=0}^{+\infty} x_n\right| < \sum_{n=0}^{+\infty} |x_n| < +\infty \quad \text{si} \quad \lim_{n\to\infty} |x_n|^{\frac{1}{n}} < 1$$

Séquences causales :

Une séquence x_n est dite causale si elle est nulle pour $n < 0$.

Pour une séquence causale, la transformée en z est monolatérale :

$$X(z) = \sum_{n=0}^{+\infty} x_n z^{-n}$$

X(z) existe si :

$$\lim_{n\to\infty} \left|x_n z^{-n}\right|^{\frac{1}{n}} < 1 \quad \text{c'est à dire si} \quad \lim_{n\to\infty} |x_n|^{\frac{1}{n}} |z|^{-1} < 1 \quad \text{soit} \quad |z| > R_+ = \lim_{n\to\infty} |x_n|^{\frac{1}{n}}$$

La transformée en z d'une suite causale est donc définie à l'extérieur d'un cercle de rayon R_+.

Séquences anticausales :

Une séquence x_n est dite anticausale si elle est nulle pour $n \geq 0$.

Pour une séquence anticausale, la transformée en z s'écrit:

$$X(z) = \sum_{n=-\infty}^{-1} x_{-n} z^{-n} = \sum_{n=1}^{+\infty} x_{-n} z^{n}$$

X(z) existe si :

$$\lim_{n\to\infty} |x_{-n} z^{n}|^{\frac{1}{n}} < 1 \quad \text{c'est à dire si} \quad \lim_{n\to\infty} |x_{-n}|^{\frac{1}{n}} |z| < 1 \quad \text{soit} \, |z| < R_- = \lim_{n\to\infty} |x_{-n}|^{-\frac{1}{n}}$$

La transformée en z d'une suite anticausale est donc définie à l'intérieur d'un cercle de rayon R_-.

Suite quelconque

Une suite quelconque x_n est la somme d'une suite causale x^+ et d'une suite anticausale x^- :

$$x_n = x_n^+ + x_n^- \quad \text{avec :}$$

$$\begin{array}{lll} x_n^+ = x_n & \text{si} & n \geq 0 \\ x_n^+ = 0 & \text{si} & n < 0 \end{array} \quad \text{et} \quad \begin{array}{lll} x_n^- = x_n & \text{si} & n < 0 \\ x_n^+ = 0 & \text{si} & n \geq 0 \end{array}$$

La transformée en z d'une suite quelconque est donc définie dans une couronne :$R_+ < |z| < R_-$, quand cette couronne existe.

2.2 Linéarité

$$\forall \lambda_1 \quad \text{et} \quad \forall \lambda_2 \quad TZ(\lambda_1 x_1(n) + \lambda_2 x_2(n)) = \lambda_1 TZ(x_1(n)) + \lambda_2 TZ(x_2(n))$$

2.3 Théorème du retard

Cas de la transformée en z bilatérale $X(z) = \sum_{n=-\infty}^{+\infty} x_n z^{-n}$

$$TZ(x_n) = X(z) \Rightarrow TZ(x_{n-k}) = z^{-k} X(z)$$

Cas de la transformée en z monolatérale $X(z) = \sum_{n=0}^{+\infty} x_n z^{-n}$

Il faut alors tenir compte des conditions initiales.

$$\begin{aligned} TZ(x_n) &= X(z) \Rightarrow \\ TZ(x_{n-k}) &= z^{-k}\left(\sum_{n=0}^{+\infty} x_{n-k} z^{-(n-k)}\right) = z^{-k}\left(\sum_{m=-k}^{+\infty} x_m z^{-m}\right) = z^{-k}X(z) + z^{-k}\left(\sum_{m=-k}^{-1} x_m z^{-m}\right) \\ TZ(x_{n-k}) &= z^{-k}X(z) + \left(\sum_{n=1}^{k} x_{-n} z^{n-k}\right) \end{aligned}$$

2.4 Théorème de la convolution

Soit z_n la suite obtenue par convolution de 2 suites x_n et y_n : $z_n = x_n * y_n$

on utilise les notations : $X(z) = TZ(x_n) \quad Y(z) = TZ(y_n) \quad Z(z) = TZ(z_n)$ Où TZ signifie transformée en z monolatérale.

Et on montre facilement que :

$$Z(z) = X(z)Y(z)$$

2.5 Théorème de Parseval

$$\sum_{n=-\infty}^{+\infty} |x_n|^2 = \frac{1}{2j\pi} \int_{\text{Cercle unité}} X(z)X(z^{-1})z^{-1}dz = \frac{1}{f_e} \int_{-\frac{f_e}{2}}^{\frac{f_e}{2}} |X(f)|^2 df$$

2.6 Théorème de la valeur initiale et de la valeur finale

$$\begin{aligned} \lim_{n\to 0} x_n &= \lim_{z\to+\infty} X(z) \text{pour une suite causale} \\ \lim_{n\to+\infty} x_n &= \lim_{z\to 1}(1 - z^{-1})X(z) \end{aligned}$$

2.7 Intégration et dérivation

$$\begin{aligned} TZ\left(\sum_{i=-\infty}^{n} x_i\right) &= \frac{1}{1-z^{-1}} X(z) \\ TZ\left(x_n - x_{n-1}\right) &= (1 - z^{-1})X(z) \end{aligned}$$

2.8 Inversion de la transformée en z

L'inverse de la transformée en z est donnée par :

$$\boxed{x_n = \frac{1}{2j\pi} \int_{\text{Cercle unité}} X(z) z^{n-1} dz}$$

Pour montrer cette relation, on calcule d'abord l'intégrale :

$$\int_{\mathcal{C}} z^{-n} dz \text{ où } \mathcal{C} \text{ représente le cercle unité.}$$

En remplaçant z par $z = re^{j\theta}$, avec r=1, et dz par $dz = e^{j\theta} d\theta$, l'intégrale devient :

$$\int_{\mathcal{C}} z^{-n} dz = j \int_0^{2\pi} e^{-jn\theta} e^{j\theta} d\theta$$

Si $n \neq 1$ alors :

$$j \int_0^{2\pi} e^{-jn\theta} e^{j\theta} d\theta = j \frac{e^{-j(n-1)2\pi} - 1}{j(1-n)} = 0$$

Et si n=1

$$j \int_0^{2\pi} e^{-j\theta} e^{j\theta} d\theta = j \int_0^{2\pi} d\theta = 2j\pi$$

En conclusion :

$$\boxed{\begin{array}{lcl} \int_{\mathcal{C}} z^{-n} dz & = & 2j\pi \quad \text{si} \quad n = 1 \\ \int_{\mathcal{C}} z^{-n} dz & = & 0 \quad \text{si} \quad n \neq 1 \end{array}}$$

D'où pour X(z) :

$$\int_{\text{cercle unité}} X(z) z^{n-1} dz = \int_{\text{Cercle unité}} \sum_{k=-\infty}^{+\infty} x_k z^{-k} z^{n-1} dz = \sum_{k=-\infty}^{+\infty} x_k \int_{\text{Cercle unité}} z^{n-k-1} dz = 2j\pi \, x_n$$

D'où :

$$x_n = \frac{1}{2j\pi} \int_{\text{cercle unité}} X(z) z^{n-1} dz$$

3 Fonctions de transfert rationnelles en z, FIR, IIR

En général, on utilisera des filtres dont la fonction de transfert est rationnelle. H(z) s'écrit alors comme le rapport d'un numérateur N(z) et d'un dénominateur D(z) polynŞmes en z^{-1}.

$$\boxed{H(z) = \frac{N(z)}{D(z)} = \frac{\sum_{i=0}^{Q} b_i z^{-i}}{1 + \sum_{k=1}^{P} a_k z^{-k}}}$$

On a normalisé a_0 à 1.

Dans la suite de ce polycopié, on se limitera, sauf exception, à ces filtres.

Ces filtres sont simples à réaliser. La sortie y_n s'écrit en fonction des entrées et des sorties précédentes à l'aide d'une équation de récurrence à coefficients constants.

En effet :

$$H(z) = \frac{N(z)}{D(z)} = \frac{Y(z)}{X(z)} \Rightarrow Y(z) D(z) = X(z) Nz)$$

$$Y(z) + \sum_{k=1}^{P} a_k z^{-k} Y(z) = \sum_{i=0}^{Q} b_i z^{-i} X(z)$$

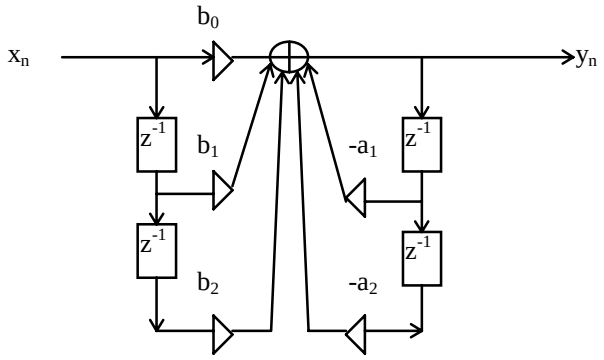
D'où par transformée en z inverse :

$$y_n + \sum_{k=1}^{P} a_k y_{n-k} = \sum_{i=0}^{Q} b_i x_{n-i}$$

$$\boxed{y_n = \sum_{i=0}^{Q} b_i x_{n-i} - \sum_{k=1}^{P} a_k y_{n-k}}$$

Cette équation de récurrence à coefficients constants permet de réaliser simplement le filtre. Il y a bien sûr d'autres structures de réalisation possibles, mais elles utilisent, de même, des multiplieurs, des additionneurs et des mémoires.

La figure suivante représente une cellule de filtrage d'ordre 2 (P=Q=2). Sur la figure, les rectangles entourant z^{-1} représentent un retard d'un échantillon.



On aurait pu considérer, de façon plus générale des fonctions de transfert H(z) rationnelles contenant des puissances de z positives, ce qui correspondrait à des filtres non causaux (voir définition ci-dessous).

On distingue 2 types de filtres : les filtres RII à Réponse Impulsionnelle Infinie et les filtres RIF à Réponse Impulsionnelle Finie.

Les filtres RII ont la forme générale $H(z) = \frac{N(z)}{D(z)}$ avec $D(z) \neq 1$.

Les filtres RII sont dits récursifs car le calcul de la sortie y_n à l'instant n fait intervenir les valeurs de P sorties précédentes y_{n-k}.
Les filtres RII sont caractérisés par les pôles et les zéros de leur fonction de transfert H(z).

Les filtres RIF ont un dénominateur D(z) égal à 1. Ils sont caractérisés par leurs seuls zéros.
La réponse impulsionnelle des filtres RIF est de longueur finie (d'où leur nom) puisque :

$$H(z) = \sum_{i=0}^{Q} b_i z^{-i} = \sum_{n=-\infty}^{+\infty} h_n z^{-n} \Rightarrow \begin{cases} \forall n \notin [0,Q-1] & h_n = 0 \\ \forall n \in [0,Q-1] & h_n = b_n \end{cases}$$

La réponse impulsionnelle h_n d'un filtre RIF n'a donc qu'un nombre fini de valeurs non nulles, et h_n coïncide avec les coefficients b_n de l'équation de récurrence. En conclusion, pour un filtre RIF :

$$\boxed{\begin{array}{rcl} H(z) & = & \sum_{i=0}^{Q} b_i z^{-i} \\ y_n & = & \sum_{i=0}^{Q} b_i x_{n-i} \end{array}}$$

3.1 Calcul de la réponse impulsionnelle d'un filtre RII

On peut, quand le filtre est de type RII, calculer sa réponse impulsionnelle de différentes façons :
- soit en appliquant la formule de la transformée en z inverse et utilisant le théorème des résidus,
- soit en effectuant une division polynomiale de N(z) par D(z),

– soit, quand les pôles de D(z) sont simples, en utilisant la somme de la série géométrique

$$\sum_{n=0}^{+\infty} q^n = \frac{1}{1-q} \quad \text{pour} \quad |q| < 1$$

3.1.1 Rappel sur le théorème des résidus

L'intégrale sur un contour fermé C d'une fonction complexe holomorphe F(z) rationnelle vaut :

$$\int_{\mathcal{C}} F(z)dz = 2j\pi \sum_{\text{pôles } z_i \text{ dans } \mathcal{C}} \text{Résidu}(z_i)$$

Où z_i est un pôle de F(z).

Et pour $F(z) = \frac{N(z)}{D(z)} = \frac{N(z)}{\prod_{i=1}^{P}(1-z_i z^{-1})}$

Si z_i est un pôle simple :

$$\text{résidu}(z_i) = \lim_{z \to z_i} (z - z_i)F(z)$$

Si z_i est un pôle multiple d'ordre k :

$$\text{résidu}(z_i) = \lim_{z \to z_i} \frac{1}{(k-1)!} \frac{\partial^{k-1}\left((z-z_i)^k F(z)\right)}{\partial z^{k-1}}$$

Exemple :

Soit $H(z) = \frac{1}{1+a_1 z^{-1}}$

Calcul de h_n par identification avec une série géométrique
Lorsque $|a_1| < 1$ (on verra par la suite que cette condition est nécessaire pour que le filtre soit stable) on peut écrire :

$$H(z) = \frac{1}{1+a_1 z^{-1}} = \sum_{n=0}^{+\infty} \left(-a_1 z^{-1}\right)^n = \sum_{n=-\infty}^{+\infty} h_n z^{-n}$$

Et par identification :

$$\begin{cases} h_n = 0 \quad \text{pour } n < 0 \\ h_n = (-a_1)^n \quad \text{pour } n \geq 0 \end{cases}$$

Calcul de h_n par transformée en z inverse
Pour cet exemple, la formule de la transformée en z inverse nous donne :

$$h_n = \frac{1}{2j\pi} \int_{\mathcal{C}} H(z) z^{n-1} dz = \text{Résidu}_{\text{pour } -a_1}\left(H(z)z^{n-1}\right) = \lim_{z \to -a_1} (z+a_1)H(z)z^{n-1} = (-a_1)^n$$

4 Causalité et stabilité

4.1 Causalité

Un filtre numérique est dit causal, si sa réponse impulsionnelle h_n est nulle pour $n < 0$.
Sa transformée en z converge alors à l'extérieur d'un cercle.

Un filtre numérique est dit anticausal, si sa réponse impulsionnelle h_n est nulle pour $n \geq 0$.
Sa transformée en z converge alors à l'intérieur d'un cercle.

4.2 Stabilité

On dira qu'un filtre numérique est stable, si à toute entrée bornée x_n correspond une sortie y_n bornée.

Pour toute suite (x_n) bornée :

$\exists M \quad \forall \text{ n} \quad |x_n| < M \qquad \Rightarrow \qquad \exists M' \quad \forall \text{ n} \quad |y_n| < M'$

4.2.1 $1^{ère}$ condition nécessaire et suffisante de stabilité

> Une condition nécessaire et suffisante de stabilité s'écrit :
> $\sum_{n=-\infty}^{+\infty} |h_n| < A$

C'est une condition suffisante car :

$$y_n = \sum_{k=-\infty}^{+\infty} h_k x_{n-k}$$

$$|y_n| = \left| \sum_{k=-\infty}^{+\infty} h_k x_{n-k} \right| \leq \sum_{k=-\infty}^{+\infty} |h_k| \, |x_{n-k}|$$

$$|x_n| < M \quad \Rightarrow \quad |y_n| < M \sum_{k=-\infty}^{+\infty} |h_k| < MA$$

C'est une condition nécessaire. Un contre exemple suffit à le montrer. Ainsi l'entrée bornée x_n, égale à 1 si h_{-n} est positif et à -1 si h_{-n} est négatif, génère une sortie y_0 non bornée :

Si $x_n = signe(h_{-n})$ alors $y_0 = \sum_{n=-\infty}^{+\infty} x_n h_{-n} = \sum_{n=-\infty}^{+\infty} |h_n|$ non bornée

4.2.2 $2^{ème}$ condition nécessaire et suffisante de stabilité

> Lorsque le filtre est causal et que sa fonction de transfert est rationnelle,
> il existe une autre condition nécessaire et suffisante de stabilité :
> il faut et il suffit que tous les pôles du H(z) soient à l'intérieur du cercle unité.
>
> Réciproquement pour un filtre anticausal de fonction de transfert rationnelle,
> une condition nécessaire et suffisante de stabilité est que
> tous les pôles du H(z) soient à l'extérieur du cercle unité.

Démonstration dans le cas causal :

H(z) rationnelle peut se décomposer en K éléments simples correspondant à des pôles simples ou multiples.

La réponse impulsionnelle h(n) est la somme de K termes $h_i(n)$ égaux aux transfomées en z inverses des éléments de la décomposition pour les pôles z_i.

Pour un pôle simple z_i, en notant $\frac{A_i}{1-z_i z^{-1}}$ l'élément correspondant de la décomposition de H(z) en éléments simples, on a :

$$h_i(n) = \lim_{z \to z_i} (z - z_i) \frac{A_i z^{n-1}}{1 - z_i z^{-1}} = A_i z_i^n$$

Et la somme des valeurs absolues de $h_i(n)$ est bornée si et seulement si $|z_i| < 1$.

De même pour un pôle z_i multiple d'ordre k, en notant $\frac{A_i(z)}{(1-z_i z^{-1})^k}$ l'élément correspondant de la décomposition de H(z) en éléments simples, on a :

$$h_i(n) = \frac{1}{(k-1)!} \frac{\partial^{k-1} \left(A_i(z) z^{n+k-1} \right)}{\partial z^{k-1}} = B_i(z_i)$$

Où $B_i(zi)$ est un pôlynome de degré n+k-1 au maximum.

Et la somme des valeurs absolues de $h_i(n)$ est bornée si et seulement si $|z_i| < 1$.

4.2.3 Stabilité des FIR

On remarquera que les filtres FIR ont comme seul pôle z=0 et que ce pôle est à l'intérieur du cercle unité. Les filtres FIR sont donc toujours stables. C'est une des raisons pour lesquelles ils sont très utilisés en filtrage adaptatif. En effet, les coefficients d'un filtre adaptatif varient à chaque nouvel échantillon, il faut donc vérifier la stabilité du filtre à chaque nouvel échantillon. Mais pour un filtre RIF, cette vérification est inutile.

5 Etude des filtres numériques élémentaires

5.1 Introduction

Cette section présente les caractéristiques des cellules élémentaires de filtrage FIR et IIR d'ordre un et deux. La connaissance du comportement de ces filtres est importante car très souvent, les filtres d'ordre élevé sont réalisés par l'association en cascade ou en parallèle de ces cellules élémentaires.

Pour cette section, dans les tracés de fonction de transfert, l'axe des fréquences sera limité à la plage $[0,f_e/2]$ et on normalisera $f_e/2$ à 1.

Avant d'étudier les cellules élémentaires de filtrage le chapitre commence par une analyse des zéros de transmission.

5.2 Etude des zéros de transmission

Les filtres FIR et les filtres IIR possédant un numérateur peuvent présenter des zéros de transmission, c'est à dire que leur fonction de transfert fréquentielle peut s'annuler pour certaines valeurs de fréquence.

La fonction de transfert en fréquence correspondant aux valeurs prises par la fonction de transfert en z lorsque z évolue sur le cercle unité, un zéro de transmission correspond aussi à un zéro de la fonction de transfert H(z) et ce zéro, notéz_i, est situésur le cercle unité.

Son module vaut un et son argument θ_i :

$$z_i = e^{j\theta_i}$$

Cet argument θ_i correspond à un zéro de transmission pour une pulsation ω et une fréquence f telles que $\omega = 2\pi f$ et :

$$\begin{aligned} z &= e^{j\omega T_e} = e^{j2\pi f T_e} \\ z &= z_i \Rightarrow f = \frac{\theta_i}{2\pi} f_e \end{aligned}$$

5.2.1 Cas d'une cellule FIR d'ordre 1

Dans le cas d'une cellule FIR d'ordre 1 ou d'un numérateur d'ordre 1 pour un filtre IIR, le polynôme d'ordre un en z^{-1} s'écrit :

$$N(z) = b_0 + b_1 z^{-1}$$

Le zéro de N(z) vaut $z_0 = -\frac{b_1}{b_0}$, il est réel. Comme le module de z_0 vaut 1, les seules solutions sont : $z_0 = 1$ et $z_0 = -1$. Les zéros de transmission ne peuvent avoir lieu qu'à la fréquence nulle ou à la fréquence $f_e/2$.

Les seuls polynômes d'ordre un présentant des zéros de transmissions sont de la forme:

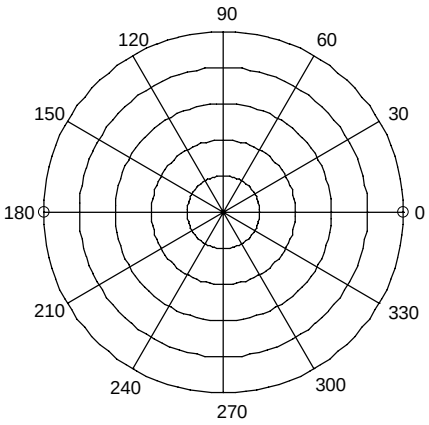
$$\begin{aligned} N(z) &= b_0(1 + z^{-1}) \\ N(z) &= b_0(1 - z^{-1}) \end{aligned}$$

C'est à dire que b_0 et b_1 sont égaux ou opposés.

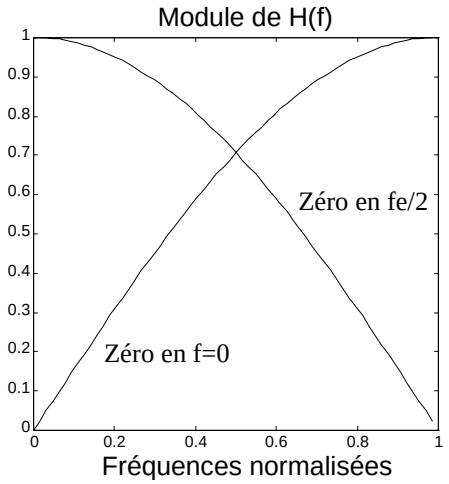
Les figures suivantes représentent :

- Les zéros de H(z) dans le plan z, pour les deux types de fonctions à zéro de transmission à l'ordre un. Les zéros sont matérialisés par des cercles sur cette figure.

– Les deux fonctions de transfert possibles (en module) à l'ordre un, possédant un zéro de transmission soit en $f = 0$ soit en $f = f_e/2$.



positions possibles des zéros de H(z)



5.2.2 Cas d'une cellule FIR d'ordre deux

Dans le cas d'une cellule FIR d'ordre deux ou d'un numérateur d'ordre deux pour un filtre IIR, le polynôme d'ordre deux en z^{-1} s'écrit :

$$N(z) = b_0 + b_1 z^{-1} + b_2 z^{-2}$$

Les coefficients de N(z) sont réels, les zéros sont donc imaginaires conjugués et s'écrivent :

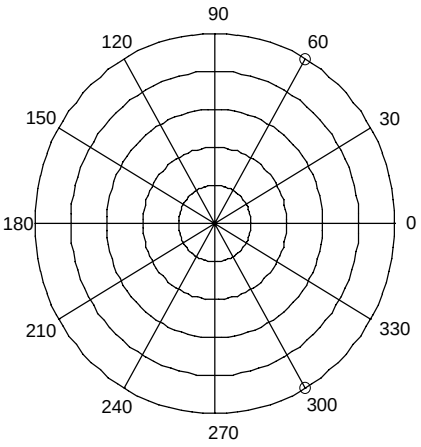
$$\begin{aligned} z_1 &= e^{j\theta_1} \\ z_2 &= \bar{z}_1 = e^{-j\theta_1} \end{aligned}$$

Et

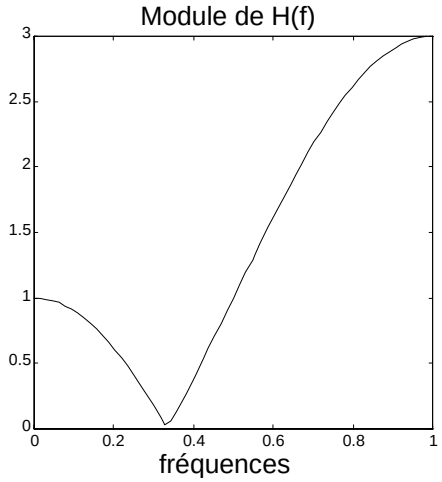
$$\begin{aligned} N(z) &= b_0(1 - z_1 z^{-1})(1 - \bar{z}_1 z^{-1}) \\ N(z) &= b_0(1 - 2\cos\theta_1 z^{-1} + z^{-2}) \end{aligned}$$

Les coefficients b_0 et b_2 sont donc égaux.

La figure suivante représente le cas de 2 zéros d'arguments $\theta_1 = \frac{\pi}{3}$ et $\theta_2 = \frac{-\pi}{3}$, ce qui correspond à un zéro de transmission en $f_e/6$.



positions possibles des zéros de H(z)



5.3 Cellule FIR d'ordre un

5.3.1 Généralités

La cellule causale FIR d'ordre un a pour fonction de transfert en z, le polynôme $H(z)$ suivant :

$$H(z) = b_0 + b_1 z^{-1}$$

La fonction H(z) n'a pas de pôle à l'extérieur du cercle unité, elle est donc stable.

Comme pour tous les FIR, la réponse impulsionnelle est facile à calculer par identification avec la définition : $H(z) = \sum_{n=0}^{n=P} h_n z^{-n} = b_0 + b_1 z^{-1}$ on déduit que $\forall n\ h_n = b_n$ et en particulier ici:

$$\begin{aligned} h_0 &= b_0 \\ h_1 &= b_1 \\ h_n &= 0 \quad \text{si } n \notin [0,1] \end{aligned}$$

La fonction de transfert en fréquence est obtenue à partir de H(z) par la relation :
$H(e^{j2\pi fT_e}) = \left[H(z)/z = e^{j2\pi fT_e}\right]$. Cette fonction sera notée un peu abusivement H(f) par la suite.

$$H(f) = b_0 + b_1 e^{-j2\pi fT_e}$$

Son module, sa phase et son temps de propagation de groupe $\tau(f)$ valent :

$$\begin{aligned} |H(f)|^2 &= b_0^2 + b_1^2 + 2b_0 b_1 \cos(2\pi fT_e) \\ \Phi(f) &= -\text{arctg}\left(\frac{b_1 \sin(2\pi fT_e)}{b_0 + b_1 \cos(2\pi fT_e)}\right) \\ \tau(f) &= -\frac{1}{2\pi}\frac{\partial\Phi(f)}{\partial f} = b_1 \frac{b_1 + b_0 \cos(2\pi fT_e)}{b_0^2 + b_1^2 + 2b_0 b_1 \cos(2\pi fT_e)} \end{aligned}$$

Le module est une fonction monotone sur la demi période $[0,f_e/2]$ et les valeurs extrêmes sont :

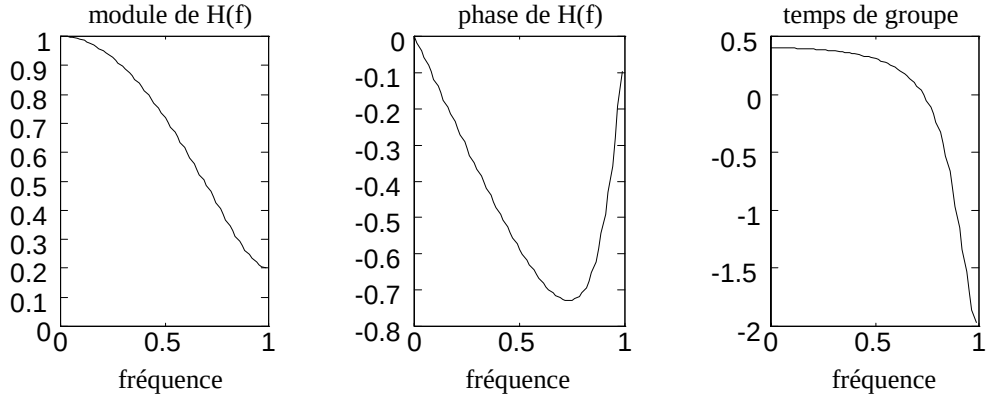
$$\begin{aligned} |H(0)| &= |b_0 + b_1| \quad \text{avec} \quad \Phi(0) = 0 \\ \left|H(\tfrac{f_e}{2})\right| &= |b_0 - b_1| \quad \text{avec} \quad \Phi\left(\tfrac{f_e}{2}\right) = 0 \end{aligned}$$

Si les deux coefficients sont de mêmes signes, le filtre est un passe bas. Et réciproquement, si les deux coefficients sont de signes opposés, le filtre est un passe haut. En effet ce changement de signe revient à changer z en - z, c'est à dire f en $\frac{f_e}{2} - f$.

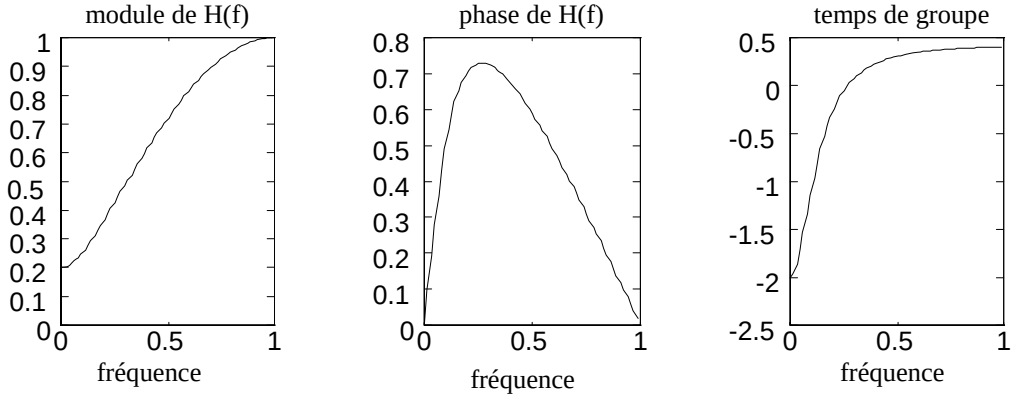
L'étude de la fonction $\phi(f)$ montre que $\phi(f) \leq \frac{\pi}{2}$.

5.3.2 Exemple

Les courbes suivantes illustrent une cellule d'ordre un de coefficients $b_0 = 0.6$ et $b_1 = 0.4$.



Les courbes suivantes illustrent une cellule d'ordre un de coefficients $b_0 = 0.6$ et $b_1 = -0.4$.



5.3.3 Cellules spéciales

Deux cellules d'ordre un sont particulièrement intéressantes parce qu'elles possèdent des zéros de transmission :
$H(z) = b_0(1 + z^{-1})$ qui s'annule pour la fréquence $f_e/2$ et $H(z) = b_0(1 - z^{-1})$ qui s'annule pour la fréquence 0.

D'autre part si $b_0 = -b_1$ la phase s'écrit:

$$\Phi(f) = \frac{\pi}{2} - \pi f T_e$$

et le temps de propagation de groupe est constant. Il vaut $T_e/2$.

Si $b_0 = b_1$ la phase s'écrit :

$$\Phi(f) = -\pi f T_e$$

Cette phase est linéaire et le temps de propagation de groupe est constant. Il vaut $T_e/2$.

5.4 Cellule FIR d'ordre 2

5.4.1 Généralités

La cellule causale FIR d'ordre deux a pour fonction de transfert en z, le polynôme H(z) suivant :

$$H(z) = b_0 + b_1 z^{-1} + b_2 z^{-2}$$

La fonction H(z) n'a pas de pôle à l'extérieur du cercle unité elle est donc stable.

Comme pour tous les FIR, la réponse impulsionnelle est facile à calculer par identification avec la définition : $H(z) = \sum_{n=0}^{n=p} h_n z^{-n} = b_0 + b_1 z^{-1} + b_2 z^{-2}$ on déduit que $\forall n \quad h_n = b_n$ et en particulier ici:

$$\begin{aligned} h_0 &= b_0 \\ h_1 &= b_1 \\ h_2 &= b_2 \\ h_n &= 0 \quad \text{si} \quad n \notin [0,1,2] \end{aligned}$$

La fonction de transfert en fréquence est obtenue à partir de H(z) par la relation :
$H(e^{j2\pi f T_e}) = \left[H(z)/z = e^{j2\pi f T_e}\right]$. Cette fonction sera notée un peu abusivement H(f) par la suite :

$$H(f) = b_0 + b_1 e^{-j2\pi f T_e} + b_2 e^{-j4\pi f T_e}$$

Le polynôme H(z) en z^{-1} possède deux zéros. Si les deux zéros sont réels, le filtre peut être considéré comme formé de deux cellules d'ordre un en cascade. Nous nous intéressons, dans ce chapitre seulement

au cas d'une "vraie" cellule d'ordre deux possédant deux zéros complexes. Comme les coefficients b_i sont réels, ces zéros sont complexes conjugués. Appelons z_1 et z_2 ces zéros. On peut les écrire en coordonnées polaires sous la forme :

$$\begin{aligned} z_1 &= r_1 e^{j\theta_1} \\ z_2 &= \bar{z}_1 = r_1 e^{-j\theta_1} \\ H(z) &= b_0(1 - z_1 z^{-1})(1 - z_2 z^{-1}) \end{aligned}$$

En identifiant les deux écritures de H(z), on déduit les relations liant les coefficients b_i aux coordonnées polaires des zéros.

$$\begin{aligned} b_1 &= -b_0(z_1 + \bar{z}_1) = -2b_0 r_1 \cos(\theta_1) \\ b_2 &= b_0 z_1 \bar{z}_1 = b_0 r_1^2 \end{aligned}$$

Le module, la phase et le temps de propagation de groupe de la fonction de transfert en fréquence peuvent s'exprimer en fonction des coordonnées polaires des zéros. Ils valent :

$$\begin{aligned} |H(f)|^2 &= b_0^2 \left|1 - z_1 e^{-j2\pi f T_e}\right|^2 \left|1 - \overline{z}_1 e^{-j2\pi f T_e}\right|^2 \\ |H(f)|^2 &= b_0^2 \left|1 - r_1 e^{-j(2\pi f T_e - \theta_1)}\right|^2 \left|1 - r_1 e^{-j(2\pi f T_e + \theta_1)}\right|^2 \\ |H(f)|^2 &= b_0^2 \left(1 + r_1^2 - 2r_1 \cos(2\pi f T_e - \theta_1)\right) \left(1 + r_1^2 - 2r_1 \cos(2\pi f T_e + \theta_1)\right) \end{aligned}$$

$$\Phi(f) = \operatorname{arctg}\left(\frac{r_1 \sin(2\pi f T_e - \theta_1)}{1 - r_1 \cos(2\pi f T_e - \theta_1)}\right) + \operatorname{arctg}\left(\frac{r_1 \sin(2\pi f T_e + \theta_1)}{1 - r_1 \cos(2\pi f T_e + \theta_1)}\right)$$

$$\begin{aligned} \tau(f) &= -\frac{1}{2\pi}\frac{\partial \Phi(f)}{\partial f} \\ \tau(f) &= -2r_1 \frac{(1 + r_1^2)\left[\cos(\theta_1)\cos(2\pi f T_e) - r_1\right] - r_1\left[\cos(4\pi f T_e) + \cos(2\theta_1) - 2r_1\cos(\theta_1)\cos(2\pi f T_e)\right]}{\left[1 + r_1^2 - 2r_1\cos(2\pi f T_e - \theta_1)\right]\left[1 + r_1^2 - 2r_1\cos(2\pi f T_e + \theta_1)\right]} \end{aligned}$$

Pour une cellule d'ordre 2, la fonction $\phi(f)$ est en valeur absolue infÉrieure à π.

5.4.2 Etude des extréma du module de la fonction de transfert en fréquence

Pour étudier les extrémas de $|H(f)|$, il faut étudier sa dérivée par rapport à f. Lorsque cette dérivée s'annule, la fonction $|H(f)|$ passe par un extrémum. Pour trouver les valeurs de f où $\frac{\partial H(f)}{\partial f} = 0$, on dérive $\|H(f)\|^2$ et on étudie pour quelle valeur de f, cette dérivée s'annule sans que $H(f)$ s'annule.

$$\frac{\partial |H(f)|^2}{\partial f} = 2|H(f)|\frac{\partial |H(f)|}{\partial f} = b_0^2 4\pi T_e r_1 \sin(2\pi f T_e)\left[(1 + r_1^2)\cos(\theta_1) - 2r_1\cos(2\pi f T_e)\right]$$

Dans l'intervalle [0,f_e/2], cette dérivée s'annule lorsque :

- Soit $\sin(2\pi f Te)$ s'annule, c'est à dire pour $f = 0$ et $f = f_e/2$.
- Soit $(1 + r_1^2)\cos(\theta_1) - 2r_1\cos(2\pi f T_e)$ s'annule, c'est à dire pour une fréquence f_R telle que :

$$\cos(2\pi f_R T_e) = \frac{(1 + r_1^2)\cos(\theta_1)}{2r_1} = -\frac{(b_0 + b_2)b_1}{4b_2 b_0}$$

Ce dernier cas n'est possible que si : $\left|\frac{(1+r_1^2)\cos(\theta_1)}{2r_1}\right| \leq 1$ ou, ce qui revient au même : $\left|\frac{(b_0+b_2)b_1}{4b_2b_0}\right| \leq 1$.
Ces zéros sont des zéros de $\frac{\partial H(f)}{\partial f}$ et non de $\|H(f)\|$, sauf dans le cas trivial où $r_1 = 1$ et $\theta_1 = 0$.

D'autre part la dérivée de $|H(f)|$ par rapport à f est négative pour $f < f_R$ puis positive pour $f > f_R$. La fréquence f_R, quand elle existe, correspond donc forcément à un minimum de $|H(f)|$. Une cellule FIR d'ordre deux peut donc présenter une **antirésonance**.

Si r_1 est proche de 1, la fréquence f_R est proche de $\frac{\theta_1 fe}{2\pi}$. Il y a égalité si $r_1 = 1$.

La valeur de $|H(f)|$ pour $f = f_R$ vaut :

$$|H(f_R)| = \left| b_0(1 - r_1^2)\sin\theta_1 \right|$$

Comme on l'a vu précédemment si $r_1 = 1$, la cellule possède un zéro de transmission qui correspond au minimum de $|H(f)|$.

Soit B_R la largeur de bande à mi-hauteur de cette antirésonance.

$$B_R = |f_+ - f_-|$$

où f_- et f_+ sont les fréquences pour lesquelles $|H(f)|^2 = 2\,|H(f_R)|^2$, lorsque ces frquences existent. On peut écrire $|H(f)|^2$ sous la forme d'un polynôme d'ordre 2 en $cos(2\pi f T_e)$:

$$|H(f)|^2 = 4r_1^2 cos(2\pi f T_e)^2 - 4r_1(1 + r_1^2)cos(\theta_1)cos(2\pi f T_e) + 2{r_1}^2 cos(2\theta_1) + (1 + r_1^2)^2$$

Pour $|H(f)|^2 = (1 - r_1^2)^2 \sin(\theta_1)^2$ le discrimant du polynôme est nul et on retrouve la solution correspondant à la fréquence de résonance $cos(2\pi f_R T_e = \frac{(1+r_1^2)cos(\theta_1)}{2r_1}$.

Pour $|H(f)|^2 = 2\,|H(f_R)|^2 = 2(1 - r_1^2)^2 \sin(\theta_1)^2$, le polynôme s'écrit :

$$4r_1^2 cos(2\pi f T_e)^2 - 4r_1(1 + r_1^2)cos(\theta_1)cos(2\pi f T_e) + cos(2\theta_1) + (1 + r_1^4) + 2r_1^2 = 0$$

le discriminant s'écrit :

$$\Delta = 16r_1^2(1 + r_1^2)^2 cos(\theta_1)^2 - 16r_1^2\left(cos(2\theta_1) + (r_1^4 + 1) + 2r_1^2\right) = 16r_1^2(1 - r_1^2)^2 \sin(\theta_1)^2$$

Le discriminant est positif, on en déduit 2 expressions pour les racines potentielles :

$$\begin{aligned} \cos(2\pi f_+ T_e) &= \cos(2\pi f_R T_e) - \frac{(1 - r_1^2)\sin\theta_1}{2r_1} \\ \cos(2\pi f_- T_e) &= \cos(2\pi f_R T_e) + \frac{(1 - r_1^2)\sin\theta_1}{2r_1} \end{aligned}$$

Ces 2 racines n'existent que si :

$$\begin{aligned} \left|\cos(2\pi f_R T_e) - \frac{(1 - r_1^2)\sin\theta_1}{2r_1}\right| &\leq 1 \\ \left|\cos(2\pi f_R T_e) + \frac{(1 - r_1^2)\sin\theta_1}{2r_1}\right| &\leq 1 \end{aligned}$$

Toutes les situations sont possibles : les 2 racines existent, une seule racine existe, ou bien aucune des 2 n'existe. Dans le cas où les 2 conditions sont vérifiées, c'est à dire où les 2 racines existent, on peut calculer une approximation de la largeur de bande de l'antirésonance. A partir des valeurs de $cos(2\pi f_+ T_e)$ et de $cos(2\pi f_- T_e$, on déduit que, pour $r_1 \approx 1$:

$$\cos(2\pi f_+ T_e) - \cos(2\pi f_+ T_e) = -2\sin(2\pi(f_R + B/2)T_e)\sin(\pi B T_e) \approx -2\sin(\theta_1)\pi B T_e \approx -2\frac{(1 - r_1^2)\sin(\theta_1)}{2r_1}$$

D'où :

$$B_R \approx \frac{(1 - r)}{\pi} f_e$$

5.4.3 inversion du module des zéros, polynôme réciproque de H(z)

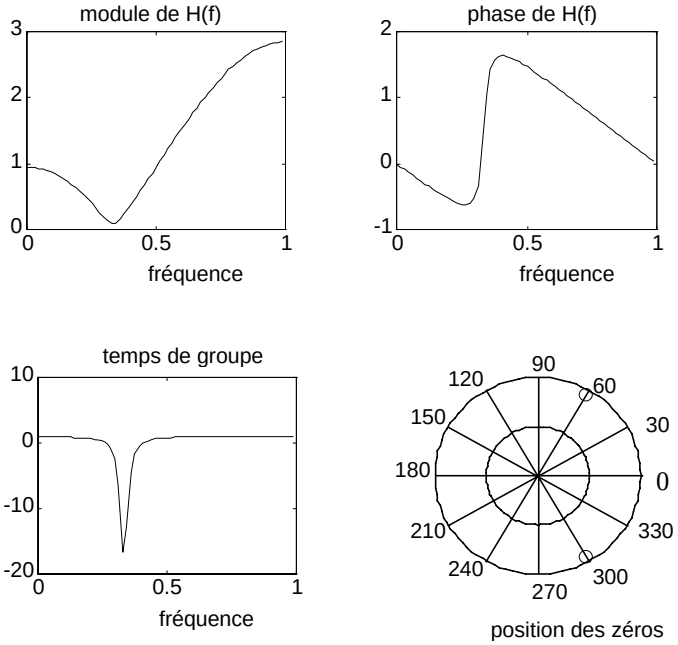
Les cellules FIR d'ordre 2 définies par les fonctions de transfert H(z) et $H_r(z) = z^{-2}H(z^{-1})$ possèdent des pôles de mêmes arguments θ_1 et $-\theta_1$ mais de modules inverses r_1 et $1/r_1$. Elles présentent (si les conditions d'antirésonance sont vérifiées) une antirésonance pour la même fréquence f_R. Ces deux fonctions de transfert ont le même module mais :

$$\begin{aligned} \Phi_r(f) &= -4\pi f T_e - \Phi(f) \\ \tau_r(f) &= 2T_e - \tau(f) \end{aligned}$$

Les filtres numériques dont les zéros sont à l'intérieur du cercle unité sont dits à **phase minimale**.

5.4.4 Exemple

Les figures suivantes représentent la fonction de transfert H(f) en module, en phase, en temps de propagation de groupe ainsi que la position des zéros de H(z), pour les valeurs : $r_1 = 0.95$ $\theta_1 = \pi/3$, ce qui correspond à $H(z) = 1 - 0.95z^{-1} + 0.9025z^{-2}$.
La fréquence d'antirésonance f_R est telle que :
$2\pi f_R T_e = 1.0464$, ce que l'on peut comparer à $\theta_1 = \frac{\pi}{3} = 1.0472$.



5.4.5 Changement du signe du coefficient b_1, changement de z en -z

Si l'on change le signe de b_1, ou ce qui revient au même z en -z, le filtre change de type, les valeurs de $H(0)$ et de $H(f_e/2)$ sont inversées. Dans l'exemple ci-dessus, le filtre est plutôt un passe haut (dans le sens que $H(f = 0) < H(f = f_e/2)$. Si on remplace $b_1 = -0.95$ par $b_1 = 0.95$, le filtre résultant est un passe bas. Changer le signe de b_1 revient à remplacer θ_1 par $\pi - \theta_1$. Si les angles des zéros (en valeur absolue) sont inférieurs à $\pi/2$ le filtre est un passe haut. Si les angles des zéros (en valeur absolue) sont supérieurs à $\pi/2$ le filtre est un passe bas.
En fait changer z en -z, revient à multiplier z par $e^{j\pi}$. Dans le domaine fréquentiel, multiplier $e^{j2\pi fT_e}$ par $e^{j\pi}$, revient à remplacer $e^{j2\pi fT_e}$ par $e^{j2\pi(f+f_e/2)T_e}$. Remplacer z par -z est donc équivalent à une translation de $-f_e/2$ dans le domaine fréquentiel, et ceci quelque soit H(z). Un passe-bas devient un passe-haut et réciproquement.

5.5 Cellule IIR d'ordre 1

5.5.1 Généralités

La cellule causale IIR d'ordre 1 a pour fonction de transfert en z, le polynôme H(z) suivant :

$$H(z) = \frac{1}{a_0 + a_1 z^{-1}}$$

Par la suite, on normalise le terme constant du dénominateur de H(z) à 1, $a_0 = 1$.

Pour que la fonction H(z) soit stable et causale, il faut qu'elle n'ait pas de pôle à l'extérieur du cercle unité. Or le pôle de cette cellule vaut $-a_1$. Une condition nécessaire et suffisante de stabilité de cette cellule causale est que $|a_1| < 1$.

La réponse impulsionnelle peut se calculer par la formule d'inversion de la transformée en z, à savoir:

$$h_n = \frac{1}{2j\pi} \int_{\text{cercle unité}} H(z) z^{n-1} dz$$

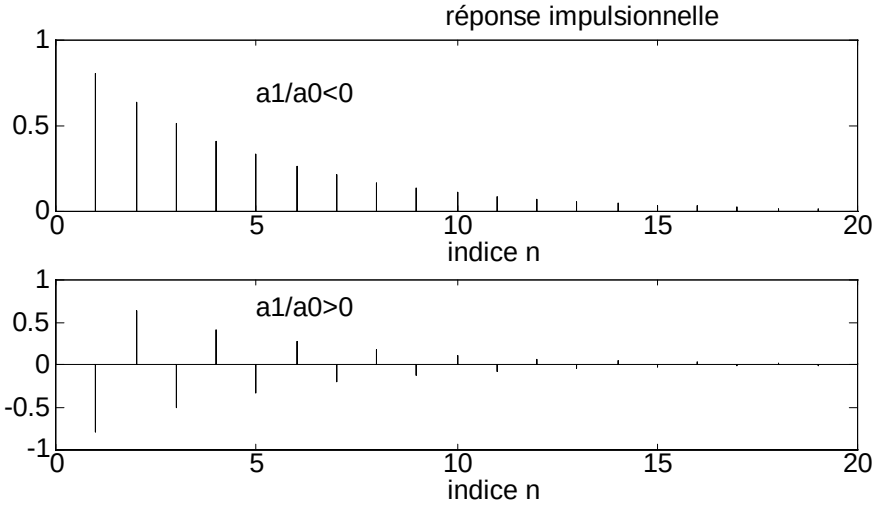
ou par identification de H(z) avec la somme d'une série géométrique de raison $q = -a_1 z^{-1}$:

$$H(z) = \sum_{n=0}^{\infty} h_n z^{-n} = \sum_{n=0}^{\infty} q^n = \sum_{n=0}^{\infty} (-a_1)^n z^{-n}$$

D'où l'on déduit que :

$$\begin{aligned} \forall n \geq 0 \qquad & h_n = (-a_1)^n \\ \forall n < 0 \qquad & h_n = 0 \end{aligned}$$

La figure suivante visualise l'allure de h_n pour a_1 positif et a_1 négatif.



La fonction de transfert en fréquence est obtenue à partir de $H(z)$ par la relation :

$$H(e^{j2\pi f T_e}) = \left[H(z)/z = e^{j2\pi f T_e} \right]$$

. Cette fonction sera notée un peu abusivement H(f) par la suite.

$$H(f) = \frac{1}{1 + a_1 e^{-j2\pi f T_e}}$$

Pour le calcul de ses caractéristiques principales, il suffit de reprendre les résultats obtenus pour la cellule FIR. Son module, sa phase et son temps de propagation de groupe $\tau(f)$ valent :

$$|H(f)|^2 = \frac{1}{1 + a_1^2 + 2a_1 \cos(2\pi f T_e)}$$

$$\begin{aligned}\Phi(f) &= \text{arctg}\left(\frac{a_1 \sin(2\pi f T_e)}{1 + a_1 \cos(2\pi f T_e)}\right)\\ \tau(f) &= -\frac{1}{2\pi}\frac{\partial \Phi(f)}{\partial f} = -a_1 \frac{a_1 + \cos(2\pi f T_e)}{1 + a_1^2 + 2a_1 \cos(2\pi f T_e)}\end{aligned}$$

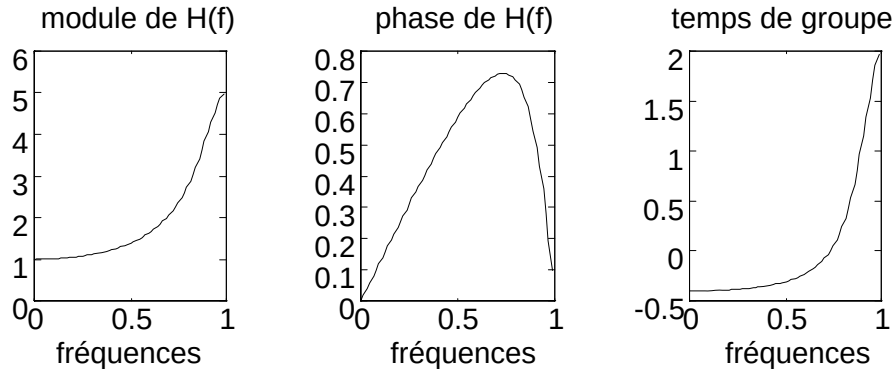
Le module est une fonction monotone sur la demi période $[0,f_e/2]$ et les valeurs extrêmes sont :

$$\begin{aligned}|H(0)| &= \frac{1}{|1+a_1|} \quad \text{avec} \quad \Phi(0) = 0\\ \left|H\left(\frac{f_e}{2}\right)\right| &= \frac{1}{|1-a_1|} \quad \text{avec} \quad \Phi\left(\frac{f_e}{2}\right) = 0\end{aligned}$$

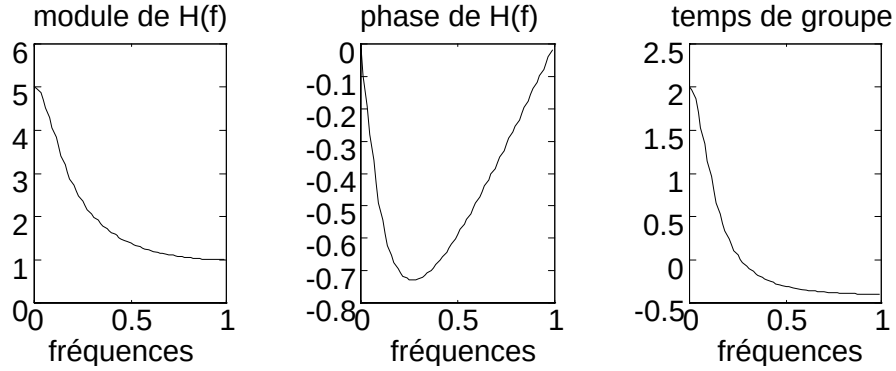
Si les deux coefficients sont de mêmes signes, le filtre est un passe-haut. Et réciproquement, si les deux coefficients sont de signes opposés, le filtre est un passe-bas.
On remarque que pour une cellule d'ordre 1, le déphasage introduit par la cellule est inférieur ou égal à $\pi/2$.

5.5.2 Exemple

Les courbes suivantes illustrent une cellule d'ordre un de coefficients $a_0 = 0.6$ et $a_1 = 0.4$:



Les courbes suivantes illustrent une cellule d'ordre un de coefficients $a_0 = 0.6$ et $a_1 = -0.4$:



5.6 Cellule IIR d'ordre 2

5.6.1 Cellule d'ordre purement récursive, généralités

La cellule causale IIR d'ordre deux a pour fonction de transfert en z, le polynôme H(z) suivant :

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

On a, comme pour la cellule d'ordre 1, normalisé a_0 à 1.
La fonction $H(z)$causale est stable si et seulement si ses pôles sont à l'intérieur du cercle unité, c'est à dire sont de module inférieur à 1.

La fonction $H(z)$ possède deux pôles. Si les deux pôles sont réels, le filtre peut être considéré comme formé de deux cellules d'ordre un en cascade. Nous nous intéressons, dans ce chapitre seulement au cas

d'une « vraie » cellule d'ordre deux possédant deux pôles complexes. Comme les coefficients a_i sont réels ces pôles sont complexes conjugués. Appelons z_1 et z_2 ces pôles. On peut les écrire en coordonnées polaires sous la forme :

$$\begin{aligned} z_1 &= r_1 e^{j\theta_1} \\ z_2 &= \bar{z}_1 = r_1 e^{-j\theta_1} \end{aligned}$$

$$H(z) = \frac{1}{(1 - z_1 z^{-1})(1 - z_2 z^{-1})}$$

En identifiant les deux écritures de $H(z)$, on déduit les relations liant les coefficients a_i aux coordonnées polaires des pôles.

$$\begin{aligned} a_1 &= (z_1 + \bar{z}_1) = -2r_1 \cos(\theta_1) \\ a_2 &= z_1 \bar{z}_1 = r_1^2 \end{aligned}$$

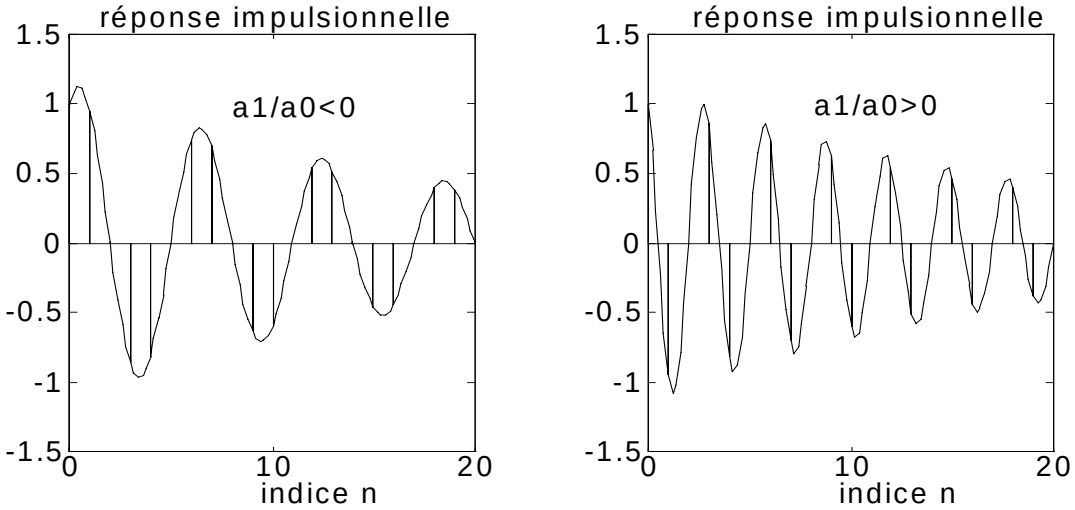
La réponse impulsionnelle peut se calculer par la formule d'inversion de la transformée en z, ou par identification de $H(z)$ avec des sommes de séries géométriques.

$$h_n = \frac{1}{2j\pi} \int_{\text{tiny cercle unité}} H(z) z^{n-1} dz = \frac{1}{2j\pi} \int_{\text{tiny cercle unité}} \frac{z^{n-1}}{(1 - z_1 z^{-1})(1 - \bar{z}_1 z^{-1})}$$

$$h_n = \sum_{\text{pôles de H(z)}} \text{résidus}(H(z) z^{n-1})$$

$$\begin{aligned} \forall n \geq 0 \quad h_n &= \left(\frac{z_1^{n+1}}{z_1 - \bar{z}_1} + \frac{\bar{z}_1^{n+1}}{\bar{z}_1 - z_1} \right) = \left(\frac{r_1^n \sin\left[(n+1)\theta_1\right]}{\sin \theta_1} \right) \\ \forall n < 0 \quad h_n &= 0 \end{aligned}$$

La figure suivante représente les réponses impulsionnelles pour $r_1 = 0.95$ et $\theta_1 = \pi/3$ (ce qui correspond à un coefficient $a_1 < 0$) ainsi que pour $r_1 = 0.95$ et $\theta_1 = 2\pi/3$ (ce qui correspond à un coefficient $a_1 > 0$).



La fonction de transfert en fréquence est obtenue à partir de H(z) par la relation :

$$H(e^{j2\pi f T_e}) = \left[H(z)/z = e^{j2\pi f T_e} \right].$$

Comme précédemment cette fonction sera notée un peu abusivement H(f).

$$H(f) = \frac{1}{1 + a_1 e^{-j2\pi f T_e} + a_2 e^{-j4\pi f T_e}}$$

Le module, la phase et le temps de propagation de groupe de la fonction de transfert en fréquence peuvent s'exprimer en fonction des coordonnées polaires des pôles, il suffit de reprendre et de modifier correctement les résultats obtenus pour la cellule FIR. Ils valent :

$$
\begin{aligned}
|H(f)|^2 &= \frac{1}{\left|1 - z_1 e^{-j2\pi f T_e}\right|^2 \left|1 - \overline{z}_1 e^{-j2\pi f T_e}\right|^2} \\
|H(f)|^2 &= \frac{1}{\left|1 - r_1 e^{-j(2\pi f T_e - \theta_1)}\right|^2 \left|1 - r_1 e^{-j(2\pi f T_e + \theta_1)}\right|^2} \\
|H(f)|^2 &= \frac{1}{(1 + r_1^2 - 2r_1\cos(2\pi f T_e - \theta_1))\,(1 + r_1^2 - 2r_1\cos(2\pi f T_e + \theta_1))} \\
\Phi(f) &= -\text{arctg}\left(\frac{r_1 \sin(2\pi f T_e - \theta_1)}{1 - r_1\cos(2\pi f T_e - \theta_1)}\right) - \text{arctg}\left(\frac{r_1 \sin(2\pi f T_e + \theta_1)}{1 - r_1\cos(2\pi f T_e + \theta_1)}\right) \\
\tau(f) &= -\frac{1}{2\pi}\frac{\partial \Phi(f)}{\partial f} \\
\tau(f) &= 2r_1 \frac{(1 + r_1^2)\left[\cos(\theta_1)\cos(2\pi f T_e) - r_1\right] - r_1\left[\cos(4\pi f T_e) + \cos(2\theta_1) - 2r_1\cos(\theta_1)\cos(2\pi f T_e)\right]}{\left[1 + r_1^2 - 2r_1\cos(2\pi f T_e - \theta_1)\right]\left[1 + r_1^2 - 2r_1\cos(2\pi f T_e + \theta_1)\right]}
\end{aligned}
$$

5.6.2 Etude des extréma du module de la fonction de transfert en fréquence pour une cellule purement récursive

Les calculs qui ont déjà été faits pour la cellule FIR dans le paragraphe précédent peuvent être repris ici. La dérivée de $|H(f)|$ par rapport à f s'écrit:

$$
\frac{\partial |H(f)|^2}{\partial f} = 2\,|H(f)|\,\frac{\partial |H(f)|}{\partial f} = -\frac{4\pi T_e r_1 \sin(2\pi f T_e)\left[2(1 + r_1^2)\cos(\theta_1) - 2r_1\cos(2\pi f T_e)\right]}{(1 + r_1^2 - 2r_1\cos(2\pi f T_e - \theta_1))^2\,(1 + r_1^2 - 2r_1\cos(2\pi f T_e + \theta_1))^2}
$$

Dans l'intervalle [0,f_e], cette dérivée s'annule lorsque :

- Soit $\sin(2\pi f Te)$ s'annule, c'est à dire pour $f = 0$ et $f = f_e/2$.
- Soit $2(1 + r_1^2)\cos(\theta_1) - 2r_1\cos(2\pi f T_e)$ s'annule, c'est à dire pour une fréquence f_R telle que :

$$
\cos(2\pi f_R T_e) = \frac{(1 + r_1^2)\cos(\theta_1)}{2r_1} = -\frac{(b_0 + b_2)b_1}{4b_2 b_0}
$$

Ce dernier cas n'est possible que si :

$$
\left|\frac{(1 + r_1^2)\cos(\theta_1)}{2r_1}\right| \leq 1
$$

ou, ce qui revient au même :

$$
\left|\frac{(b_0 + b_2)b_1}{4b_2 b_0}\right| \leq 1
$$

D'autre part, inversement au cas des FIR la dérivée de $|H(f)|$ par rapport à f est positive pour $f < f_R$ puis négative pour $f > f_R$. La fréquence f_R, quand elle existe, correspond donc forcément à un maximum de $|H(f)|$. Une cellule IIR d'ordre deux ne peut donc présenter qu'une **résonance**, alors qu'une cellule FIR d'ordre deux ne peut présenter qu'une antirésonance. La fréquence f_R est appelée fréquence de résonance.

Si r_1 est proche de 1, la fréquence f_R est proche de $\frac{\theta_1 fe}{2\pi}$. Il y a égalité si $r_1 = 1$.

La valeur de $|H(f)|$ pour $f = f_R$ vaut :

$$
|H(f_R)| = \frac{1}{|(1 - r_1^2)\sin\theta_1|} = \frac{1}{(1 - r_1^2)\sin\theta_1}
$$

Si $r_1 = 1$, la cellule est un résonateur pur ou oscillateur, à la limite de stabilité (cette stabilité étant définie par le fait qu'à une entrée bornée correspond une sortie bornée).

Soit B_R la largeur de bande à mi-hauteur de cette antirésonance :

$$B_R = |f_+ - f_-|$$

où f_- et f_+ sont les fréquences pour lesquelles $|H(f)|^2 = |H(f_R)|^2/2$.

$$\begin{aligned} \cos(2\pi f_+ T_e) &= \cos(2\pi f_R T_e) - \frac{(1-r_1^2)\sin\theta_1}{2r_1} \\ \cos(2\pi f_- T_e) &= \cos(2\pi f_R T_e) + \frac{(1-r_1^2)\sin\theta_1}{2r_1} \end{aligned}$$

D'où l'on déduit (comme pour les FIR d'ordre 2) que, pour $r_1 \approx 1$:

$$B_R \approx \frac{(1-r)}{\pi} f_e$$

5.6.3 Inversion du module des zéros, polynôme réciproque de H(z)

La cellule IIR d'ordre deux ayant des pôles de même argument et de module $1/r_1$ présentera une résonance pour la même fréquence f_R. En fait cette cellule a une fonction de transfert $H_r(z)$ en z^{-1} qui est le polynôme réciproque de H(z).

$$H_r(z) = \frac{1}{a_2 + a_1 z^{-1} + z^{-2}} = z^{-2} H(z^{-1})$$

$H(f)$ et $H_r(f)$ ont le même module mais :

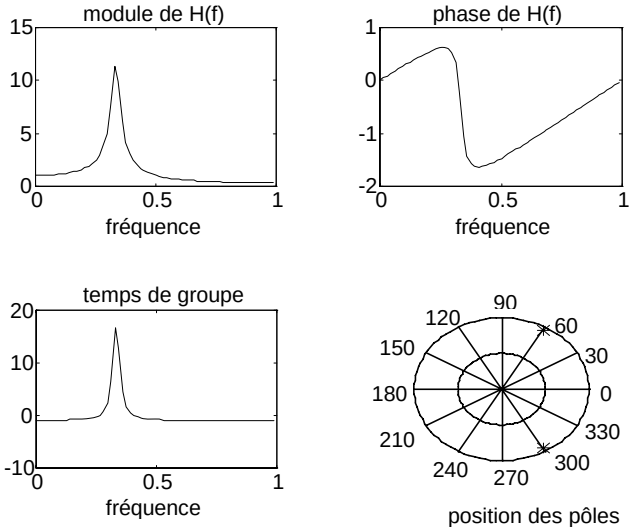
$$\begin{aligned} \Phi_r(f) &= -4\pi f T_e - \Phi(f) \\ \tau_r(f) &= 2T_e - \tau(f) \end{aligned}$$

5.6.4 Exemple

Les figures suivantes représentent la fonction de transfert $H(f)$ en module, en phase, en temps de propagation de groupe ainsi que la position des zéros de $H(z)$, pour les valeurs : $r_1 = 0.95$ $\theta_1 = \pi/3$, ce qui correspond à :

$$H(z) = \frac{1}{1 - 0.95z^{-1} + 0.9025z^{-2}}$$

La fréquence de résonance f_R est telle que :
$2\pi f_R T_e = 1.0464$, ce que l'on peut comparer à $\theta_1 = \frac{\pi}{3} = 1.0472$



5.6.5 Changement du signe du coefficient a_1, changement de z en -z

Si l'on change le signe de a_1, le filtre change de type, les valeurs de $H(0)$ et de $H(f_e/2)$ sont inversées. Dans l'exemple ci-dessus, le filtre est plutôt un passe-bas (dans le sens que $H(f=0) > H(f=f_e/2)$). Si on remplace $a_1 = -0.95$ par $a_1 = 0.95$, le filtre résultant est un passe-haut. Changer le signe de a_1 revient à remplacer θ_1 par $\pi - \theta_1$. Si les angles des zéros (en valeur absolue) sont inférieurs à $\pi/2$ le filtre est un passe bas. Si les angles des zéros (en valeur absolue) sont supérieurs à $\pi/2$ le filtre est un passe haut.
Changer le signe de a_1, revient à remplacer z par -z. Et comme on l'a vu précédemment pour la cellule FIR d'ordre 2, cela revient à effectuer une translation de $-f_e/2$ dans le domaine fréquentiel.

5.6.6 Cellule IIR d'ordre 2 générale

La fonction de transfert en z^{-1} s'écrit dans le cas général :

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}}$$

Les cellules les plus intéressantes possèdent des zéros de transmission. Dans ce cas:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_0 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}}$$

On peut poser, sans perte de généralité $a_0 = 1$. En gardant les notations utilisées précédemment pour représenter les pôles, et en notant les zéros sous la forme:

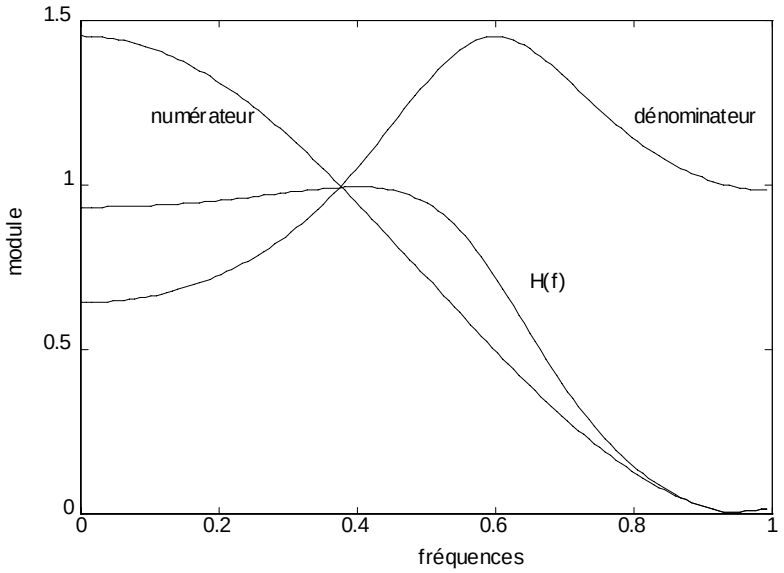
$$z_0 = e^{j\theta_0}$$
$$\bar{z}_0 = e^{-j\theta_0}$$

On déduit l'expression du module de la fonction de transfert:

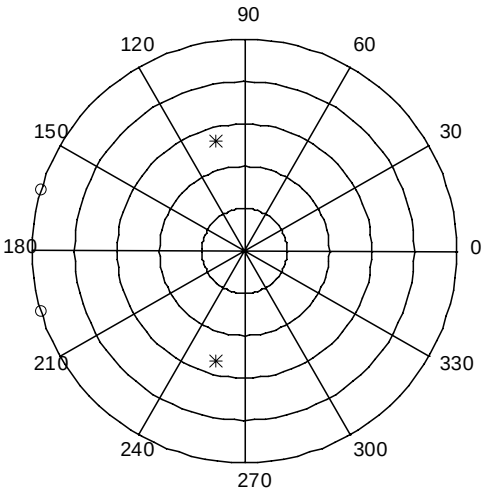
$$|H(f)|^2 = \frac{16 b_0^2 \sin^2\left(\frac{\theta_0 + 2\pi f T_e}{2}\right) \sin^2\left(\frac{\theta_0 - 2\pi f T_e}{2}\right)}{\left(1 + r_1^2 - 2r_1\cos(\theta_1 + 2\pi f T_e)\right)\left(1 + r_1^2 - 2r_1\cos(\theta_1 - 2\pi f T_e)\right)}$$

La figure suivante représente les modules du dénominateur, du numérateur et de la fonction $H(f)$ elle même, pour une cellule d'ordre 2 elliptique définie par:

$$H(f) = \frac{0.3758 + 0.7194 z^{-1} + 0.3758 z^{-2}}{1 + 0.2706 z^{-1} + 0.2876 z^{-2}}$$



La figure suivante représente la position des pôles (*) et des zéros (o) dans le plan z.



5.6.7 Cellule d'ordre 2 déphaseur pur

Compte tenu de ce qui a été vu précédemment sur les pôlynomes réciproques, il est facile de trouver la forme générale de la fonction de transfert déphaseur pur.

$$H(z) = \frac{z^{-2}D(z^{-1})}{D(z)} = \frac{a_2 + a_1 z^{-1} + a_0 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}}$$

Cette forme peut d'ailleurs se généraliser à un ordre quelconque.

En utilisant les résultats du paragraphe 5.3, la phase s'écrit:

$$\Phi(f) = \Phi_N(f) - \Phi_D(f)$$

où N représente le numérateur et D le dénominateur

$$\Phi(f) = -4\pi f T_e - 2\Phi_D(f)$$

L'expression de $\Phi_D(f)$ a déjà été donnée pour la cellule purement récursive.

De plus le temps de propagation de groupe s'écrit:

$$\tau(f) = 2T_e - 2\tau_D(f)$$

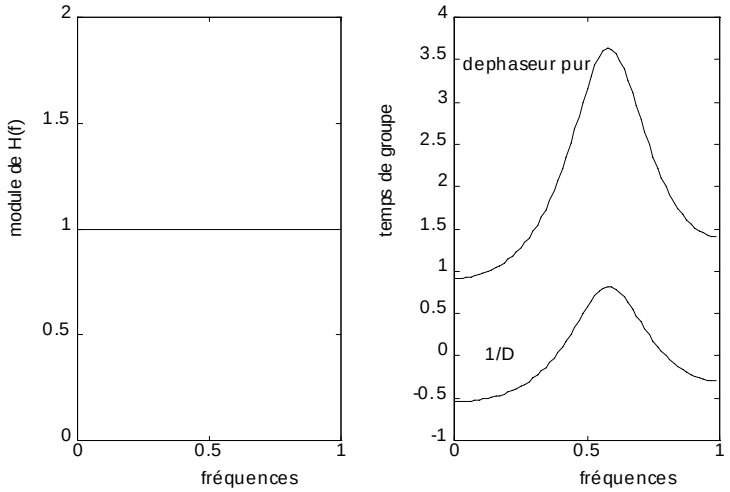
L'expression de $\tau_D(f)$ a déjà été donnée pour la cellule purement récursive.

La figure suivante représente le module (qui vaut 1) et le temps de propagation de groupe pour la cellule :

$$H(z) = H(f) = \frac{0.2876 + 0.2706z^{-1} + z^{-2}}{1 + 0.2706z^{-1} + 0.2876z^{-2}}$$

Sur la même figure est aussi représenté le temps de propagation de groupe de la cellule purement

récursive correspondante (noté 1/D).



Chapitre II

ÉTUDE DES FILTRES FIR À PHASE LINÉAIRE

1 Rappel de la définition des filtres FIR

Un filtre FIR (Finite Impulse Response) est un filtre dont la réponse impulsionnelle ne comporte qu'un nombre fini d'éléments non nuls.

Pour un filtre causal, l'équation de récurrence temporelle liant l'entrée x_n et la sortie y_n du filtre s'écrit :

$$y_n = \sum_{i=0}^{N-1} b_i x_{n-i}$$

N est le nombre de coefficients du filtre.

La fonction de transfert en z a pour valeur :

$$H(z) = \sum_{i=0}^{N-1} b_i z^{-i} = \sum_{n=0}^{+\infty} h_n z^{-n}$$

La réponse impulsionnelle est liée aux coefficients du filtre par :

$$\begin{aligned} 0 \leq n \leq N-1 \qquad & h_n = b_n \\ n < 0 \qquad & h_n = 0 \\ n > N-1 \qquad & h_n = 0 \end{aligned}$$

2 Propriétés des filtres FIR

- Pour les filtres FIR causaux, la fonction de transfert en z étant un polynôme en z^{-1}, la condition de stabilité est toujours vérifiée.
- Il est possible de réaliser des filtres FIR à temps de retard de groupe constant $\tau(f) = \tau$.

Les filtres à temps de retard de groupe constant sont souvent utilisés en communications numériques. En effet, lorsque le temps de retard de groupe est constant et les ondulations en bande passante sont faibles, le filtre ne déforme pratiquement pas les signaux qui traversent sa bande passante. En particulier, les fronts des impulsions sont bien conservés.

Ainsi pour un signal d'entrée x_n sinusoïdal de fréquence f_0 le signal de sortie du filtre y_n est une sinusoïde de même fréquence et retardée de τ, indépendant de f_0, par rapport à l'entrée:

$$\begin{aligned} x_n &= \sin(2\pi f_0 n T_e) \\ y_n &= |H(f_0)| \sin\left(2\pi f_0 \left(nT_e - \tau\right)\right) \end{aligned}$$

Ce retard τ ne dépend pas de la fréquence.

Dans le cas d'un filtre à temps de retard de groupe constant égal à τ la phase s'écrit :

$$\forall \omega \geq 0 \quad \phi(\omega) = -\tau\omega + \phi_0$$

La phase est donc linéaire ou affine.
La constante ϕ_0 ne peut prendre que les valeurs :

$$\phi_0 = \begin{cases} 0 \\ \pi/2 \\ -\pi/2 \end{cases} \text{ modulo } \pi$$

3 Différents types de filtres FIR à temps de retard de groupe constant

3.1 Conditions pour que le temps de retard de groupe soit constant

La fonction de transfert en fréquence s'écrit :

$$H(\omega) = |H(\omega)|\, e^{j\phi(\omega)}$$

Le module $|H(\omega)|$ est une fonction paire de ω.
La phase $\phi(\omega)$ est une fonction impaire de ω.

$$\begin{aligned} H(-\omega) &= |H(\omega)|\, e^{-j\phi(\omega)} \\ H(\omega) &= H(-\omega)e^{2j\phi(\omega)} \end{aligned}$$

$$H(\omega) = \sum_{i=0}^{N-1} b_i e^{-j\omega i T_e} = \sum_{n=0}^{N-1} b_n e^{j\omega n T_e} e^{2j\phi(\omega)} = \sum_{n=0}^{N-1} b_n e^{j\omega(nTe-2\tau)+2j\phi_0}$$

Une solution triviale au problème est le filtre dont tous les coefficients sont nuls. Pour qu'il existe une solution non triviale possédant N coefficients, il faut que pour chaque indice i de la première somme il existe un indice n de la deuxième somme tel que :

$$iT_e = 2\tau - nT_e$$

$$\text{C'est à dire : } i + n = 2\frac{\tau}{T_e}$$

Ces relations sont à comprendre modulo 2.
Cette relation n'est possible que si le temps de retard de groupe τ vérifie :

$$2\tau = kT_e \quad \text{où k est un nombre entier}$$

Et comme i et n sont compris entre 0 et N-1, la seule valeur possible pour τ est :

$$\tau = \frac{N-1}{2} T_e.$$

Pour cette valeur de τ, la condition exprimant un retard de groupe constant s'écrit :

$$\forall \omega \geq 0 \quad \sum_{i=0}^{N-1} \left(b_i - b_{N-1-i} e^{-2j\phi_0} \right) e^{-j\omega i T_e} = 0$$

Pour que cette relation soit vérifiée il faut que :

$$\forall i \in [0,N-1] \quad b_i = b_{N-1-i} e^{2j\phi_0}$$

Et comme les coefficients sont réels, cette équation entraîne que $e^{2j\phi_0}$ est réel.
C'est à dire qu'il existe deux types de filtre à temps de retard de groupe constant, correspondant à :

$$\phi_0 = \begin{cases} 0 \\ \pm\frac{\pi}{2} \end{cases} \text{ modulo } \pi$$

$$\begin{aligned} \text{Si } \phi_0 = 0 \quad & \forall i \in [0,N-1] \qquad b_i = b_{N-1-i} \\ \text{Si } \phi_0 = \pm\frac{\pi}{2} \quad & \forall i \in [0,N-1] \qquad b_i = -b_{N-1-i} \end{aligned}$$

En conclusion, lorsque le retard de groupe est constant, il vaut :

$\tau = \frac{N-1}{2} T_e$

La phase peut s'écrire de 2 façons:

Le 1er type de phase est défini par : $\forall f \geq 0 \quad \phi(f) = -\pi f (N-1) T_e$

Et dans ce cas : $\forall$ i $\in [0,N-1] \quad b_i = b_{N-1-i}$

On dit alors que la réponse impulsionnelle est **symétrique.**

Le 2ème type de phase est défini par : $\forall f \geq 0 \quad \phi(f) = -\pi f (N-1) T_e \pm \frac{\pi}{2}$

Et dans ce cas : $\forall i \in [0,N-1] \quad b_i = -b_{N-1-i}$

On dit alors que la réponse impulsionnelle est **antisymétrique.**

3.2 Etude des filtres FIR à réponse impulsionnelle symétrique

Ces filtres sont tels que :

$$\forall i \in [0,N-1] \quad b_i = b_{N-1-i}$$

Leur temps de propagation de groupe est constant. Leur phase est linéaire.

$$\tau = \frac{N-1}{2} T_e$$

$$\forall f \geq 0 \quad \phi(f) = -\pi f (N-1) T_e$$

Deux cas sont possibles. Soit N (le nombre de coefficients) est pair, soit N est impair.

3.2.1 Cas réponse symétrique et N pair

Lorsque la réponse impulsionnelle h_n est symétrique et que le nombre de coefficients N est pair, la fonction de transfert en fréquence $H(f)$ s'écrit :

$$\begin{aligned}
H(f) &= \sum_{n=0}^{N-1} b_n e^{-2j\pi f n T_e} = \sum_{n=0}^{N/2-1} b_n \left(e^{-2j\pi f n T_e} + e^{-2j\pi f (N-1-n) T_e} \right) \\
H(f) &= \sum_{n=0}^{N/2-1} 2 b_n e^{-j\pi f (N-1) T_e} \cos\left(2\pi f \left(\frac{N-1}{2} - n \right) T_e \right) \\
H(\omega) &= e^{-j\omega \frac{N-1}{2} T_e} \sum_{k=1}^{N/2} 2 b_{\frac{N}{2}-k} \cos\left(\omega \left(k - \frac{1}{2} \right) T_e \right)
\end{aligned}$$

Comme :

$$\cos\left(\left(k - \frac{1}{2} \right) \omega T_e \right) = \frac{\cos\left(\frac{\omega}{2} T_e \right)}{1 + \cos(\omega T_e)} \left(\cos(k\omega T_e) + \cos\left((k-1)\, \omega T_e \right) \right)$$

On peut écrire :

$$H(\omega) = e^{-j\omega \frac{N-1}{2} T_e} \cos\left(\frac{\omega T_e}{2} \right) \sum_{l=0}^{N/2-1} c_l \cos(l \omega T_e)$$

Les coefficients c_l dépendent simplement des coefficients b_n. Cette forme sera utile par la suite.
On peut noter que, pour $\omega = \pi f_e$, $\cos\left(\frac{\omega T_e}{2} \right) = 0$.

C'est à dire que :

$$H\left(\frac{f_e}{2} \right) = 0$$

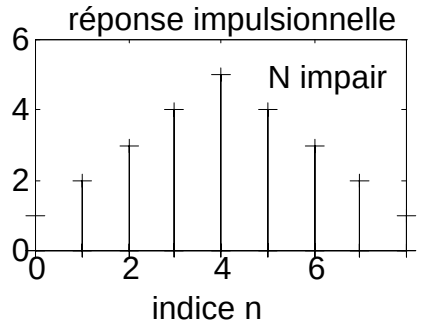
3.2.2 Cas réponse symétrique et N impair

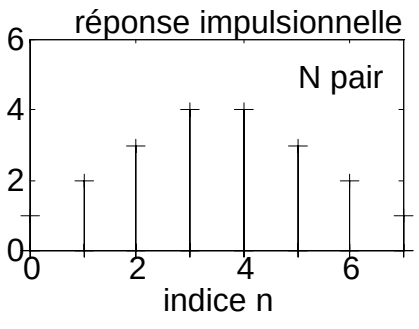
Lorsque le nombre de coefficients N est impair, et la réponse impulsionnelle symétrique, la fonction de transfert s'écrit :

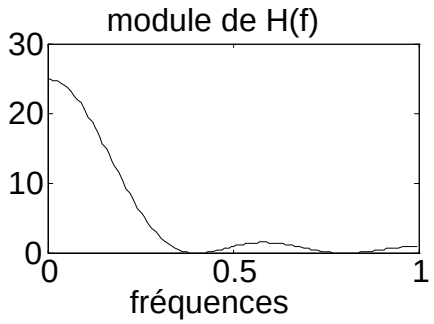
$$\begin{aligned} H(f) &= \sum_{n=0}^{N-1} b_n e^{-2j\pi f n T_e} = \sum_{n=0}^{(N-3)/2} b_n \left(e^{-2j\pi f n T_e} + e^{-2j\pi f (N-1-n) T_e} \right) + b_{(N-1)/2} e^{-2j\pi f \frac{N-1}{2} T_e} \\ H(f) &= e^{-j\pi f (N-1) T_e} \left[b_{(N-1)/2} + \sum_{n=0}^{(N-3)/2} 2 b_n \cos\left(2\pi f \left(\frac{N-1}{2} - n \right) T_e \right) \right] \\ H(f) &= e^{-j\pi f (N-1) T_e} \left[b_{(N-1)/2} + \sum_{k=1}^{(N-1)/2} 2 b_{\frac{N-1}{2}-k} \cos\left(2\pi f k T_e \right) \right] \end{aligned}$$

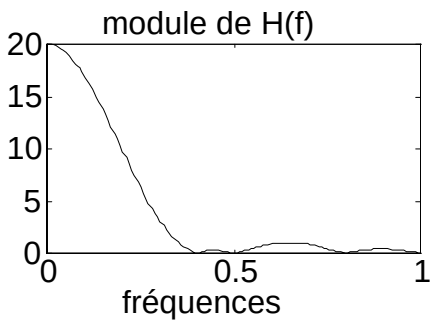
3.2.3 Exemples de FIR à réponse impulsionnelle symétrique

Les figures suivantes représentent le cas d'un filtre FIR à phase linéaire et réponse impulsionnelle symétrique, pour un nombre de coefficients N pair et un nombre de coefficients N impair.









3.3 Etude des filtres FIR à réponse impulsionnelle antisymétrique

3.3.1 Cas réponse antisymétrique et N pair

Lorsque le nombre de coefficients N est pair, et la réponse impulsionnelle antisymétrique, la fonction de transfert s'écrit :

$$\begin{aligned} H(f) &= \sum_{n=0}^{N/2-1} 2jb_n e^{-j\pi f (N-1) T_e} \sin\left(2\pi f \left(\frac{N-1}{2} - n \right) T_e \right) \\ H(\omega) &= e^{-j\omega \frac{N-1}{2} T_e} \sum_{k=1}^{N/2} 2jb_{\frac{N}{2}-k} \sin\left(\omega \left(k - \frac{1}{2} \right) T_e \right) \end{aligned}$$

Comme :

$$\sin\left(\left(k - \frac{1}{2} \right) \omega T_e \right) = \frac{\sin\left(\frac{\omega}{2} T_e \right)}{1 - \cos(\omega T_e)} \left(\cos(k\omega T_e) + \cos\left((k-1)\, \omega T_e \right) \right)$$

On peut écrire :

$$H(\omega) = e^{-j\omega\frac{N-1}{2}T_e} \sin\left(\frac{\omega T_e}{2}\right) \sum_{l=0}^{N/2-1} c_l \cos(l\omega T_e)$$

les coefficients c_l dépendent simplement des coefficients b_n.
On peut noter que :

$$\text{Pour } \omega = 0 \quad \sin\left(\frac{\omega T_e}{2}\right) = 0$$

C'est à dire que :

$$H(0) = 0$$

3.3.2 Cas réponse antisymétrique et N impair

Lorsque le nombre de coefficients N est impair, le coefficient $b_{(N-1)/2}$ vérifie :

$$b_{(N-1)/2} = -b_{(N-1)/2}$$

donc $b_{(N-1)/2} = 0$.
la fonction de transfert s'écrit :

$$\begin{aligned}
H(f) &= \sum_{n=0}^{N-1} b_n e^{-2j\pi f n T_e} = \sum_{n=0}^{(N-3)/2} b_n \left(e^{-2j\pi f n T_e} - e^{-2j\pi f (N-1-n) T_e}\right) \\
H(f) &= e^{-j\pi f (N-1) T_e} \left[\sum_{n=0}^{(N-3)/2} 2jb_n \sin\left(2\pi f \left(\frac{N-1}{2} - n\right) T_e\right)\right] \\
H(f) &= e^{-j\pi f (N-1) T_e} \left[\sum_{k=1}^{(N-1)/2} 2jb_{\frac{N-1}{2}-k} \sin\left(2\pi f k T_e\right)\right]
\end{aligned}$$

Comme :

$$\sin(k\omega T_e) = \frac{\sin(\omega T_e)}{1 - \cos(2\omega T_e)} \left(\cos\left((k-1)\omega T_e\right) - \cos\left((k+1)\omega T_e\right)\right)$$

On peut écrire :

$$H(\omega) = e^{-j\omega\frac{N-1}{2}T_e} \sin(\omega T_e) \sum_{l=0}^{(N-3)/2} c_l \cos(l\omega T_e)$$

Les coefficients c_l dépendent simplement des coefficients b_n.
On peut noter que :

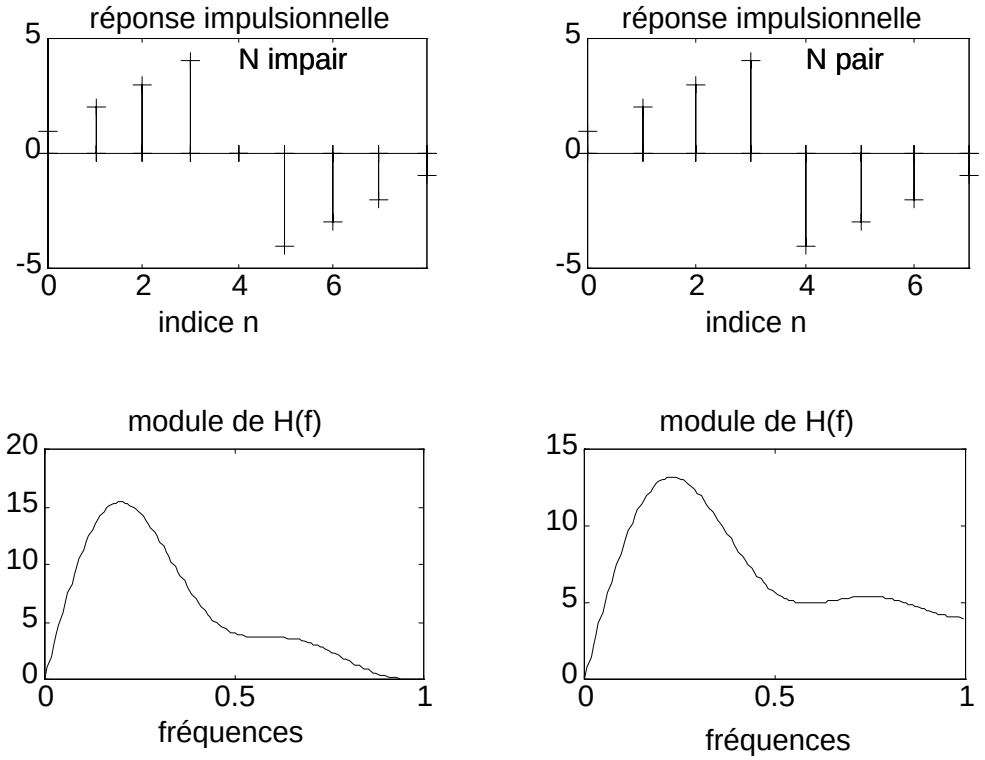
$$\text{Pour } \omega = 0 \text{ et } \omega = \pi f_e \quad \sin\left(\frac{\omega T_e}{2}\right) = 0$$

C'est à dire que :

$$\begin{aligned}
H(0) &= 0 \\
H\left(\frac{f_e}{2}\right) &= 0
\end{aligned}$$

3.3.3 Exemples de FIR à réponse impulsionnelle antisymétrique

Les figures suivantes représentent le cas d'un filtre FIR à retard de groupe constant et réponse antisymétrique, pour un nombre de coefficients N pair et un nombre de coefficients N impair.



Chapitre III

CALCUL DES FILTRES IIR ET FIR

1 Introduction, généralités

Ce chapitre présente différentes méthodes de calcul des coefficients des filtres numé riques IIR et FIR.

Pour satisfaire à certaines spécifications, par exemple un gabarit fréquentiel sur l'atténuation ou sur le temps de groupe, et/ou des contraintes sur la réponse impulsionnelle, il faut :

- Premièrement chercher un vecteur de coefficients (ai et bi par exemple) et un ordre k qui permettront de satisfaire les spécifications souhaitées et correspondront à un filtre réalisable (stable et causal). Selon la méthode utilisée, le filtre pourra être optimum au sens d'un certain critère.
- Deuxièmement, réaliser le filtre, étape dans laquelle on s'efforcera de minimiser l'influence de la limitation du nombre de bits des coefficient et des données. La connaissance du nombre de bits utilisé dans l'implantation peut être prise en compte lors du calcul du filtre, dans certains programmes d'optimisation sous contraintes.

Les spécifications pour un filtre numérique sont tout à fait semblables aux spécifications d'un filtre analogique. Toutefois pour un filtre numérique on précise un paramètre supplémentaire : la fréquence d'échantillonnage fe.

2 Calcul des filtres IIR

Deux démarches sont possibles :

- **L'approche indirecte** utilise les résultats et méthodes connus du filtrage analogique en ce qui concerne le calcul des fonctions d'approximation. Le principe consiste à calculer d'abord un filtre analogique puis à le transformer en un filtre numérique.
- **L'approche directe** utilise des méthodes d'optimisation qui effectuent un calcul direct des coefficients du filtre numérique sans passer par l'intermédiaire d'un filtre analogique.

2.1 Méthodes indirectes

Il existe un grand nombre de telles méthodes qui diffèrent par la transformation permettant de passer du filtre analogique au filtre numérique.

Ces transformations doivent :

- transformer une fonction de transfert rationnelle en p : $H_A(p)$ en une fonction de transfert rationnelle en z : $H_N(z)$.
- Conserver les propriétés de stabilité et de causalité du filtre.
- Assurer que : si le filtre analogique vérifie certaines spécifications dans le domaine analogique, le filtre numérique vérifiera les spécifications souhaitées.
- Transformer une fonction de transfert fréquentielle en ω_A non périodique, définie de $-\infty$ à $+\infty$, en une fonction de transfert en ω_N périodique de période $\omega_e = 2\pi f_e$.

– Conserver, si possible certains critères d'optimalité : ainsi pour un filtre analogique optimum au sens du critère des moindres carrés, il est intéressant d'obtenir un filtre numérique optimum pour le même critère.

2.1.1 Méthode de l'invariance impulsionnelle

Le principe de la méthode est très simple mais peu efficace. On le décrira pour mémoire.

On calcule un filtre analogique qui vérifie les spécifications numériques au moins dans l'intervalle $[0,\omega_e = 2\pi f_e]$. Soit $H_A(p)$ la fonction de transfert en p de ce filtre analogique. On calcule la réponse impulsionnelle $h_A(t)$ du filtre analogique, puis on l'échantillonne à la fréquence f_e. On obtient finalement la réponse impulsionnelle du filtre numérique en posant:

$$h_n = T_e h_A(nT_e)$$

La fonction de transfert en z s'écrit :

$$H_N(z) = T_e \sum_{\text{Pôles de } H_A(p)} \text{Résidus}\left[H_A(p)\frac{1}{1-z^{-1}e^{T_e p}}\right]$$

$H_N(z)$ est donc bien une fonction de transfert rationnelle en z.

Au point de vue fréquentiel, il y a repliement de spectre :

$$H_N(\omega) = \sum_{k=-\infty}^{+\infty} H_A(\omega + k\omega_e)$$

De ce fait les caractéristiques fréquentielles du filtre numérique sont très différentes de celle du filtre analogique et donc de celles cherchées. Ceci se comprend bien dans le cas d'un filtre passe haut, où le repliement de spectre détruit la nature « passe-haut » du filtre analogique. Il est clair qu'il ne sert à rien d'augmenter la fréquence d'échantillonnage.

Cette méthode a toutefois comme intérêt de fournir un filtre numérique dont la réponse impulsionnelle a la forme de la réponse impulsionnelle d'un filtre analogique, ce qui est parfois l'objectif cherché.

2.1.2 Méthode de la transformation bilinéaire

De nombreuses autres méthodes indirectes sont possibles, telles que celle qui consiste à transformer les pôles et les zéros p_i de la fonction de transfert analogique en pôles et zéros z_i pour la fonction de transfert numérique, par la relation :

$$z_i = e^{p_i T_e}$$

Une classe de méthodes consiste à transformer une équation différentielle à coefficients constants correspondant au filtre analogique en une équation de récurrence correspondant au filtre numérique. Cette approche revient à approximer la dérivation ou l'intégration analogique par un calcul numérique à base d'équations aux différences et/ou de sommes discrètes.

Ainsi on peut approximer une dérivation analogique par une différence d'ordre un numérique:

$$\frac{\partial y(t)}{\partial t} \Rightarrow y_n - y_{n-1}$$

C'est à dire que l'on substitue à une multiplication par p dans le plan p, une multiplication par $(1-z^{-1})$ dans le plan z. Les méthodes LDI appartiennent à ce type.

On peut aussi approcher une intégration analogique par une somme d'aires de trapèzes:

$$z(t) = \int_{-\infty}^{t} y(u)du \Rightarrow z(n) = \sum_{k=-\infty}^{n} T_e \frac{(y_k + y_{k-1})}{2}$$

On effectue donc la correspondance suivante entre les plans p et z :

$$Z(p) = \frac{1}{p}Y(p) \Rightarrow Z(z) = \frac{T_e}{2}\frac{1+z^{-1}}{1-z^{-1}}Y(z)$$

On remplace donc l'opérateur d'intégration analogique par un opérateur numérique:

$$\frac{1}{p} \Rightarrow \frac{T_e}{2}\frac{1+z^{-1}}{1-z^{-1}}$$

La méthode décrite ci-dessus qui effectue la substitution :

$$\boxed{p \Rightarrow \frac{2}{T_e}\frac{1-z^{-1}}{1+z^{-1}}}$$

Porte le nom de **transformation bilinéaire**.

Soit une fonction de transfert $H_A(p)$, la transformation bilinéaire remplace cette fonction rationnelle en p en une fonction rationnelle en z:

$$H_N(z) = H_A\left(\frac{2}{T_e}\frac{1-z^{-1}}{1+z^{-1}}\right)$$

Dans cette transformation, l'image de l'axe imaginaire dans le plan p est le cercle unité dans le plan z (et réciproquement) :

$$[\Re(p) = 0 \Leftrightarrow p = j\omega_A] \Leftrightarrow [z \in \text{cercle unité } \Leftrightarrow |z| = 1]$$

$$z = \frac{2/T_e + p}{2/T_e - p} = \frac{2/T_e + j\omega_A}{2/T_e - j\omega_A p} \Rightarrow |z| = 1 = e^{j\omega_N T_e}$$

D'un point de vue fréquentiel, on effectue la substitution suivante:

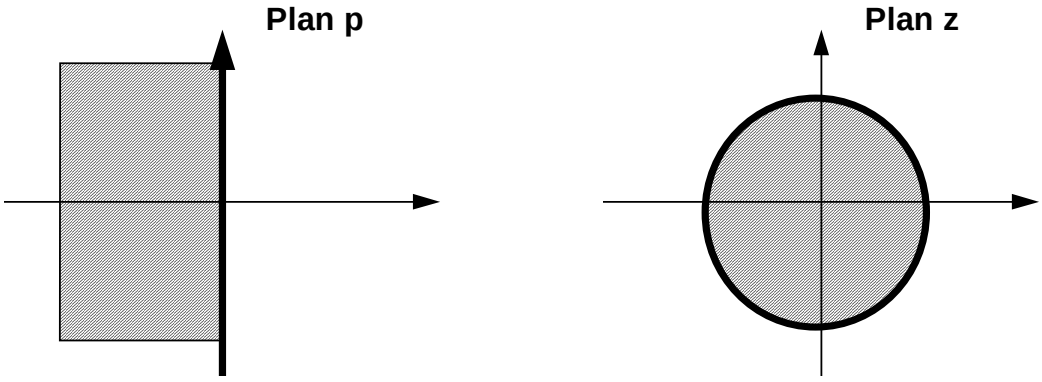
$$\begin{aligned} j\omega_A &\rightarrow \frac{2}{T_e}\frac{1-e^{-j\omega_N T_e}}{1+e^{-j\omega_N T_e}} = \frac{2j}{T_e}\frac{\sin(\omega_N T_e/2)}{\cos(\omega_N T_e/2)} \\ \omega_A &\rightarrow \frac{2}{T_e} tg\,(\omega_N T_e/2) \\ H_A(\omega_A) &= H_N\,(\omega_N) \end{aligned}$$

On peut remarquer que la relation liant les fréquences des deux plans n'est pas linéaire. L'image de la fréquence nulle est la fréquence nulle. Pour les fréquences petites devant f_e, la tangente peut s'assimiler à son argument et $\omega_A \approx \omega_N$. L'image de la fréquence infinie est $f_e/2$. L'image de l'axe des fréquences analogiques de $-\infty$ à $+\infty$ est l'intervalle $[-f_e/2, f_e/2]$.

Par ailleurs l'image de la partie gauche du plan p est l'intérieur du cercle unité :

$$[\Re(p) < 0 \Leftrightarrow p = a + jb \text{ et } a < 0] \Leftrightarrow [z \in \text{ intérieur du cercle unité } \Leftrightarrow |z| < 1]$$

Ces relations sont résumées sur le schéma suivant:



Une des conséquences de ces relations est qu'une fonction de transfert analogique $H_A(p)$ stable et causale a pour transformée une fonction de transfert $H_N(z)$ stable et causale.

D'autre part :

$$|H_A(\omega_A)| < A \quad \forall\omega_A \in [\omega_{A1},\omega_{A2}] \quad \Rightarrow |H_N(\omega_N)| < A \quad \forall\omega_N \in [\omega_{N1},\omega_{N2}]$$

De même :

$$|H_A(\omega_A)| > A \quad \forall \omega_A \in [\omega_{A1},\omega_{A2}] \quad \Rightarrow |H_N(\omega_N)| > A \forall \omega_N \in [\omega_{N1},\omega_{N2}]$$

Où les pulsations ω_A et ω_N sont images par la transformation bilinéaire.
Ainsi, lorsque le filtre analogique vérifie un gabarit d'atténuation défini par des fréquences caractéristiques ω_{Ai} et des atténuations A_k, le filtre numérique correspondant vérifie un gabarit d'atténuation défini par des fréquences caractéristiques ω_{Ni} (images des ω_{Ai}) et les mêmes atténuations A_k.

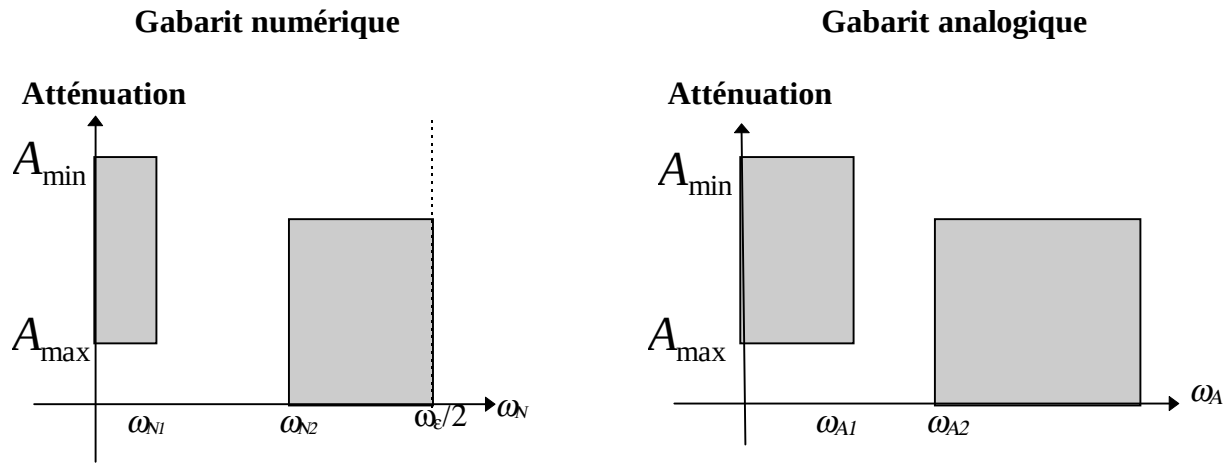
Cette méthode est tout à fait adaptée au calcul d'un filtre devant satisfaire un gabarit d'atténuation constant par morceaux. Etant donné un gabarit d'affaiblissement constant par morceaux défini par des fréquences caractéristiques ω_{Ni}, des atténuations A_k. et une fréquence d'échantillonnage f_e, la démarche est la suivante :

- Prédistordre le gabarit numérique de façon à obtenir un gabarit analogique tel qu'en appliquant la transformation bilinéaire au filtre analogique on retrouve le bon gabarit numérique. Dans cette prédistortion les atténuations sont conservées mais les fréquences caractéristiques Ni sont transformées par la relation:

$$\omega_{Ni} \rightarrow \omega_{Ai} \quad \omega_{Ai} = \frac{2}{T_e} \mathrm{tg}\,(\omega_{Ni} T_e/2)$$

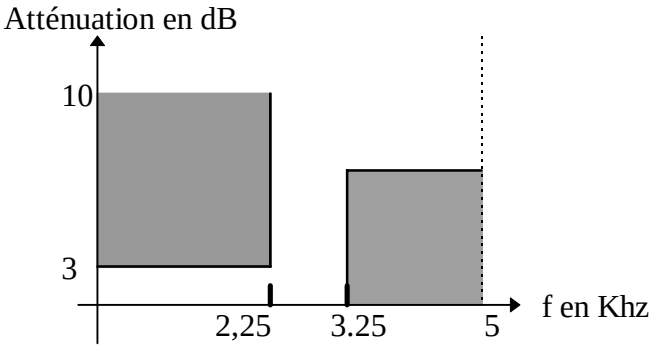
- Calculer un filtre analogique vérifiant le gabarit analogique ainsi obtenu. On utilise pour calculer ce filtre analogique les fonctions d'approximation classiques (Buterworth, tchebychef, Cauer) ou une méthode (placeur de pôles par exemple) permettant d'optimiser le critère souhaité (moindres carrés, tchebychef,...). On obtient ainsi une fonction de transfert $H_A(p)$.
- Appliquer la transformation bilinéaire à la fonction $H_A(p)$. On obtient alors la fonction de transfert numérique cherchée $H_N(z)$.

La méthode de la transformation bilinéaire est la plus utilisée parmi les méthodes indirectes. En effet, pour des gabarits d'affaiblissement, elle fournit une solution numérique qui a les mêmes qualités que le filtre analogique dont elle est issue. La stabilité et la causalité sont conservées. Si le filtre analogique entre dans le gabarit analogique, le filtre numérique entre dans le gabarit numérique de départ. Si le filtre analogique est optimum au sens du critère des moindres carrés ou du minimax, il en est de même du filtre numérique. Par contre la transformation de l'axe des fréquences n'étant pas linéaire, la forme du temps de retard de groupe est modifiée.
La figure suivante représente la prédistortion du gabarit numérique.

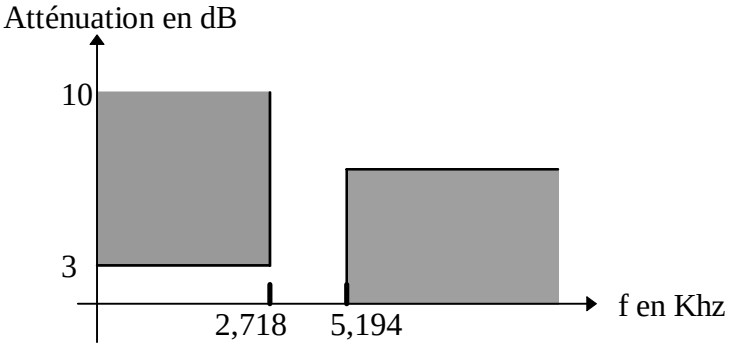


Exemple numérique :

Soit le gabarit numérique passe - bas suivant :



La fréquence d'échantillonnage vaut 10 Khz.
Le gabarit analogique obtenu par prédistortion du gabarit analogique est caractérisé par les fréquences $f_{A1} = 2{,}718$ Khz et $f_{A2} = 5.194$ Khz. Il est représenté sur la figure suivante.



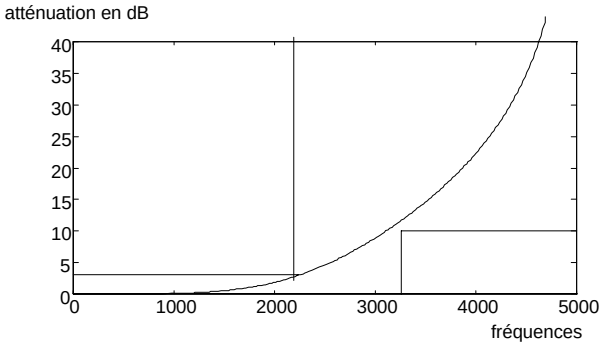
Un filtre de Butterworth d'ordre 2 suffit pour ce gabarit.
La fonction de transfert analogique $H_A(p)$ vaut:

$$H_A(p) = \frac{1}{\frac{p^2}{\omega_n^2} + \sqrt{2}\frac{p}{\omega_n} + 1}$$

avec : $\omega_n = 2718{,}63$ Hz.
La fonction de transfert numérique HN(z) vaut:

$$\frac{0.2483 - 0.4967z^{-1} + 0.2483z^{-2}}{1 - 0.1842z^{-1} + 0.1775z^{-2}}$$

La courbe suivante représente la fonction de transfert fréquentielle correspondante:



2.2 Méthodes directes

Les méthodes directes consistent à calculer directement le filtre numérique, sans passer par l'intermédiaire d'un filtre analogique.

Ce sont principalement des méthodes d'optimisation. Elles s'appliquent aussi bien pour des spécifications fréquentielles que pour des spécifications temporelles.

Le principe de ces méthodes est le suivant : on cherche le jeu de paramètres a_i, b_i tel que la distance entre une caractéristique du filtre correspondant aux ai, bi et la caractéristique désirée soit minimum.

Ces méthodes diffèrent selon la norme utilisée pour mesurer cette distance. Les normes les plus utilisées sont les normes Lp, et plus particulièrement les normes L_2 et L_∞ (Tchebycheff, ou minimax). Divers algorithmes d'optimisation peuvent être utilisés selon le critère à minimiser.

D'autre part, les paramètres à optimiser ne sont pas forcément les coefficients a_i, b_i vus jusqu'à maintenant. On peut préférer travailler avec d'autres paramètres : comme la valeur des pôles et des zéros, ou les paramètres d'une représentation d'état du système, ou les coefficients d'une structure treillis équivalente. Le choix du jeu de paramètres n'est pas anodin, il détermine en partie le comportement des algorithmes et il influence la difficulté de contrôle de la stabilité du filtre.

Norme L_2 : minimisation de l'erreur quadratique

On cherche à minimiser une erreur quadratique . Supposons que l'on cherche à optimiser le module de la fonction de transfert en fréquence, le critère s'écrit alors:

$$\min \epsilon = \int_{-\frac{f_e}{2}}^{\frac{f_e}{2}} (|H(f)| - |H_D(f)|)^2 \, df$$

Où $H_D(f)$ est la fonction à approcher.

En fait comme les calculs sont numériques , on travaille sur une grille de fréquence f_n, et on cherche à minimser J:

$$\min J = \sum_{n=1}^{P} (|H(f_n)| - |H_D(f_n)|)^2$$

Il s'agit d'un problème d'optimisation non linéaire, compliqué par la contrainte de stabilité du filtre.

De façon à ne pas être piégé dans de mauvais minima locaux, il est important de bien déterminer le point de départ de ces algorithmes itératifs.

En ce qui concerne la contrainte de stabilité 2 approches sont possibles. On peut introduire cette contrainte dans l'algorithme d'optimisation. Ou bien, simplement vérifier à la fin de chaque itération que le jeu de paramètres obtenus correspond à un filtre stable. Si certains pôles sont à l'extérieur du cercle unité, on peut rendre stable le filtre sans changer le module de la fonction de transfert fréquentielle. Il suffit, pour cela, de remplacer les pôles instables par des pôle de même argument mais de module inverse:

$$(r,\theta) \Rightarrow \left(\frac{1}{r},\theta\right)$$

3 Calcul des filtres FIR

On utilise uniquement des méthodes directes pour le calcul des filtres FIR.

Il existe des méthodes sous-optimales très simples pour le calcul des filtres FIR. On peut citer dans cette catégorie, les méthodes de la fenêtre et de l'échantillonnage en fréquence. Elles sont toutefois suffisamment performantes pour être suffisantes dans de nombreux cas pratiques, et pour servir d'initialisation dans les algorithmes d'optimisation.

3.1 Méthode de la fenêtre

Considérons le cas de spécifications en fréquence.

La fonction de transfert désirée $H_D(f)$ est périodique. Elle est définie de $-f_e/2$ à $+f_e/2$.

Soit la fonction $H_D(f)$ égale à la fonction désirée dans l'intervalle $[-f_e/2, +f_e/2]$ et nulle en dehors de cet intervalle.

Supposons que l'on cherche un FIR à temps de retard de groupe constant. Dans cet objectif, on suppose que la phase de $H_D(f)$ est soit nulle soit égale à $\pm\pi/2$.

Soit N le nombre de coefficients du filtre FIR recherché.

On notera $H(z)$ la fonction de transfert en z :

$$H(z) = \sum_{n=0}^{N-1} h_n z^{-n}$$

Le principe de la méthode consiste à :

1. Calculer la réponse impulsionnelle $h_D(t)$ correspondant à $H_D(f)$. Cette réponse temporelle est de durée infinie et non discrète. Elle est en général non causale. Mais compte tenu de la contrainte de phase, elle est paire ou impaire.
2. Echantillonner $h_D(t)$ à la période d'échantillonnage T_e, de façon à obtenir la réponse impulsionnelle numérique de durée infinie $h_D(n)$. On multiplie par T_e cette réponse pour des raisons de normalisation. L'échantillonnage se fait sur les multiples de T_e si N est impair, ou aux instants $t_n = nT_e + T_e/2$ si N est pair. On obtient de cette façon : $h_D(n) = T_e h_D(nT_e)$ si N est impair et $h_D(n) = T_e h_D(0.5Te + nT_e)$ si N est pair.
3. Limiter en durée, cette réponse $h_D(n)$ en la multipliant par une fenêtre $w(n)$ de longueur finie N, symétrique par rapport à l'origine des temps. cette symétrie est nécessaire pour conserver la contrainte de temps de retard de groupe. Soit $h_{nc}(n)$ le résultat : $h_{nc}(n) = w(n)h_D(n)$. Cette fonction $h_{nc}(n)$ n'est pas causale.
4. Rendre causale cette réponse impulsionnelle, en la retardant de $\frac{(N-1)T_e}{2}$. La fonction retardée est le résultat de la méthode: $h_n = h_{nc}(n - (N-1)/2)$.

En conclusion que N soit pair ou impair :

$$\begin{aligned} h_n &= T_e h_D\left(nT_e - \frac{(N-1)T_e}{2}\right) w\left(nT_e - \frac{(N-1)T_e}{2}\right) \quad \text{pour} \quad n \in [0,N-1] \\ h_n &= 0 \quad \text{pour} \quad n \notin [0,N-1] \end{aligned}$$

Dans le domaine fréquentiel, l'échantillonnage de $h_D(t)$ a rendu périodique $H_D(f)$. La limitation de la durée temporelle par la multiplication avec w(n), a convolué la fonction de transfert fréquentielle avec la transformée de Fourier de la fenêtre. Enfin le retard de $(N-1)T_e/2$ n'a pas changé le module du spectre mais a ajouté un déphasage proportionnel à la fréquence.

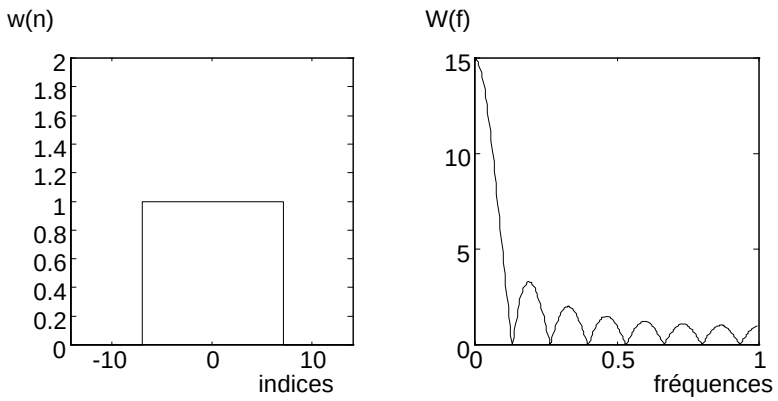
La fenêtre $w(n)$ introduit des ondulations sur la fonction de transfert et limite la raideur du filtre obtenu. Le choix de cette fenêtre se fait de façon à obtenir un compromis satisfaisant entre la hauteur des ondulations (phénomène de Gibbs) et la raideur du filtre. Evidemment plus la fenêtre temporelle est large, plus son spectre est étroit et se rapproche d'une impulsion.

3.1.1 Exemples de fenêtres

On considère des fenêtres centrées et pour des raisons de simplicité de notations on se limite à des longueurs N impaires, les résultats se généralisant très simplement au cas d'une longueur N paire.

Fenêtre rectangulaire de longueur N impaire:

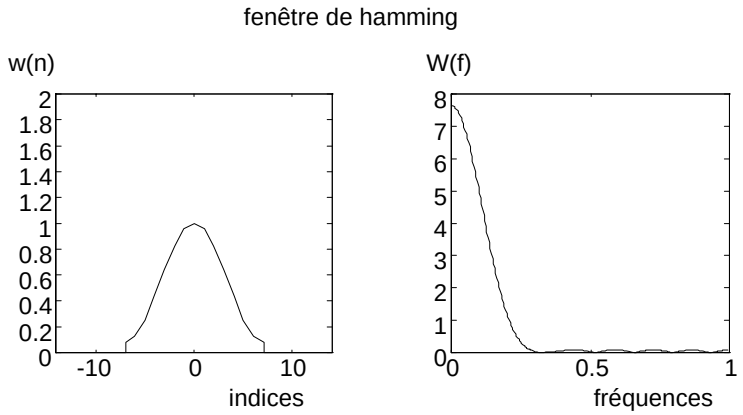
$$\begin{aligned} w(n) &= 1 \quad \in \left[-\frac{(N-1)}{2}, \frac{(N-1)}{2}\right] \\ w(n) &= 0 \quad n \notin \left[-\frac{(N-1)}{2}, \frac{(N-1)}{2}\right] \\ W(f) &= \frac{\sin(\pi f N T_e)}{\sin(\pi f T_e)} \end{aligned}$$



Fenêtre de Hamming :

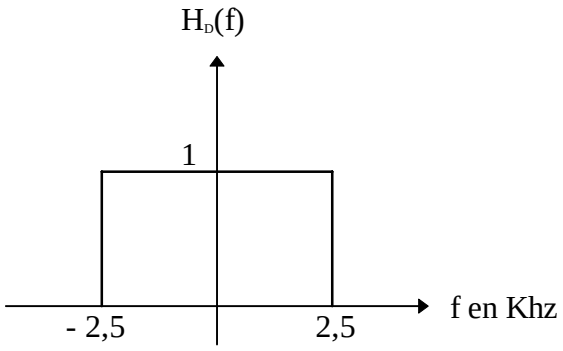
$$w(n) = 0.54 - 0.46\cos\left(2\pi\left(n - \frac{N-1}{2}\right)/N-1\right) \quad \text{si } n \in \left[-\frac{(N-1)}{2}, \frac{(N-1)}{2}\right]$$
$$w(n) = 0 \quad \text{si } n \notin \left[-\frac{(N-1)}{2}, \frac{(N-1)}{2}\right]$$

Il existe de nombreuses autres fenêtres. La fenêtre de Dolb-tchebycheff est celle qui présente le lobe principal le plus étroit pour une hauteur donnée des ondulations.



Exemple numérique :
Calcul d'un passe - bas de bande de transition allant de 2 Khz à 3 Khz, pour une fréquence d'échantillonnage de 10 Khz et un nombre N de coefficients.

On choisit comme fonction de transfert désirée, une fonction $H_D(f)$ paire, égale à 1 de 0 à 2,5 Khz et nulle à partir de 2,5 Khz.



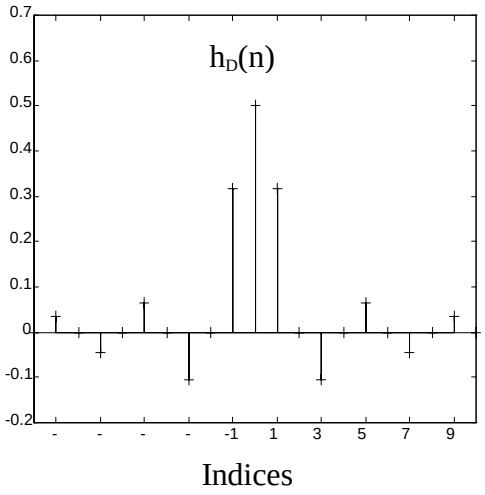
La fonction désirée a pour transformée de Fourier inverse $h_D(t)$:

$$h_D(t) = 5000\frac{\sin(\pi t 5000)}{(\pi t 5000)}$$

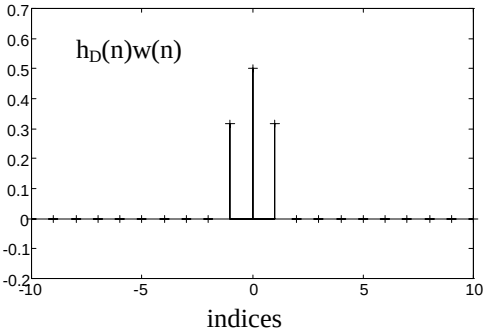
Cas N impair N=3

- On échantillonne $h_D(t)$ à f_e et on multiplie par T_e. Le résultat $h_D(n)$ est représenté sur la courbe

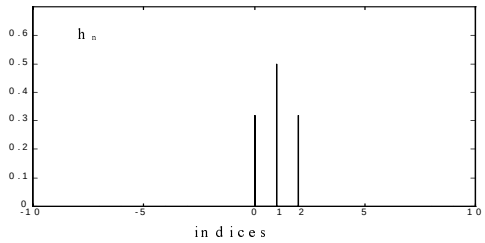
ci-dessous.



– Puis on limite la durée en multipliant $h_D(n)$ par une fenêtre $w(n)$. Pour une fenêtre rectangulaire, de longueur 3, on obtient les 3 échantillons de la figure suivante :



– Puis on rend causal le résultat en introduisant un retard de 1 $(1 = (N-1)/2)$ échantillon. La fonction ainsi obtenue est le résultat final h_n et est représentée sur la figure suivante.



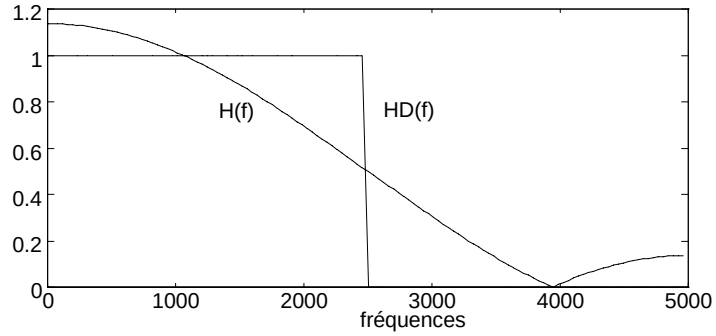
En conclusion : pour N=3 et une fenêtre rectangulaire, on trouve:

$$\begin{aligned} h_0 &= 0.3183 = 1/\pi \\ h_1 &= 0.5 \\ h_2 &= 0.3183 = 1/\pi \end{aligned}$$

D'où:

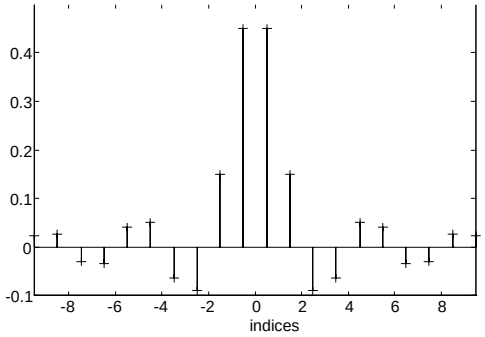
$$\begin{aligned} H(z) &= h_0 + h_1 z^{-1} + h_2 z^{-2} = \frac{1}{\pi}\left(1 + z^{-2}\right) + 0.5 z^{-1} \\ H(f) &= \exp\left(-j2\pi f T_e\right)\left[0.5 + \frac{1}{\pi} 2\cos\left(2\pi f T_e\right)\right] \end{aligned}$$

La figure suivante représente le module de la fonction $H(f)$ et de la fonction $H_D(f)$.

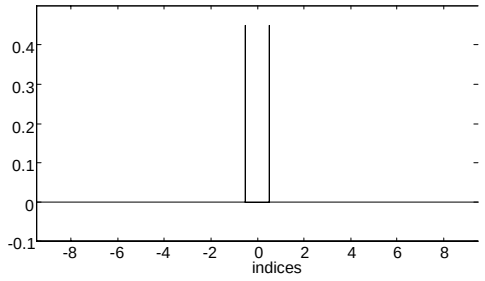


Cas N pair N=2

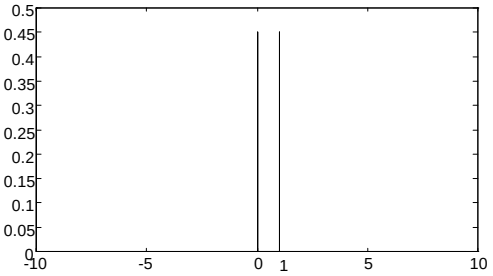
- On échantillonne $h_D(t)$ à f_e et on multiplie par T_e. Le résultat $h_D(n)$ diffère du résultat précédent car la grille d'échantillonnage ne comprend pas l'instant $t = 0$, les instants d'échantillonnage sont du type:
  $t = nT_e + \frac{T_e}{2}$
  La figure suivante représente $h_D(n)$.



- Puis on limite la durée en multipliant $h_D(n)$ par une fenêtre $w(n)$. Pour une fenêtre rectangulaire, de longueur 2, on obtient les 2 échantillons de la figure suivante :



  Puis on rend causal le résultat en introduisant un retard de $T_e/2$ ($T_e/2 = (N-1)T_e/2$). La fonction ainsi obtenue est le résultat final h_n et est représentée sur la figure suivante.



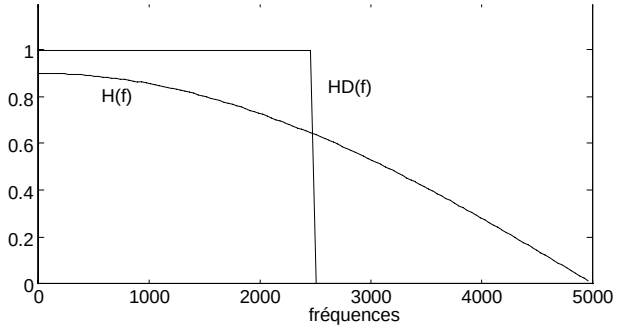
En conclusion : pour $N = 2$ et une fenêtre rectangulaire, on trouve:

$$\begin{aligned} h_0 &= 0.4502 = 2\sin(\pi/4)/\pi \\ h_1 &= h_0 \end{aligned}$$

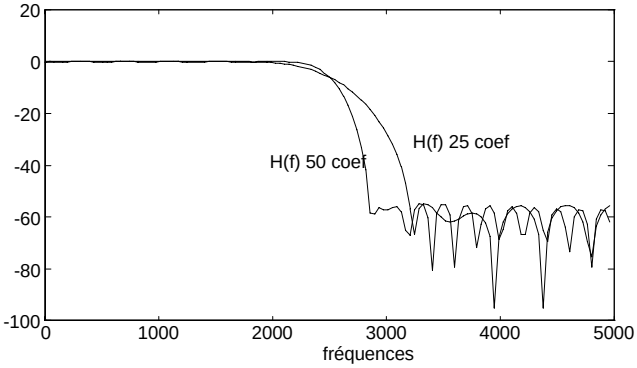
D'où :

$$\begin{aligned} H(z) &= h_0 + h_1 z^{-1} = \frac{\sqrt{2}}{\pi}\left(1 + z^{-1}\right) \\ H(f) &= \exp\left(-j\pi f T_e\right)\frac{\sqrt{2}}{\pi}\cos\left(\pi f T_e\right) \end{aligned}$$

La figure suivante représente le module de la fonction $H(f)$ et de la fonction $H_D(f)$ sur une échelle linéaire.



Si on augmente le nombre de coefficients N, la raideur du filtre augmente mais la hauteur maximale des ondulations ne diminue pas (phénomène de Gibbs). La figure suivante illustre ce phénomène, elle représente les fonctions de transfert (en dB) obtenues avec une fenêtre rectangulaire et $N = 50$ coefficients ou $N = 25$ coefficients.



3.2 Méthode de l'échantillonnage en fréquence

Un filtre FIR à N coefficients ayant comme son nom l'indique, une durée limitée dans le domaine temporel, le théorème de Shannon permet de dire que ce filtre est parfaitement défini :

– soit par sa réponse impulsionnelle :

$$h_i \quad i \in [0,N-1],$$

– soit par N valeurs H_k dans le domaine fréquentiel :

$$H_k = H\left(f = k\frac{f_e}{N}\right) = \sum_{n=0}^{N-1} h_n \exp\left(-2\pi j\frac{nk}{N}\right).$$

Les valeurs h_i sont liées aux H_k par la relation :

$$h_i = \frac{1}{N}\sum_{k=0}^{N-1} H_k \exp\left(2\pi j\frac{ik}{N}\right).$$

Les N valeurs H_k correspondent à un échantillonnage de la première période de $H(f)$ avec un pas d'échantillonnage égal à la limite de Shannon : f_e/N.

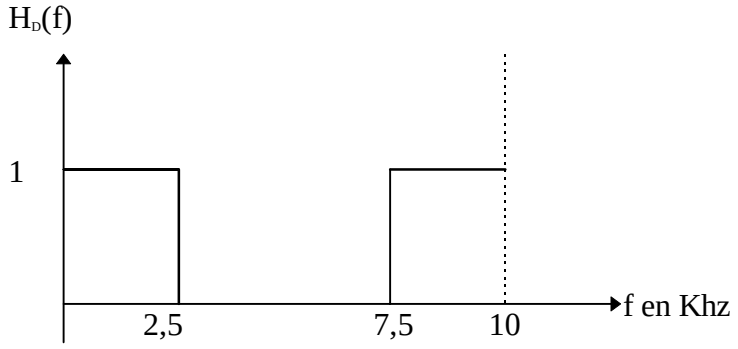
Par ailleurs la fonction $H(f)$ peut s'exprimer en fonction de ces valeurs H_k par la relation :

$$H(f) = \frac{1}{N}\exp\left(-j\pi f\left(N-1\right)T_e\right)\sum_{k=0}^{N-1} H_k \exp\left(-j\pi k/N\right)\frac{\sin\left(\pi f N T_e\right)}{\sin\left(\pi f T_e - k\pi/N\right)}.$$

Soit $H_D(f)$ la fonction de transfert que l'on désire approcher.
Une manière très simple de déterminer un filtre FIR approchant $H_D(f)$, consiste à échantillonner $H_D(f)$ dans le domaine fréquentiel avec un pas d'échantillonnage égal à f_e/N. Cet échantillonnage fournit N valeurs H_k à partir desquelles on peut calculer N valeurs h_i par une Transformée de Fourier Discrète Inverse. Le filtre FIR ainsi obtenu, présente une erreur $e(f) = H(f) - H_D(f)$ nulle aux points d'échantillonnage et finie entre ces points.

Exemple numérique :

Considérons le même exemple que pour la méthode précédente:
Calcul d'un passe - bas de bande de transition allant de 2 Khz à 3 Khz, pour une fréquence d'échantillonnage de 10 Khz et un nombre N de coefficients.
On choisit comme fonction de transfert désirée, une fonction $H_D(f)$ paire de période $f_e = 10Khz$. Sur la première période $H_D(f)$ est égale à 1 de 0 à 2,5 Khz, nulle de 2,5 Khz à 7,5 Khz, puis égale à 1 de 7,5 à 10 Khz.



Cas N=5
On échantillonne $H_D(f)$ avec un pas en fréquence égal à $f_e/5$, soit 2 Khz. On obtient 5 valeurs H_k :

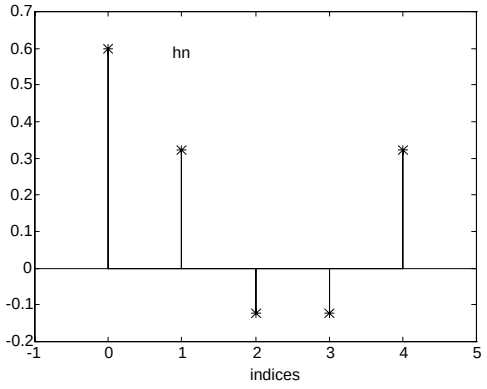
$$H_0 = 1,\ H_1 = 1,\ H_2 = 0,\ H_3 = 0,\ H_4 = 1.$$

De ces 5 valeurs fréquentielles on calcule 5 valeurs temporelles h_i correspondant à la réponse impulsionnelle du filtre FIR cherché.

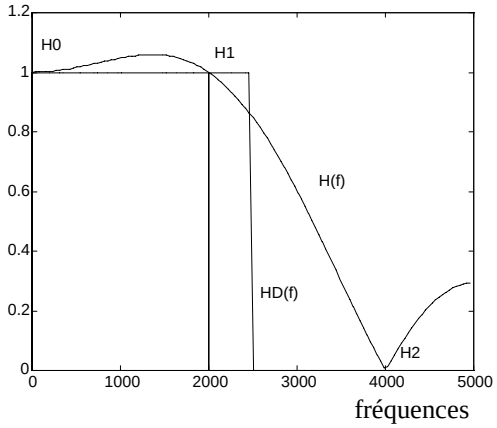
$$h_i = \frac{1}{N} \sum_{k=0}^{4} H_k \exp\left(2\pi j \frac{ik}{5}\right).$$

D'où :

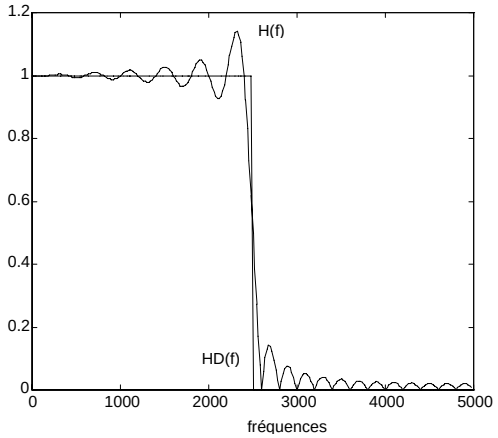
$$\begin{aligned} h_0 &= \frac{3}{5} = 0{,}6 \\ h_1 &= 0{,}3236 \\ h_2 &= -0{,}1236 \\ h_3 &= -0{,}1236 \\ h_4 &= 0{,}3236 \end{aligned}$$



La fonction de transfert fréquentielle de ce filtre passe par les 5 points H_k. Elle est représentée sur la figure suivante.



Avec $N = 50$ coefficients, on obtient la fonction de transfert fréquentielle représentée sur la figure suivante (l'amplitude est sur une échelle linéaire) :



Les deux approches présentées ci-dessus : méthode de la fenêtre et méthode de l'échantillonnage en fréquence, sont sous-optimales. Dans le cas du calcul d'un filtre FIR on peut sans difficulté calculer un filtre optimum au sens de la norme L_2 ou L_∞.

3.3 Calcul d'un filtre FIR optimum pour la norme L_2

Cette méthode calcule le filtre de fonction de transfert $H(f)$ telle que :

$$J = \int_0^{f_e} \left(H(f) - H_D(f)\right)^2 df \quad \text{soit minimum.}$$

Si les coefficients optimisés sont les b_i, il s'agit d'un critère quadratique et d'un simple problème d'optimisation linéaire. Le filtre optimal d'ordre $N-1$ est le filtre de coefficients b_i (rappelons que pour un filtre FIR $b_i = h_i$) tels que :

$$\forall i \in [0,N-1] \qquad \frac{\partial J}{\partial b_i} = 0.$$

Il suffit donc de résoudre un système de N équations linéaires.
En pratique le critère J est calculé sur une grille de fréquence, avec un pas d'échantillonnage inférieur à f_e/N. Typiquement : $\frac{f_e}{NM}$ et :

$$J = \sum_{n=0}^{NM-1} \left(H(f_n) - H_D(f_n)\right)^2 = \sum_{n=0}^{NM-1} e\left(n\frac{f_e}{NM}\right)^2.$$

On peut d'autre part attacher une pondération $W(n)$ à chaque échantillon fréquentiel.

$$J = \sum_{n=0}^{NM-1} e\left(n\frac{f_e}{NM}\right)^2 W(n)^2.$$

Le problème peut être résolu soit de manière directe, soit en utilisant des méthodes itératives, comme l'algorithme du gradient.

3.4 Calcul d'un filtre FIR optimum pour la norme L_∞

On cherche le filtre de fonction de transfert H(f) qui minimise le maximum de l'erreur, c'est à dire optimum pour le critère J :

$$\min\left(\max\left(|H(f) - H_D(f)|\right)\right)$$
$$\min\left(\max\left(|e(f)|\right)\right) \text{e(f)} = H(f) - H_D(f)$$

Ce critère est évalué sur une grille de fréquences.

Une méthode efficace pour résoudre ce problème, dans le cas de filtres FIR à temps de retard de groupe constant, consiste à utiliser l'algorithme de Remez.
En effet, la fonction de transfert fréquentielle des filtre FIR à temps de retard de groupe constant, peut toujours s'écrire sous la forme:

$$H(f) = Q\left(f\right) P\left(f\right).$$

où :

- $Q(f)$ est une fonction de f qui ne dépend pas des coefficients du filtre,
- $P(f)$ est une combinaison linéaire de fonctions de type $\cos(2\pi f n T_e)$, les coefficients de cette combinaison linéaire dépendant des coefficients du filtre de manière simple.

Les diverses formes possibles de $Q(f)$ et de $P(f)$ ont été vues dans le chapitre sur l'étude des filtres FIR à temps de retard de groupe constant. Elles sont rappelées dans le tableau suivant, où N représente le nombre de coefficients du filtre :

Type du filtre	$Q(f)$	$P(f)$
N impair. Réponse impulsionnelle symétrique	$e^{(-\pi j f (N-1) T_e)}$	$\sum_{n=0}^{(N-1)/2} r_n \cos\left(2\pi f n T_e\right)$
N pair. Réponse impulsionnelle symétrique	$\cos\left(\pi f T_e\right)\ e^{(-\pi j f (N-1) T_e)}$	$\sum_{n=0}^{N/2-1} g_n \cos\left(2\pi f n T_e\right)$
N impair. Réponse impulsionnelle antisymétrique	$\sin\left(2\pi f T_e\right)\ e^{(-\pi j f (N-1) T_e)}$	$\sum_{n=0}^{(N-3)/2} c_n \cos\left(2\pi f n T_e\right)$
N pair. Réponse impulsionnelle antisymétrique	$\sin\left(\pi f T_e\right)\ e^{(-\pi j f (N-1) T_e)}$	$\sum_{n=0}^{N/2-1} d_n \cos\left(2\pi f n T_e\right)$

Les coefficients r_n, g_n, c_n, d_n sont liés simplement aux coefficients b_n du filtre.
On note $e(f)$ l'erreur pondérée entre $H(f)$ et $H_D(f)$:

$$e(f) = W(f)\left(H(f) - H_D(f)\right).$$

$W(f)$ est la fonction de pondération.
En remplaçant $H(f)$ par son expression en fonction de $P(f)$ et de $Q(f)$, Cette erreur s'exprime par :

$$e(f) = W(f)Q(f)\left(P(f) - \frac{H_D(f)}{Q(f)}\right) = \tilde{W}(f)\left(P(f) - \tilde{H}_D(f)\right).$$

Le problème revient donc à approcher au sens d'un critère de tchebycheff pondéré une fonction $\tilde{H}_D(f)$ par une fonction $P(f)$ qui s'exprime comme une combinaison linéaire de fonctions de type $\cos\left(2\pi f n T_e\right)$.

Le critère est optimisé sur une grille de fréquences f_n.

Le théorème d'alternance peut être utilisé ici.

Théorème d'alternance

Si $P(f)$ est une combinaison linéaire de K fonctions $\cos\left(2\pi f n T_e\right)$, une condition nécessaire et suffisante pour que $P(f)$ soit la meilleure approximation au sens de Tchebycheff pondéré de la fonction

$\tilde{H}_D(f)$ sur un intervalle I sous ensemble compact de $[0,f_e/2]$ est que la fonction d'erreur $e(f)$ présente $K+1$ fréquences dans l'intervalle pour lesquelles le module de $P(f)$ passe par un extrémum dont le signe alterne. C'est à dire qu'il existe $K+1$ fréquences f_i telles que :

$$f_1 < f_2 < \cdots < f_{K+1}.$$

$$e(f_i) = -e(f_{i+1}) = (-1)^i \max_{f\in I}(e(f)).$$

Dans la suite on note : $\delta = \max_{f\in I} e(f)$
L'algorithme de Remez (ou de Parks McClellan) permet de déterminer de façon itérative les fréquences f_i :

- On part d'un ensemble quelconque de K+1 valeurs f_i et on calcule la valeur associée.
  En effet lorsque la position des fréquences extrémales est fixée, la valeur de δ est automatiquement déterminée par les relations:
  $$\forall i \in [0,K] \quad e(f_i) = (-1)^i \delta$$
  Ces relations forment un système de $(K+1)$ équations linéaires à $(K+1)$ inconnues (r_n, g_n, c_n, d_n) et δ.
- Les valeurs de la fonction $P(f)$ sont obtenues par la formule d'interpolation de Lagrange.
  On utilise les valeurs extrémales de ce polynôme comme départ pour une nouvelle itération.
  La valeur de δ augmente à chaque itération.
  On interrompt le processus lorsque la variation de δ d'une étape à l'autre tombe en dessous d'un seuil fixé à priori.

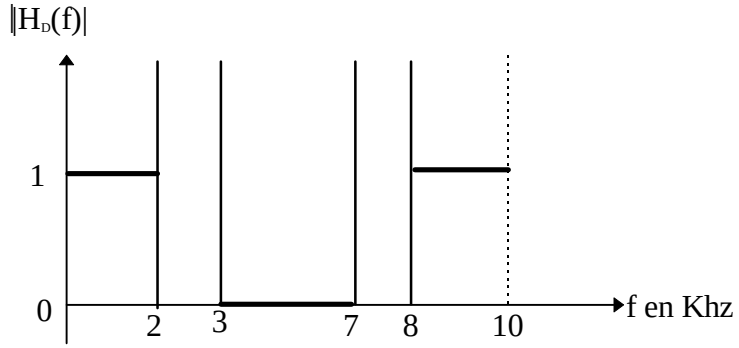
Lorsque l'on connait les $K+1$ fréquences optimales et la valeur de l'extrémum, la valeur des coefficients (r_n, g_n, c_n ou d_n) se détermine en résolvant un simple système linéaire. A partir de ces K coefficients on peut déterminer les K coefficients b_i du filtre.

Remarque sur le nombre d'extréma de $e(f)$:
La fonction $P(f)$ est une combinaison linéaire de K fonctions cosinus. Il est clair que sur l'intervalle $[0,f_e/2]$, la fonction $\cos(2\pi f(K-1)T_e)$ (et donc la fonction $e(f)$ possède au plus K extréma. Or il y a $K+1$ inconnues. Mais en général le gabarit est spécifié sur des intervalles disjoints(zones de transition) et de ce fait on aura un extremum en chaque fréquence limite. Ainsi, un passe-bas est défini par une bande passante de fréquences limites $[0,f_1]$ et une bande atténuée de fréquences limites $[f_2,f_e/2]$. Les fréquences 0 et $f_e/2$ correspondent à des extréma des fonctions cos, mais la fonction $e(f)$ présentera de plus des extrémas en f_1 et f_2. Le nombre maximum d'extréma de $e(f)$ pour un passe bas est donc $K+2$.

Exemple numérique

Calcul d'un FIR à phase linéaire, passe - bas de bande de transition allant de 2 Khz à 3 Khz, pour une fréquence d'échantillonnage de 10 Khz et un nombre N de coefficients.

On choisit comme fonction de transfert désirée, une fonction $H_D(f)$ paire de période $f_e = 10Khz$. Sur la première période |HD(f)| est égale à 1 de 0 à 2 Khz, nulle de 3 Khz à 7 Khz, puis égale à 1 de 8 à 10 Khz.



Pour un filtre à $N = 5$ coefficients:
On note $H(z)$ la fonction de transfert recherchée.

$$\begin{aligned} H(z) &= h_o + h_1 z^{-1} + h_2 z^{-2} = h_0\left(1 + z^{-4}\right) + h_1\left(z^{-1} + z^{-3}\right) + h_2 z^{-2}. \\ H(z) &= z^{-2}\left[h_o\left(z^{-2} + z^{+2}\right) + h_1\left(z^{-1} + z^{+1}\right) + h_2\right]. \\ H(f) &= \exp\left(-4\pi j f T_e\right)\left[2h_0\cos\left(4\pi f T_e\right) + 2h_1\cos\left(2\pi f T_e\right) + h_2\right]. \\ P(f) &= \left[h_2 + 2h_1\cos\left(2\pi f T_e\right) + 2h_0\cos\left(4\pi f T_e\right)\right]. \end{aligned}$$

Itération n^{o}1
On choisit, comme fréquences extrémales f_i , 4 fréquences quelconques entre 0 et $f_e/2$.

$$f_0 = 0,\ f_1 = 1{,}5Khz,\ f_2 = 3Khz,\ f_3 = 4{,}5Khz.$$

La valeur δ doit vérifier les équations suivantes :

$$\begin{aligned} P(f_0) &= 1 + \delta \\ P(f_1) &= 1 - \delta \\ P(f_2) &= +\delta \\ P(f_3) &= -\delta. \end{aligned}$$

C'est à dire :

$$\begin{aligned} 2h_0 + 2h_1 + h_2 &= 1 + \delta \\ 2h_0\cos\left(4\pi f_1 T_e\right) + 2h_1\cos\left(2\pi f_1 T_e\right) + h_2 &= 1 - \delta \\ 2h_0\cos\left(4\pi f_2 T_e\right) + 2h_1\cos\left(2\pi f_2 T_e\right) + h_2 &= \delta \\ 2h_0\cos\left(4\pi f_3 T_e\right) + 2h_1\cos\left(2\pi f_3 T_e\right) + h_2 &= -\delta \end{aligned}$$

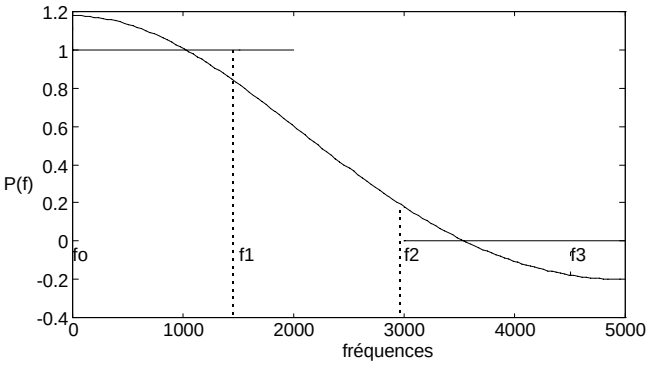
Et finalement :

$$\begin{pmatrix} 2 & 2 & 1 & -1 \\ 2\cos\left(4\pi f_1 T_e\right) & 2\cos\left(2\pi f_1 T_e\right) & 1 & 1 \\ 2\cos\left(4\pi f_2 T_e\right) & 2\cos\left(2\pi f_2 T_e\right) & 1 & -1 \\ 2\cos\left(4\pi f_3 T_e\right) & 2\cos\left(2\pi f_3 T_e\right) & 1 & 1 \end{pmatrix} \begin{pmatrix} h_0 \\ h_1 \\ h_2 \\ \delta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

D'où :

$$\begin{aligned} h_0 &= 0{,}0271 \\ h_1 &= 0{,}3446 \\ h_2 &= 0{,}4342 \\ \delta &= 0{,}1775 \end{aligned}$$

La fonction $P(f)$ correspondante est représentée dans la figure suivante :



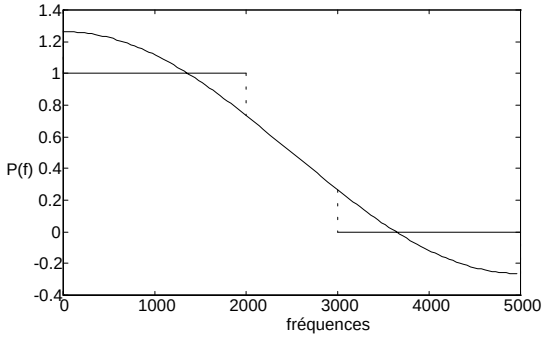
Itération n^{o}2
Comme on le remarque sur les 3 fréquences f_i seules f_0 et f_2 sont des extréma. Pour la deuxième itération on choisit 4 nouvelles fréquences f_i :
f_0=0(max), f_1=2 Khz (min), f_2=3 Khz(max), f_3=5 Khz(min).

Et on réitère le processus précédent.
La résolution du système à 4 équations linéaires nous donne :

$$h_0 = 0$$
$$h_1 = 0{,}3820$$
$$h_2 = 0{,}5$$
$$\delta = 0{,}2639$$

On peut observer que δ a augmenté. D'autre part le théorème de convergence étant vérifié l'algorithme a convergé.

La courbe $P(f)$ correspondante est représentée sur la figure suivante.



Le filtre $H(z)$ s'écrit :

$$H(z) = 0{,}382z^{-1} + 0{,}5z^{-2} + 0{,}382z^{-3}$$

Pour un filtre à $N = 7$ coefficients

On reprend le même algorithme, pour les mêmes spécifications mais pour calculer un filtre à 7 coefficients.

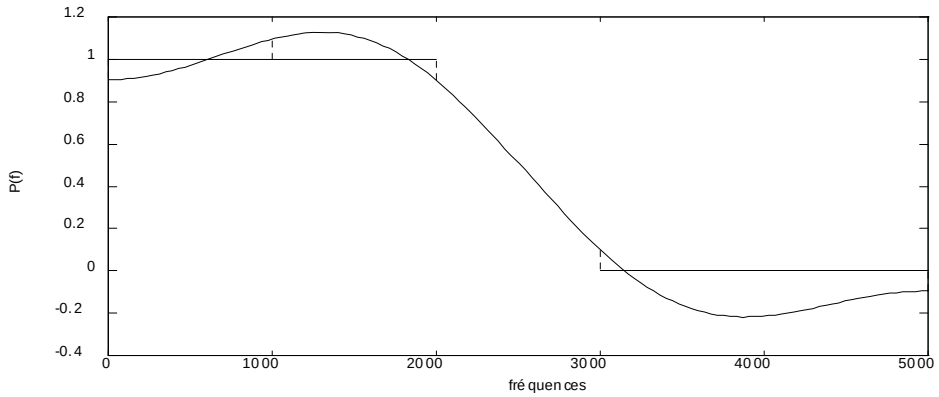
$$P(f) = [h_3 + 2h_2 \cos(2\pi f T_e) + 2h_1 \cos(4\pi f T_e) + 2h_0 \cos(6\pi f T_e)]$$

Itération n^{o}1
on choisit 5 fréquences extrémales : f_0=0 Khz, f_1=1 Khz, f_2=2 Khz, f_3=3 Khz, f_4=5 Khz.

La résolution du système à 4 équations linéaires nous donne :

$$h_0 = -0{,}1118$$
$$h_1 = -0{,}0264$$
$$h_2 = 0{,}3618$$
$$h_3 = 0{,}4573$$
$$\delta = -0{,}0955$$

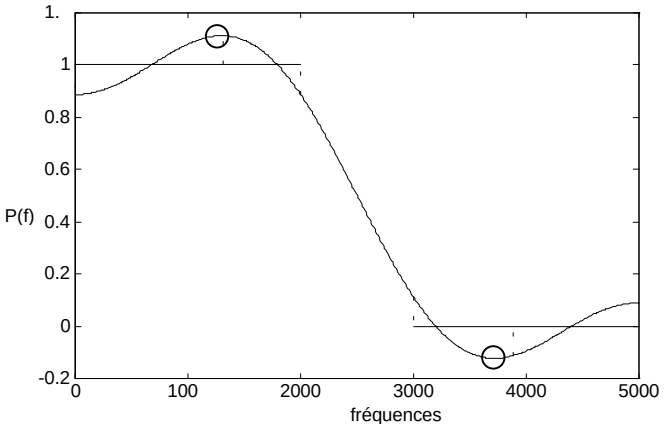
La courbe $P(f)$ correspondante est représentée sur la figure suivante.



Itération n^{o}2
On observe sur la courbe $P(f)$ que la fréquence f_1 ne correspond pas à un extrémum. On remplace donc f_1 par la fréquence f_1=1308 hz qui correspond à un maximum. De même on remplace f_4 par f_4=3877 qui correspond à un minimum. Et on itère les calculs. La solution du système linéaire donne :

$$h_0 = -0{,}1182$$
$$h_1 = -0{,}0031$$
$$h_2 = 0{,}3176$$
$$h_3 = 0{,}4950$$
$$\delta = -0{,}1124$$

On remarque que δ a augmenté. La courbe $P(f)$ est représentée sur la figure suivante.

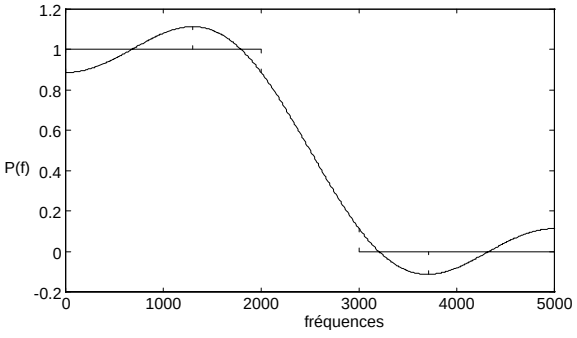


Les cercles sur les figure représentent la position des extréma autour de f_1 et f_4.

Itération n^{o}3
On remplace les anciennes valeurs de f_1 et f_4 par 2 nouvelles valeurs correspondant aux extréma : f_1=1298,8 Hz et f_4=3720,7 Hz. La solution du système linéaire est :

$$h_0 = -0{,}1196$$
$$h_1 = -0{,}0001$$
$$h_2 = 0{,}3132$$
$$h_3 = 0{,}4999$$
$$\delta = -0{,}1130$$

δ a augmenté, mais assez peu.
La courbe $P(f)$ est représentée sur la figure suivante.



Itération n^{o}4

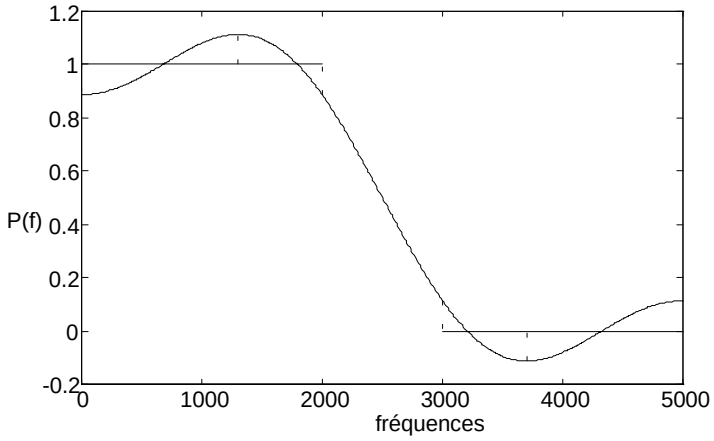
On remplace l'anciennes valeur de f_4 par une nouvelle valeur correspondant à un minimum de $P(f)$:

f_4=3701,2 Hz. La solution du système linéaire est:

$$h_0 = -0{,}1196$$
$$h_1 = -0{,}0000$$
$$h_2 = 0{,}3131$$
$$h_3 = 0{,}5000$$
$$\delta = -0{,}1130$$

δ n'a pratiquement pas augmenté, on décide d'arrêter l'algorithme. On peut remarquer qu'il y a bien 5 extréma qui alternent.

La fonction $P(f)$ est représentée sur la figure suivante.



Le filtre $H(z)$ s'écrit:

$$H(z) = -0.1196 + 0.3131z^{-2} + 0{,}5z^{-3} + 0.3131z^{-4} - 0{,}1196z^{-6}.$$

On peut noter que les coefficients d'indice impair sont nuls, sauf le coefficient central. C'est une propriété générale des filtres FIR demi-bande à phase linéaire et nombre impair de coefficients.

Les calculs ont été faits sans pondération. L'amplitude maximale des ondulations finales en bande passante a donc la même valeur qu'en bande atténuée. Si le gabarit du filtre est tel que les amplitudes maximales des ondulations en bande passante et en bande atténuée doivent être respectivement p et a, il suffit d'utiliser l'algorithme de Remez avec des pondérations w_p en bande passante et w_a en bande atténuée, telles que :

$$\frac{w_p}{w_a} = \frac{\delta_a}{\delta_p}.$$

Exemple de calcul d'un filtre de Hilbert:

La fonction de transfert théorique d'un filtre de Hilbert est :

$$H(f) = -j\text{sign}(f).$$

C'est à dire que son module est constant et égal à 1, et que sa phase vaut $\pm\pi/2$.

Quand on veut calculer un filtre FIR à temps de retard de groupe constant, approchant un filtre de Hilbert, on choisit pour des raisons évidentes un filtre à réponse impulsionnelle antisymétrique. En effet, on a vu précédemment que pour les filtres à réponse antisymétrique de longueur N, la phase est de la forme :

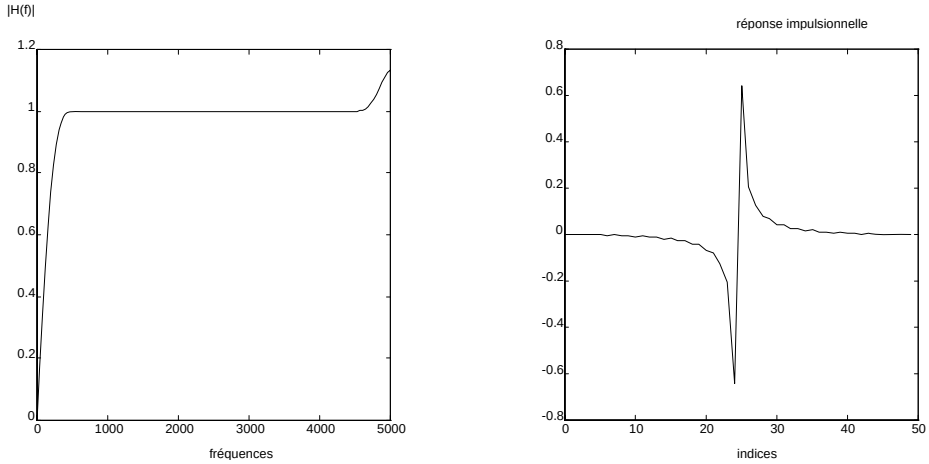
$$\phi(f) = \pm\frac{\pi}{2} - \pi j f(N-1)T_e.$$

pour f appartenant à l'intervalle $[0, f_e/2]$.

On définit un intervalle de fréquence dans lequel on souhaite que le filtre présente les caractéristiques de filtre de Hilbert. Il est intéressant que cet intervalle soit symétrique par rapport à $f_e/4$, car dans ce cas un coefficient sur deux est nul.

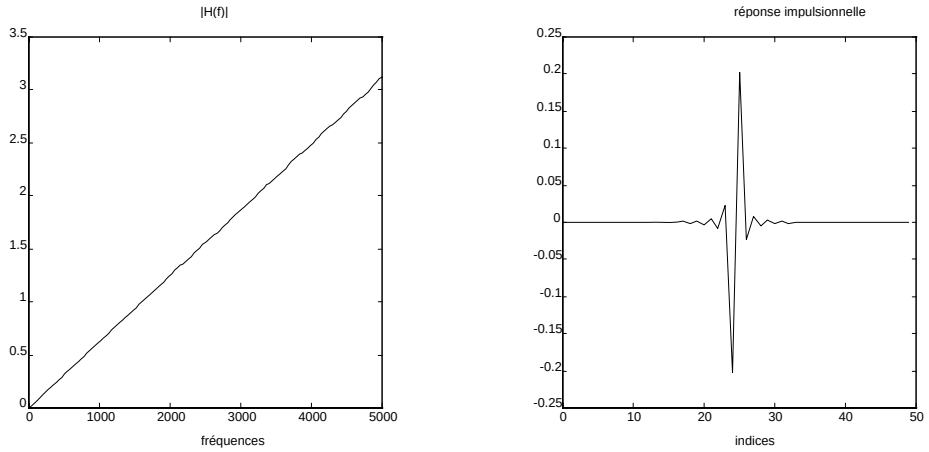
Exemple numérique :
Les spécifications du filtre sont les suivantes :
N= nombre de coefficients du filtre=50
f_e=10 Khz
f_1=0.5 Khz f_2=4,5 Khz
Le module de la fonction de transfert désirée vaut 1 entre f_1 et f_2, et la phase $-\pi/2$.

La figure suivante représente le module de la fonction de transfert ainsi que la réponse impulsionnelle du filtre obtenu en utilisant l'algorithme de Remez.



Exemple de calcul d'un filtre différentiateur :
La fonction de transfert théorique d'un différentiateur vaut $j2\pi fT_e$ dans l'intervalle $[0,f_e/2]$. Pour les mêmes raisons que dans le cas d'un filtre de Hilbert on choisira un filtre FIR à réponse impulsionnelle antisymétrique. Ce filtre introduit par nature un déphasage de $\pm\pi/2$. Il suffira donc d'optimiser le module de la fonction de transfert pour qu'il s'approche de $2\pi fT_e$.
Exemple numérique :
Les spécifications du filtre sont les suivantes:
N= nombre de coefficients du filtres=50
f_e=10 Khz
f_1=0.5 Khz f_2=4,5 Khz.

La figure suivante représente le module de la fonction de transfert ainsi que la réponse impulsionnelle du filtre obtenu en utilisant l'algorithme de Remez.



3.5 Norme L_∞ et programmation linéaire

Les méthodes de programmation linéaire permettent de résoudre les problèmes d'optimisation linéaires sous des contraintes linéaires.

Un exemple de tel problème est le calcul d'un FIR vérifiant des spécifications sur le module de la fonction de transfert sous des contraintes temporelles sur la réponse impulsionnelle. Les contraintes sur la réponse impulsionnelle sont souvent du type : h_n s'annule périodiquement.

Le problème s'écrit alors:

$$\begin{array}{l} -\delta \leq \tilde{W}(f_i)\left[\tilde{H}(f_i) - P(f_i)\right] \leq +\delta \\ f_i \in [0,f_e/2] \end{array}$$

et

$$h_n = 0 \text{ si } n = iK \text{ et } n \neq 0$$

Les fréquences f_i constituent la grille sur laquelle se fait l'optimisation.

On peut reformuler le problème par:

$$\begin{array}{l} \min \delta \\ \left\{ \begin{array}{l} -\tilde{W}(f_i)P(f_i) - \delta \leq -\tilde{W}(f_i)\tilde{H}_D(f_i) \\ \tilde{W}(f_i)P(f_i) - \delta \leq \tilde{W}(f_i)\tilde{H}_D(f_i) \\ f_i \in [0,f_e/2] \\ n = iK, n \neq 0 \Rightarrow h_n = 0 \end{array} \right. \end{array}$$

Un algorithme efficace dans cette situation est l'algorithme du simplex.

Il est plus lent que l'algorithme de Remez, mais on ne peut pas utiliser Remez lorsqu'il y a des contraintes fréquentielles et temporelles.

Chapitre IV

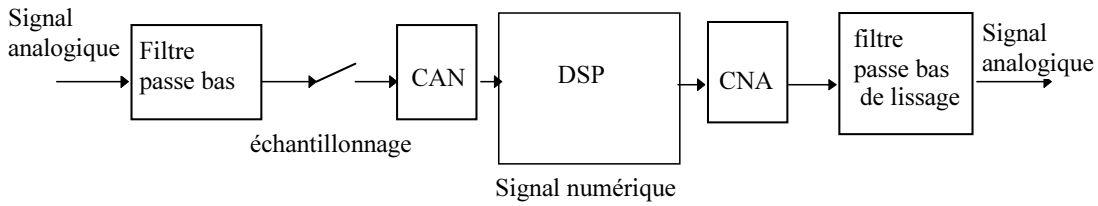
NUMÉRISATION ET REPRÉSENTATIONS BINAIRES

Dans les systèmes de traitement de signal, la numérisation des signaux intervient dans les dispositifs de conversion analogique numérique ainsi que dans les processeurs (DSP). Lorsque le convertisseur et le processeur utilisent des représentations différentes, la type de quantification est imposée par le système de traitement.

1 Représentation numérique d'un signal

Les signaux physiques sont en général analogiques. Pour les traiter avec un DSP ou un autre système de traitement numérique des signaux (ASIC par exemple), il faut les échantillonner et convertir chaque échantillon en une donnée numérique. Cette conversion est réalisée par un convertisseur analogique numérique CAN. La conversion se décompose en une quantification et numérisation (codage) de la valeur quantifiée. Réciproquement les résultats numériques fournis par un DSP pourront être convertis en signaux analogiques à l'aide d'un convertisseur numérique analogique CNA.

1.1 Interface CAN - DSP - CNA



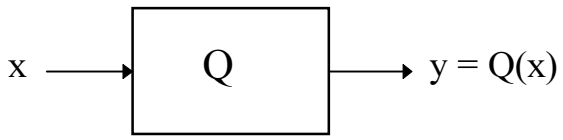
Les convertisseurs CAN fournissent une représentation numérique de chaque échantillon. Cette représentation peut ensuite être transformée par le DSP selon le type d'arithmétique utilisée.

1.2 Quantification

On se limite ici à la quantification scalaire, c'est à dire à la quantification d'un échantillon isolé. On distingue plusieurs types de quantification scalaire:

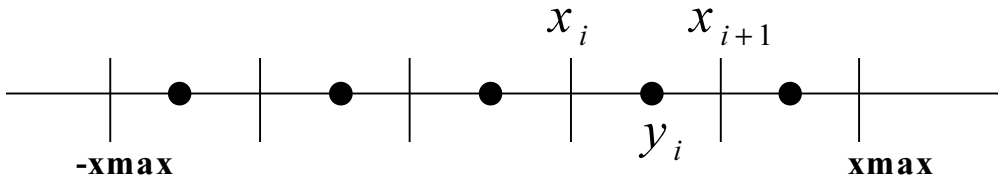
- La quantification uniforme.
- La quantification non uniforme, en particulier la conversion de type logarithmique.
- La quantification adaptative.

1.2.1 Quantification scalaire



Quantifier une grandeur x, pouvant prendre une valeur quelconque dans un intervalle continu compris entre $-x_{max}$ et x_{max}, consiste à remplacer x par une valeur quantifiée $Q(x) = y_i$ choisie parmi un ensemble fini (ou dénombrable) de N valeurs possibles. On appelle les valeurs y_i les valeurs de quantification. Le choix de la valeur de quantification pour un x donné est déterminé en fonction de $N+1$ valeurs de décision x_i, par la règle suivante:

$$x \in [x_i, x_{i+1}[\Rightarrow Q(x) = y_i$$



Quantification uniforme

Lorsque la largeur des intervalles de décision est constante on parle de quantification uniforme ou linéaire. La largeur des intervalles est appelée pas de quantification et notée q.

$$x_{i+1} - x_i = q = \text{pas de quantification constant}$$

Le pas de quantification q peut s'exprimer en fonction des valeurs extrémes par :

$$q = \frac{2x_{\max}}{2^N}$$

On définit le facteur de surcharge, noté Γ, comme le rapport entre la valeur maximale du convertisseur x_{max} et l'écart type des échantillons à convertir, noté σ_x.

$$\Gamma = \frac{x_{\max}}{\sigma_x}$$

Lors de la quantification, deux types d'erreur peuvent être commises : l'erreur de granulation et l'erreur de saturation.

Une erreur de saturation se produit lorsque l'amplitude de l'échantillon à convertir est supérieure à x_{max}. Cete erreur est d'autant plus gênante qu'elle n'est pas bornée, on cherche donc à minimiser la probabilité de saturation. La probabilité de saturation p_D dépend de la valeur de Γ.

Pour des échantillons gaussiens :

$$\Gamma = 2 \Rightarrow p_D = 0.045$$
$$\Gamma = 4 \Rightarrow p_D = 0.00006$$

Une erreur de granulation se produit sur les échantillons d'amplitude inférieure à x_{max} en valeur absolue. C'est la différence entre l'échantillon et l'échantillon quantifié. Cette erreur est bornée. Si la quantification s'effectue par arrondi au plus proche voisin, l'erreur de granulation e_g en valeur absolue est inférieure à $q/2$.

$$e_g = x - Q(x)$$
$$|e_g| \leq \frac{q}{2}$$

Sous certaines hypothèses relativement générales, on peut calculer la valeur moyenne et l'écart type de cette erreur :

$$E(e_g) = 0$$
$$E\left(e_g^e\right) = \sigma_g^2 = \frac{q^2}{12}$$

Avec les mêmes hypothèses, Le rapport (noté RSB_{dB}) entre la puissance du signal σ_x^2 et la puissance de l'erreur de granulation σ_g^2 peut s'exprimer en décibels par la relation suivante où N représente le nombre de bits du convertisseur :

$$RSB_{dB} = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_g^2}\right)$$
$$RSB_{dB} \approx 10\log_{10}\left(\sigma_x^2\right) + 6N - 10\log_{10}\left(x_{\max}^2\right) + 10\log_{10}\left(\frac{3}{2}\right)$$
$$RSB_{dB} \approx +6N + 10\log_{10}\left(\frac{3}{2}\right) - 20\log_{10}\left(\Gamma\right)$$

Le rapport signal sur bruit en dB dépend donc de façon linéaire de la puissance du signal en dB. Il est d'autant plus grand que la puissance du signal est grande.

Quantification logarithmique

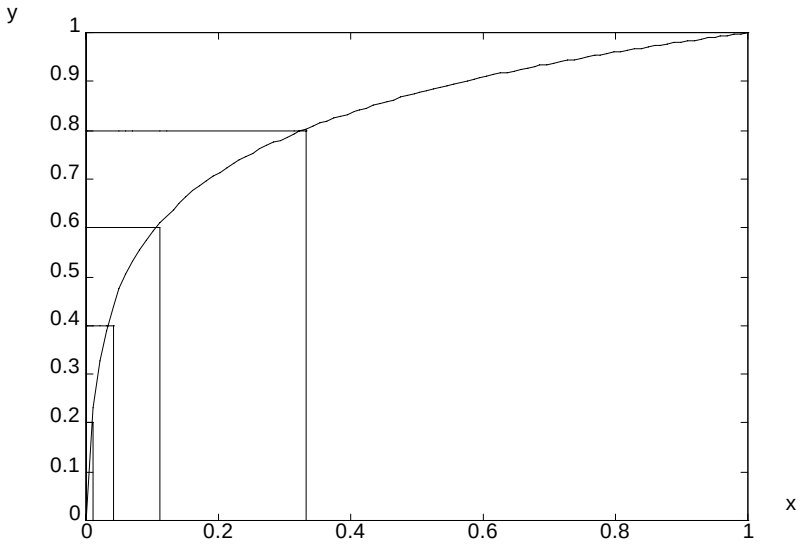
La quantification de type logarithmique permet d'obtenir un rapport signal sur bruit de quantification à peu près constant quelque soit la puissance du signal. L'écart entre les seuils de décision n'est pas constant. Il croît logarithmiquement en fonction de l'amplitude du signal à quantifier. Une quantification logarithmique peut se réaliser par une compression des amplitudes suivie d'une quantification uniforme, puis d'une expansion des amplitudes.

Lois de compression expansion :

La loi de compression est notée $C(x)$. La loi d'expansion est l'opération inverse.

$$-x_{\max} \leq x \leq x_{\max} \qquad y = C(x) \qquad x = C^{-1}(y)$$

La loi de compression doit approcher d'une fonction logarithme. Deux lois sont très utilisées en pratique: la loi A et la loi μ. Ces deux lois sont appliquées dans les codecs : circuits de conversion analogiques numériques en téléphonie. La loi A est appliquée en Europe, la loi μ aux USA et au Japon.



Description de la loi A

La définition de la fonction de compression $C(x)$ fait intervenir une constante appelée A.

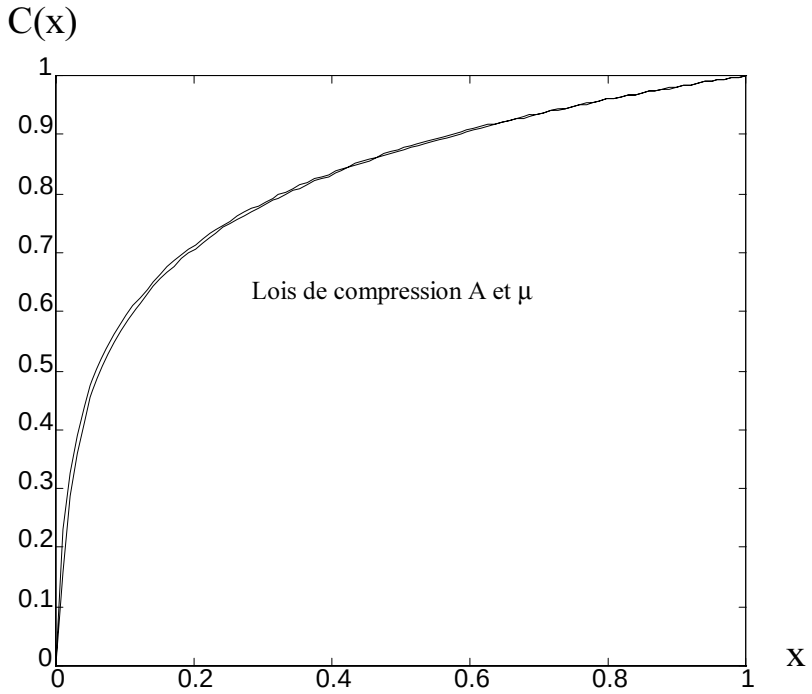
$$\begin{aligned} C(x) &= \frac{A\,|x|}{1+\log(A)}\text{sign}(x) \qquad 0 \leq \frac{|x|}{x_{\max}} < \frac{1}{\text{A}} \\ C(x) &= x_{\max}\frac{1+\log\left(\frac{A|x|}{|x_{\max}|}\right)}{1+\log(A)}\text{sign}(x) \qquad \frac{1}{A} \leq \frac{|x|}{x_{\max}} \leq 1 \\ A &= 87.56. \end{aligned}$$

Description de la loi μ

La définition de la loi μ fait intervenir une constante appelée μ :

$$C(x) = x_{\max}\frac{\log\left(1+\frac{\mu|x|}{x_{\max}}\right)}{\log(1+\mu)} \quad \text{avec } \mu = 255$$

La figure suivante représente les deux fonctions de compression ainsi obtenues. On s'aperçoit qu'elles sont partiquement superposées. Sur la figure, on a normalisé x_{max} à 1.

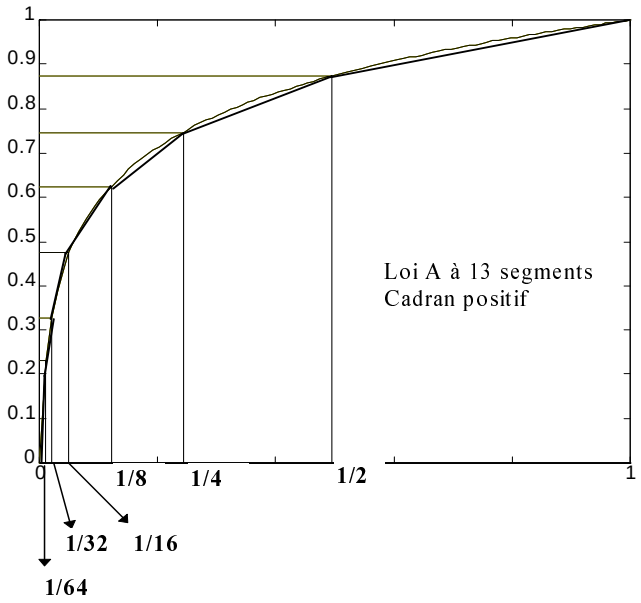


Approximations par segments des lois de compression A et μ

Pour leur réalisation matérielle les lois A et μ sont approximées par des segments de droite. La loi A est approximée par une courbe à à 13 segments, et la loi μ par une courbe à 15 segments. Elles sont appliquées dans ce cas là avec une numérisation sur 8 bits.

En ce qui concerne la loi A, la pente du premier segment passant par l'origine, est de 16. Puis les pentes des segments successifs sont obtenus par division par deux. La pente du dernier segment vaut donc 1/4.

La courbe suivante représente la loi A à 13 segments. Là encore x_{max} est normalisé à 1.



Le rapport signal sur bruit de quantification obtenu avec la loi A est constant sur une large plage de signal. Dans le cas de la conversion sur 8 bits, on peut remarquer que les petits signaux sont amplifiés par un facteur 16 avant d'être convertis, ce qui revient à diviser par 16 le pas de quantification, c'est à

dire à utiliser 12 bits de quantification (gain de 4 bits). Par contre pour les grands signaux, le pas de quantification est multiplié par 4 par rapport à un convertisseur 8 bits uniforme, on perd donc 2 bits.

Le rapport signal sur bruit de quantification en décibels pour la loi A s'écrit :

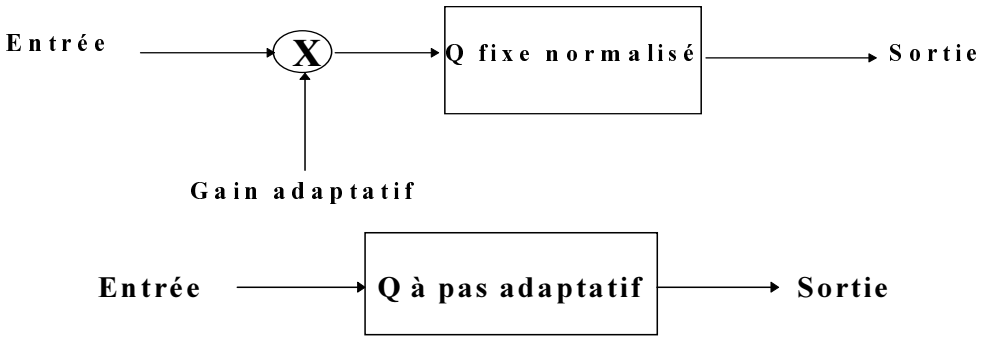
$$\begin{aligned} x \text{ grand} : RSB_{dB} &\approx 6N + 4.77 - 20\log(1 + \ln A) \\ x \text{ petit} : RSB_{dB} &\approx 6N + 4.77 + 10\log\left(\frac{A}{1+\ln A}\right) - 20\log(\Gamma) \\ \Gamma &= \frac{x_{\max}}{\sigma_x} \end{aligned}$$

N représente le nombre de bits utilisés et Γ le facteur de charge.

Pour le signal téléphonique, la qualité subjective obtenue avec une conversion selon la loi A sur 8 bits est équivalente à celle obtenue avec une conversion uniforme sur 12 bits. Le rapport signal sur bruit maximal est meilleur pour la conversion uniforme sur 12 bits puisqu'il est de l'ordre de 70 dB au lieu de 38 dB pour la conversion logarithmique sur 8 bits. Mais dans le cas de la conversion logarithmique sur 8 bits la dynamique de signal pour laqeul le rapport signal sur bruit maximal est obtenu, est grande (une trentaine de dB), alors que pour la conversion uniforme le RSB en dB décroît proportionnellement avec la puissance adu signal en dB.

Quantification adaptative

La quantification adaptative est utilisée dans de nombreux codeurs de parole ou d'image. Elle consiste à faire varier au cours du temps les caractéristique du quantificateur pour les adapter au mieux à l'évolution de la puissance du signal. Plus précisément, un quantificateur donné travaillant sur N bits sera d'autant mieux utilisé que sa tension pleine échelle x_{max} sera adaptée à l'écart type du signal σ_x. Dans un quantificateur adaptatif, on peut faire varier x_{max} au cours du temps de façon à suivre l'évolution de σ_x et à maintenir un RSB à peu près constant. Au lieu de faire varier les caractéristiques du quantificateur, on peut normaliser en amplitude le signal à convertir en le multipliant par un gain adaptatif dépendant de la puissance du signal. Ces deux approches sont représentée sur la figure c-dessous:



L'estimation de la puissance du signal pourra se faire sur le signal à l'entrée ou à la sortie du convertisseur. On parle respectivement d'adaptation directe ou rétrograde. Dans le premier cas il sera nécessaire de transmettre l'estimation de puissance au décodeur, ce qui augmente le débit binaire mais permet une bonne estimation. Dans le deuxième cas, il n'est pas utile de transmettre l'estimation au décodeur, car il peut la réaliser lui même sur les données quantifiées.

D'autre part, l'estimation de puissance peut se faire sur des durée de signal plus ou moins longues, et se renouveler plus ou moins souvent.

2 Représentation des données et arithmétique en précision finie

Les processeurs de traitement de signal peuvent travailler sur des données représentées en **virgule fixe** ou **en virgule flottante**. Pour certains algorithmes, il peut être intéressant de les faire fonctionner en **virgule flottante par bloc**. Avant de préciser ces différentes représentations, il est utile de rappeler les représentations binaires les plus courantes des entiers relatifs.

2.1 Représentation binaire des entiers relatifs

Plusieurs représentations binaires des entiers relatifs sont couramment utilisées, en particulier dans les convertisseurs analogiques numériques. On peut citer : ů

- Complément à 2,
- Complément à 1,
- Signe, valeur absolue,
- Binaire décalé.

Pour illustrer ces diverses représentations on donne un exemple sur 3 bits:
Exemple sur 3 bits

Entiers positifs	Binaire pur
7	111
6	110
5	101
4	100
3	011
2	010
1	001
0	000

Entiers relatifs	Binaire décalé	Signe plus valeur absolue	Complément à 1	Complément à 2
+3	111	011	011	011
+2	110	010	010	010
+1	101	001	001	001
0		000	000	
0	100	100	111	000
-1	011	101	110	111
-2	010	110	101	110
-3	001	111	100	101
-4	000			100

Dans la représentation en complément à 1, un entier négatif x est codé par la représentation binaire pure de l'entier positif y égal au complément à 1 de x :

$$y = 2^N - |x| - 1$$

Le terme complément à 1 représente en fait le complément à $2^N - 1$, entier positif qui sur N bits s'écrit avec tous les bits à 1. Dans la représentation en complément à 2 un entier négatif x est codé par la représentation binaire pure de l'entier positif y égal au complément à 2 de x :

$$y = 2^N - |x|$$

Le terme complément à 2 représente en fait le complément à 2^N.

2.2 Entiers relatifs en complément à 2

La représentation des nombres entiers relatifs utilisée dans les DSP comme dans la plupart des microprocesseurs est la représentation en complément à deux.

Soit un entier relatif x. Sa représentation binaire en complément à 2 sur N bits est constituée de la suite de bits b_i :

$$x \rightarrow b_{N-1} b_{N-2} \cdots b_k \cdots b_1 b_0$$

La relation entre x et les valeurs des bits b_i est la suivante :

Pour x positif, la représentation en complément à deux correspond à la représentation binaire pure, c'est à dire à une représentation pondérée en base 2 :

$$x \geq 0 \Rightarrow x = \sum_{i=0}^{N-1} b_i 2^i \text{ et } b_{N-1} = 0.$$

Pour x négatif, la représentation en complément à 2 de x est la représentation binaire pure du complément à 2^N de x :

$$x < 0 \Rightarrow y = 2^N - |x| \Rightarrow y = \sum_{i=0}^{N-1} b_i 2^i$$

Dans tous les cas on peut écrire la relation suivante :

$$x = -2^{N-1} b_{N-1} + \sum_{i=0}^{N-2} b_i 2^i$$

Le bit de poids le plus fort vaut 0 pour les entiers positifs et 1 pour les entiers négatifs.

Quelques propriétés de la représentation en complément à 2

valeurs extrêmes

Les valeurs extrêmes représentables en complément à 2 sur N bits sont :

$$\begin{aligned} \max &= 2^{N-1} - 1 \\ \min &= -2^{N-1} \end{aligned}$$

Mode d'overflow

La représentation en complément à 2 est une représentation circulaire. C'est à dire que lorsqu'on ajoute 1 à la plus grande valeur positive, on obtient la valeur négative de plus grande valeur absolue. de même, si on enlève 1 à la valeur négative de plus grande valeur absolue, on obtient la plus grande valeur positive.

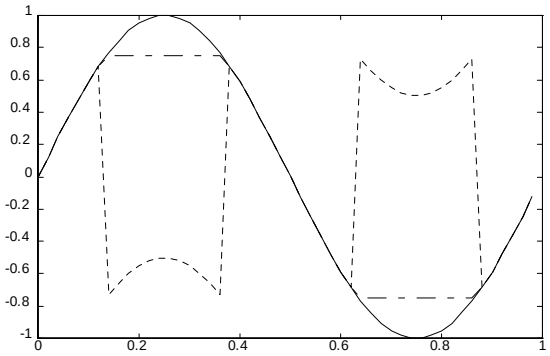
$$\left(2^{N-1} - 1\right) + 1 = 2^{N-1} \Leftrightarrow -2^{N-1}$$

Cette particularité peut avoir des conséquences fâcheuses. Ainsi lorsque le gain d'un filtre numérique est trop grand pour que la sortie puisse s'exprimer sur N bits, au lieu de saturer comme en analogique, la sortie oscille entre des valeurs de grandes amplitudes positives et négatives. On parle de cycles limites de grande amplitude. Les débordements, en complément à 2, génèrent ainsi des pics difficiles à filtrer.

Aussi, les DSP disposent-ils en général d'une arithmétique de saturation. Ils peuvent être configurés soit pour travailler en vrai complément à 2 soit pour travailler en complément à deux avec saturation arithmétique. Dans ce dernier cas, si le résultat d'un calcul est en valeur absolue supérieur à la plus grande valeur représentable en complément à 2 dans l'accumulateur, l'arithmétique de saturation détecte le débordement et le processeur remplace ce résultat par une valeur d'écrêtage : la plus grande valeur positive ou négative représentable, c'est à dire qu'il réalise une saturation. Dans les DSP le mode de fonctionnement est déterminé par la valeur d'un bit de mode (souvent appelé OVM= Overflow Mode).

La figure suivante représente une sinusoïde d'amplitude supérieure à ce qui peut être représenté avec les N bits. On a supposé que la plus grande amplitude représentable était 0.75, alors que l'amplitude de la sinusoïde était de 1. La sinusoïde est en trait plein, la sinusoïde écrêtée par une arithmétique de saturation est représentée en tirets longs - tirets courts, la sinusoïde en complément à 2 avec overflow

est en pointillés.



Extension du bit de signe

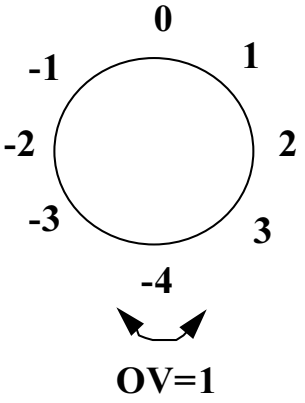
Lorsque l'on connaît la représentation en complément à 2 sur N bits d'un nombre x, pour obtenir la représentation en complément à 2 du même nombre sur M (M>N) bits, il suffit d'étendre le bit de signe, c'est à dire de conserver les N bits en bits de poids faibles et de remplir les bits de poids forts supplémentaires en répétant la valeur du bit de signe (le MSB sur N bits).

Lorsque l'on charge une données 16 bits dans un accumulateur 32 bits (ou plus), le bit de signe est automatiquement étendu. Dans certains cas cette extension du bit de signe n'est pas souhaitable, c'est le cas par exemple pour une donnée 16 bits représentant une adresse, grandeur forcément positive. Il est en général possible de configurer les DSP pour qu'il effectue ou non une extension du bit de signe lors d'un chargement dans l'accumulateur. Dans les DSP un bit de mode (souvent appelé SXM (Sign eXtension Mode)) permet généralement de choisir le mode désiré.

Addition/Soustraction en complément à 2

En complément à 2 les additions et les soustractions sont simples à réaliser : pour ajouter 2 nombres entiers signés N bits, avec un résultat sur N bits, quelque soit le signe des nombres, il suffit d'ajouter les codes en complément à deux. La figure suivante donne quelques exemples d'addition avec N=3 bits. On y a mis en évidence la circularité de cette représentation, ainsi que le bit de retenue ou bit de carry.

Lorsque le résultat d'une addition est supérieur au plus grand nombre représentable, il y a débordement ou overflow. Dans les DSP, lorsqu'il y a overflow, un bit d'état (souvent appelé OV) est positionné.



Le bit de retenue (Carry) est souligné

111	(-1)	010	(2)	110	(-2)	110	(-2)
+ 111	(-1)	+ 001	(1)	+ 011	(3)	+ 001	(1)
1 110	(-2)	0 011	(3)	1 001	(1)	0 111	(-1)

De plus, si le résultat d'une série d'additions et/ou de soustractions peut s'exprimer sur les N bits de l'accumulateur, le résultat est correct même si des débordements intermédiaires se produisent.

Exemple sur 3 bits:

Soit à calculer 3 + 1 - 2, le résultat théorique vaut 2, ce qui peut s'exprimer sur 3 bits. Mais si on

ajoute les données deux par deux, on effectue successivement les opérations (3+1) dont le résultat est noté R, puis (R-2). Le résultat de 3+1 vaut 4 en théorie, mais sur 3 bits on obtient : $R = 011 + 001 =$ 100 c'est à dire R = -4 car il y a overflow. Ce résultat intermédiaire est faux, mais lors de l'opération suivante $R - 2$ on obtient 100 -010 = 010 c'est à dire 2 ce qui est bien le résultat correct.

Pour exprimer sans risque d'overflow le résultat de l'addition de 2 nombres N bits , il faut au moins N+1 bits.

La simplicité de l'addition est la raison principale pour laquelle la représentation en complément à 2 est utilisée dans la majorité des processeurs numériques.

Multiplication et décalage en complément à 2

En complément à 2, les multiplications sont plus difficiles qu'avec une représentation en signe + valeur absolue.

le résultat de la multiplication de 2 nombres N bits s'écrit sur $2N$-1 bits

En général le registre produit est sur $2N$ bits. De ce fait, les deux bits de poids forts du résultat d'un produit sont identiques (extension du bit de signe quand on passe de $2N - 1$ bits à $2N$ bits). Le MSB est ici inutile. Aussi est - il souvent possible dans les DSP de décaler d'un bit à gauche le résultat des produits de façon systématique, sans temps de cycle supplémentaire ni instruction particulière, il suffit de configurer correctement le mode produit à l'aide des bits de mode associés.

De nombreux processeurs travaillant en complément à 2, effectuent la multiplication de façon efficace en utilisant l'algorithme de Booth. Cet algorithme est décrit par l'exemple suivant de multiplication de 2 nombres A et B exprimés sur 3 bits.

Exemple d'algorithme de Booth

$$A = (a_2\, a_1\, a_0) = -4a_2 + 2a_1 + a_0$$

$$B = (b_2\, b_1\, b_0) = -4b_2 + 2b_1 + b_0$$

$$AB = -4A(b_2 - b_1) - 2A(b_1 - b_0) - A(b_0 - 0)$$

Dans un DSP travaillant en complément à 2, la multiplication est câblée, par contre les divisions doivent en général se faire par logiciel. Les multiplications ou divisions par une puissance de 2 peuvent être effectuées par des décalages arithmétiques à gauche ou à droite respectivement. Ces décalages de quelques bits peuvent en général s'effectuer en un seul temps de cycle. Lors des décalages arithmétiques à droite le bit de signe est étendu. Lors des décalages à gauche des zéros sont introduits dans les bits de poids faibles.

2.3 Représentation binaire des nombres réels en précision finie

Deux approches sont utilisées pour la représentation binaire des nombres réels en précision finie :

- La représentation en virgule fixe,
- La représentation en virgule flottante.

Les DSP sont adaptés à l'une ou à l'autre de ces représentations. Toutefois, un DSP travaillant en virgule fixe pourra aussi effectuer des calculs en virgule flottante, mais de manière peu efficace et réciproquement.

2.3.1 Représentation binaire des nombres fractionnaires en format (ou virgule) fixe

On appelle représentation en format (ou virgule) fixe des nombres fractionnaires, ou plus généralement des nombres réels avec une précision finie, une représentation comprenant une partie entière suivie d'une partie fractionnaire correspondant à des bits après la virgule.

On utilise souvent le terme Format Q_k pour indiquer une représentation comportant k bits derrière la virgule.

Soit un nombre x réel quelconque, sa représentation binaire en virgule fixe en précision finie sur N bits en format Q_k (c'est à dire avec k bits derrière la virgule) s'écrira :

$$x \rightarrow \overbrace{b_{N-1-k}\cdots b_1 b_0}^{\text{Partie entière}}, \quad \overbrace{b_{-1}b_{-2}\cdots b_{-k}}^{\text{Partie fractionnaire}}$$

Elle correspond à la représentation du nombre entier y obtenu en arrondissant à l'entier le plus proche le nombre réel formé du produit de x par 2^k.

$$y = \left[2^k x\right]$$

[...] signifiant arrondi au plus proche voisin.
Par la suite on suppose que cette représentation de y est faite en complément à 2.

$$\overbrace{b_{N-1-k}\cdots b_1 b_0}^{\text{Partie entière}}, \quad \overbrace{b_{-1}b_{-2}\cdots b_{-k}}^{\text{Partie fractionnaire}}$$

correspond au nombre fractionnaire :

$$z = -b_{N-1-k}2^{N-1-k} + b_{N-2-k}2^{N-2-k} + \cdots b_0 + b_{-1}2^{-1} + \cdots + b_{-k}2^{-k}$$

qui est une approximation en précision finie du réel x, sur N bits avec k bits derrière la virgule.

Un nombre réel étant rarement de précision finie, la représentation sur un nombre fini de bits introduit une erreur. En virgule fixe sur N bits avec format Q_k, cette erreur est inférieure à 2^{-k}, si N-k bits suffisent pour la partie entière.

Valeurs extrêmes en virgule fixe sur N bits avec format Q_k
Les valeurs extrêmes représentables sont:

$$\begin{aligned} \max &= 2^{N-1-k} - 2^{-k} \\ \min &= -2^{N-1-k} \end{aligned}$$

Dynamique et précision en virgule fixe sur Nbits avec format Q_k
En virgule fixe sur N bits avec k bits de partie fractionnaire, il est possible de représenter les réels compris entre -2^{N-1-k} et $2^{N-1-k} - 2^{-k}$ avec une erreur inférieure à 2^{-k} en valeur absolue. Pour les réels en dehors de cette plage, on dit qu'il y a :
Overflow si le nombre est trop grand en valeur absolue. Il y a débordement.
Underflow si le nombre est trop petit. Il est alors représenté par zéro.

En conclusion, l'erreur absolue est inférieure à 2^{-k} sur une plage de valeurs correspondant à une dynamique de $6N$ dB. La dynamique est ici définie comme 2 fois le rapport entre la plus petite grande et la plus petite des valeurs positive exprimables.

Exemple de représentation en virgule fixe sur $N = 8$ bits en format $Q5$

Le terme format Q5 signifie qu'il y a 5 bits derrière la virgule.

On travaille ici en complément à deux.

La partie entière est formée de 3 bits (8-5). Elle permet de représenter des entiers relatifs compris entre 3 et -4.

La partie fractionnaire, sur 5 bits permet de représenter des nombres compris entre 0 et 0,96875 ce qui correspond à la somme des 5 premières puissances de 2 négatives.

Le tableau suivant donne quelques représentations et leurs équivalences décimales :

représentation binaire virgule fixe, format Q5	Valeur décimale
011 10000	3,5
001 10100	1,625
110 10001	-1.46875
100 00000	-4
011 11111	3,96875

Pour cette représentation les valeurs extrêmes représentables sont -4 et 3,96875.

Par ailleurs, la représentation sur un nombre fini de bits introduit une erreur. cette erreur est inférieure à 2^{-k} soit 1/32 dans l'exemple. Le tableau suivant donne quelques exemples de nombres réels, leur représentation sur 8 bits en format Q5 avec l'équivalence décimale, et l'erreur commise par cette représentation :

Valeur réelle	représentation binaire virgule fixe, format Q5	Equivalence décimale	Erreur commise
1/3	000 01011	0.34375	0,010416666...
$\sqrt{2}$	001 01101	1,40625	0,007963562...
π	011 00101	3,15625	0.014657346...

Addition de nombres fractionnaires en virgule fixe

Lors de l'addition de nombres fractionnaires en virgule fixe, il faut comme en décimal aligner les virgules. La somme de 2 nombres en format Q_k donne un résultat en format Q_k :

$$Q_k + Q_k \Rightarrow Q_k$$

Multiplication de 2 nombres fractionnaires en virgule fixe

Le produit de nombres en virgule fixe sur N bits donne un résultat sur $2N-1$ bits. Et comme en décimal, le nombre de bits fractionnaires (après la virgule) du résultat est égal à la somme des nombres de bits fractionnaires des 2 opérandes :

$$Q_k x Q_{k'} \Rightarrow Q_{k+k'}$$

2.3.2 Représentation binaire des nombres fractionnaires en virgule flottante

Dans la représentation binaire en virgule flottante en précision finie sur N bits, un nombre x est représenté par une mantisse M et un exposant E.

$$x = M2^E.$$

Pour une représentation sur N bits, La mantisse M est exprimée sur m bits, el l'exposant E sur e bits, avec $N = m + e$.

La mantisse M est en général normalisée, par exemple :

$$\frac{1}{2} \leq |M| < 1$$

Plage de nombres représentables

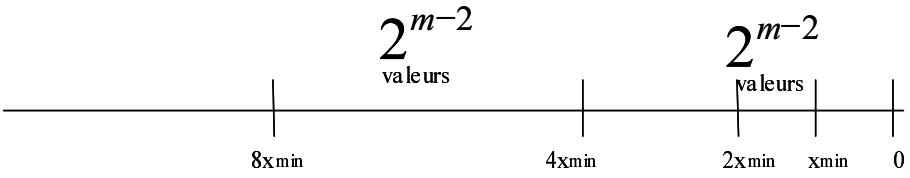
Sur N bits, avec m bits pour la mantisse, e bits pour l'exposant et une mantisse normalisée entre 0.5 et 1, on peut représenter des nombres dont la valeur absolue est comprise dans l'intervalle :

$$\left[\frac{1}{2}2^{-2^{e-1}}, \left(1-2^{1-m}\right)2^{2^{e-1}-1}\right]$$

$$\textit{Overflow} \text{ si } \quad |x| > \left(1 - 2^{1-m}\right) 2^{\left(2^{e-1}-1\right)}$$
$$\textit{Underflow} \text{ si } \quad |x| < \tfrac{1}{2} 2^{-\left(2^{e-1}\right)}$$

Interprétation de la représentation binaire virgule flottante

Les nombres représentés dans une représentation en virgule flottante sont répartis sur une échelle non linéaire, comme indiquée sur la figure ci-dessous, pour les nombres positifs :



Ainsi pour les nombres positifs, la plage de valeurs représentables comprises entre $x_{\min} = \frac{1}{2} 2^{-2^{e-1}}$ et $x_{\max} = \left(1 - 2^{1-m}\right) 2^{2^{e-1}-1}$ est partagée en intervalles de largeur en progression géométrique de raison égale à 2. Chaque intervalle correspond à une valeur de l'exposant E et contient 2^{m-2} valeurs associées aux différentes valeurs positives possibles de la mantisse entre 0.5 et 1 sur m bits. La largeur du premier intervalle est égale à x_{min}.

On peut remarquer que la précision absolue de la représentation est meilleure pour les valeurs de faible amplitude que pour les valeurs de forte amplitude.

Dans le premier intervalle [x_{min},$2x_{min}$] l'erreur de représentation est inférieure à $2^{-2^{e-1}} 2^{1-m}$. Dans le dernier intervalle [$0.5x_{max}$,x_{max}] l'erreur de représentation est inférieure à $2^{2^{e-1}-1} 2^{1-m}$.
La précision relative est à peu près constante.
Pour un nombre de bits donné, le nombre de bits de la mantisse détermine la précision, le nombre de bits de l'exposant détermine la dynamique. La dynamique D, ou plus précisément le rapport entre la plus grande et la plus petite des valeurs positives exprimables exprimé en dB, vaut:

$$D = 20 \log_{10} \left(2 \left(1 - 2^{1-m}\right) 2^{2^e}\right) \approx 6 \left(2^e + 1\right) dB$$

Exemple de représentation en virgule flottante sur 8 bits avec $m = 5$ et $e = 3$
On suppose que la mantisse et l'exposant sont exprimés en complément à 2 et que la mantisse M est normalisée $\frac{1}{2} \leq |M| < 1$. On suppose de plus que la mantisse est écrite avant l'exposant : ME.

Le tableau suivant donne quelques représentations et leurs équivalences décimales :

Représentation binaire : Mantisse M, Exposant E	Valeur décimale $M2^E$
01110 010	1,75
01100 100	0,046875
10010 011	-7
01000 100	0,03125
01111 011	7,5

Pour cette représentation les valeurs positives extrêmes représentables sont 0,03125 et 7,5.

Par ailleurs, la représentation sur un nombre fini de bits introduit une erreur. Le tableau suivant donne quelques exemples de nombres réels, leurs représentation sur 8 bits en virgule flottante (m=5 , e=3) avec l'équivalence décimale, et l'erreur commise par cette représentation:

Valeur réelle	M	E	Equivalence décimale	Erreur commise
1/3	01011 (0,6875)	111 (-1)	0,6875 x 2^{-1} = 0,34375	0,0134166.... 4%
$\sqrt{2}$	01011 (0,6875)	001 (1)	0,6875 x 2^1 = 1,375	-0,0392135... 2,77%
π	01101 (0,8125)	010 (2)	0,8125 x 2^2 = 3,25	0.10840734... 3,4%
$\sqrt{50}$	01110 (0,875)	011 (3)	0,875 x 2^3 = 7	0,07106781... 1%

Addition en virgule flottante

Pour ajouter 2 nombres A et B en virgule flottante, Il faut dénormaliser le nombre le plus petit (B) :

$$A + B = M_A 2^{E_A} + M_B 2^{E_B} = \left(M_A + M_B 2^{E_B - E_A}\right) 2^{E_A}$$

Cette dénormalisation complique les opérations par rapport à une addition en virgule fixe et fait perdre de la précision sur la représentation du plus petit nombre à cause de l'arrondi de la mantisse.

Multiplication en virgule flottante

Soit 2 nombres A et B en virgule flottante, on peur écrire :

$$AB = M_a M_b 2^{E_a + E_b} = M 2^E$$

Il faut normaliser le produit des mantisses $M_a M_b$ et corriger l'exposant, pour obtenir M et E.

Il faut $2m-1$ bits pour exprimer exactement M. Il faut $e+1$ bits pour exprimer E. Si on tronque M à m bits, l'erreur absolue augmente vite.

2.3.3 Comparaison virgule fixe virgule flottante

La numérisation virgule fixe sur N bits correspond à une échelle linéaire et à une quantification uniforme. La numérisation virgule flottante sur N bits correspond à une échelle non linéaire en progression géométrique de raison 2 et à une quantification non uniforme de type quasi logarithmique.

Erreur de quantification en virgule fixe

Soit x une valeur réelle et sa représentation en virgule fixe sur N bits en format Q_k. On suppose que x est inférieur à x_{max} la plus grande valeur représentable. Sous certaines hypothèses assez générales sur la distribution de x, on peut considérer que l'erreur d de représentation est une variable uniforme et on peut écrire :

$$\begin{aligned}
d &= \hat{x} - x \text{ (arrondi)} \\
|d| &\leq \frac{q}{2} \text{ Pour } |x| \leq x_{\max} \\
q &= 2^{-k} \\
x_{\max} &= 2^{N-1-k} - 2^{-k} = q\left(2^{N-1} - 1\right) \\
E(d) &= 0 \\
E\left(d^2\right) &= \sigma_d^2 = \frac{q^2}{12} \\
RSB_{dB} &= 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_d^2}\right) \\
RSB_{dB} &\approx 10\log_{10}\left(\sigma_x^2\right) + 6N - 10\log_{10}\left(x_{\max}^2\right) + 10\log_{10}\left(\frac{3}{2}\right)
\end{aligned}$$

Erreur de quantification en virgule flottante

Soit x une valeur réelle et $\hat{x}$ sa représentation en virgule flottante sur N bits avec m bits de mantisse et e bits d'exposant. On suppose que x est inférieur à x_{max} la plus grande valeur représentable. On peut écrire :

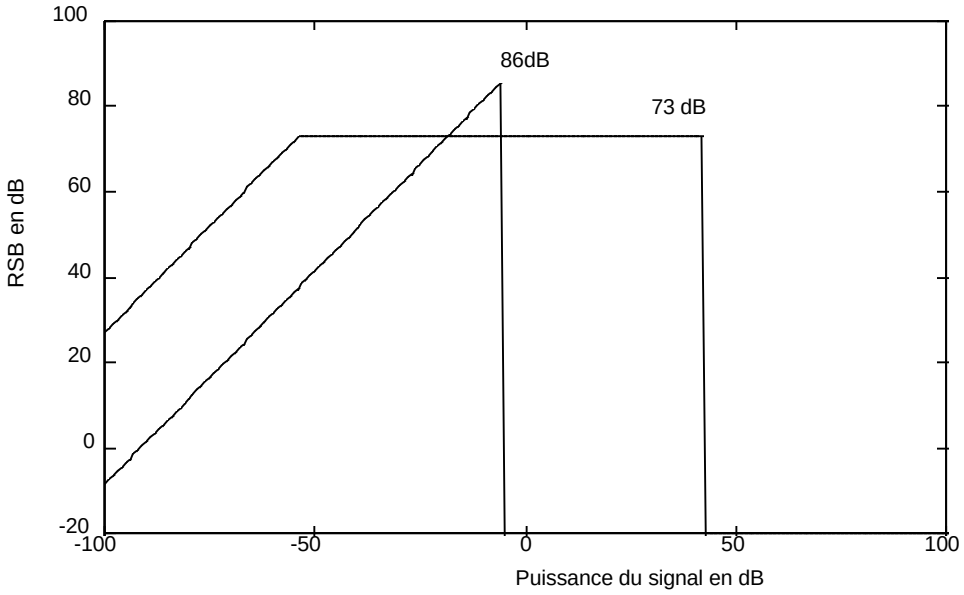
$$\begin{aligned}
d &= \hat{x} - x \text{ (arrondi)} \\
d_m &= \text{ erreur d'arrondi sur la mantisse} \\
0 &\leq |d_m| \leq \frac{1}{2} 2^{-(m-1)} \\
d_r &= \text{ erreur relative sur x} = \frac{d}{x} = \frac{d_m}{M} \\
|d_r| &\leq 2^{-(m-1)}
\end{aligned}$$

Sous certaines hypothèses assez générales sur la distribution de x, on peut montrer que :

$$\begin{aligned} d_r & \quad \text{est un bruit blanc non corrélé avec } x \\ d &= \hat{x} - x = x d_r \\ \sigma_d^2 &= \sigma_x^2 \sigma_{d_r}^2 \\ RSB_{dB} &= 6m + 1.44 \end{aligned}$$

Ici le RSB ne dépend pas de la puissance du signal, à la différence de la représentation en virgule fixe.

La figure suivante illustre le rapport signal sur bruit obtenu avec N =16 bits en virgule fixe, et en virgule flottante pour m=12 et e=4.



Comparaison des dynamiques

On appelle ici dynamique 2 fois le rapport en dB entre la plus grande et la plus petite amplitude non nulle représentables.

	Virgule fixe N bits Format Q_k	Virgule flottante N bits, M sur m bits et E sur e bits
Dynamique	$6N$ dB	$20\log_{10}\left(2\left(1-2^{1-m}\right)2^{2^e}\right)$ $\approx 6\left(2^e+1\right)dB$

Exemple numérique pour N =32 bits

L'exemple suivant compare la dynamique et la précision obtenues en virgule fixe et virgule flottante pour N = 32 bits.

	Virgule fixe N =32	Virgule flottante N =32 m = 24 e = 8
Dynamique	$>10^9$	$> 10^{77}$
Précision	Précision max > 9 digits	précision toujours >7 digits

Conclusion

Pour un nombre de bits N donné, la représentation des nombres en virgule flottante réalise un compromis dynamique (E), précision (M).

En virgule fixe, les opérateurs de traitement sont simples mais il faut surveiller le cadrage des données pour éviter les débordements tout en conservant un maximum de précision.

En virgule flottante les opérateurs sont plus complexes mais on dispose d'une plus grande dynamique pour une précision minimale donnée et le cadrage des données est moins critique.

2.3.4 Format IEEE 754, virgule flottante

Le format IEEE 754 de représentation des nombres en virgule flottante possède les principales caractéristiques suivantes :

Pour N =32 bits
Un bit de signe S
Un exposant sur 8 bits
Une fraction sur 23 bits

L'exposant est représenté en binaire décalé avec un biais égal à 127. La mantisse constituée du bit de signe et de la fraction et exprimée en signe plus valeur absolue. La valeur absolue est normalisée en binaire entre 1.00...00 et 1.11...11, et comme le premier bit vaut toujours 1, il est caché (non représenté), on stocke seulement la partie fractionnaire.

Exemple pour N =32 bits :
Le nombre x=28 est représenté de la façon suivante :

x = 28 = 1,75 2^4 → 0 10000011 1100...0

C'est à dire:

S = 0	nombre positif
E = 4	est représenté en binaire décalé avec un biais de 127. Il est donc exprimé par la représentation binaire pure de 127+4 = 131, soit en binaire 10000011
M = 1,75	La partie fractionnaire de la mantisse vaut 0.75. La représentation binaire correspondante est 1100...0, compte tenu du bit caché.

Le format IEEE 754 définit aussi la double précision étendue sur 64 bits, la simple précision étendue sur 43 bits et la double précision étendue sur 79 bits. Les caractéristiques correspondantes sont résumées dans le tableau suivant :

Nombre de bits:	Signe	Exposant	fraction	total
Double précision 64 bits	1	11	52	64
Simple précision étendue 43 bits	1	11	31	43
Double précision étendue 79 bits	1	15	63	79

Les DSP ne respectent pas forcément le format IEEE 754 de représentation des nombres en virgule flottante.

2.3.5 Virgule flottante par bloc

Dans certains cas, en particulier lorsqu'on utilise un DSP virgule fixe pour effectuer des calculs nécessitant à la fois une grande précision et une grande dynamique (une FFT par exemple), il peut être intéressant de travailler en virgule flottante par bloc.

Dans la représentation en virgule flottante par bloc, on utilise un registre qui contient la valeur de l'exposant à appliquer à un bloc de données : EXPOSANT DE BLOC. Cet exposant de bloc est constant pour un bloc de données. Chaque bloc de données est testé et mis à l'échelle par l'exposant de façon à éviter les débordements.

Le processeur travaillant sur des mots de N bits, la mantisse conserve N bits.

En fait chaque mot est donc représenté par une matisse sur N bits et un exposant, ou facteur d'échelle qui est stocké dans un registre séparé, en général sur N bits. Cet exposant reste constant pour un bloc de données. Les calculs se font en virgule fixe sur les mantisses du bloc, puis les résultats sont mis à l'échelle en fonction de l'exposant.

Cette représentation est utile quand N est petit (EX : 16 bits) par rapport aux contraintes de dynamique et de précision du problème. Elle limite la perte de précision due à l'augmentation de la dynamique en virgule flottante, pour un nombre de bits fixé. La complexité des opérations reste raisonnable.

Chapitre V

IMPLÉMENTATION DES FILTRES NUMÉRIQUES

Ce chapitre est consacré à l'implémentation des filtres numériques. Il commence par la présentation des principales structures de filtrage puis il traite des aspects de précision finie.

1 Structures des filtres numériques

Il n'est pas dans l'objectif de ce document de présenter de façon très approfondie les structures possibles d'implémentation de filtres numériques. On se limite aux approches les plus classiques.

1.1 Structures directes

Les structures directes correspondent à des implémentations dans lesquelles les valeurs des coefficients de l'équation de récurrence interviennent explicitement. Pour des raisons de simplicité, on se limite ici à l'ordre 2, mais les concepts présentés s'étendent sans difficulté à un ordre quelconque. Soit l'équation de récurrence d'un filtre numérique IIR d'ordre 2 :

$$y_n = \sum_{k=0}^{2} b_k x_{n-k} - \sum_{k=1}^{2} a_k y_{n-k}$$

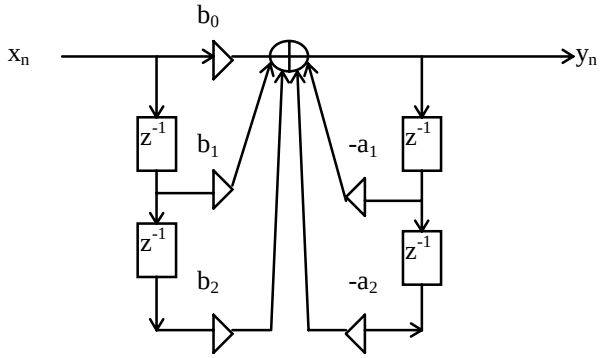
Pour implanter ce filtre, il suffit que le système de traitement soit capable d'effectuer des multiplications, des additions et des retards ou mise en mémoire. Les retards dans le cas d'un système temps réel peuvent être mis en oeuvre par l'intermédiaire de registres à décalage cadencés à la fréquence d'échantillonnage des signaux.

1.2 Structures directes non canoniques

La structure directe la plus simple consiste à mémoriser 4 échantillons : les 2 derniers échantillons des suites d'entrée et de sortie (x_{n-1}, x_{n-2}, y_{n-1}, y_{n-2}) et , pour chaque échantillon x_n de la suite d'entrée effectuer les opérations suivantes, pour calculer l'échantillon correspondant de la suite de sortie y_n :

- effectuer les produits des 3 valeurs x_n, x_{n-1}, x_{n-2} par les coefficients du numérateur de $H(z)$ b_0, b_1, b_2,
- effectuer les produits des 2 valeurs y_{n-1}, y_{n-2} par les coefficients du dénominateur de $H(z)$ a_1, a_2,
- puis cumuler les 5 produits, avec un signe positif pour les coefficients b_i et un signe négatif pour les coefficients a_i
- enfin mettre à jour les mémoires x_{n-1}, x_{n-2},y_{n-1}, y_{n-2} pour préparer le calcul de la sortie suivante y_{n+1}

La figure suivante représente cette structure directe :



1.3 structures directes canoniques DN et ND

La structure directe présentée précédemment n'est pas canonique, dans le sens où elle n'utilise pas un nombre minimum de mémoires.
Il est possible de réaliser le même filtre d'ordre 2 avec seulement 2 cases mémoire. Les structures canoniques DN et ND sont des structures directes utilisant explicitement les coefficients a_i, b_i et ne nécessitant que 2 cases mémoire pour 1 filtre d'ordre 2.

Structure canonique DN

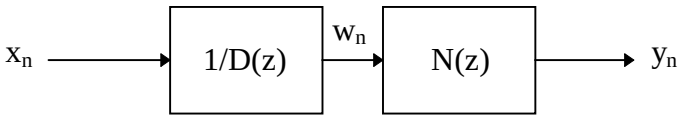
La structure canonique DN réalise le filtre en calculant d'abord la sortie du filtre de fonction de transfert $\frac{1}{D(z)}$ puis la sortie du filtre de fonction de transfert $N(z)$. D'où le nom destructure DN.
On peut écrire :

$$\begin{aligned} Y(z) &= H(z)X(z) \\ Y(z) &= \frac{N(z)}{D(z)}X(z) \\ Y(z) &= \frac{X(z)}{D(z)}N(z) \end{aligned}$$

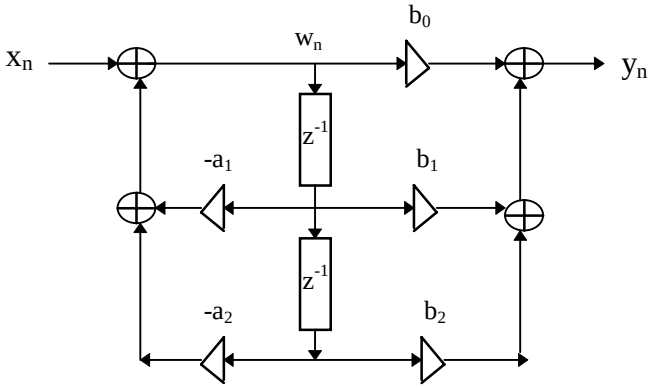
Soit :

$$\begin{aligned} W(z) &= \frac{X(z)}{D(z)} \\ Y(z) &= W(z)N(z) \end{aligned}$$

Ces relations peuvent se résumer par le schéma suivant :



La structure canonique DN correspond finalement à l'implémentation suivante :



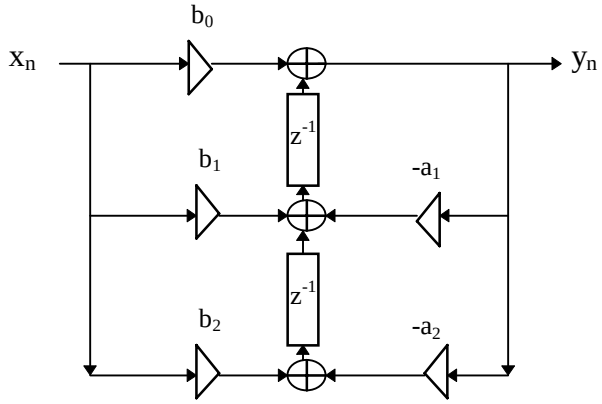
Structure canonique ND

La structure canonique ND est la structure duale (au sens de la théorie des graphes) de la structure DN. Elle doit son nom au fait que les coefficients du numérateur b_i interviennent en amont des coeffcients

du dénominateur a_i.
Les équations qui caractérisent cette structure sont :

$$\begin{aligned} u_n &= b_1 x_n - a_1 y_n \\ v_n &= b_2 x_n - a_2 y_n \\ y_n &= b_0 x_n + u_{n-1} + v_{n-2} \end{aligned}$$

La structure canonique ND correspond finalement à l'implémentation suivante :



Cette structure est la plus fréquemment utilisée dans les circuits intégrés spécifiques pour le filtrage numérique.

1.4 Structures directes pour les filtres FIR symétriques ou antisymétriques

Les filtres FIR à temps de propagation de groupe constant ont une réponse impulsionnelle symétrique ou antisymétrique. Cette propriété peut être exploitée pour l'implémentation de façon à diviser par deux le nombre de multiplications à effectuer.
La relation de récurrence entrée-sortie s'écrit dans la cas symétrique :

$$y(n) = \sum_{i=0}^{\frac{N}{2}-1} b(i)\left(x(n-i) + x(n-N+1+i)\right) \text{ N pair}$$

$$y(n) = \sum_{i=0}^{\frac{N-1}{2}-1} b(i)\left(x(n-i) + x(n-N+1+i)\right) + b\left(\tfrac{N-1}{2}\right) x\left(n - \tfrac{N-1}{2}\right) \text{ N impair}$$

Cette relation se généralise de manière évidente au cas antisymétrique.

1.5 Structures décomposées

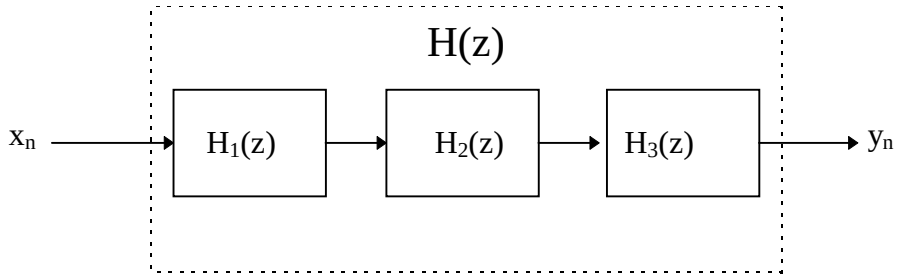
Les structures décomposées n'utilisent pas explicitement les coefficients a_i, b_i dans l'implémentation du filtre. elles se caractérisent par une décomposition de la fonction de transfert $H(z)$ de degré N, en éléments de degré plus faible, 1 ou 2 en général.
Ces structures peuvent permettre d'obtenir de meilleures performances pour une implémentation en précision finie, en terme de sensibilité à la quantification des coefficients ou en terme de bruit de calcul.

1.5.1 Structures cascade

La structure cascade se caractérise par une décomposition de H(z) en un produit de termes $H_i(z)$ d'ordre 1 ou 2, selon que les pôles sont réels ou complexes. La fonction globale de filtrage H(z) est réalisée par une cascade de cellules de filtrage $H_i(z)$ d'odre 1 ou 2.

$$\begin{aligned} H(z) &= \prod_{i=0}^{K} H_i(z) \quad \text{avec} \\ H_i(z) &= \frac{b_{i,0} + b_{i,1} z^{-1} + b_{i,2} z^{-2}}{1 + a_{i,1} z^{-1} + a_{i,2} z^{-2}} \end{aligned}$$

Les cellules $H_i(z)$ est implantée sous une forme canonique DN ou ND. La figure suivante représente l'implémentation cascade d'un filtre d'ordre 5, dont le dénominateur possède 1 pôle réel et 2 pôles complexes conjugués.



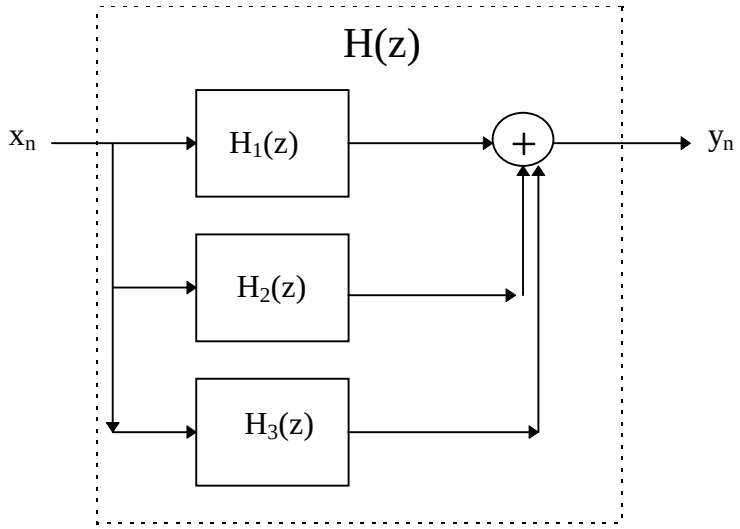
Pour une fonction H(z) donnée, il existe plusieurs implémentations cascade possibles, selon la façon dont on groupe les pôles et les zéros pour former les cellules $H_i(z)$ et selon la façon dont on ordonne ces cellules. Ainsi pour la fonction H(z) d'ordre 5 du schéma précédent, existe-t-il 12 réalisations cascade différentes (si on impose de grouper le zéro réel avec le pôle réel). Ces réalisations ne sont pas équivalentes en terme de performances pour une implémentation en précision finie.

1.5.2 Structures parallèles

La structure parallèle se caractérise par une décomposition de H(z) en un une somme d'éléments simples $H_k(z)$ d'ordre 1 ou 2, selon que les pôles sont réels ou complexes. La fonction globale de filtrage H(z) est réalisée par une somme de sortie de cellules de filtrage $H_k(z)$ d'odre 1 ou 2, attaquées par la même entrée x_n.

$$\begin{aligned} H(z) &= \sum_{k=0}^{L} H_k(z) \quad \text{avec} \\ H_k(z) &= \frac{b_{k,0} + b_{k,1}z^{-1}}{1 + a_{k,1}z^{-1} + a_{k,2}z^{-2}} \end{aligned}$$

Les cellules $H_k(z)$ est implantée sous une forme canonique DN ou ND. La figure suivante représente l'implémentation cascade d'un filtre d'ordre 5, dont le dénominateur possède 1 pôle réel et 2 pôles complexes conjugués.



Les structures cascade et parallèles ont des perfomances assez proches pour une implémentation en précision finie. Toutefois, on peut remarquer que la structure parallèle ne permet pas aussi facilement que la structure cascade de conserver les zéros de transmission après quantification des coefficients des cellules $H_i(z)$.
La structure cascade se prête bien à une implémentation séquentielle. La structure parallèle permet une implémentation parallèle des opérateurs de calcul conduisant à une plus grande vitesse d'exécution.

1.6 Autres structures

1.6.1 Structure de l'échantillonnage en fréquence pour les FIR

La structure de l'échantillonnage en fréquence s'appuie sur sur la formule d'interpolation de Lagrange pour la représentation d'un polynôme, en l'occurence du polynôme $H(z)$.
On exprime $H(z)$ à l'aide des valeurs $H(z_n)$ prises en des points équirépartis sur le cercle unité :

$$z_n = e^{\frac{j2\pi n}{N}} \quad n = 0,1,\ldots,N-1.$$

La formule d'interpolation de Lagrange permet d'exprimer $H(z)$ en fonctions des N valeurs $H(z_n)$:

$$H(z) = \sum_{n=0}^{N} H(z_n) \prod_{i \neq n} \frac{\left(1 - z_i z^{-1}\right)}{\left(1 - z_i z_n^{-1}\right)}$$

Comme les z_n sont les racines n$\overset{\text{ème}}{=}$ de l'unité :

$$1 - z^{-N} = \prod_{n=0}^{N-1} \left(1 - z_n z^{-1}\right).$$

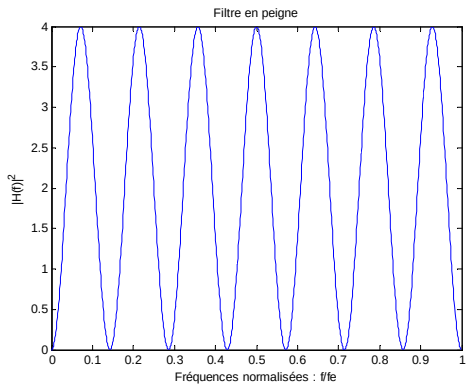
D'où :

$$\begin{aligned} \prod_{i \neq n} \left(1 - z_i z^{-1}\right) &= \frac{1 - z^{-N}}{1 - z_n z^{-1}}. \\ \prod_{i \neq n} \left(1 - z_i z_n^{-1}\right) &= N \end{aligned}$$

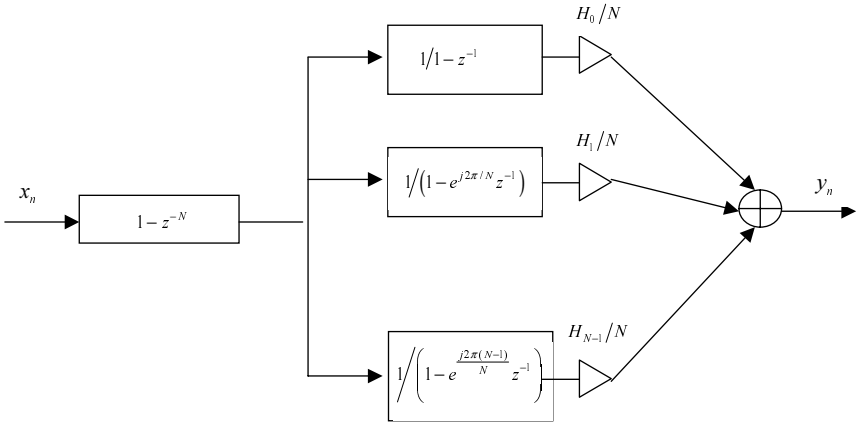
Finalement :

$$\begin{aligned} H(z) &= \prod_{n=0}^{N-1} \left(1 - z^{-1} z_n\right) \sum_{m=0}^{N} \frac{A_m}{1 - z^{-1} z_m}, \\ A_m &= \frac{1}{N} H\left(e^{\frac{j2\pi m}{N-1}}\right) \\ H(z) &= \frac{1}{N}\left(1 - z^{-N}\right) \sum_{m=0}^{N-1} \frac{H(z_m)}{1 - z^{-1} z_m} \end{aligned}$$

Les filtres de fonction de transfert $\frac{1}{1-z_n z^{-1}}$ sont des résonateurs purs (pôle sur le cercle unité). Pour éviter l'instabilité, on amortit généralement leurs pôles. On regroupe deux à deux les pôles complexes conjugués pour former des cellules de fonction de transfert : $\frac{1}{1-2\cos(2\pi k/N)z^{-1}+z^{-2}}$
La fonction $1 - z^{-N}$ correspond à un filtre en peigne. Dans le domaine fréquentiel, la fonction de transfert correspondante présente $(N-1)/2$ maxima uniformément répartis entre 0 et $f_e/2$. La figure suivante représente le module de cette fonction de transfert pour N=15.



La figure suivante illustre cette structure de l'échantillonnage en fréquence pour les FIR (on n'a pas regroupé les pôles complexes conjugués sur ce schéma).



Cette structure est intéressante quand un nombre important de valeurs H_k sont nulles. Par exemple, c'est une manière efficace de réaliser un filtre passe-bande étroit avec peu de multiplieurs.

1.6.2 Autres structures, rappel sur la représentation d'état

Ils existent de nombreuses autres structures décomposées, comme la structure treillis, ou les filtres d'ondes. Ces structures ne seront pas développées ici. La théorie de la représentation d'état permet de comprendre le passage d'une structure à une autre et de comparer ces structures. On pourra se reporter aux ouvrages de référence [3].

Pour conserver les mêmes notations que celles utulisées dans le reste du document, on note x_n l'entrée du filtre, alors que généralement cette notation est utilisée pour représenter le vecteur d'état. On note E_n le vecteur d'état.

La représentation d'état comprend les 2 équations suivantes (équation de transition et équation de mesure) :

$$\begin{aligned} E_{n+1} &= AE_n + Bx_n \\ y_n &= CE_n + Dx_n \end{aligned}$$

Par exemple, pour le filtre de fonction de transfert :

$$H(z) = \frac{1}{1 - 0{,}7683z^{-1} + 0{,}4248z^{-2}}$$

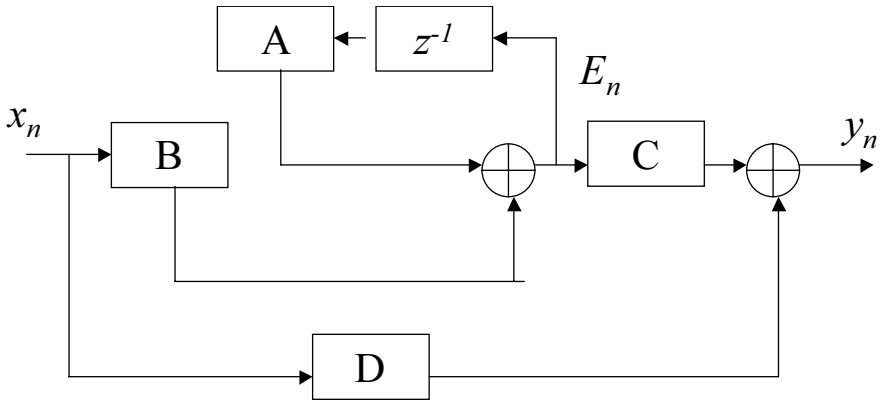
En prenant un vecteur d'état En construit sur les variables internes de la structure DN et défini par :

$$E_n = \begin{pmatrix} w_{n-1} \\ w_{n-2} \end{pmatrix} \text{ avec } y_n = x_n - \sum_{i=1}^{P} a_i w_{n-i}.$$

On obtient la représentation d'état avec les matrices :

$$\begin{aligned} A &= \begin{pmatrix} 0.7683 & -0.4248 \\ 1 & 0 \end{pmatrix} \\ B &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ C &= \begin{pmatrix} 0.3374 & 0.1157 \end{pmatrix} \\ D &= \begin{pmatrix} 0.2011 \end{pmatrix}. \end{aligned}$$

La figure suivante représente ces relations :



Pour illuster la représentation d'état, considérons l'exemple d'une cellule d'ordre 2 purement récursive caractérisée par la fonction de transfert en z $H(z)$ suivante :

$$H(z) = \frac{0{,}2011 + 0{,}1829z^{-1} + 0{,}2011z^{-2}}{1 - 0{,}7683z^{-1} + 0.4248z^{-2}}.$$

L'équation de récurrence correspondante est :

$$y_n = 0{,}2011x_n + 0{,}1829x_{n-1} + 0{,}2011x_{n-2} + 0{,}7683y_{n-1} - 0{,}4248y_{n-2}.$$

2 Implémentation en précision finie

Cette section traite des question relatives à l'implémentation en précision finie. Elle aborde la limitation de la précision des coefficients du filtre et des données.

2.1 Limitation de la précision des coefficients

On s'intéresse maintenant au bruit de calcul introduit par la limitation du nombre de bits pour la représentation des coefficients. On se limite au cas d'une implémentation en format fixe qui est le cas le plus critique.
On note :

- B_C le nombre total de bits pour la représentation des coefficients,
- B_{CE} le nombre de bits pour la partie entière des coefficients,
- B_{CF} le nombre de bits pour la partie fractionnaire des coefficients.

La sensibilité du filtre à la précision des coefficients dépend de la structure du filtre. De manière générale, les structures directes seront plus sensibles que les structures décomposées. En effet, la dynamique des coeffcients pour une structure directe n'est pas limitée et le nombre de bits nécessaire pour la partie entière pourra être important, limitant ainsi la précision (nombre de bits disponibles pour la partie fractionnaire).

2.1.1 Principe de la quantification en format fixe sur B_C bits

Soit un jeu de N coefficients c_n à coder en format fixe sur B_C bits.
On commence par chercher la valeur absolue maximale pour déterminer le nombre de bits nécessaires à la partie entière :

$$\begin{aligned} c_{\max} &= \max_n |c_n| \\ B_{CE} &= [\log_2(c_{\max})] \text{ où } [\ldots] \text{ représente l'arrondi à l'entier le plus proche} \\ B_{CF} &= B_C - B_{CE} \end{aligned}$$

En supposant une quantification uniforme, l'erreur de quantification sur les coeffcients e est bornée par :

$$|e| < \frac{q}{2}, \qquad q = 2^{-B_{CF}}$$

2.1.2 Cas des filtres FIR réalisés avec une structure directe

Pour un filtre FIR réalisé avec structure directe, l'erreur sur les coefficients est bornée par $q/2$. Après quantification des N coefficients b_k du filtre, la nouvelle fonction de transfert en z $\tilde{H}(f)$, peut s'écrire :

$$\begin{aligned} \tilde{H}(f) &= \sum_{k=0}^{N-1} (b_k + e_k) z^{-k} = H(z) + dH(z), \\ dH(z) &= \sum_{k=0}^{N-1} e_k z^{-k} \end{aligned}$$

où e_k représente l'erreur de quantification sur b_k.
La fonction d'erreur $dH(z)$ est bornée dans le domaine fréquentiel par :

$$|dH(f)| \leq N\frac{q}{2}$$

On peut utiliser une borne moins large, en faisant l'hypothèse que les erreurs de quantification e_k sont des variables uniformes centrées comprises entre $-q/2$ et $+q/2$ et indépendantes entre elles. Sous cette hypothèse, la variance de e est égale à $q^2/12$ et on peut utliser la borne :

$$E\left(|dH(f)|^2\right) \leq N\frac{q^2}{12}.$$

On considère ainsi que $|dH(f)|$ est une variable alétoire d'écart-type σ avec :

$$\sigma \leq \frac{q}{2}\sqrt{\frac{N}{3}}.$$

Considérons par exemple un gabarit passe-bas, avec une atténuation minimum Rs en bande atténuée correspondant à une ondulation δ_s donnée par :

$$\delta_s = 10^{-R_s/20}.$$

On choisira, dans ce cas, un pas de quantification q tel que :

$$q < (\delta_s - \delta_0)\sqrt{\frac{3}{N}},$$

où :

- δ_s = valeur imposée par le gabarit pour l'amplitude des oscillations.
- δ_0 = amplitude des ondulations du filtre avant quantification des coeffcients.

Il faut donc B_C bits :

$$B_C \geq \log_2\left(\frac{1}{q}\max_{k\in[0,N-1]}|b_k|\right) \geq \log_2\left(\sqrt{\frac{N}{3}}\frac{1}{(\delta_s-\delta_0)}\max_{k\in[0,N-1]}|b_k|\right).$$

2.1.3 Cas des filtres IIR réalisés avec une structure directe

De la même façon que pour les FIR en structure directe, après quantification des coefficients b_k du numérateur et des coefficients a_j du dénominateur, la nouvelle fonction de transfert en z $\tilde{H}(f)$, peut s'écrire :

$$\begin{aligned}
\tilde{H}(f) &= \frac{\sum_{k=0}^{P-1}(b_k + e_{Nk})z^{-k}}{1 + \sum_{j=1}^{Q-1}(a_j + e_{Dj})z^{-j}} = \frac{N(z) + e_N(z)}{D(z) + e_D(z)}, \\
e_N(z) &= \sum_{k=0}^{P-1} e_{Nk} z^{-k} \\
e_D(z) &= \sum_{j=0}^{Q-1} e_{Dj} z^{-j}
\end{aligned}$$

On peut avec une approximation à l'ordre un, approcher $\tilde{H}(z)$ par :

$$\tilde{H}(z) \approx \frac{N(z) + e_N(z) - e_D(z)H(z)}{D(z)}.$$

On peut appliquer la même approche que dans le cas filtres FIR. En particulier, on peut sous certaines hypothèses considérer que $|e_N(f)|$ et $|e_D(f)|$ sont des variables aléatoires d'écart-type respectifs σ_N et σ_D :

$$\sigma_N = \frac{q}{2}\sqrt{\frac{P}{3}}, \quad \sigma_D = \frac{q}{2}\sqrt{\frac{Q-1}{3}}.$$

Analyse en bande atténuée
Si on analyse la situation en bande atténuée, $|H(f)| \approx 0$ et en première approximation :

$$\tilde{H}(f) \approx \frac{N(f) + e_N(f)}{D(f)}.$$

L'erreur $e(f)$ sur la fonction de transfert en fréquence dans la bande atténuée peut être bornée par :

$$|e(f)| = |H(f) - \tilde{H}(f)| < \frac{\sigma_N}{|D(f|}.$$

Et pour une ondulation maximale souhaitée δ_s en bande atténuée, on fera en sorte que :

$$\frac{\sigma_N}{|D(f|}. < \frac{\delta_s}{2}.$$

Analyse en bande passante
En bande passante, $|H(f)| \approx 1$, on peut alors approcher $|\tilde{H}(f)|$ par :

$$\tilde{H}(f) \approx \frac{N(f) + e_N(f) - e_D(f)}{D(f)}.$$

L'erreur $e(f)$ sur la fonction de transfert en fréquence dans la bande passante peut donc être approchée par :

$$|e(f)| = |H(f) - \tilde{H}(f)| \approx \frac{|e_N(f) - e_D(f)|}{|D(f)|}.$$

D'où, en notant δ_p l'ondulations maximale autorisée en bande passante :

$$\frac{\sqrt{\sigma_N^2 + \sigma_D^2}}{|D(f)|} < \frac{\delta_p}{2}$$

Généralement cette contrainte est plus forte que la contrainte en bande atténuée car $|D(f)|$ peut être très faible.

La quantification des coefficients a pour conséquence que les zéros et les pôles de $\tilde{H}(z)$ ne peuvent prendre qu'un nombre fini de valeurs. La quantification des coefficients du dénominateur peut rendre le filtre instable.

Les dynamique des valeurs des coefficients b_k, a_j peut être très importante. Il faut donc un grand nombre de bits pour les représenter avec une bonne précision.

Kaiser a montré que les structures directes présentent une très grande sensibilité aux valeurs des coefficients. C'est une des raisons pour lesquelles elles sont peu employées (en dehors des cellules élementaires d'ordre 1 ou 2).

2.1.4 Cas des filtres IIR réalisés avec une structure décomposée

Pour les différentes raisons vues précédemment, quand on travaille en précision finie en format fixe, on utilise peu les structures directes. on préfère en général les structures décomposées formées de cellules élémentaires d'ordre un ou deux. En effet, pour les cellules d'ordre un ou deux, la dynamique des coefficients est bornée. Pour une cellule d'ordre 2 ayant deux pôles complexes conjugués, les coefficients du dénominateur sont inférieurs à 2 en valeur absolue. Si la cellule a un zéro de transmission correspondant à deux zéros complexes conjugués de module égal à 1, les coefficients du numérateur sont inférieurs ou égaux à 2.

Dans le cas des structures cascade, on peut faire en sorte de conserver les zéros de transmission : il suffit de garder la propriété $b_0 = b_2$ pour une cellule d'ordre. Par contre, c'est plus difficile pour une structure parallèle.

Quantifications des coefficients - Cas d'une structure cascade

Dans le cas d'une structure cascade, on peut écrire :

$$\begin{aligned}
H(f) &= \prod_i \frac{N_i(f)}{D_i(f)} \\
\tilde{H}(f) &= \prod_i \frac{N_i(f) + e_{Ni}(f)}{D_i(f) + e_{Di}(f)} \\
&\approx \prod_i \frac{N_i(f) + e_{Ni}(f) - e_{Di}(f)H_i(f)}{D_i(f)} \\
&\approx H(f) + H(f)\left(\sum_i \frac{e_{Ni}(f)}{N_i(f)} - \sum_i \frac{e_{Di}(f)}{D_i(f)}\right)
\end{aligned}$$

En bande passante, $|H(f)| \approx 1$ et $N_i(f) \neq 0$. L'erreur sur la fonction de transfert est donc liée aux termes sous forme de somme :

$$\sum_i \frac{1}{|D_i(f)|} \text{ et } \sum_i \frac{1}{|N_i(f)|}$$

alors que pour la structure directe, intervient plutôt un terme sous forme de produit :

$$\frac{1}{|D(f)|} = \prod_i \frac{1}{|D_i(f)|}.$$

Et généralement, en bande passante près des pôles :

$$\sum_i \frac{1}{|D_i(f)|} << \prod_i \frac{1}{|D_i(f)|} = \frac{1}{|D(f)|}.$$

D'où l'intérêt de la structure cascade.

2.2 Limitation de la précision des données

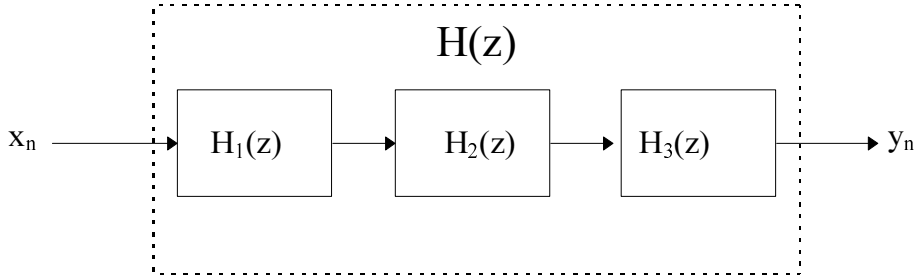
On s'intéresse maintenant au bruit de calcul introduit par la limitation du nombre de bits pour la représentation des données. On se limite au cas de la représentation des données en format fixe.

On note :

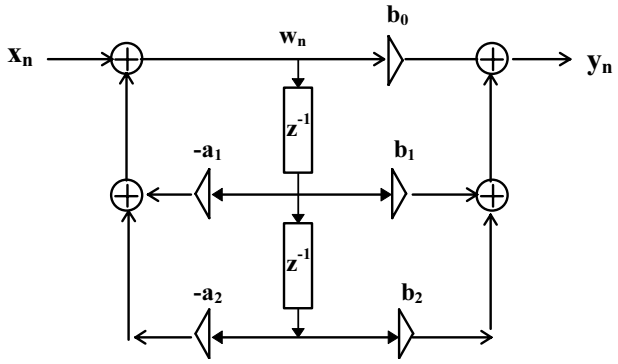
- B_D = le nombre total de bits pour les données,
- B_{DE} = le nombre de bits de la partie entière des données,
- B_{DF} = le nombre de bits de la partie fractionnaire.

On étudie en détail le cas de la structure cascade qui est une des structures les plus courantes. On suppose que les différentes cellules élémentaires qui constituent la strcuture cascade sont réalisées en structure canonique DN.

Les figures suivantes rappellent respectivement la réalisation d'un filtre $H(z)$ par une structure cascade et la structure canonique DN pour une cellule d'ordre 2.



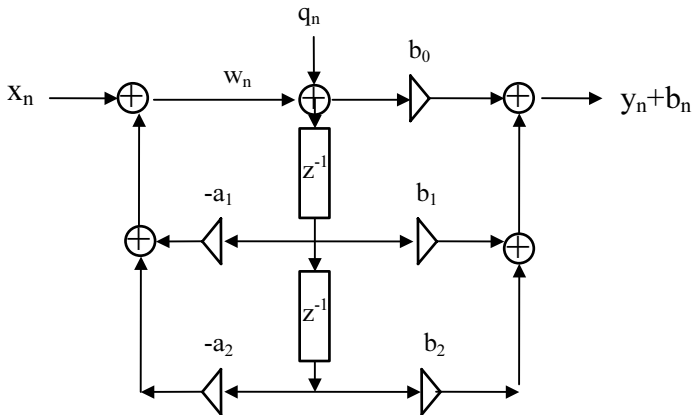
Structure cascade.



Structure canonique DN pour une cellule de la structure cascade.

On considère une implémentation de type DSP dans laquelle la quantification a lieu au moment de mettre en mémoire les variables w_n. On rappelle que w_n correspond à x_n filtré par la fonction de transfert $\frac{1}{D(z)}$.

La figure suivante illustre l'introduction du bruit de quantification dans une cellule DN.



Bruit dans une structure canonique DN.

Par la suite, pour représenter les données (ou coefficients) correspondant à la i$^{\text{ème}}$ cellule, on ajoutera un indice i. Par exemple, $w_{i,n}$ représente la variable w_n dans la i$^{\text{ème}}$ cellule et $q_{i,n}$ le bruit généré dans la i$^{\text{ème}}$ cellule.

Le bruit de quantification q_n sur w_n s'ajoute directement à x_n et subit donc la fonction de transfert $H(f)$ globale de la cellule. En faisant l'hypothèse que q_n est un bruit blanc (dans les bandes passantes du filtre), de densité de probabilité uniforme et de variance $q^2/12$, la densité spectrale $S_b(f)$ de puissance du bruit b_n qui en résulte en sortie du filtre dans le cas d'une cellule d'ordre 2, peut s'écrire :

$$S_b(f) = \frac{q^2}{2}|H(f)|^2.$$

Et plus généralement, si on note $H_{Di}(f)$ la fonction de transfert entre le point de génération du bruit $q_{i,n}$ et la sortie du filtre, la densité spectrale $S_{b_i}(f)$ de puissance du bruit $b_{i,n}$ qui en résulte en sortie du filtre, peut s'écrire :

$$S_{b_i}(f) = \frac{q^2}{2}|H_{Di}(f)|^2.$$

2.2.1 Contrôle des débordements - Facteurs d'échelle

Le format fixe choisi pour les données $w_{i,n}$ doit avoir un nombre de bits pour la partie entière suffisant pour éviter le débordement. On considère par la suite (sans perte de généralité) que toutes les données (x_n, y_n, w_n) sont comprises dans l'intervalle $[-1,1[$ et sont représentées avec 1 seul bit de partie entière. Pour éviter le débordement dans la cellule i, il faut donc que $w_{i,n} < 1$ quelque soit l'entrée x_n. Et dans ce cas :

$$\begin{aligned} B_{DF} &= B_D - 1 \\ q &= 2^{-B_{DF}} \end{aligned}$$

Par la suite, on note $G_i(f)$ la fonction de transfert entre l'entrée du filtre x_n et $w_{i,n}$. Par exemple, dans le cas d'une seule cellule d'ordre 2, $G(f) = \frac{1}{D(f)}$.

En pratique, on peut utiliser différents critères pour ce débordement. Les critères les plus utilisés sont les suivants :

- On peut exiger qu'il n'y ait pas de débordement sur w_n quelque soit l'entrée x_n bornée. Ce critère correspond à une fonction $G_i(f)$ bornée pour la norme L_1.

$$\int_{-\infty}^{\infty} |G_i(f)|df < 1.$$

- On peut se contenter du fait que pour toute entrée x_n bornée et de fréquence pure, la suite w_n soit bornée. Ce critère correspond à une fonction $\frac{1}{D(f)}$ bornée pour la norme L_∞.

$$\max_f |G_i(f)| < 1.$$

Pour satisfaire cette contrainte (norme L_1 ou L_∞ ou plus généralement une norme L_p), on introduit un facteur d'échelle en entrée de la cellule. On note sc_i le facteur d'échelle pour la cellule i.

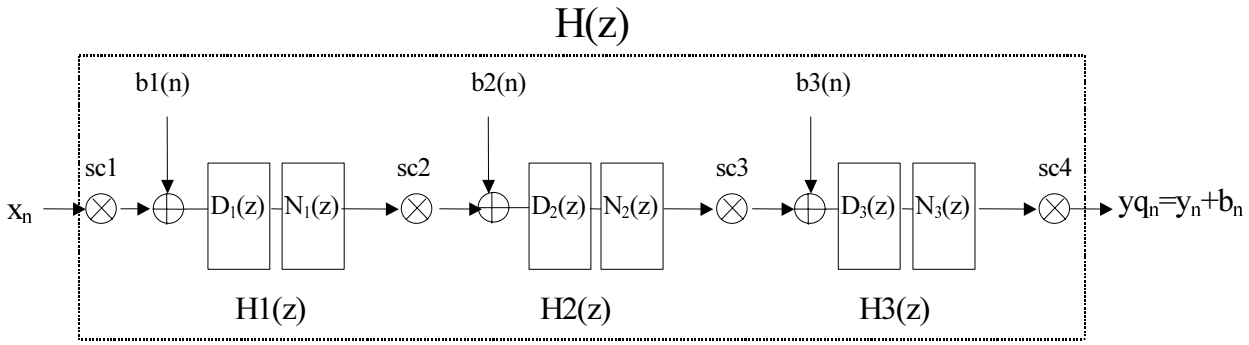
La dernier facteur d'échelle (en sortie de la dernière cellule) est calculé pour compenser les autres. Pour un filtre constitué de K cellules en cascade (H_1, H_2, ..., H_i, ..., H_K), si l'on veut un gain global égal à G, on a la relation :

$$H(z) = G\prod_{i=1}^{K} H_i(z).$$

Et on en déduit :

$$sc_{K+1} = \frac{G}{\prod_{i=1}^{K} sc_i}.$$

La figure suivante représente pour une structure cascade composée de 3 cellules en cascade, la position des facteurs d'échelle et les points d'entreé des bruits de quantification.



La fonction de transfert $G_i(f)$ s'écrit :

$$G_i(f) \quad = \quad \frac{1}{D_i(f)} \prod_{k=1}^{i-1} H_k(f).$$

On note $\|A\|_p$ la norme L_p de la fonction A. On calcule les facteurs d'échelle :

$$sc_i = \frac{1}{\|G_i(f)\|_p} \frac{1}{\prod_{k=1}^{i-1} sc_k} = \frac{\|G_{i-1}(f)\|_p}{\|G_i(f)\|_p} \quad \forall i \in [1,K].$$

$$sc_{K+1} = G\|G_K\|_p.$$

2.2.2 Calcul du bruit en sortie du filtre

On suppose que les bruits de quantification introduits dans les différentes cellules sont indépendants, la densité spectrale de puisance du bruit total b_n en sortie du filtre est la somme des densités spectrales de puissance des différents bruits filtrés $b_{i,n}$ plus le bruit ajouté sur la sortie par le convertisseur numérique analogique (ce dernier bruit est supposé blanc de densité spectrale de puissance $\frac{q^2}{12}$). Pour un filtre formé de K cellules, $S_{b_i}(f)$ en sortie du filtre vaut :

$$S_{b_i}(f) = \frac{q^2}{2}|H_{Di}(f)|^2 = \frac{q^2}{12} \prod_{k=i+1}^{K+1} sc_k^2 \prod_{k=i}^{K} \|H_k(f)\|^2 = \frac{q^2}{12} G^2 |G_i(f)\|_p^2 \prod_{k=i}^{K} \|H_k(f)\|^2$$

La densité spectrale de puissance du bruit total b_n en sortie vaut :

$$S_b(f) = \frac{q^2}{12} + \sum_{i=1}^{K} Sb_i(f) = \frac{q^2}{12} + G^2 \frac{q^2}{12} \sum_{i=1}^{K} \left(|G_i(f)\|_p^2 \prod_{k=i}^{N} \|H_k(f)\|^2 \right). \tag{V.1}$$

On peut aussi écrire cette puissance sous la forme :

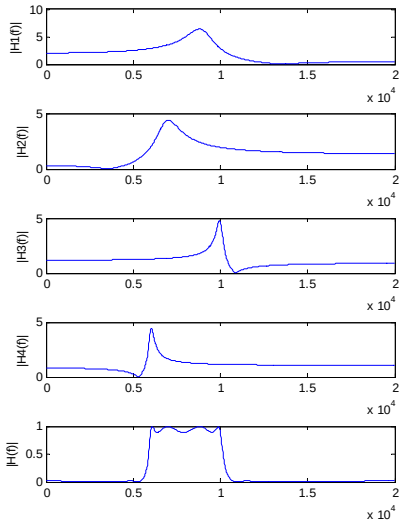
$$S_b(f) = \frac{q^2}{12} + \frac{q^2}{12} G^2 \sum_{i=1}^{K} \left(\frac{1}{\prod_{k=1}^{i} sc_k^2} \prod_{k=i}^{K} \|H_k(f)\|^2 \right). \tag{V.2}$$

Pour minimiser la puissance du bruit due à la limitation de la précision des données en sortie du filtre, il faut apparier les pôles et les zéros et ordonner les cellules de manière optimale. La solution optimale peut être obtenue en testant tous les appariements et ordonnancements possibles et en choisissant la configuration qui minimise la puissance de bruit donnée par la relation V.1. Le nombre de possibilités à tester est important ($(K!)^2$) pour les ordres de filtres un peu élevé. Par exemple, pour un ordre 10 et un filtre contitué de 5 cellules d'ordre 2, il faut tester14400 configurations.

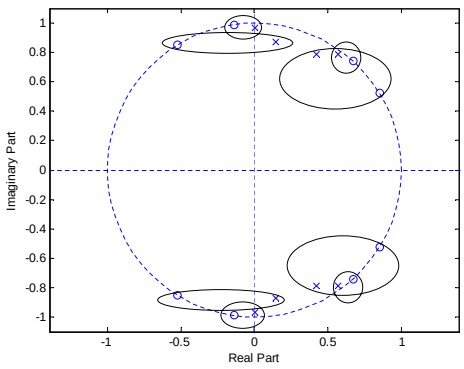
En partique, on utilise souvent des règles empiriques. Ainsi en vue de minimiser la puissance du bruit de calcul en sortue du filtre, tout en évitant les débordements pour les fréquences pures (choix de la norme L_∞ pour les débordements), la règle empirique consiste à :

- Apparier les pôles et les zéros en regroupant chaque pôle avec le zéro qui lui est le plus proche, en commançant par le pôle le plus proche du cercle unité et en continuant par ordre de module décroissant pour les pôles. L'objectif est de former des cellules $H_i(f)$ ayant le moins possible un aspect résonant.
- Ordonner les cellules en commençant par le pôle le plus proche de l'origine et en continuant par ordre de module croissant pour les pôles. En effet, l'équation V.2 montre que le premier facteur d'échelle joue sur K les termes du bruit, le deuxième facteur d'échelle ne joue que sur les $K-1$ derniers termes du bruit et ainsi de suite jusqu'au dernier facteur qui ne joue que sur le dernier terme du bruit.

Les quatre premières courbes de la figure suivante représentent, pour un filtre passe-bande d'ordre 8, les 4 fonctions de transfert élémentaires de la cascade $|H_i(f)|, \quad i \in [1,4]$ obtenues en regroupant les pôles et les zéros comme indiqué précédemment. Les fonctions élémentaires sont indiquées dans l'ordre de la cascade (de la moins résonnante à la plus résonnante. La dernière coubre représente la fonction de transfert globale $|H(f)$.

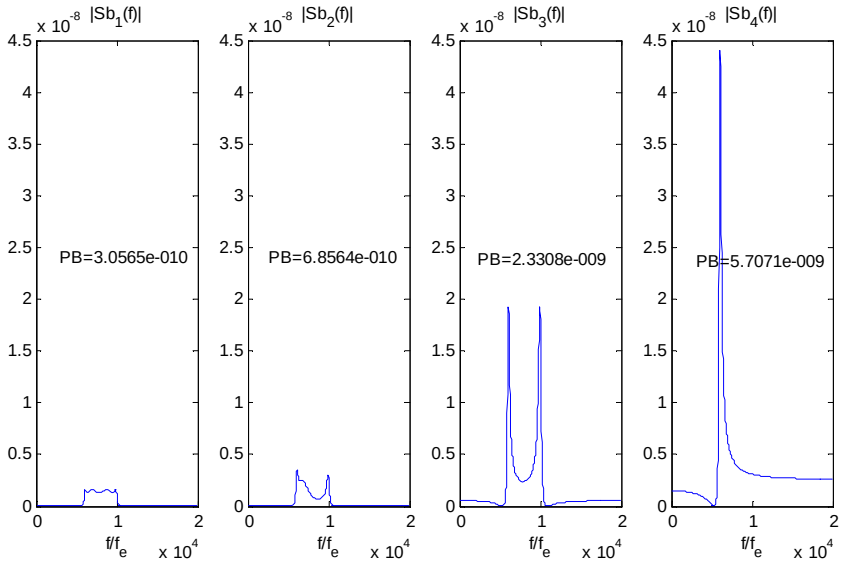


La figure suivante représente la façon dont les pôles et les zéros ont été appariés.



La figure suivante montre les contributions des différentes cellules à la densité spectrale de bruit en sortie. Elle représente les 4 fonctions $|S_{b_i}(f)|$. La grandeur PB indiquée sur les tracés est la puissance

totale correspondant à chaque courbe.



2.2.3 Cycles limites

Les cycles limites sont des oscillations qui apparaissent à cause de non-linéarités dans un filtre normalement stable. Ces non-linéarités sont générées par des débordements dans les calculs (données trop grandes en valeur absolue) ou par des troncatures (données petites). On distingue deux types de cycles limites : les cycles limites de grande amplitude et les cycles limites de petites amplitudes.

Cycles limites de grande amplitude

Ils sont dûs aux débordement dans les calculs.

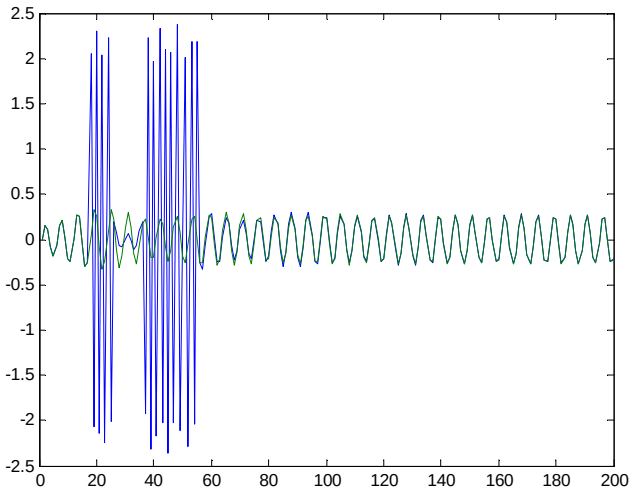
Comme on l'a vu dans le chapitre IV, dans la section 2.2, quand on travaille en complément à deux, on peut considérer deux modes de travail : le mode de débordement circulaire naturel (overflow en anglais) et le mode de saturation arithmétique.

Le mode de débordement circulaire naturel correspond à la proporiété de circularité de la représentation en complément à deux. C'est à dire que lorsqu'on ajoute 1 à la plus grande valeur positive, on obtient la valeur négative de plus grande valeur absolue. de même, si on enlève 1 à la valeur négative de plus grande valeur absolue, on obtient la plus grande valeur positive.

Le mode saturation arithmétique correspond à détecter les résultats de valeur absolue supérieure à la plus grande valeur représentable dans le format utilisé et à la remplacer par la valeur maximale de même signe.

En mode de débordement circulaire, la sortie du filtre oscille avec de grandes amplitudes.

La figure suivante illustre ce phénoméne pour une entrée sinusoïdale.



On travaille donc plutôt en mode de saturation arithmétique pour éviter ces oscillations. Sur la figure précédente est superposée la sortie pour ce mode de saturation. On peut constater que la saturation est à peine visible.

Cycles limites de petite amplitude

Les cycles limites de petite amplitude, comme leur nom l'indique, sont des oscillations dont les amplitude sont faibles, typiquement quelques pas de quantification q.

Ils apparaissent par création d'un pôle sur le cercle unité dû à la quantification des calculs intermédiares. Ils se manifestent par le fait que pour une entrée nulle, on obtient une sortie qui ne tend pas vers zéro et qui est périodique de petite amplitude.

La théorie relative à ce phénomène est complexe. On l'illustre ici pour les cellules d'ordre 1 ou 2.

Cas de la cellule d'ordre 1

Soit le filtre d'équation de récurrence :

$$y_n = x_n - a_1 y_{n-1}.$$

Dire qu'il y a un pôle sur le cercle unité, signifie à l'ordre un, que $|a_1| = 1$, c'est-à-dire que $y_n = y_{n-1}$ si $a_1 < 0$ ou que $y_n = -y_{n-1}$ si $a_1 > 0$. Les cycles limites apparaissent donc quand :

$$[a_1 y_{n-1}] = \pm y_{n-1}.$$

où $[x]$ représente la valeur de x quantifiée avec la précision de la représentation binaire utilisée dans le filtre. C'est-à-dire que la quantification a la même conséquence qu'une valeur a_1 égale à un en valeur absolue, ce qui correspond à un pôle sur le cercle unité.

On note la condition initiale $y_{-1} = y_{ini}$ et q la valeur du pas de quantification.

Toute valeur quantifiée est égale à un nombre entier de fois le pas de quantification. Donc La valeur initiale $y_{-1} = y_{ini}$ peut s'écrire :

$$y_{-1} = y_{ini} = kq \text{ avec } k \text{ entier.}$$

Pour qu'il y ait cycle limite, il faut que :

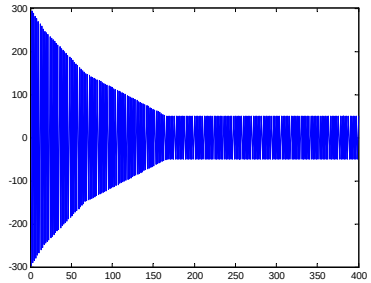
$$\begin{aligned}
[a_1 y_{ini}] &= \pm y_{ini} \\
[a_1 k] &= \pm k \\
|k| - |a_1 k| &< \frac{1}{2} \\
|k| &< \frac{1}{1-|a_1|}. \\
k_{\max} &= \frac{1}{(1-|a_1|)} \\
\max |y_{ini}| &= q k_{\max}
\end{aligned}$$

Il peut donc y avoir cycle limite pour des conditions initiales d'amplitude en valeur absolue inférieure à la limite :

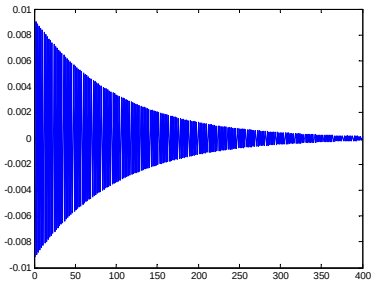
$$|y_{ini}| < q k_{\max} = \frac{q}{2(1-|a_1|)}.$$

La figure suivante illustre le phénomène pour $a_1 = 0.99$. On part d'une condition initiale supérieure à $k_{\max} q$, la sortie commence donc à décroître en module et lorsque la valeur y_n arrive à $k_{\max} q$, le cycle limite apparait : la sortie oscille entre + et -$k_{\max} q$. L'axe des ordonnées est graduée en pas de

quantification. Ici $k_{\max} = \frac{1}{2(1-0.99)} = 50$.



La figure suivante illustre la sortie obtenue avec une précision bien supérieure pour la représentation des données. On ne voit pas apparaître de cycle limite.



Cas de la cellule d'ordre 2

Dans le cas d'une cellule d'ordre 2 avec des pôles complexes conjugués, dire que les pôles sont sur le cercle unité revient à dire que $|a_2| = 1$, c'est-à-dire, pour une implémentation avec une structure canoniqu DN que :

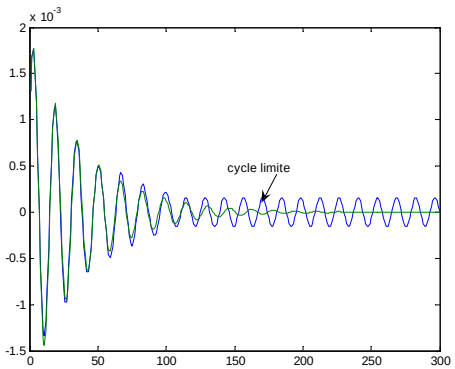
$$[a_2 w_{n-2}] = \pm w_{n-2},$$

en notant w_n la variable interne de la structure DN.
La valeur de a_1 ($a_1 = -2r\cos(\theta)$) conditionne la période des oscillations du cycle limite.
Avec le même raisonnement que pour la cellule d'ordre 1, on peut dire que por qu'il y ait des cycles limites, il suffit que :

$$\begin{aligned} w_{2ini} &= kq \\ k_{\max} &< \frac{1}{2(1-|a_2|)}. \end{aligned}$$

La figure suivante (axe des ordonnées gradué en pas de quantification) illustre le phénomène de cycle limite pour la cellule d'ordre 2 de fonction de transfert :

$$H(z) = \frac{1}{1 - 1{,}8z^{-1} + 0{,}95z^{-2}}.$$



EXERCICES ET PROBLÈMES

Exercice 1 : On considère un système linéaire discret stationnaire de réponse impulsionnelle:

$$h_n = \frac{1}{n+1} \quad \text{pour} \quad 0 \leq n \leq 3 \quad \text{et h(n)=0 ailleurs}$$

1. Donnez l'équation aux différences permettant de calculer la sortie y(n) pour une entrée quelconque x(n).

2. Calculez la sortie du filtre pour $x(n) = a^n$ pour a=0.5, $0 \leq n \leq 3$ et x(n)=0 ailleurs.

3. Calculez la sortie stationnaire correspondant à l'entrée $x(n) = \cos(2\pi n/4)$..

Exercice 2 : Soit les systèmes de fonction de transfert :

$$\begin{aligned} H_1(z) &= \frac{1}{1 - 1.16z^{-1} + 0.92z^{-2}} \\ H_2(z) &= 1 - 1.16z^{-1} + 0.92z^{-2} \end{aligned}$$

Donnez les équations aux différences correspondantes, et calculez les réponses impulsionnelles associées.

Exercice 3 : On considère le système d'équation aux différences :

$$y(n) = x(n) + y(n-1) - 0.5y(n-2)$$

1. Calculez la transformée en z Y(z) monolatérale de y(n), et démontrez la relation:

   $$Y(z) = H(z)X(z) + P(z)$$

   où P(z) dépend des conditions initiales. On posera y(-1)=a et y(-2)=b.

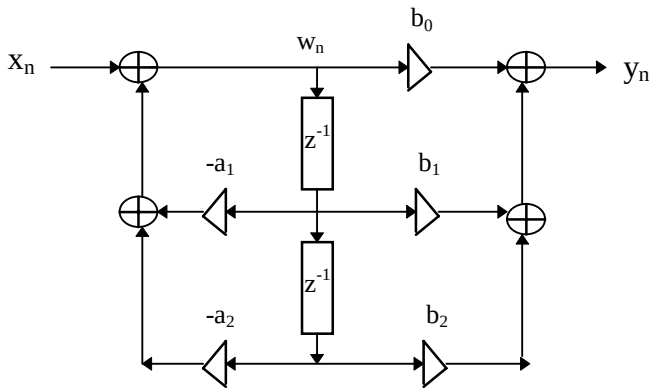
2. Calculez les pôles et les zéros de H(z). Décomposez H(z) en une somme de deux fractions rationnelles du premier ordre. Déduisez en la réponse impulsionnelle du système (vous supposerez que celui-ci est causal). Le système est-il stable ?

3. Donnez la sortie y(n) du système soumis à l'entrée $x(n) = \exp(j2\pi f_0 n T_e)$, pour $n \geq 0$ et 0 si $n \leq 0$. Vous commencerez par calculer la TZ X(z) de x(n), puis vous rechercherez la transformée inverse de H(z)X(z) (on supposera les conditions initiales nulles).

4. Donnez le module de la réponse en fréquence $H(f)$. Calculez la valeur de la fréquence de résonnance et tracez l'allure de la réponse en fréquence.

5. Donnez une structure d'implantation de ce filtre.

Exercice 4 : Donnez l'équation aux différences réalisée par la structure suivante. Vous utiliserez pour cela la variable intermédiaire $w(n)$.



Problème I :

Question 1 : On considère l'équation aux différences

$$y(n) = x(n-1) + x(n) + x(n+1).$$

1-a Mettre cette équation aux différences sous la forme d'une convolution discrète, donnez la rĆponse impulsionnelle $h(n)$ du filtre d'entrĆe $x(n)$ et de sortie $y(n)$.

1-b Donnez la fonction de transfert $H(z)$ de ce filtre.

1-c Donnez la réponse en fréquence $H(f)$. Vous pourrez chercher à faire apparaŇtre un cosinus.

1-d DĆduisez en le module de $H(f)$ et sa phase $\phi(f)$. ReprĆsentez $|H(f)|$, en précisant clairement le domaine de variation de f.

Question 2 : Si l'entrée est

$$x(n) = A\cos(2\pi f_0 n + \phi), \text{ avec } f_0 = 1/3,$$

quelle est la sortie du filtre ?

Question 3 : L'entrée est maintenant un bruit blanc à temps discret, centré, d'autocorrélation

$$\begin{cases} R_{XX}(k) = P_B & \text{si } k = 0, \\ R_{XX}(k) = 0 & \text{sinon.} \end{cases}$$

3-a Quelle est la moyenne statistique m_Y de la sortie $y(n)$ du filtre ?

3-b Calculez l'autocorrélation de la sortie $R_{YY}(k)$.

Question 4 : La sortie du filtre $y(n)$ est maintenant bruitée par un bruit additif $b(n)$, centré et d'autocorrélation $R_{BB}(k)$, indépendant de $x(n)$, et par conséquent de $y(n)$

$$z(n) = y(n) + b(n).$$

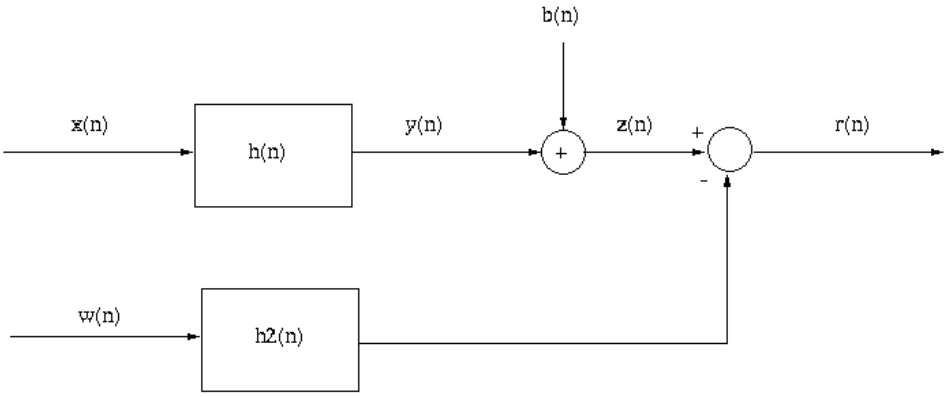
On dispose par ailleurs d'un signal de référence (sortie d'un capteur) $w(n)$, centré, corrélé à $b(n)$ et indépendant de $y(n)$. On peut donc interpréter $b(n)$ comme la sortie d'un filtre $h_2(n)$ d'entrée $w(n)$:

$$b(n) = (h_2 * w)(n).$$

4-a Donnez la relation qui relie l'intercorrélation $R_{BW}(k)$, l'autocorrélation $R_{WW}(k)$ et la réponse impulsionnelle $h_2(k)$ (*cours*).

4-b Calculez l'intercorrélation R_{ZW}, et montrez que si $w(n)$ est blanc centré de variance 1, on en déduit alors la réponse impulsionnelle $h_2(n)$.

4-c Expliquez alors à quoi sert le dispositif



et donnez l'expression de $r(n)$.

4-d On a considéré ci-dessus que $w(n)$ était un bruit blanc. Si tel n'est pas le cas, et si sa densité spectrale de puissance $S_{WW}(f)$ est connue, montrez que le signal $w'(n)$ obtenu comme la sortie d'un filtre de fonction de transfert en fréquence $G(f) = 1/\sqrt{S_{WW}(f)}$ (on notera $g(n)$ sa réponse impulsionnelle) est un bruit blanc de variance unité. Comment faut-il alors modifier le dispositif précédent ?

Chapitre VI

Références

[1] Maurice Bellanger. *Traitement numérique du signal, Théorie et pratique* . Dunod 8e édition ISBN : 2100501623 (2006) (1ère édition en 1981).

[2] LR Rabiner and B. Gold. *Theory and Application of Digital Signal Processing.* Prentice Hall, 1986. ISBN 0-13-914101-4. (1975)

[3] Sanjit K. Mitra. *Digital Signal Processing.* Mc Graw Hill, ISBN : 978-0-07-124467-1 (07/2005).

Index