Travail préparatoire séance 3: quantification des coefficients, étude des sensibilités

*mêmes règles que pour les travaux préparatoires des séances 1 et 2 => la partie **exercices** est à rédiger individuellement, manuscrite, et sera contrôlée par l'enseignant. La lecture des parties 1,2,3 (quoique conseillée) n'est pas indispensable pour répondre aux exercices. Ces parties ont pour but de présenter les techniques classiques d'étude de sensibilité des filtres numériques, et les conclusions qui en découlent.*

1 sensibilité d'une quantité vis à vis d'un ou plusieurs paramètres

1.1 définitions

1.1.1 définition

Soit une quantité *y* dépendant d'une quantité *x*. La sensibilité de y par rapport à x a pour but de déterminer si une variation de *x* va entraîner une petite, ou une grande variation de *y...*

*L*a sensibilité de *y* par rapport à *x*, autour d'une valeur x0 , peut donc être définie de la façon suivante :

pour $x=x_0$, on a $y=y(x_0)$

si on ajoute une petite quantité Δx à $x=x_0$

$x=x_0+\Delta x$

alors y variera d'une petite quantité Δy correspondante

$y=y(x_0+\Delta x)=y(x_0)+\Delta y$

la sensibilité de *y* par rapport à *x*, en $x=x_0$ est mesurée par le rapport $\frac{\Delta y}{\Delta x}$.

cette quantité dépend de la valeur de Δx retenue, pour lui donner le sens d'une sensibilité ''au point $x=x_0$ '', on suppose de plus que Δx est infiniment petit :

$$\underbrace{S_{yx}|_{x=x_0}}_{\text{sensibilité de } y \text{ par rapport à } x, \text{ en } x=x_0} = \frac{\lim_{\Delta x \to 0} y(x_0+\Delta x)-y(x_0)}{[x_0+\Delta x]-x_0}$$

=> la sensibilité est la dérivée (*partielle si y dépend de plusieurs variables*) de *y* par rapport à x, évaluée en $x=x_0$.

$$\underbrace{S_{yx}|_{x=x_0}}_{\text{sensibilité de } y \text{ par rapport à } x, \text{ en } x=x_0} = \left.\frac{\partial y}{\partial x}\right|_{x=x_0}$$

notation : il est courant, pour simplifier les notations, d'exprimer directement cette quantité en fonction de *x* , au lieu de x_0 . La variable *x* a alors le sens du point x_0 autour duquel on évalue la sensibilité (ou la dérivée partielle), et on note :

$$\underbrace{S_{yx}}_{\text{sensibilité de } y \text{ par rapport à } x, \text{ au point } x} = \frac{\partial y}{\partial x}$$

1.2 exemple d'utilisation (correcteur avance de phase)

on considère la fonction de transfert en z^{-1} du correcteur à avance de phase, du travail préparatoire

séance 2 : $F(z)=\frac{b_0+b_1.z^{-1}}{1+a_1.z^{-1}}\approx\frac{22{,}35-20{,}79.z^{-1}}{1-0{,}569.z^{-1}}$,

les sensibilités de $F(z)$ par rapport aux coefficients b_i, a_j , et à la variable z s'écrivent (*on reviendra plus tard sur le calcul des dérivées partielles, ne pas vous focaliser sur la façon dont on obtient ces expressions...*)

$$S_{Fb_0}=\frac{\partial F}{\partial b_0}=\frac{1}{1+a_1.z^{-1}}\approx\frac{1}{1-0{,}569z^{-1}}$$

$$S_{Fb_1}=\frac{\partial F}{\partial b_1}=\frac{z^{-1}}{1+a_1.z^{-1}}\approx\frac{z^{-1}}{1-0{,}569.z^{-1}}$$

$$S_{Fa_1}=\frac{\partial F}{\partial a_1}=-z^{-1}.\frac{b_0+b_1.z^{-1}}{\left(1+a_1.z^{-1}\right)^2}\approx-z^{-1}.\frac{22{,}35-20{,}79.z^{-1}}{\left(1-0{,}569.z^{-1}\right)^2}$$

$$S_{Fz}=\frac{\partial F}{\partial z}=-z^{-2}.\frac{b_1.\left(1+a_1.z^{-1}\right)-a_1.\left(b_0+b_1.z^{-1}\right)}{\left(1+a_1.z^{-1}\right)^2}=\frac{\left(a_1.b_0-b_1\right).z^{-2}}{\left(1+a_1.z^{-1}\right)^2}\approx\frac{8{,}07.z^{-2}}{\left(1-0{,}569.z^{-1}\right)^2}$$

1.2.1 Intérêt des sensibilités (dérivées partielles)

les sensibilités permettent de quantifier la façon dont les paramètres influent sur les caractéristiques de la fonction considérée.

Si on opère une petite variation Δb_0 du coefficient b_0 , autour de sa valeur nominale $b_0=22{,}35$, sans modifier les autres coefficients, la sensibilité S_{Fb_0} nous permet de calculer la variation ΔF correspondante de $F(z)$

- $\left[\Delta F \text{ due aux variations de } b_0\right]\approx S_{Fb_0}.\Delta b_0\approx\frac{\partial F}{\partial b_0}.\Delta b_0\approx\frac{1}{1-0{,}569z^{-1}}.\Delta_{b0}$

le même raisonnement peut être appliqué pour les autres coefficients :

- $\left[\Delta F \text{ due aux variations de } b_1\right]\approx S_{Fb_1}.\Delta b_1\approx\frac{\partial F}{\partial b_1}.\Delta b_1\approx\frac{z^{-1}}{1-0{,}569.z^{-1}}.\Delta b_1$
- $\left[\Delta F \text{ due aux variations de } a_1\right]\approx S_{Fa_1}.\Delta a_1\approx\frac{\partial F}{\partial a_1}.\Delta a_1\approx-z^{-1}.\frac{22{,}35-20{,}79.z^{-1}}{\left(1-0{,}569.z^{-1}\right)^2}\Delta a_1$

on peut ainsi en déduire les variations des caractéristiques importantes de la fonction de transfert, en fonction des variations des coefficients.

- 1- variation du gain statique , en posant $z=1$
- 2- variation du gain à $f=\frac{f_e}{2}$ en posant $z=-1$
- 3- variation du gain (complexe) pour n'importe quelle fréquence *f*, en posant $z=e^{j2\pi\frac{f}{f_e}}$

Par exemple , les variations du gain statique $G_s=F(z=1)$ s'écrivent

$\left[\Delta G_s \text{ due aux variations de } b_0\right]\approx\frac{1}{1-0{,}569}.\Delta_{b0}\approx 2{,}3\,\Delta_{b0}$

- $\left[\Delta G_s \text{ due aux variations de } b_1\right]\approx\frac{1}{1-0{,}569}.\Delta b_1\approx 2{,}3\,\Delta b_1$

$$\left[\Delta G_s \text{ due aux variations de } a_1\right] \approx -1.\frac{22{,}35-20{,}79}{\left(1-0{,}569\right)^2}\Delta a_1 \approx -8{,}4\,\Delta a_1$$

1.3 sensibilité(s) d'une fonction de plusieurs variables

Dans l'exemple que l'on vient de traiter (j'espère qu')il est clair que les coefficients b_i, a_j vont varier simultanément (lorsqu'on va les quantifier sur un nombre fini de bits). La quantité que l'on désire analyser dans ce cas est la variation totale de *F* (due à la variation simultanée de tous les coefficients), et non pas la variation individuelle de *F* (due à la variation d'un seul coefficient, lorsque tous les autres sont fixes). Cette variation totale est la somme de toutes les variations élémentaires (sous l'hypothèse que la fonction est différentiable), on a donc

$$\left[\Delta F \text{ totale}\right] \approx \underbrace{\frac{\partial F}{\partial b_0}.\Delta b_0}_{\Delta F \text{ due à } \Delta b_0} + \underbrace{\frac{\partial F}{\partial b_1}.\Delta b_1}_{\Delta F \text{ due à } \Delta b_1} \underbrace{\frac{\partial F}{\partial a_1}.\Delta a_1}_{\Delta F \text{ due à } \Delta a_1}$$

soit $x = \begin{bmatrix} b_0 \\ b_1 \\ a_1 \end{bmatrix}$ le vecteur des coefficients ,

et le vecteur $\frac{\partial F}{\partial x}$ des dérivées partielles de *F* par rapport à chacune des composantes de x : $\frac{\partial F}{\partial x} = \begin{bmatrix} \frac{\partial F}{\partial b_0} \\ \frac{\partial F}{\partial b_1} \\ \frac{\partial F}{\partial a_1} \end{bmatrix}$ (appelé **gradient de la fonction *F* par rapport au vecteur x**)

la variation totale de *F* s'écrit également (sous forme matricielle plus compacte) :

$$\left[\Delta F \text{ totale}\right] \approx \underbrace{\begin{bmatrix} \frac{\partial F}{\partial b_0} & \frac{\partial F}{\partial b_1} & \frac{\partial F}{\partial a_1} \end{bmatrix}}_{\text{transposée de } \frac{\partial F}{\partial x}} . \underbrace{\begin{bmatrix} \Delta b_0 \\ \Delta b_1 \\ \Delta a_1 \end{bmatrix}}_{\text{vecteur } \Delta x} \approx \left[\frac{\partial F}{\partial x}\right]^T . \Delta x$$

Dans le cas de variations infinitésimales , la variation Δx est notée dx , et la variation infinitésimale correspondante de F est notée dF , et appelée **différentielle de la fonction** $F(x)$ **:** $dF = \left[\frac{\partial F}{\partial x}\right]^T . dx$.

On remarquera que dans ce cas l'égalité approximative $\approx$ devient une égalité vraie =

1.4 rappel, (j'espère...) calcul de dérivées partielles

Loi de composition des dérivées partielles :

soit $f(g(x))$ une fonction dépendant de x au travers d'une fonction $g(x)$, d'un vecteur x.

La dérivée partielle $\frac{\partial f}{\partial x}$ s'écrit alors $\frac{\partial f}{\partial x} = \frac{\partial g}{\partial x}.\frac{\partial f}{\partial g}$.

dans le cas de fonctions f,g scalaires, il est plus intuitif d'écrire $\frac{\partial f}{\partial x}=\frac{\partial f}{\partial g}.\frac{\partial g}{\partial x}$

démonstration : il suffit de raisonner sur les différentielles (accroissements infinitésimaux)

$df=\left[\frac{\partial f}{\partial g}\right]^T.dg$, et $dg=\left[\frac{\partial g}{\partial x}\right]^T.dx$, donc $df=\left[\frac{\partial f}{\partial g}\right]^T.\left[\frac{\partial g}{\partial x}\right]^T.dx=\underbrace{\left[\frac{\partial g}{\partial x}.\frac{\partial f}{\partial g}\right]^T}_{\left[\frac{\partial f}{\partial x}\right]^T}.dx$

cas où $f(g(x),h(x))$ dépend de plusieurs fonctions de *x* :

$df=\left[\frac{\partial f}{\partial g}\right]^T.dg+\left[\frac{\partial f}{\partial h}\right]^T.dh$

et donc : $\frac{\partial f}{\partial x}=\frac{\partial g}{\partial x}.\frac{\partial f}{\partial g}+\frac{\partial h}{\partial x}.\frac{\partial f}{\partial h}$

dans le cas de fonctions f,g,h scalaires, il est plus intuitif d'écrire :

$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial g}.\frac{\partial g}{\partial x}+\frac{\partial f}{\partial h}.\frac{\partial h}{\partial x}$

exemples :

1- **dérivée d'un produit**

soit la fonction $f(u(x),v(x))=u(x).v(x)$, on a alors :

$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial u}.\frac{\partial u}{\partial x}+\frac{\partial f}{\partial v}.\frac{\partial v}{\partial x}=v(x).\frac{\partial u}{\partial x}+u(x).\frac{\partial v}{\partial x}$,

que vous connaissez sous la forme : $[uv]'=u'.v+u.v'$

2- **dérivée d'un rapport:** $f(u(x),v(x))=\frac{u(x)}{v(x)}=u(x).\underbrace{v(x)^{-1}}_{g(v(x))}$

$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial u}.\frac{\partial u}{\partial x}+\frac{\partial f}{\partial g}.\frac{\partial g}{\partial v}.\frac{\partial v}{\partial x}=\left[\frac{1}{v(x)}\right].\frac{\partial u}{\partial x}+u(x).\left[\frac{-1}{v^2(x)}\right].\frac{\partial v}{\partial x}$

que vous connaissez sous la forme :

$\left[\frac{u}{v}\right]'=\frac{u'}{v}-\frac{u.v'}{v^2}=\frac{u'.v-u.v'}{v^2}$

3-**comment le prof. a -t-il calculé en 2 temps 3 mouvements la quantité** :

$\frac{\partial F}{\partial a_1}=-z^{-1}.\frac{b_0+b_1.z^{-1}}{\left(1+a_1.z^{-1}\right)^2}$, avec $F=\frac{b_0+b_1.z^{-1}}{\left(1+a_1.z^{-1}\right)}$

réponse :

1-**il a pensé :** $F=N.\frac{1}{h(a1)}=N.g(h(a1))$,

avec $N=b_0+b_1.z^{-1}$, $g(h)=\frac{1}{h}$, et $h=(1+a_1.z^{-1})$

2- **et donc il en a déduit**

$\frac{\partial F}{\partial a_1}=N.\frac{\partial g}{\partial h}.\frac{\partial h}{\partial a_1}=N.-\frac{1}{h^2}.z^{-1}=(b_0+b_1.z^{-1}).-\frac{1}{\left[1+a_1.z^{-1}\right]^2}.z^{-1}$

1.5 variation maximale de ΔF , due à la quantification des coefficients

1.5.1 cas où chaque coefficient est quantifié à sa propre échelle

En supposant que chaque coefficient $x_i=b_j,a_j$ est quantifié à sa propre échelle, sur N bits, l'erreur $\Delta x_i=x_{iq}-x_i$ maximale que l'on commettra en quantifiant x_i

vérifie :

$|\Delta x_i| = \rho . |x_i|$ avec $\rho = 2^{-[N-1]}$

En étant pessimiste, la pire erreur que l'on peut commettre sur

$[\Delta F \text{ totale}] \approx \sum_i \frac{\partial F}{\partial x_i} . \Delta x_i$ vérifie

$$|\Delta F \text{ totale}| \approx \sum_i \left|\frac{\partial F}{\partial x_i}\right| . |\Delta x_i| \leq \rho . \sum_i \left|\frac{\partial F}{\partial x_i}\right| . |x_i| = \rho . \underbrace{\left|\frac{\partial F}{\partial x}\right|^T . |x|}_{\text{abus de langage}}$$

Par exemple, la pire erreur que l'on peut commettre sur le gain statique de $G_s = F(z=1)$, si on quantifie séparément les coefficients sur $N = 16$ bits, vérifie :

$$|\Delta G_s| \approx 2^{-15} . \left[\underbrace{\left|\frac{\partial G_s}{\partial b_0}\right|}_{2,3} , \underbrace{\left|\frac{\partial G_s}{\partial b_1}\right|}_{2,3} , \underbrace{\left|\frac{\partial G_s}{\partial a_1}\right|}_{8,4} \right] . \begin{bmatrix} \underbrace{|b_0|}_{22,35} \\ , \underbrace{|b_1|}_{20,79} \\ , \underbrace{|a_1|}_{0,569} \end{bmatrix} \approx 2^{-15} . 104 \approx 0,003$$

Dans le cas de caractéristiques fréquentielles, on préfère raisonner en valeurs relatives =>

$$\left|\frac{\Delta G_s}{G_s}\right| \approx 2^{-15} . \frac{104}{3,62} \approx 2^{-15} . 28,7 \approx 8,8\,10^{-4}$$

1.5.2 cas où tous les coefficients sont quantifiés à la même échelle

dans ce cas, l'erreur maximale sur chaque coefficient est la même, et vérifie

$|\Delta x_i| = \rho . \max_j |x_j|$ avec $\rho = 2^{-[N-1]}$

En étant pessimiste, la pire erreur que l'on peut commettre sur

$[\Delta F \text{ totale}] \approx \sum_i \frac{\partial F}{\partial x_i} . \Delta x_i$ devient alors :

$|\Delta F \text{ totale}| \approx \sum_i \left|\frac{\partial F}{\partial x_i}\right| . |\Delta x_i| \leq \rho . \max_j |x_j| . \sum_i \left|\frac{\partial F}{\partial x_i}\right|$, avec $\rho = 2^{-[N-1]}$

Par exemple, la pire erreur que l'on peut commettre sur le gain statique de $G_s = F(z=1)$, lorsque les coefficients sont quantifiés sur $N = 16$ bits avec le même facteur d'échelle ,vérifie :

$$|\Delta G_s| \approx 2^{-15} . \underbrace{|b_0|}_{22,35} . \left[\underbrace{\left|\frac{\partial G_s}{\partial b_0}\right|}_{2,3} + \underbrace{\left|\frac{\partial G_s}{\partial b_1}\right|}_{2,3} + \underbrace{\left|\frac{\partial G_s}{\partial a_1}\right|}_{8,4} \right] . \approx 2^{-15} . 291 \approx 8,8\,10^{-3}$$

2 sensibilité des pôles et des zéros

l'analyse précédente permet d'analyser la variation de la réponse fréquentielle induite par la quantification des coefficients. Il est également possible d'analyser la variation des pôles et des zéros de la fonction de transfert, comme présenté dans cette partie

2.1 variation des zéros d'un polynôme, due aux variations de coefficients

soit un polynôme

$P(z)=\sum_{k=0}^{N} p_k.z^k=p_N.\prod_{j=1}^{N}(z-z_j)$, (où les z_j sont les zéros du polynôme, et les p_k sont les coefficients du polynôme).

on cherche la dérivée partielle du $j^{ème}$ zéro z_j par rapport au $k^{ème}$ coefficient p_k , soit encore
$\frac{\partial z_j}{\partial p_k}$.

Pour cela on opère de petites variations $(\delta z, \delta p_k)$ de (z, p_k) , et on écrit le développement en série de Taylor correspondant en $z=z_j$:

[1]- $P(z+\delta z, p_k+\delta p_k)\Big|_{z=z_j} \approx P(z, p_k)\Big|_{z=z_j} + \left.\frac{\partial P(z, p_k)}{\partial z}\right|_{z=z_j}.\delta z + \left.\frac{\partial P(z, p_k)}{\partial p_k}\right|_{z=z_j}.\delta p_k$

or z_j est une racine du polynôme, on a donc : $P(z, p_k)\Big|_{z=z_j}=0$

Pour une variation infiniment petite dp_k du coefficient p_k du polynôme *P*, on peut choisir la variation dz de telle façon que le nouveau polynôme s'annule en $z=z_j+dz$.
Dans ce cas, la quantité dz représente la variation dz_j de la racine z_j du polynôme.
De plus, ce choix particulier de dz correspond au cas pour lequel :
$P(z+dz_j, p_k+dp_k)\Big|_{z=z_j}=0$ (car z_j+dz est une racine du polynôme de coefficient p_k+dp_k)
L'équation [1] se réécrit donc, pour ce choix particulier dz_j de dz :

$$0=0+\left.\frac{\partial P(z, p_k)}{\partial z}\right|_{z=z_j}.dz_j+\left.\frac{\partial P(z, p_k)}{\partial p_k}\right|_{z=z_j}.dp_k$$

On peut donc en déduire l'expression de la variation infinitésimale(ou différentielle) dz_j de la $j^{ème}$ racine, en fonction de la variation infinitésimale(ou différentielle) dp_k du $k^{ème}$ coefficient :

$$dz_j=-\frac{\left.\frac{\partial P(z, p_k)}{\partial p_k}\right|_{z=z_j}}{\left.\frac{\partial P(z, p_k)}{\partial z}\right|_{z=z_j}}.dp_k=\frac{\partial z_j}{\partial p_k}.dp_k$$

la **dérivée partielle de la $j^{ème}$ racine par rapport au $k^{ème}$ coefficient** s'écrit donc :

$$\frac{\partial z_j}{\partial p_k}\approx-\frac{\left.\frac{\partial P(z)}{\partial p_k}\right|_{z=z_j}}{\left.\frac{\partial P(z)}{\partial z}\right|_{z=z_j}}$$

avec le même raisonnement, on peut montrer que dans le cas où tous les coefficients varient, la différentielle de la $j^{ème}$ racine s'écrit :

$$dz_j=\sum_{k=0}^{N}\frac{\partial z_j}{\partial p_k}.dp_k=-\sum_{k=0}^{N}\frac{\left.\frac{\partial P(z,\,p_k)}{\partial p_k}\right|_{z=z_j}}{\left.\frac{\partial P(z,\,p_k)}{\partial z}\right|_{z=z_j}}.dp_k$$

.

ATTENTION : cette formule ne fonctionne pas pour des racines multiples!...

1-calculatoirement, elle conduit dans ce cas à une forme du type $\frac{\partial z_j}{\partial p_k}=\pm\infty$.

2- fondamentalement, cela traduit le fait qu'une seule racine multiple se transforme alors en plusieurs racines simples (dz_j n'est pas unique=> la fonction $z_j(p_k)$ n'est pas différentiable)

2.2 implications pour le codage des filtres iir

considérons un polynôme $P(z)$ représentant le dénominateur en z d'une fonction de transfert $F(z)$. $P(z)$ possède alors les 2 propriétés suivantes

_1- son coefficient de plus haut degré p_N est normalisé à 1

_2- ses zéros z_n sont tous de module inférieur strictement à 1 { sinon $F(z)$ est instable }

_On peut donc l'écrire sous la forme suivante :

$$P(z)=\sum_{k=0}^{N}p_k.z^k=\prod_{n=1}^{N}(z-z_n),\left|z_n\right|<1\,,\,p_N=1$$

_la sensibilité du $j^{\text{ème}}$ zéro z_j par rapport au $k^{\text{ème}}$ coefficient p_k s'écrit :

$$\frac{\partial z_j}{\partial p_k}\approx-\frac{\left.\frac{\partial P(z)}{\partial p_k}\right|_{z=z_j}}{\left.\frac{\partial P(z)}{\partial z}\right|_{z=z_j}}$$

_- la quantité $\left.\frac{\partial P(z)}{\partial p_k}\right|_{z=z_j}$ s'évalue facilement à partir de l'expression $P(z)=\sum_{k=0}^{N}p_k.z^k$:-

1- $$\left.\frac{\partial P(z)}{\partial p_k}\right|_{z=z_j}=z^k\Big|_{z=z_j}=z_j^k$$

_- de façon à mettre en évidence les problèmes, on évalue la quantité $\left.\frac{\partial P(z)}{\partial z}\right|_{z=z_j}$ à partir de l'expression factorisée du polynôme :

$$P(z)=\prod_{n=1}^{N}(z-z_n)=\underbrace{(z-z_j)}_{u(z)}.\underbrace{\prod_{n\neq j}(z-z_n)}_{v(z)}=u(z).v(z)$$

_on en déduit : $$\left.\frac{\partial P(z)}{\partial z}\right|_{z=z_j}=\underbrace{\frac{\partial u(z)}{\partial z}}_{=1}.v(z)\Bigg|_{z=z_j}+\underbrace{u(z)}_{=z-z_j}.\left.\frac{\partial v(z)}{\partial z}\right|_{z=z_j}$$

_et donc , en $z=z_j$, on obtient :

_2- $$\left.\frac{\partial P(z)}{\partial z}\right|_{z=z_j}=1.v(z)+0.\frac{\partial v(z)}{\partial z}=v(z_j)=\prod_{n\neq j}(z_j-z_n)$$

on en déduit finalement l'expression du module de la sensibilité de z_j par rapport à p_k

$$\left|\frac{\partial z_j}{\partial p_k}\right| = \left| -\frac{\frac{\partial P(z)}{\partial p_k}}{\frac{\partial P(z)}{\partial z}} \right|_{z=z_j} = \frac{|z_j|^k}{\prod_{n\neq j} |z_j - z_n|}$$

conséquence 1 : comme $|z_j|<1$, la plus grande sensibilité est atteinte pour $k=0$, qui correspond à la sensibilité par rapport au coefficient de plus bas degré p_0 .

conséquence 2 : la sensibilité peut devenir gigantesque pour des filtres d'ordre élevé, dont les pôles ou zéros (les zéros de $D(z), N(z)$) sont très proches.

2.3 pôles en w et en z, pseudo-pulsation propre v_n et facteur d'amortissement ξ

Pour expliquer l'implication de la conséquence 2, on va étudier le lien entre les pôles dans le plan *W*, et les pôles correspondants en *z*. Ce paragraphe a pour objet de montrer que des pôles en *Z* correspondant

1-à des fréquences propres petites devant la fréquence d'échantillonnage $f_c \ll f_e$

2- à des facteurs d'amortissement ξ faibles

sont forcément très proches du point -1

2.4 pôles-zéros en w, en fonction de v_n et ξ

soient deux pôles complexes conjugués dans le plan w : $p_w = r_w \pm i.i_w$.

le polynôme associé à ces pôles s'écrit, sous forme normalisée :

$$D(w) = [1 - \frac{w}{r_w - i.i_w}].[1 - \frac{w}{r_w + i.i_w}] = 1 + 2\xi\left[\frac{w}{v_n}\right] + \left[\frac{w}{v_n}\right]^2$$

avec

1- $v_n = \sqrt{r_w^2 + i_w^2} = |p_w|$, la pseudo-pulsation propre non amortie associée aux pôles

2- $\xi = \frac{-r_w}{v_n} = \frac{-r_w}{|p_w|} = -\cos(Arg(p_w))$, le facteur d'amortissement associé au pôle

---

p_w peut s'écrire directement en fonction de v_n et ξ sous la forme :

$$p_w = \underbrace{-\xi.v_n}_{r_w} \pm i \underbrace{\sqrt{1-\xi^2}.v_n}_{i_w}$$

- Le lieu des pôles en W pour $v_n = cte$ est un cercle de rayon v_n et de centre 0.

- Le lieu des pôles en W pour $\xi = cte$ est une demi-droite de pente $\pm\frac{\sqrt{1-\xi^2}}{\xi}$

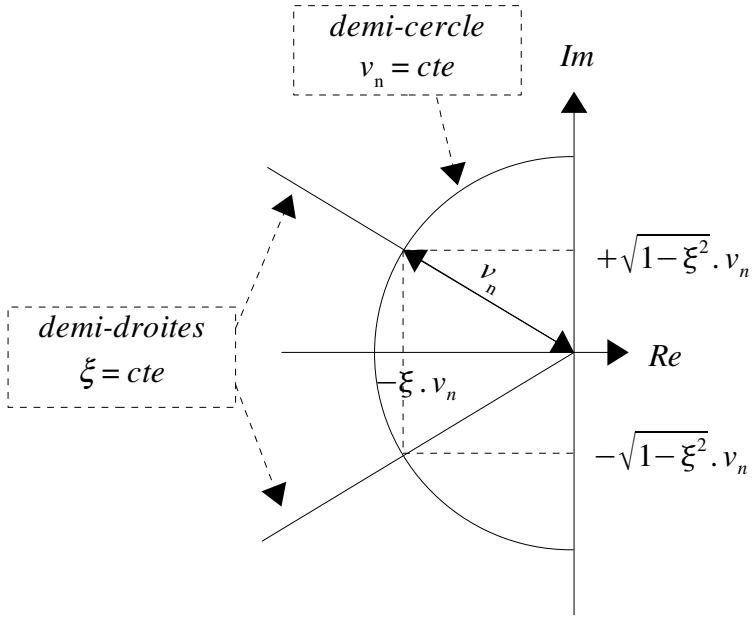


figure 1: lien entre pôle p_w , pulsation propre v_n et facteur d'amortissement ξ

La réponse fréquentielle associée à *D*(*w*) s'écrit :

$$D(iv)=1+2\xi\left[\frac{iv}{v_n}\right]+\left[\frac{iv}{v_n}\right]^2 = \underbrace{1-\left[\frac{v}{v_n}\right]^2}_{\text{partie réelle de } D(iv)} + i \underbrace{2\xi\left[\frac{v}{v_n}\right]}_{\text{partie imaginaire de } D(iv)}$$

Elle présente les caractéristiques suivantes :

- 1- gain statique *D*(*0*) = 1 (0 décibels)
- 2- gain à v_n $\left|D\left(iv_n\right)\right|=2\xi$ (*argument* = 90°)
- 3- pour $\xi<\frac{1}{\sqrt{2}}\approx 0{,}7$, gain minimum $\frac{1}{Q}=2\xi.\sqrt{1-\xi^2}$ obtenu à $v_r=\sqrt{1-2\xi^2}v_n$.
- *Q* est appelé *facteur de résonance associé du pôle*
- v_r est la pulsation de résonance, et est proche de v_n pour des grandes valeurs de Q

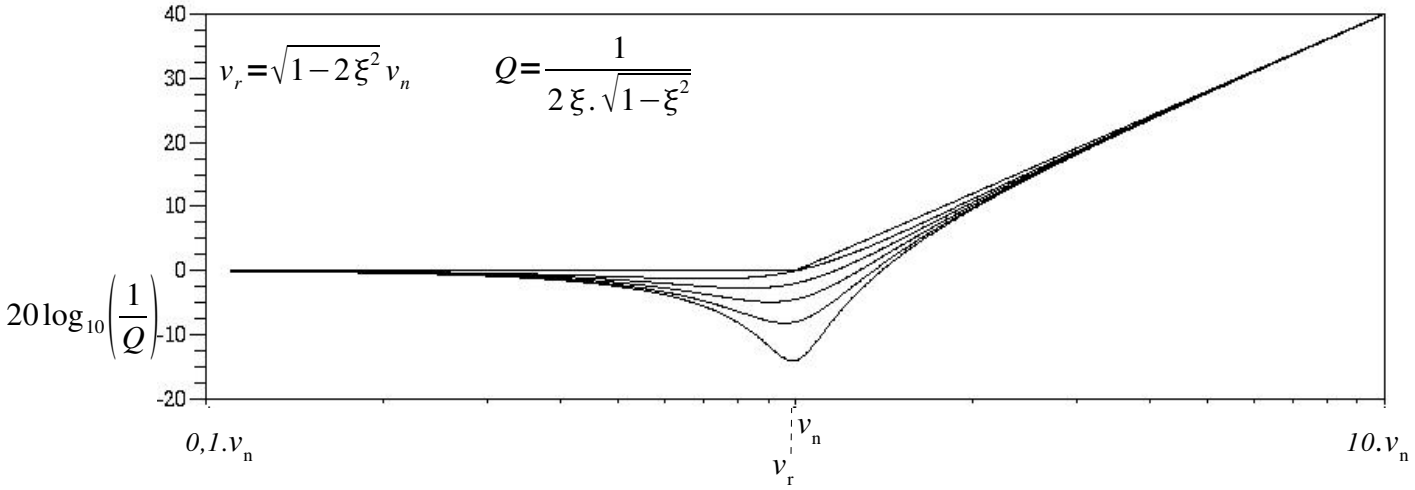


figure 2: diagrammes de bode de D(jv) pour $\xi\in\{0.1, 0.2, 0.3, 0.4, 0.5\}$

2.4.1 pôles-zéros en z, en fonction de v_n et ξ

en appliquant la transformation bilinéaire $z=\frac{1+w}{1-w}$ au pôle p_w , on obtient le pôle correspondant en z : $p_z=\frac{1+p_w}{1-p_w}$

Il est connu que les transformations bilinéaires (transformations de Möbius), du type : $p_w \to p_z=\frac{a.p_w+b}{c.p_w+d}$ *ont pour propriété de transformer les droites et les cercles du plan complexe, en d'autres droites ou cercles. Le calcul complet étant assez pénible, seuls les résultats sont résumés ici,uniquement dans le cas de la transformation bilinéaire :*

- le l**ieu des points** p_z **lorsque** $v_n=cte$ *(image par la transformation bilinéaire du cercle de rayon* $v_n=cte$ *et de centre 0 dans le plan w)* est un cercle ,

  de centre $z_0=\frac{1+v_n^2}{1-v_n^2}$, et de rayon $R=\left|\frac{2.v_n}{1-v_n^2}\right|$

- lorsque $v_n=1$, ce cercle devient une ligne droite verticale d'abscisse 0...
- pour $v_n<1$ l'intérieur du cercle correspond aux points p_z tels que $|v_n|<cte$

- le l**ieu des points** p_z **lorsque** $\xi=cte$ *(image par la transformation bilinéaire de la droite de pente* $\frac{\pm\xi}{\sqrt{1-\xi^2}}$ *passant par 0 dans le planW) est un cercle*

  de centre $z_0=\pm i.\frac{\xi}{\sqrt{1-\xi^2}}$, et de rayon $R=\frac{1}{\sqrt{1-\xi^2}}$

- pour $\xi<1$ l'extérieur du cercle correspond aux points p_z tels que $|\xi|<cte$
- pour $\xi=1$ ce lieu correspond à l'axe horizontal

2.4.2 conséquences quand à la sensibilité des pôles par rapport aux coefficients

1- les pseudos-pulsations v_n sont directement liées au fréquences réelles f_n par :

- $v_n = \tan\left(\pi . \frac{f_n}{f_e}\right)$, f_e étant la fréquence d'échantillonnage.

2- de plus le lieu des points p_z vérifiant simultanément
1- $v_n < v_{n0}$, 2- $\xi < \xi_0$, et 3- $|p_z| < 1$ (système stable)
correspond à l'intersection des 3 lieux correspondants :

——— 1- intérieur du cercle de centre $z_0 = \frac{1+v_{n0}^2}{1-v_{n0}^2}$, et de rayon $R = \left|\frac{2.v_{n0}}{1-v_{n0}^2}\right|$

——— 2- extérieur du cercle de centre $z_0 = \frac{\pm\xi_0}{\sqrt{1-\xi_0^2}}$, et de rayon $R = \frac{1}{\sqrt{1-\xi_0^2}}$

——— 3- intérieur du cercle de centre $z_0 = 0$, et de rayon $R = 1$

Ce lieu devient extrêmement petit lorsque

——— 1- les dynamiques à coder sont lentes : $\frac{f_{n0}}{f_e} \ll 1 \Leftrightarrow v_{n0} = \tan\left(\pi . \frac{f_0}{f_e}\right) \ll 1$

——— 2- les facteurs d'amortissement sont faibles $\xi \ll 1$

par exemple on a représenté figure 3 le lieu des pôles dans le plan z , correspondant simultanément à

1- une fréquence de coupure $\frac{f_n}{f_e} < 0{,}05 \Leftrightarrow v_{n0} < 0{,}16$

2- un facteur d'amortissement $\xi < 0{,}3$

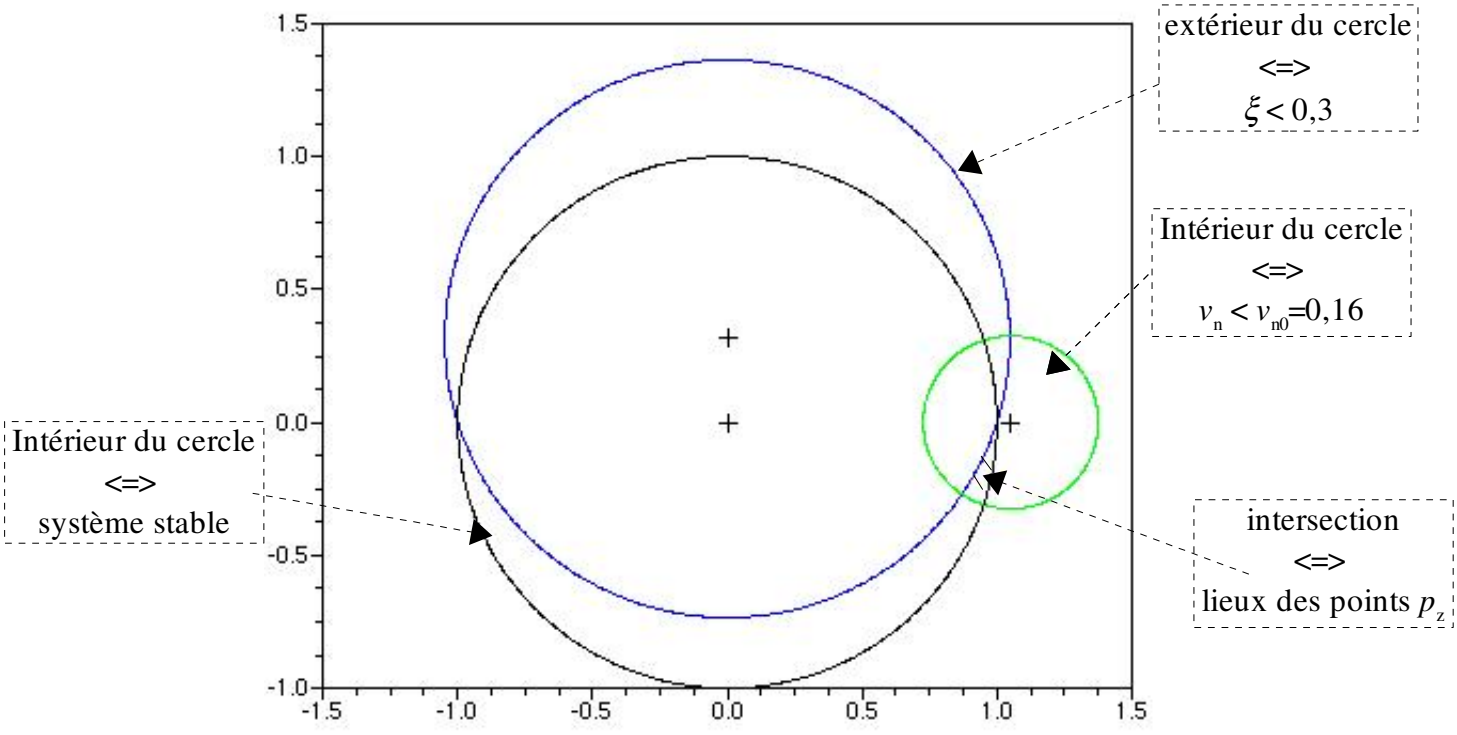


figure 3: lieu des pôles pour ξ <0,3 , v_n <0,16

en conclusion: *les pôles z_n d'un filtre*
——— 1-dont les fréquences de coupure sont petites
——— 2-et-ou les facteurs d'amortissement sont petits
sont très proches du point 1, et très proches les uns des autres...

Dans ce cas, la sensibilité des pôles par rapport au plus petit coefficient $\left|\frac{\partial z_j}{\partial p_0}\right| = \frac{1}{\prod_{n \neq j} |z_j - z_n|}$ *devient gigantesque (pôles très proches).*

Cela implique un très grand nombre de bits de codage pour rendre compte avec suffisamment de précision de la dynamique (des pôles-zéros) du filtre...

2.4.3 exemple : fitre de Butterworth passe-bas d'ordre N

Soit un filtre passe-bas d'ordre N, de type : Butterworth , de fréquence de coupure à 3db f_n

La figure 4 représente, en fonction de f_n *et N, la plus grande sensibilité des pôles en z, par rapport aux coefficients* d_i *du dénominateur en z:*

$$S = max\left|\frac{\partial p_{zk}}{\partial d_0}\right| = \frac{1}{\prod_{n \neq k} |p_{zk} - p_{zn}|} \quad .$$

cette figure met en évidence les 2 phénomènes suivants

1- la sensibilité augmente très vite en fonction de l'ordre N
2- la sensibilité augmente très vite au fur et à mesure que la fréquence diminue

1- Les filtres d'ordre N élevé, et de fréquence de coupure f_n ***petite devant la fréquence d'échantillonnage*** f_e ***demandent un nombre de bits de codage très grand, si on les programment directement!...***
2- Cette constatation reste valable pour les filtres programmés avec des nombres réels en double précision (dont les coefficients ont une précision relative de $\approx \frac{10^{-16}}{1}$ ***)...***

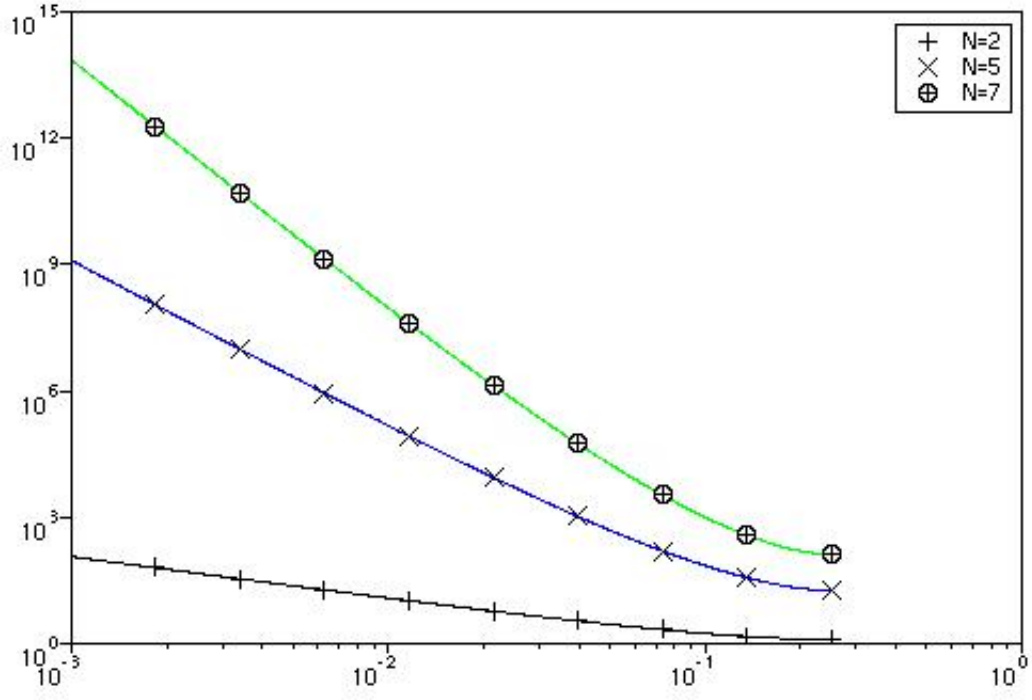


figure 4: sensibilité des pôles en fonction de l'ordre N, et de $\frac{f_n}{f_e}$

2.4.4 règles de codage pour améliorer les sensibilités

Pour les raisons qui viennent d'être présentées ,
Règle 1- les filtres IIR $F(z)$ sont toujours codés sous la forme
1- soit d'un produit (décomposition cascade, factorisation)
2- soit d'une somme (décomposition parallèle, éléments simples)
de filtres $F_i(z)$ d'ordre 1 (pôles réels) ou 2 (pôles complexes conjugués).
Les filtres $F_i(z)$ sont appelés 'cellules élémentaires'.
Cette technique permet d'améliorer grandement la sensibilité des pôles-zéros (et donc de la réponse fréquentielle), vis à vis des coefficients, lorsque l'ordre *N* du filtre est élevé...
Règle 2- On ne doit pas sur-échantillonner un filtre numérique:
1-les fréquences de coupure f_n correspondent aux fréquences réelles, ou encore à la dynamique réelle, désirée pour le filtre.

2- La fréquence d'échantillonnage f_e fixe la bande totale de fréquences $\left[0, f_{maxi}=\frac{f_e}{2}\right]$ traitées par le filtre.

3- Augmenter inutilement la fréquence f_e diminue les rapports $\frac{f_n}{f_e}$ et les pseudo-pulsations associées $v_n=\tan\left(\pi\frac{f_n}{f_e}\right)$, ce qui induit des conséquences néfastes quant à la sensibilité des pôles (et donc de la réponse fréquentielle) vis à vis des coefficients...
=> On devra donc choisir f_e aussi petite que possible, pour une fréquence f_{maxi} donnée.

Remarque :
Ces règles sont à l'opposé de celles appliquées par les personnes ''*qui ne veulent pas s'embêter avec*'' les filtres numériques

1- elles codent directement l'équation récurrente d'ordre élevé (obtenue par approximation discrète d'une fonction de transfert $F(p)$, *delta* ou *tustin*)

2- pour que leur approximation soit correcte, elles choisissent $f_e \gg f_n$

3- ensuite elles se demandent pourquoi le filtre numérique ne fonctionne pas comme prévu(le plus souvent il est instable, certains pôles se sont déplacés à l'extérieur du cercle unité dans le plan *Z*)...

3 Analyse de l'effet de la quantification des coefficients dans le plan W

L'analyse des sensibilités des caractéristiques essentielles d'un filtre vis à vis des coefficients permet d'acquérir l'intuition pour améliorer le comportement vis à vis d'un codage 'naïf'. Toutefois elle apporte peu d'intuition quant aux méthodes qui permettraient de 'vraiment désensibiliser' ces caractéristiques...

En effet, comme le montre la figure 4, même pour des filtres d'ordre 2, les caractéristiques se dégradent rapidement en fonction du rapport $\frac{f_n}{f_e}$.

L'analyse qui suit a pour but de fournir au lecteur l'intuition nécessaire pour motiver les techniques de type 'changement d'opérateur', qui vont rendre les sensibilités indépendantes du rapport $\frac{f_n}{f_e}$.

3.1 réponse fréquentielle d'un polynôme en z^{-1}, analyse en w

soit $F(z)=\frac{N(z)}{D(z)}=\frac{n_0+n_1.z^{-1}+n_2.z^{-2}}{1+d_1.z^{-1}+d_2.z^{-2}}$ un filtre numérique (supposé d'ordre 2, les ordres supérieurs étant factorisés ou décomposés en éléments simples)

Soient $N(w), D(w), F(w)=\frac{N(w)}{D(w)}$ les transformées en w respectives de $N(z), D(z), F(z)$.

la première constatation surprenante est que $N(w), D(w)$ ne sont pas des polynômes, mais des fractions rationnelles de la variable w, dont tous les pôles sont situés en $w=-1$.

$$N(w)=n_0+n_1.\frac{1-w}{1+w}+n_2.\left[\frac{1-w}{1+w}\right]^2=\frac{\overbrace{\left[n_2+n_1+n_0\right]}^{n_{0w}}+\overbrace{\left[2.n_2-2.n_0\right]}^{n_{1w}}.w+\overbrace{\left[n_2-n_1+n_0\right]}^{n_{2w}}.w^2}{\left[1+w\right]^2}$$

$$D(w)=\frac{\overbrace{\left[d_2+d_1+1\right]}^{d_{0w}}+\overbrace{\left[2.d_2-2\right]}^{d_{1w}}.w+\overbrace{\left[d_2-1+d_0\right]}^{d_{2w}}.w^2}{\left[1+w\right]^2} \quad ,$$

Remarque: le fait que le coefficient d_0 de $D(z)$ soit égal à *1*, se traduit dans le plan *W* par la condition de normalisation : $D(w)\big|_{en\,w=1}=1$

la fonction de transfert $F(w)=\frac{N(w)}{D(w)}$ est réalisée comme le rapport des 2 fractions rationnelles $N(w)=\frac{n_{0w}+n_{1w}.w+n_{2w}.w^2}{\left[1+w\right]^2}$, $D(w)=\frac{d_{0w}+d_{1w}.w+d_{2w}.w^2}{\left[1+w\right]^2}$, dont les

dénominateurs se simplifient.

Faisons apparaître explicitement le gain statique *K*, et les pôles -zéros de $F(w)$, en écrivant les pseudo-pulsations propres v_n, v_d et facteurs d'amortissement ξ_n, ξ_d associés :

$$F(w)=K.\frac{1+2.\xi_n.\frac{w}{v_n}+\left[\frac{w}{v_n}\right]^2}{1+2.\xi_d.\frac{w}{v_d}+\left[\frac{w}{v_d}\right]^2} \quad .$$

Les fractions rationnelles $N(w), D(w)$ s'écrivent alors forcément sous la forme :

$$N(w)=A.K.\frac{1+2.\xi_n.\frac{w}{v_n}+\left[\frac{w}{v_n}\right]^2}{\left[1+w\right]^2} \quad \text{, et} \quad D(w)=A.\frac{1+2.\xi_d.\frac{w}{v_d}+\left[\frac{w}{v_d}\right]^2}{\left[1+w\right]^2} \quad ,$$

où la constante A est telle que l'on respecte la condition de normalisation $D(w)\big|_{en\,w=1}=1$, ce qui donne : $A=\frac{4}{1+2\frac{\xi_d}{v_d}+\frac{1}{v_d^2}}$

cas général

soit $P(z)=p_0+p_1.z^{-1}+p_{2.}z^{-2}$ un polynôme en z^{-1} ,

et $P(w)=\frac{p_{0w}+p_{1w}.w+p_{2w}.w^2}{\left[1+w\right]^2}$ sa transformée en W.

lorsque $P(w)$ a deux zéros complexes conjugués, de pulsation propre v_p , et de facteur d'amortissement ξ_p , alors $P(w)$ peut toujours s'écrire sous la forme :

$$P(w)=A.\frac{1+2.\xi.\frac{w}{v_p}+\left[\frac{w}{v_p}\right]^2}{\left[1+w\right]^2} \quad \text{, avec} \quad A=\frac{4}{1+2\frac{\xi}{v_p}+\frac{1}{v_p^2}}.p_0 \quad .$$

lorsque $v_p<1$ la réponse fréquentielle $P(j.v)$ est celle d'un ***filtre passe-haut***,

de gain statique $P(j.o)=A$

de gain maximum obtenu en haute fréquence $P(j.\infty)=\frac{1}{v_p^2}.A$

avec une anti - résonance lorsque $\xi<\sqrt{\frac{1+v_p^2}{2}}$,

à la pulsation $v_r=v_p.\sqrt{\frac{1-2\xi^2+v_p^2}{\left[1-2\xi^2\right].v_p^2+1}}$,

de valeur $\frac{A}{Q}=2.\sqrt{\frac{1-\xi^2}{\left[1+v_p^2\right]^2-\left[2\xi v_p\right]^2}}.A\approx\frac{2}{1+v_p^2}.\xi.A,\ \text{lorsque}\ \xi\ll1$

on peut également évaluer le rapport entre le gain maximum (obtenu lorsque $v\to\infty$) et le gain minimum (obtenu à $v=v_r$, ou à $v=0$) :

1- Pour de petites valeurs de ξ , ce rapport s'écrit approximativement :

$$\frac{max|P(jv)|=|P(j\infty)|}{min|P(jv)|=|P(jv_r)|}\approx\frac{1+v_p^2}{2.\xi.v_p^2}$$

2- Lorsqu'il n'y a pas d' anti-résonance: $\xi\geq\sqrt{\frac{1+v_p^2}{2}}$ ce rapport s'écrit exactement

$$\frac{|P(j\infty)|}{min|P(jv)|=|P(j0)|}=\frac{1}{v_p^2}$$

On voit donc que ce rapport croît proportionnellement à $\frac{1}{v_p^2}$ et à $\frac{1}{\xi}$

La figure 5 met en évidence ce phénomène, au travers de quelques réponses fréquentielles typiques

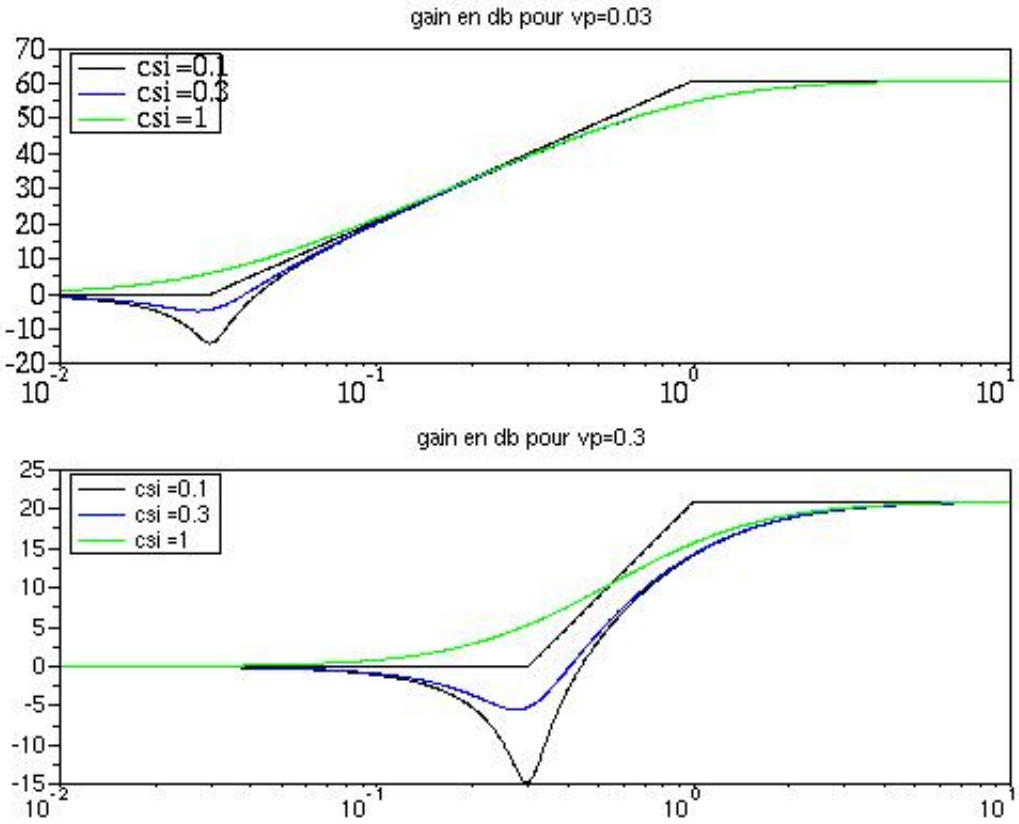


figure 5: réponses fréquentielles typiques de $P(w)=\frac{1+2\xi.\frac{w}{v_p}+\left[\frac{w}{v_p}\right]^2}{[1+w]^2}$

3.2 conséquence quant à la quantification des coefficients

soient

- $P(z)=p_0+p_1.z^{-1}+p_2.z^{-2}$ un polynôme du 2ème ordre en z^{-1}
- $P_q(z)=p_{0q}+p_{1q}.z^{-1}+p_{2q}.z^{-2}$ le même polynôme dont les coefficients ont été quantifiés sur N_B bits.
- $\Delta P(z)=P(z)-P_q(z)=\overbrace{[p_0-p_{0q}]}^{\Delta p_0}+\overbrace{[p_1-p_{1q}]}^{\Delta p_1}.z^{-1}+\overbrace{[p_2-p_{2q}]}^{\Delta p_2}.z^{-2}$ la différence entre *P* et P_q

La réponse fréquentielle de ΔP s'écrit :

- $\Delta P(e^{j2\pi\frac{f}{f_e}}) = \Delta p_0 + \Delta p_1 . e^{-j2\pi\frac{f}{f_e}} + \Delta p_2 . e^{-j4\pi\frac{f}{f_e}}$,

on peut donc majorer son module par :

- $\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| \le \left|\Delta p_0\right| + \left|\Delta p_1\right| + \left|\Delta p_2\right|$

de plus l'erreur relative maximale entre les coefficients quantifiés et les coefficients réels est inférieure à $\rho = max\left|\frac{\Delta p_i}{p_i}\right| = 2^{-[N_B - 1]}$. on aura donc : $\left|\Delta p_i\right| \le \rho . \left|p_i\right|$

On peut donc finalement en déduire

$$\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| \le \rho . \left[\left|p_0\right| + \left|p_1\right| + \left|p_2\right|\right]$$

ce qui donne une borne maximale pessimiste sur l'erreur commise sur la réponse fréquentielle, avec une quantification sur N_B bits des coefficients :

$$max\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| = \rho . \left[\left|p_0\right| + \left|p_1\right| + \left|p_2\right|\right] \text{ , avec } \rho = 2^{-[N_B - 1]}$$

{ noter que cette borne a été établie pour un polynôme d'ordre 2, mais est valable quel que soit l'ordre *N* du polynôme } :

$$max\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| = \rho . \sum_{i=0}^{N} \left|p_i\right| = \rho . \left\|P(z)\right\|_1$$

pour un polynôme de degré inférieur ou égal à 2, de pseudo-pulsation de coupure $v_p < 1$, le coefficient p_1 est de signe opposé à celui de p_0, p_2 . On a alors :

$$\rho . \left[\left|p_0\right| + \left|p_1\right| + \left|p_2\right|\right] = \rho . \left|p_0 - p_1 + p_2\right| = \rho . \left|P(z)\right|_{\text{en } z=-1}$$

or $P(z)\big|_{\text{en } z=-1} = P(w)\big|_{\text{en } w\to\infty}$, et donc :

$$max\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| = \rho . \left|P(w)\right|_{\text{en } w\to\infty}\Big|$$

Or la valeur maximale de $|P(jv)|$ est obtenue dans ce cas lorsque $v \to \infty$, on en déduit donc finalement la borne maximale sur l'erreur, directement depuis le maximum de la réponse fréquentielle :

$$max\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| = \rho . \left|P(w)\right|_{\text{en } w\to\infty}\Big| = \rho . max\left|P(j.v)\right| \text{ , avec } \rho = 2^{-[N_B - 1]}$$

Dans le cas général(*P*(*z*) de degré supérieur à 2, ou $v_p < 1$), on pourrait juste écrire

$$max\left|\Delta P(e^{j2\pi\frac{f}{f_e}})\right| \ge \rho . max\left|P(j.v)\right|$$

Cette nouvelle est très inquiétante :l'erreur maximale est reportée à n'importe quelle fréquence. Il n'est pas évident du tout que l'erreur relative sur la réponse fréquentielle reste correcte, même avec un nombre important de bits de codage :

La pire erreur relative que l'on peut commettre sur la réponse fréquentielle s'écrit en effet :

$$max\left|\frac{\Delta P}{P}\right| = \frac{max\left|\Delta P\right|}{min\left|P\right|} \underbrace{=}_{\geq \text{ dans le cas général}} \frac{\rho . max\left|P(j.v)\right|}{min\left|P(j.v)\right|} \text{ , avec } \rho = 2^{-[N_B - 1]}$$

or { voir figure 5 }, la quantité $\frac{max|P(j.v)|}{min|P(j.v)|} \approx \frac{1+v_p^2}{2\xi v_p^2}$ peut être très grande devant 1

et donc même avec $\rho = 2^{-[N_B-1]}$ petit devant 1, il n'est pas du tout certain que la réponse fréquentielle du polynôme soit correctement codée...

La figure 6 met en évidence ce phénomène, lors de la quantification sur 10 bits du polynôme $H(z^{-1})$ correspondant à $v_p \approx 0.03, \xi \approx 0.1$:

- l'erreur relative maximale sur les coefficients est $\rho = 2^{-[10-1]} \approx 0.002 \approx -54\text{dB}$
- le maximum de la réponse fréquentielle est de l'ordre de 60 dB
- l'erreur absolue maximale sur la réponse fréquentielle est de l'ordre de 60-54=6dB

La réponse fréquentielle du polynôme quantifié $H_q(z^{-1})$ diffère sensiblement de la réponse fréquentielle idéale dans la zone de fréquences pour laquelle l'erreur absolue est non-négligeable devant la réponse fréquentielle de *H*.

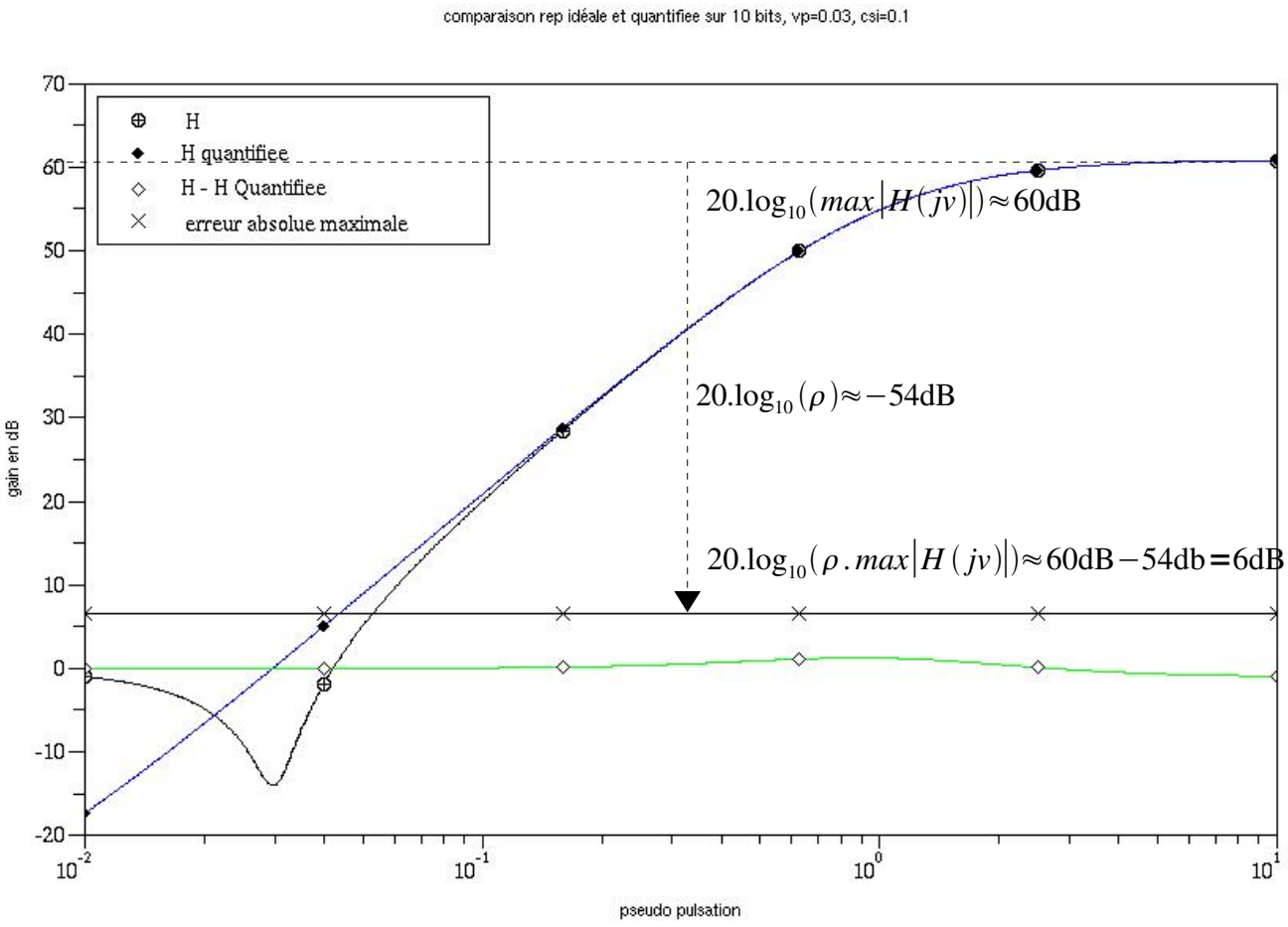


figure 6: mise en évidence des effets de la quantification sur la réponse fréquentielle

Lorsqu'on quantifie des polynômes en z^{-1} , la précision relative des coefficients $\rho = 2^{-[N_B-1]}$ est mal utilisée, parce-que la variation relative du gain de la réponse fréquentielle est également très élevée.

Par exemple pour la réponse fréquentielle figure 6, on a :

$$\frac{max|P(j.v)|}{min|P(j.v)|} \approx 60\text{db} - [-14\,db] \approx 74\text{db} \approx 10^{74/20} \approx 5000$$

Si on veut une erreur absolue sur la réponse fréquentielle située à -20db du minimum dans le pire des cas(10% d'erreur au pire sur la réponse fréquentielle), Il faudra choisir $\rho = 2^{-[N_B - 1]}$ de telle façon que :

$$\frac{\rho . max|P(j.v)|}{min|P(j.v)|} \leq -20\text{dB} \Leftrightarrow \rho \leq [-14\text{db} - 60\text{db}] - 20\text{dB} \approx 10^{-94/20} \approx 2.10^{-5}$$

Et donc :

-pour assurer $10\% = \frac{0.1}{1} = -20\,dB$ d'erreur sur la réponse fréquentielle dans le pire des cas,

-il faut coder les coefficients avec une précision meilleure que

$$\frac{2.10^{-5}}{1} = 0.002\% = -94\,dB$$

-la perte de précision relative entre les coefficients et la réponse fréquentielle est de

$$74\text{db} = 10^{\frac{74}{20}} = \frac{max|H(jv)|}{min|H(jv)|}$$

3.3 remarque générale sur le codage des filtres IIR, par rapport aux filtres FIR

Un filtre *IIR* a pour intérêt essentiel, par rapport à un filtre *FIR*, de réaliser avec peu de coefficients.des réponses temporelles à mémoire longue, soit encore des réponses fréquentielles à faible bande-passante,
Or une réponse fréquentielle à faible bande passante correspond
- soit à une petite pseudo-pulsation de coupure $v_p \ll 1$
- soit à un petit facteur d'amortissement $\xi \ll 1$
- soit aux deux simultanément

Dans ces deux cas, le nombre de bits de codage nécessaire pour rendre correctement compte de la réponse fréquentielle devient très élevé, et il devient impossible d'employer un filtre IIR(particulièrement lorsque $v_p \ll 1$).
En conclusion, et en noircissant un peu le tableau :
le codage en nombres entiers des filtres IIR en z^{-1} n'est satisfaisant que dans les cas où on aurait pu employer un filtre FIR à la place!...
Ce qui est un petit peu choquant pour l'esprit....

4 résolution par changement d'opérateur

4.1 les solutions actuelles

Les 3 grandes classes de solutions actuellement employées pour améliorer le problème de la sensibilité de la réponse fréquentielle vis à vis de la quantification des coefficients sont :
1- l'utilisation de la représentation d'état (formes normales, formes à sensibilité L2 minimales) pour rendre les pôles du filtre moins sensibles aux coefficients.
Ces techniques ne seront pas abordées ici, car
1.1- elles font appel à des mathématiques d'un niveau trop élevé.
1.2- elles demandent 9 multiplications / cellule d'ordre 2 (au lieu de 5 multiplications pour les formes directes),

1.3- elles n'améliorent que très partiellement la sensibilité (proportionnelle à $\frac{1}{v_p}$ et $\frac{1}{\xi}$)

2- la programmation du filtre sous la forme de 2 branches passe-tout de gain unité en parallèle.
Cette technique améliore la sensibilité vis à vis de faibles facteurs d'amortissement, mais ne change rien vis à vis de la pseudo-pulsation de coupure v_p . De plus elle n'est applicable que pour des filtres de formes particulières, et pas pour n'importe quelle fonction

de transfert.
3- la technique de changement d'opérateur : transformée en delta
Cette technique permet d'améliorer grandement la sensibilité de la réponse fréquentielle vis a vis de la quantification des coefficients, mais souffre d'un comportement peu satisfaisant vis a vis du bruit de signal (cycles limite, erreurs statiques).

La technique qui est présentée ici s'inscrit dans la classe des techniques de changement d'opérateur, sans souffrir des problèmes de sensibilité au bruit de la transformée en Delta.

4.1 Principe

4.1.1 propriété désirée

Notons *q* l'opérateur employé pour le codage du filtre.

Jusqu'ici nous avons employé $q(z^{-1})=z^{-1}$, soit encore $q(w)=\frac{1-w}{1+w}$

On peut résumer le paragraphe 3 par :

soit $P(q)=p_0+p_1.q+p_2.q^2$ un polynôme en $q(z^{-1})=z^{-1}$, alors l'élément correspondant dans le plan W est une fraction rationnelle , dont tous les pôles sont situés en $w=-1$.

$$P(w)=p_0+p_1.\underbrace{q(w)}_{\frac{1-w}{1+w}}+p_2.\underbrace{q^2(w)}_{\left[\frac{1-w}{1+w}\right]^2}=\frac{p_{0w}+p_{1w}.w+p_{2w}.w^2}{\left[1+w\right]^2}$$

Soit encore, en faisant apparaître la pseudo-pulsation de coupure v_p , le gain statique K et le facteur d'amortissement ξ :

$$P(w)=K.\frac{1+2.\xi.\frac{w}{v_p}+\left[\frac{w}{v_p}\right]^2}{\left[1+w\right]^2}$$

La réponse fréquentielle associée, lorsque $v_p<1$ est celle d'un ***filtre passe-haut,*** et c'est de là que viennent les problèmes : La variation relative du gain de ce filtre est proportionnelle à $\frac{1}{v_p^2}$ et croît donc très vite au fur et à mesure que v_p diminue.

La(une) solution pour pallier à ce problème consiste à choisir un opérateur $q(w)$ qui permette d'éliminer le comportement inutilement passe-haut de $P(w)$.

L'idéal serait d'obtenir , en fonction de K, v_p, ξ , l'expression suivante de $P(w)$:

$P(w)=K.\frac{1+2.\xi.\frac{w}{v_p}+\left[\frac{w}{v_p}\right]^2}{\left[1+\frac{w}{v_p}\right]^2}$, dont le gain statique et le gain à l'infini sont égaux.

4.2 Expression de l'opérateur

4.2.1 opérateur idéal

De plus, pour pouvoir être programmé dans une boucle, l'opérateur $q(z^{-1})$ doit s'annuler lorsque $z^{-1}\to 0$.

donc l'opérateur $q(w)$ doit s'annuler lorsque $w \to 1$.
En normalisant à 1 le gain statique de $q(w)$, on obtient la 'liste des opérateurs possibles ':

$$q(w) = \frac{1-w}{1+\frac{w}{v_q}} \Leftrightarrow w(q) = \frac{1-q}{1+\frac{q}{v_q}}$$,

soit encore en z^{-1} (en posant $w = \frac{1-z^{-1}}{1+z^{-1}}$ dans l'expression de $q(w)$) :

$$q(z^{-1}) = \frac{a_q . z^{-1}}{1-[1-a_q].z^{-1}}$$, avec $a_q = \frac{2.v_q}{v_q+1} \Leftrightarrow v_q = \frac{a_q}{2-a_q}$

soit $P(q) = p_0 + p_1.q + p_2.q^2$ un polynôme en q . L'élément correspondant dans le plan W est une fraction rationnelle , dont tous les pôles sont à présent situés en $w = -v_q$.

$$P(w) = p_0 + p_1 . \underbrace{q(w)}_{\frac{1-w}{1+\frac{w}{v_q}}} + p_2 . \underbrace{q^2(w)}_{\left[\frac{1-w}{1+\frac{w}{v_q}}\right]^2} = \frac{p_{0w} + p_{1w}.w + p_{2w}.w^2}{\left[1+\frac{w}{v_q}\right]^2}$$

Soit encore, en faisant apparaître la pseudo-pulsation de coupure v_p , le gain statique K et le facteur d'amortissement ξ :

$$P(w) = K . \frac{1 + 2.\xi . \frac{w}{v_p} + \left[\frac{w}{v_p}\right]^2}{\left[1+\frac{w}{v_q}\right]^2}$$

idéalement, il suffit donc de choisir $v_q = v_p$ pour éliminer le comportement passe-haut de $P(w)$.

4.2.2 opérateur techniquement réalisable, choix de v_q

techniquement,
1- la pseudo-pulsation v_q doit être codée avec une précision infinie, pour obtenir l'amélioration escomptée quant à la précision de la réponse fréquentielle. On ne peut donc pas choisir arbitrairement v_q .
2- Les multiplications par le coefficient a_q , pour la réalisation de

$$q(z^{-1}) = \frac{a_q . z^{-1}}{1-[1-a_q].z^{-1}}$$, doivent également être éliminées, si on veut un code efficace.

Pour cette raison, on se limitera à des coefficients $a_q = 2^{-L_q}$ en puissances entières de 2.
De cette façon les multiplications par a_q se traduiront par des décalages à droite.

Dans ce cas les pseudo-pulsations $v_q = \frac{a_q}{2-a_q}$ sont limitées à :

$$v_q = \frac{2^{-L_q}}{2-2^{-L_q}}$$.

On choisira L_q entier de telle façon que le rapport $\frac{v_p}{v_q}$ soit aussi proche que possible de 1.

4.2.3 exemples de choix de v_q

On va détailler ici le choix de v_q pour les exemples de la figure 5

1 - lorsque $v_p = 0.03$,

1.1- on devrait choisir $v_q = 0.03$, qui conduit à

$$a_q = \frac{2.v_q}{v_q+1} \approx 0.058 \Rightarrow L_{q\_ideal} = -\log_2(a_q) \approx 4.1$$

On choisira donc $L_q = 4$, ou $L_q = 5$

1.2-pour $L_q = 4$, on a $a_q = 2^{-4} \Rightarrow v_q = \frac{a_q}{2-a_q} \approx 0.032 \Rightarrow \ln\left(\frac{v_p}{v_q}\right) \approx -0.07$

1.3-pour $L_q = 5$, on a $a_q = 2^{-5} \Rightarrow v_q = \frac{a_q}{2-a_q} \approx 0.016 \Rightarrow \ln\left(\frac{v_p}{v_q}\right) \approx 0.63$

1.4-le rapport $\frac{v_p}{v_q}$ le plus proche de 1 correspond au logarithme le plus proche de zéro, on retiendra donc $L_q = 4 \Rightarrow v_q \approx 0.032$

2 - lorsque $v_p = 0.3$,

2.1- on devrait choisir $v_q = 0.3$, qui conduit à

$$a_q = \frac{2.v_q}{v_q+1} \approx 0.46 \Rightarrow L_{q\_ideal} = -\log_2(a_q) \approx 1.1$$

On choisira donc $L_q = 1$, ou $L_q = 2$

2.2-pour $L_q = 1$, on a $a_q = 2^{-1} \Rightarrow v_q = \frac{a_q}{2-a_q} \approx 0.33 \Rightarrow \ln\left(\frac{v_p}{v_q}\right) \approx -0.1$

2.3-pour $L_q = 2$, on a $a_q = 2^{-2} \Rightarrow v_q = \frac{a_q}{2-a_q} \approx 0.14 \Rightarrow \ln\left(\frac{v_p}{v_q}\right) \approx 0.74$

2.4-le rapport $\frac{v_p}{v_q}$ le plus proche de 1 correspond au logarithme le plus proche de zéro, on retiendra donc $L_q = 1 \Rightarrow v_q \approx 0.33$

la figure 7 correspond aux tracés des réponses fréquentielles de $P(w)$. On constate que les choix judicieux de v_q ont permis de diminuer sensiblement la variation relative du gain par rapport aux résultats précédents figure 5.

la figure 8 correspond aux tracés des résultats obtenus, lors de la quantification sur 10 bits du polynôme $H(q)$ correspondant à $v_p \approx 0.03, \xi \approx 0.1$, lorsque $L_q = 4 \Rightarrow v_q \approx 0.032$.
On constate une nette diminution de l'erreur par rapport aux résultats précédents figure 6, pour la quantification du polynôme $H(z^{-1})$
En effet la précision relative des coefficients $\rho = 2^{-[10-1]} \approx -54\,dB$ est beaucoup mieux utilisée, car la variation relative du gain $\frac{max|H(jv)|}{min|H(jv)|} \approx 20\,dB$ est beaucoup plus faible que précédemment.

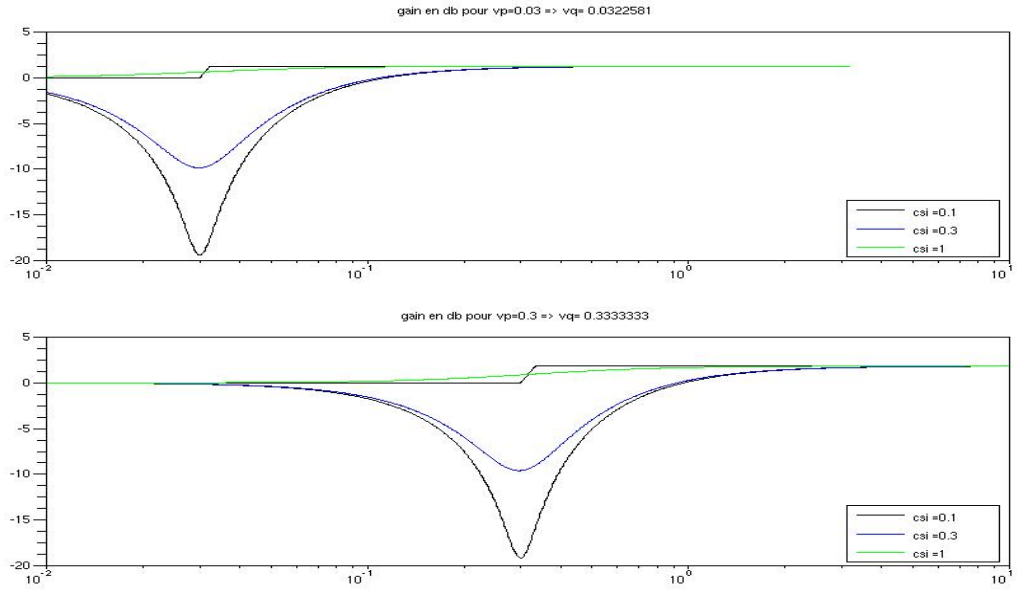


figure 7: réponses fréquentielles typiques de $P(w)=K.\dfrac{1+2.\xi.\dfrac{w}{v_p}+\left[\dfrac{w}{v_p}\right]^2}{\left[1+\dfrac{w}{v_q}\right]^2}$

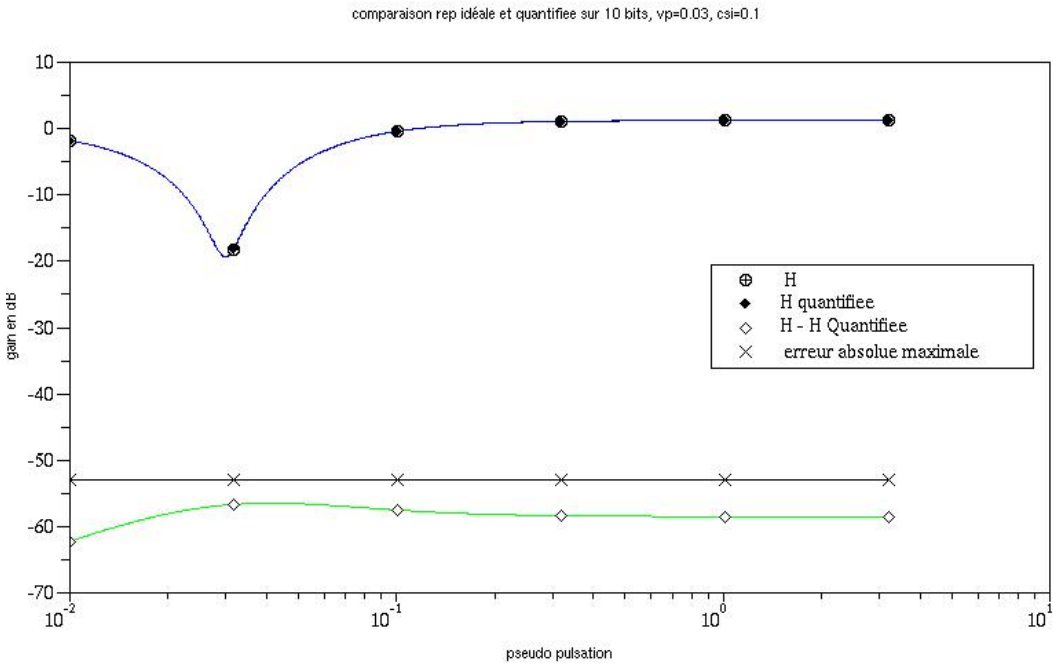


figure 8: mise en évidence des effets de la quantification sur la réponse fréquentielle

4.2.4 *programmation et schéma de principe de l'opérateur* $q\left(z^{-1}\right)=\dfrac{2^{-L_q}.z^{-1}}{1-\left[1-2^{-L_q}\right].z^{-1}}$

l'opérateur *q* peut être programmé en employant un seul décalage à droite, et une seule mémoire, conformément au schéma de principe figure 9

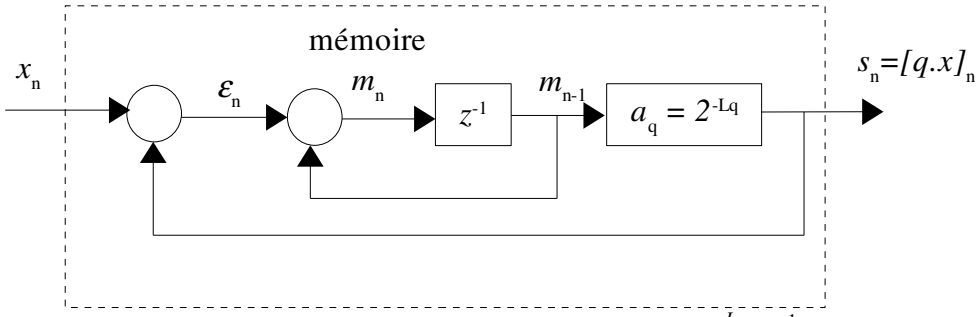


figure 9: schéma de principe de $q(z^{-1})=\frac{2^{-L_q}.z^{-1}}{1-\left[1-2^{-L_q}\right].z^{-1}}$

Le programme correspondant à chaque pas d'échantillonnage est donné ci-après :

// en entrée de programme, la mémoire m représente la quantité m_{n-1}

s = m >> Lq ; // calcul de la sortie s_n de l'opérateur à l'instant n, en fonction de la mémoire m_{n-1}

m=m-s; // mise à jour partielle de m_n , pour le moment : $m_n = m_{n-1} - s_n$

// connaissant s_n , on peut calculer m_n même s'il y a une boucle, le code dépend du schéma global

m=m+xn; // fin de la mise à jour de m_n : $m_n = m_{n-1} - s_n + x_n$

4.2.5 schéma d'analyse de l'opérateur, et analyse du bruit

le décalage à droite de L_q bits introduit un bruit de quantification b_q , et les équations correspondantes sont alors :

$$s(z^{-1})=\underbrace{\left[\frac{2^{-Lq}.z^{-1}}{1-\left[1-2^{-Lq}\right].z^{-1}}\right]}_{q(z^{-1})}.x(z^{-1})+\underbrace{\left[\frac{1-z^{-1}}{1-\left[1-2^{-Lq}\right].z^{-1}}\right]}_{H_q(z^{-1})}.b_q(z^{-1})$$

Le schéma d'analyse correspondant est représenté figure 10 :

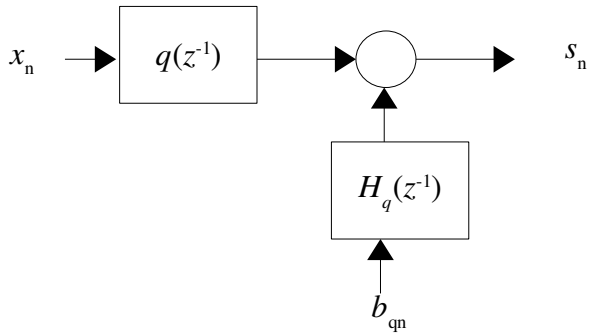


figure 10: schéma d'analyse de l'opérateur programmé

On voit donc que l'opérateur $q(z^{-1})$ introduit un bruit supplémentaire b_q , par rapport au cas de l'opérateur retard $q=z^{-1}$.

Il convient d'analyser les caractéristiques de l'effet de ce bruit sur la sortie *s* de $q(z^{-1})$, de façon à analyser la dégradation éventuelle des performances.

la fonction de transfert H_q entre le bruit b_q et la sortie *s*, s'écrit dans le plan W :

$$H_q(w)=\frac{w}{w+v_q}, v_q=\frac{2^{-L_q}}{2-2^{-L_q}}=\frac{a_q}{2-a_q} \quad .$$

Il s'agit donc d'un filtre passe-haut, de gain haute fréquence égal à 1, et de pseudo pulsation de coupure égale à v_q . On peut donc en déduire que

1- le bruit b_q n'introduira donc pas d'erreur statique(comportement dérivateur en basses fréquences)
2- l'écart-type du bruit de sortie sera au pire égal à l'écart-type de b_q (car $\|H_q\|_{H\infty}=1$)

La norme 1 de H_q peut être calculée analytiquement (fonction de transfert du premier ordre), et s'écrit $\|H_q\|_1=2, \forall L_q$.
On peut donc en déduire que le niveau maximum de bruit de sortie sera égal à 2 fois le niveau maximum de b_q , ce qui reste raisonnable (de l'ordre de 1).

4.3 exemple simple du premier ordre, filtre passe-bas

4.3.1 fonction de transfert en w

On veut réaliser un filtre dont la fonction en w est donnée par :

$F(w)=\dfrac{1+\dfrac{w}{v_1}}{1+\dfrac{w}{v_0}}$, avec $\left\{v_0\approx\tan\left(10^{-3}.\pi\right)\quad v_1\approx\tan\left(10^{-2}.\pi\right)\right.$

qui correspond à un filtre passe-bas

- de gain statique 1, de gain haute fréquence $\approx\dfrac{1}{10}$,
- de fréquence de coupure basse $f_0=\dfrac{1}{\pi}.atan(v_0).f_e=10^{-3}f_e$
- de fréquence de coupure haute $f_1=\dfrac{1}{\pi}.atan(v_1).f_e=10^{-2}f_e$

4.3.2 choix de v_q

On choisit v_q relativement au dénominateur en w, de pulsation de coupure $v_p=10^{-3}\pi$,

1.1- on devrait choisir $v_q=v_p$, qui conduit à $a_q=\dfrac{2.v_q}{v_q+1}\approx\Rightarrow L_{q\_ideal}=-\log_2(a_q)\approx 7.3$

On choisira donc $L_q=7$, ou $L_q=8$

1.2-pour $L_q=7$, on a $a_q=2^{-7}\Rightarrow v_q=\dfrac{a_q}{2-a_q}\approx 0.0039\Rightarrow\ln\left(\dfrac{v_p}{v_q}\right)\approx -0.22$

1.3-pour $L_q=8$, on a $a_q=2^{-8}\Rightarrow v_q=\dfrac{a_q}{2-a_q}\approx 0.002\Rightarrow\ln\left(\dfrac{v_p}{v_q}\right)\approx 0.47$

On retiendra donc $L_q=7$,et $v_q=\dfrac{2^{-L_q}}{2-2^{-L_q}}\approx 0.0039$

4.3.3 détermination des coefficients la fonction de transfert correspondante *F*(*q*)

on a : $q(w)=\dfrac{1-w}{1+\dfrac{w}{v_q}}\Leftrightarrow w(q)=\dfrac{1-q}{1+\dfrac{q}{v_q}}=\dfrac{v_q-v_q.q}{v_q+q}$.

En remplaçant *w* par $w(q)$ dans l'expression de $F(w)$, on obtient l'expression correspondante de $F(q)$:

$$F(w)=\frac{1+\frac{w}{v_1}}{1+\frac{w}{v_0}}=\frac{b_{0w}+b_{1w}.w}{a_{0w}+a_{1w}.w}\quad ,$$

$$F(q)=\frac{b_{0w}+b_{1w}.w(q)}{a_{0w}+a_{1w}.w(q)}=\frac{b_{0w}+b_{1w}.\left[\frac{v_q-v_q.q}{v_q+q}\right]}{a_{0w}+a_{1w}.\left[\frac{v_q-v_q.q}{v_q+q}\right]}=\frac{\left[(b_{0w}+b_{1w}).v_q\right]+\left[b_{0w}-b_{1w}.v_q\right].q}{\left[(a_{0w}+a_{1w}).v_q\right]+\left[a_{0w}-a_{1w}.v_q\right].q}$$

ce qui donne , en normalisant à 1 le coefficient de plus bas degré du dénominateur de $F(q)$:

$$F(q)=\frac{b_0+b_1.q}{1+a_1.q}=\frac{N(q)}{D(q)}\quad \text{, avec}\quad \begin{cases} b_0=\frac{b_{0w}+b_{1w}}{a_{0w}+a_{1w}}\approx 0.1028186 \\ b_1=\frac{b_{0w}-b_{1w}.v_q}{(a_{0w}+a_{1w}).v_q}\approx 0.6989104 \\ a_1=\frac{a_{0w}-a_{1w}.v_q}{(a_{0w}+a_{1w}).v_q}\approx -0.1982710 \end{cases}$$

4.3.4 quantification des coefficients de F(q) sur 8 bits

en quantifiant sur $N_B=8$ bits les coefficients du numérateur et du dénominateur de $F(q)$, on obtient la fonction de transfert en *q* du filtre quantifié :

$$F_q(q)=\frac{b_{0q}+b_{1q}.q}{1+a_{1q}.q}=\frac{N_q(q)}{D_q(q)}\quad \text{, avec}\quad \begin{cases} b_{0q}\approx 0.1025391 \\ b_{1q}\approx 0.695312 \\ a_{1q}\approx -0.1992188 \end{cases}$$

4.3.5 comparaison des réponses fréquentielles de $F(q)$ et $F_q(q)$

on peut alors comparer les réponses fréquentielles en calculant les transformées en *W* des numérateurs et dénominateurs de $F(q)$ et $F_q(q)$
par exemple :

$$N(w)=N(q(w))=b_0+b_1.q(w)=b_0+b_1.\frac{v_q-v_q.w}{v_q+w}=\frac{[(b_0+b_1)]+[\frac{b_0}{v_q}-b_1].w}{1+\frac{w}{v_q}}$$

(remarquer que $N(w)$ est (comme prévu) une fraction rationnelle, et non pas un polynôme, de la variable w, dont tous les pôles sont situés en $w=-v_q$.

Les réponses fréquentielles et erreurs sont tracées figure 11.
On remarquera que
1- les variations de gain des réponses fréquentielles des numérateur et dénominateur restent raisonnables (pas d'effet passe-haut)
2- de ce fait, la précision relative $\rho=2^{-[N_{Bits}-1]}=2^{-7}\approx -42\text{db}$ des coefficients est du même ordre de grandeur que la précision relative sur les réponses fréquentielles.

Par souci de comparaison, on a représenté figure 12 les réponses fréquentielles que l'on aurait obtenues pour un codage direct en $q=z^{-1}\Leftrightarrow v_q=1$. On remarquera que
1- les variations de gain des réponses fréquentielles des numérateur et

dénominateur sont beaucoup plus élevées (effet passe-haut jusqu'à $v_q = 1$)
2- de ce fait, la précision relative $\rho = 2^{-[N_{Bits}-1]} = 2^{-7} \approx -42\text{db}$ des coefficients est très mal utilisée (erreur relative de $-10\text{db} \approx 30\ \%$ sur le gain statique du filtre).

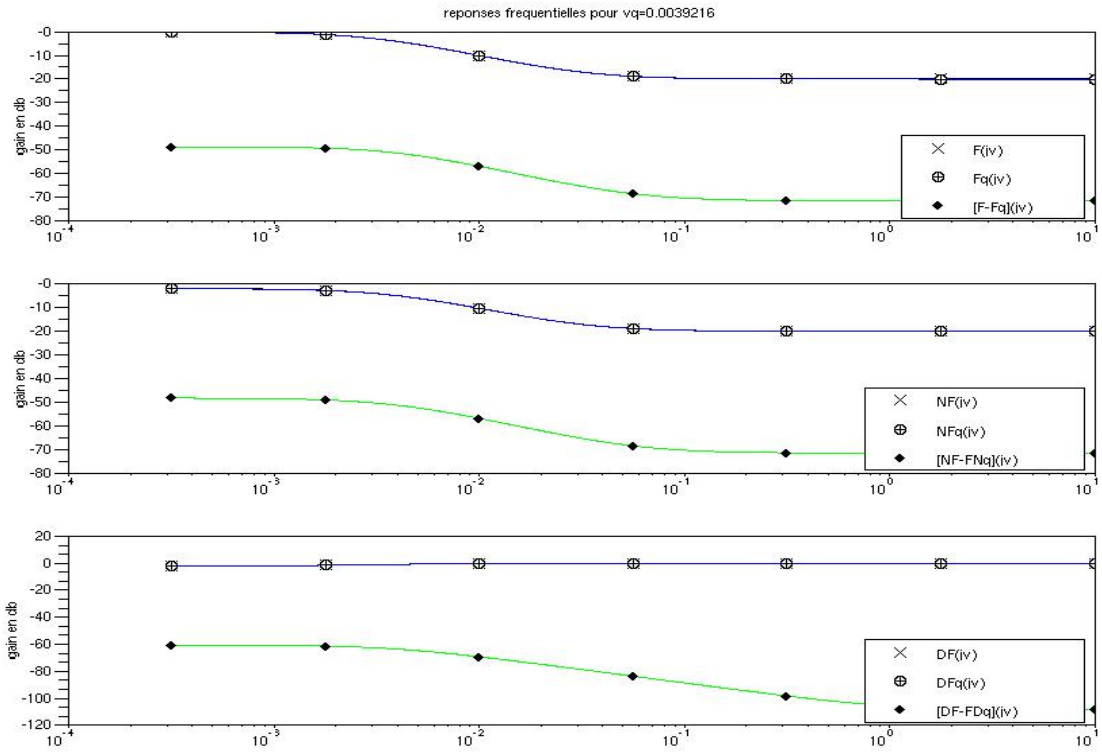


figure 11: réponses fréquentielles pour $L_q = 7 \Rightarrow v_q \approx 0.0039$

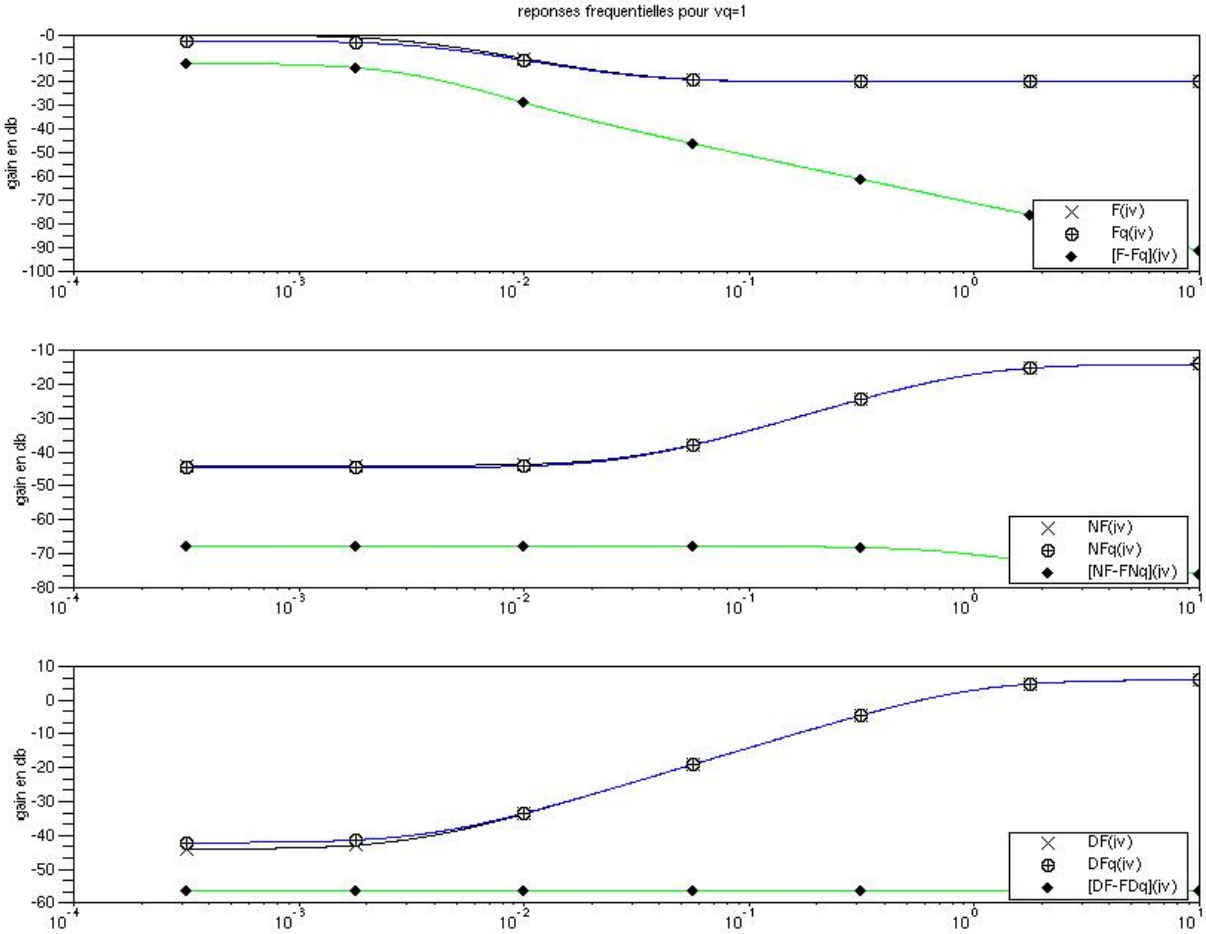


figure 12: réponses fréquentielles pour un codage en $z^{-1} \Leftrightarrow vq=1$

4.3.6 schéma de principe et d'analyse pour le codage du filtre

La fonction de transfert $F(q)$ peut être réalisée de la même façon qu'avec l'opérateur retard $q=z^{-1}$. La seule chose qui change est que l'opérateur q est réalisé conformément au schéma figure 9. Par exemple on peut réaliser $F(q)$ sous forme directe 2, en employant le schéma de principe suivant (choix du prof: on aurait pu le coder sous n'importe quelle autre forme):

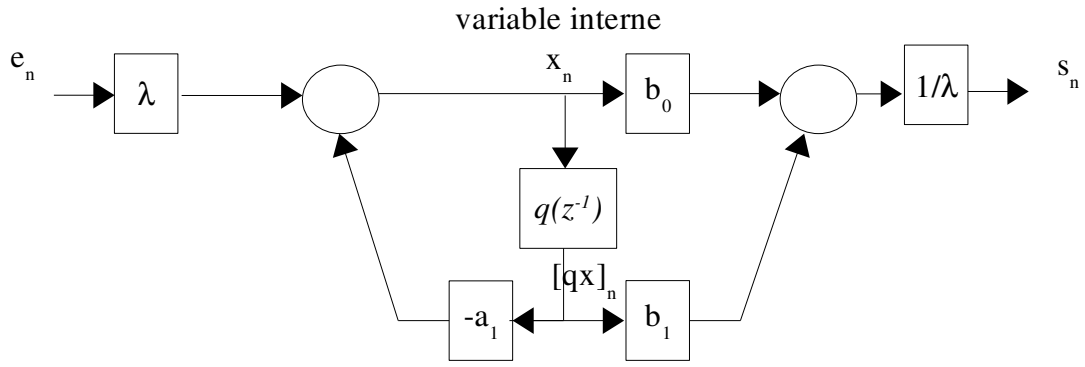


figure 13: schéma de principe sous forme df2

On va profiter de cet exemple pour montrer que les performances en termes de bruit de sortie sont nettement meilleures qu'avec une programmation en $q=z^{-1}$, et en expliquer la raison.

Le schéma d'analyse correspondant (après optimisation du code) , est représenté ci-après

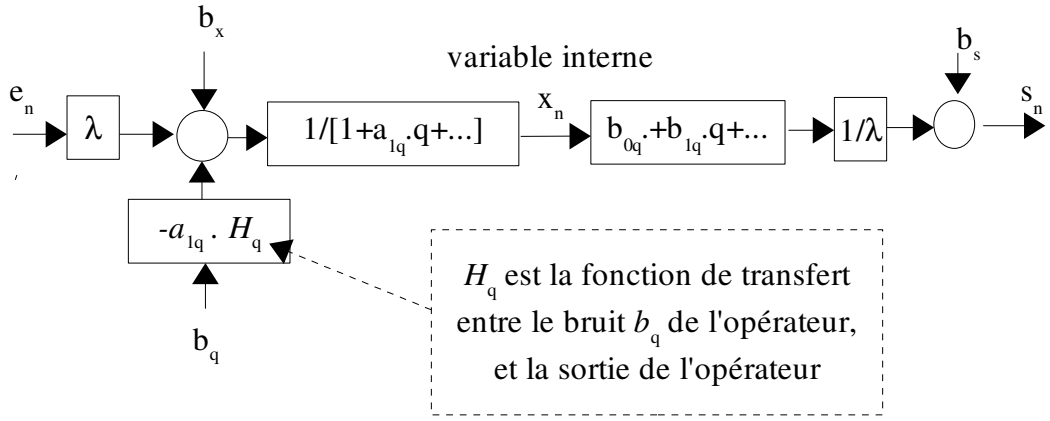


figure 14: schéma d'analyse standard de la forme df2 avec opérateur

Ce schéma d'analyse permet d'effectuer les mêmes opérations que précédemment, à savoir le calcul du plus grand facteur d'échelle λ (scaling) , et l'analyse de l'effet des bruits de quantification sur la sortie du filtre.

On remarquera l'apparition du bruit supplémentaire b_q , et de la fonction de transfert H_q , dûs à la programmation de l'opérateur $q(z^{-1})$ (voir 4.2.4 , 4.2.5)

4.3.7 *scaling, comparaison avec l'opérateur retard* $q=z^{-1}$

La fonction de transfert entre l'entrée e_n et la variable interne x_n s'écrit, en fonction de l'opérateur q :

$$H_{e\_x}(q)=\lambda.\frac{1}{1+a_{1q}.q+...}=\lambda.\frac{1}{D_{Fq}(q)} \quad ,$$

où $D_{Fq}(q)$ est le dénominateur de la fonction de transfert quantifiée en q : $F_q(q)=\frac{N_{Fq}(q)}{D_{Fq}q}$

4.3.7.1 scaling avec l'opérateur $q(z^{-1})$

Lorsqu'on emploie l'opérateur $q(z^{-1})$, avec une pseudo-pulsation v_q proche de la pseudo-pulsation de coupure du dénominateur en *w*, la réponse fréquentielle $D_{Fq}(w=j.v)$ reste de l'ordre de 1(voir figure 11 , pas de comportement passe-haut de $D_{Fq}(j.v)$ }.

Les normes de la fonction de transfert $\frac{1}{D_{Fq}(q(z^{-1}))}$ seront donc également de l'ordre de 1. On pourra donc choisir un facteur d'échelle λ qui sera également de l'ordre de 1.

techniquement, on obtient :

$$D_{Fq}(z^{-1})=1+a_{1q}.q(z^{-1})=1+a_{1q}.\frac{2^{-L_q}.z^{-1}}{1-\left[1-2^{-L_q}\right].z^{-1}}=\frac{1+\left[2^{-L_q}-a_{1q}.\left(1-2^{-L_q}\right)\right].z^{-1}}{1-\left[1-2^{-L_q}\right].z^{-1}}\approx\frac{1-0.9922\,z^{-1}}{1-0.9937\,z^{-1}}$$

Et : $H_{e\_x}(z^{-1})=\lambda.\frac{1}{D_{Fq}(z^{-1})}\approx\lambda.\frac{1-0.9937\,z^{-1}}{1-0.9922\,z^{-1}}$, dont on peut calculer analytiquement les normes :

$\left\|H_{e\_x}\right\|_1\approx 1.25\ |\lambda|$, $\left\|H_{e\_x}\right\|_2\approx 1\ |\lambda|$, $\left\|H_{e\_x}\right\|_{H\infty}\approx 1.25\ |\lambda|$

En raisonnant sur la norme 1, le plus grand facteur d'échelle $\lambda=2^{L}$ est donc $\lambda=2^{-1}$

4.3.7.2 scaling avec l'opérateur retard $q=z^{-1}$

L'emploi de l'opérateur retard $q=z^{-1}$, se ramène à utiliser une pseudo-pulsation $v_q=1$. La réponse fréquentielle $D_{Fq}(w=j.v)$ est alors celle d'un filtre passe-haut, dont le gain en haute fréquence est de l'ordre de 1, et le gain en basse-fréquence est très petit(voir figure 12 comportement passe-haut de $D_{Fq}(j.v)$). La réponse fréquentielle de la fonction de transfert $\frac{1}{D_{Fq}(q(z^{-1}))}$ sera donc celle d'un filtre passe-bas, de gain haute fréquence de l' ordre de 1, et de gain basse-fréquence très grand. On peut donc s'attendre à ce que les normes correspondantes soient très élevées, et on devra donc choisir un facteur d'échelle λ qui sera également petit devant 1.

techniquement, on obtient :

$D_{Fq}(z^{-1})=1+a_{1q}z^{-1}\approx 1-0.9922\,z^{-1}$

Et : $H_{e\_x}(z^{-1})=\lambda.\frac{1}{D_{Fq}(z^{-1})}\approx\lambda.\frac{1}{1-0.9922\,z^{-1}}$, dont on peut calculer analytiquement les normes :

$\left\|H_{e\_x}\right\|_1\approx 128\ |\lambda|$, $\left\|H_{e\_x}\right\|_2\approx 8\ |\lambda|$, $\left\|H_{e\_x}\right\|_{H\infty}\approx 128\ |\lambda|$

En raisonnant sur la norme 1, le plus grand facteur d'échelle $\lambda=2^{L}$ est donc $\lambda=2^{-7}$, beaucoup plus faible que dans le cas précédent. On peut donc s'attendre à une amplification des bruits de quantification beaucoup plus importante

4.3.7.3 *analyse, comparaison avec l'opérateur retard* $q=z^{-1}$

En raisonnant sur le schéma d'analyse, la sortie due uniquement aux bruits s'écrit :

$$S(z^{-1})=b_s(z^{-1})+\frac{1}{\lambda}.\left[F_q(z^{-1}).b_x(z^{-1})-a_1.F_q(z^{-1}).H_q(z^{-1}).b_q(z^{-1})\right]$$

en raisonnant avec la norme 1, on en déduit donc :

$$max|s|_{due\,aux\,bruits}=max|b_s|+\frac{1}{\lambda}.\left[\left\|F_q\right\|_1.max|b_x|+\left\|a_1.F_q.H_q\right\|_1.max|b_q|\right]$$

Pour interpréter cette expression,

1- on néglige l'effet du bruit de quantification final b_s

2- de plus, à la quantification des coefficients près, la fonction de transfert $F_q(z^{-1})\approx F(z^{-1})$ ne dépend pas de l'opérateur employé.

3- On majore la norme $\left\|a_1.F_q.H_q\right\|$ par $\left\|F_q\right\|.\left\|H_q\right\|$ (le coefficient a_1 est en effet de module inférieur à 1, de par la manière dont on choisit v_q)

4- on suppose que les bruits ont les mêmes valeurs maximales $max|b_x|=max|b_q|=b_{max}$

on obtient alors:

$$max|s|_{due\,aux\,bruits}\le\frac{1}{\lambda}.\left\|F\right\|_1.\left[1+\left\|H_q\right\|_1\right].b_{max}$$

de plus on a montré en 4.2.5 que $\|H_q\|_1 = 2, \forall v_q \neq 1$, et $H_q = 0,$ lorsque $v_q = 1$

En raisonnant sur la norme 1, on aura donc :

1- $max|s|_{\text{due aux bruits}} \approx \frac{1}{\lambda}.\|F\|_1.[3].b_{max}$, lorsque $v_q \neq 1$

2- $max|s|_{\text{due aux bruits}} \approx \frac{1}{\lambda}.\|F\|_1.b_{max}$, lorsque $v_q = 1$ (cas où $q = z^{-1}$)

Lorsque $v_q \ll 1$, la quantité ***1*** sera beaucoup plus petite que la quantité ***2***, et les performances en termes de bruit de sortie seront bien meilleures. (parce que le facteur d'échelle en ***1*** est beaucoup plus grand que le facteur d'échelle en ***2***)

Pour l'exemple considéré, on obtient (sans aucune approximation, et par calcul numérique)

1- $max|s|_{\text{due aux bruits}} \approx 2{,}2.b_{max}$, pour la valeur de v_q utilisée

2- $max|s|_{due\,aux\,bruits} \approx 96.b_{max}$, lorsque $v_q = 1$ (cas où $q = z^{-1}$)

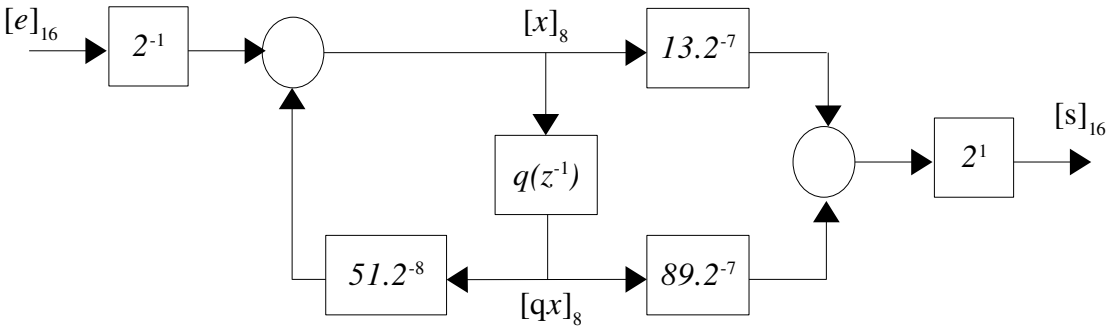
remarque: on obtient le même type de résultat pour les autres formes directes de programmation, et pour les autres normes, pour les mêmes raisons...

4.3.7.4 programmation du filtre avec l'opérateur $q(z^{-1})$

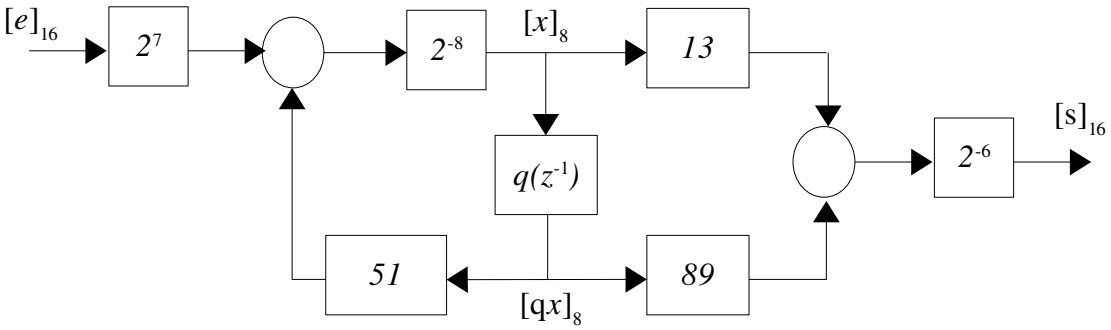
La programmation du filtre est strictement identique à une programmation avec l'opérateur retard, la seule chose qui change est la façon dont on calcule la sortie de l'opérateur, en fonction de l'entrée.

1- à partir du schéma de programmation naïf ci-dessous correspondant

- à la forme ***df2*** du filtre, avec l'opérateur $q(z^{-1})$
- avec un facteur d'échelle $\lambda = 2^{-1}$
- le même facteur d'échelle $L_{b0} = L_{b1} = 7$ pour les coefficients b_0, b_1



2- on peut appliquer les mêmes techniques d'optimisation qu'à la séance 2, pour arriver au schéma optimisé ci-après



3- il ne reste plus qu'à en déduire le programme correspondant, dont la seule nouveauté est la programmation de l'opérateur $q(z^{-1})$ conformément au paragraphe 4.2.4

```
//-------------------------------------------------------------
// calcul de la sortie [qx]8 de l'operateur q(z-1), avec Lq=7
// en fonction de sa memoire interne [mq]16
// conformement au paragraphe 5.3.4
//-------------------------------------------------------------
  qx_8 = mq_16 >> 7 ;
  mq_16-=qx_8; // mise a jour partielle de la memoire interne de l'operateur q(z-1)
//-------------------------------------------------------------
// calcul de l'entree [x]8 de l'operateur q(z-1)
// en fonction de [e]8 et [qx]8
//-------------------------------------------------------------
  acc_16 = e_8 ;
  acc_16 =acc_16 << 7 ;
  acc_16 += 51 * qx_8 ;
  x_8 =acc_16 >> 8 ;
//-------------------------------------------------------------
// mise a jour de la sortie [s]16 en fonction de [x]8 et [qx]8
//-------------------------------------------------------------
  s_16=13*x_8;
  s_16+=89*qx_8;
  s_16>>=6;
//-------------------------------------------------------------------------
// fin de la mise a jour de la memoire interne [mq]16 de l'operateur q(z-1)
//-------------------------------------------------------------------------
  mq_16+=x_8;
```

5 exercices

en vous inspirant de l'exemple traité dans la partie 4.3 , rédiger les différentes phases (scaling, analyse des bruits et rédaction du bout de programme en ***langage c*** incluses) de réalisation d'un filtre numérique passe-bas de fonction de transfert $F(w)=\frac{1}{1+\frac{w}{v_0}}$ avec

$v_0=tg\left(\pi.\frac{f_0}{f_e}\right)$, $f_0=100\text{hz}, f_e=8000\text{hz}$

-sous forme directe 2

- avec l'opérateur $q(z^{-1})$ adapté à la pseudo-pulsation propre du dénominateur de la fonction de transfert en w.
- en arithmétique 16/32 bits signée